Applied Statistics in Healthcare Research

Applied Statistics in Healthcare Research

Theoretical concepts and computational methods using SAS applications and the Webulators

WILLIAM MONTELPARE, EMILY READ, TERI MCCOMBER, ALYSON MAHAR, AND KRISTA RITCHIE

KIM MEARS

Applied Statistics in Healthcare Research by William J. Montelpare, Ph.D., Emily Read, Ph.D., Teri McComber, Alyson Mahar, Ph.D., and Krista Ritchie, Ph.D. is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, except where otherwise noted.

Contents

Acknowledgements ix

Theoretical concepts and computational methods using SAS applications

Foreword x

Preface 1

A *Note about the use of* SAS 1

Part I. Basic Principles

1. Introduction 5

Basic Principles for Applied Statistics in Healthcare 5

You got your numbers where 5

Observations 5

Surveys and Data Collection Forms 6

Risks of Bias 7

Accounting for missing data 8

A *Guiding Principle* 9

2. Measurement Concepts 10

Measurement 10

3. All About Variables 16

Defining a Variable 16

Types of Variables 17

Creating New Categorical Variables 17

Independent versus Dependent Variables 18

Here is what we covered in this section 18

4. Thinking Statistically About Your Research 19

Essential Design Considerations 19

The Research Process 20

5. The Hierarchy of Evidence 24

6. Types of Research Designs 26

Research Designs 26

Observational Research Designs 26

Experimental Research Designs 32

7. John Snow and the Natural Experiment 36
Historical Background 36

Part II. SAS Programming

8. Components of a SAS Session 45
Entering Data and Writing a SAS Program 46
An annotated practice example 47
How to Write the SAS Program 48
9. Running a SAS Program 55
10. Data Screening and Cleaning 59
11. Working with Missing Data 60
Types of Missing Data 60
12. Graphing Data for Effective Presentations 65
More Simple Barcharts – Graphing data as a Frequency Distribution Bar Chart 70
Creating a Line Graph to Summarize Data 76
Creating a Pie Chart to Represent Summary Data 80
Producing Bubble Plots 83
Producing Star Charts 88
Preparing data for graphing by transposing datasets 91

Part III. Goodness of Fit and Related Chi-Square Tests

13. Frequency Distributions 103
13.1 *Analyzing Distributions of Data* 103
Frequency & Distribution of a Count Variable 105
13.2 *Distribution for a categorical variable* 112
13.3 *Creating a Histogram in* SAS 116
Dividing a continuous variable into categories 118
13.4 *Outliers* 122
14. Percentiles 124
What is a percentile? 124
15. Introducing the Goodness of Fit Chi-Square 130
The goodness of fit chi-square for one sample with four categories 130
16. Goodness of Fit Chi-Square for k=5 137
17. The Goodness of Fit Test for Two Groups 144
The Two-Sample Chi-Square Goodness of Fit Test 144
18. Multi-way Contingency Table Chi-Square Analysis 152
Application of the Goodness of Fit Chi-square analysis to multi-way tables (3×3 and beyond) 152

19. All That From the 2 x 2 Table 157
Part 1: Introduction to the 2 x 2 Chi-Square test 157
2 x 2 CHI SQUARE WEBULATOR 161
The Case for COVID-19 Testing 163
20. Fisher's Exact and the Phi Coefficient 167
Part 2: Calculating Fisher's Exact Statistic 167
Part 3: Calculating Associations in 2 x 2 tables with the Phi Coefficient 168
21. Estimating Relative Risk, the Odds Ratio, and Attributable Risk 170
Assessing Risk 170
Relative Risk – Defining the term from the 2 x 2 table 170
Estimating the Odds Ratio 174
Estimating Attributable Risk 178
Estimating Attributable Risk Fraction and the Population Attributable Risk 181

Part IV. Analysis of Non-Parametric Outcomes

22. Calculating Probabilities 185
23. Computing the Sign Test 202
24. Computing the Wilcoxon-Mann-Whitney U Test 209
Your Turn 212
25. Computing the Z Statistic for the One Sample Runs Test 215

Part V. Parametric Statistics

26. Measures of Central Tendency 221
PART 1: *Measures of Central Tendency* 221
27. Measures of Variance 228
PART 2: *Measures of Variance* 228
PART 3: *Shapes of Distributions* 231
28. Estimating Confidence Intervals for a Sample Mean 235
Putting it all together 235
29. Applying the Student's t-test for Single and Paired Samples 238
The Pairwise t-test 244
30. The t-test for Independent Sample Means and Pooled Versus Unpooled Variance 252
31. The One Way Analysis of Variance and Post Hoc Tests 261
31.1 *Analysis of Variance* 261
31.2 *Computing the* ANOVA *by hand* 264
31.3 *Creating the* SAS PROC ANOVA *program* 266
31.6 *Determining the Location of the Difference in Means Using Post Hoc Tests or Confidence Intervals* 268

32. Research Design Applications with PROC GLM 273
33. Statistical applications with linear regression analyses 279
34. Logistic Regression Analysis using PROC LOGISTIC 287

Part VI. Measuring Correlation, Association, Reliability and Validity

35. Computing the Pearson Product Moment Correlation Coefficient 291
Defining the Correlation Coefficient for Continuous Data: the Pearson Product Moment Correlation Coefficient 291
36. Computing Correlations Based on Ranks 304
37. Demonstrating the Bland-Altman Tests for Agreement 305
38. Measures of Association - Part I: The McNemar Chi-Square 306
Part 1: The McNemar Test of Symmetry[1] 306
39. Measures of Association -- Part II: The Kappa Statistic 311
Part II: The Kappa Statistic to Measure Agreement 311

Part VII. Advanced Concepts for Applied Statistics in Healthcare

40. Computing Sample Size and Power 317
41. Using Webulator Applications to Compute Sample Size 329
1. Determining Sample Size for a Simple Random Sample to Estimate a Population Proportion 329
2. Determining Sample Size for a Comparison Study 331
3. Determining Sample Size for a Case-Control Study 333
4. Determining Sample Size for a Cohort Comparison Study 334
42. Computer Simulation and Random Number Generators 340
Consider an example using the Lotto 649 343
An Applied Health Example using Simulated Data 346
43. Survival Analysis 351
44. Repeated Measures, Split Plots, and Mixed Model ANOVAS 374

Acknowledgements

Theoretical concepts and computational methods using SAS applications

The following individuals were responsible for the production of this textbook.

Principal Author:

William J. Montelpare, Ph.D., Professor, and the Margaret and Wallace McCain Chair in Human Development and Health,
Department of Applied Human Sciences, Faculty of Science/Faculty of Nursing,
Rm 122, Health Sciences Building, University of Prince Edward Island,
550 Charlottetown, PE, Canada, C1A 4P3

Co-Authors

Emily A. Read, Ph.D., RN, University of New Brunswick (emily.read@unb.ca)
Teri McComber, Ph.D.(c), University of Prince Edward Island (tmccomber@upei.ca)
Alyson Mahar, Ph.D., University of Manitoba (Alyson_mahar@cpe.umanitoba.ca)
Krista Ritchie, Ph.D., Mount Saint Vincent University (krista.ritchie@msvu.ca)

Together we achieved the main premise of this textbook, which was to present the basic concepts of statistical methods while using SAS coding methods to evaluate data and to develop a conceptual understanding of what the results are telling us about the data. While each chapter introduces the essential theoretical foundation of statistical concepts, each concept is presented through the unpacking of relevant examples using SAS programming code, and in some cases the use of Webulators© – web-based calculators written in JavaScript and HTML.

From a pedagogical perspective, this textbook will introduce the essential elements of statistical methods applied to research questions in health using applications at an intermediate level while providing examples for the reader to relate applications of these methods to health data. The methods include but are not limited to examples from health related disciplines – healthcare, health services delivery, and health promotion with a view to understanding and implementing research design and statistical applications that researchers may use as a basis for the development of research hypotheses and a theoretical foundation for program planning, policy changes, and program modifications.

The textbook is arranged intentionally for instructors and students to work through a logical approach to statistical applications that progress from basic concepts to more complex applications of statistical methodology.

Foreword

Many students often feel a chasm between textbooks that describe research methods in a general sense and textbooks that describe statistical procedures. There is often a disconnect between the statistical concepts and skills we in healthcare aim to learn about and the contexts in which we plan to apply those concepts and skills. Our intention in producing this textbook is to facilitate a bridge that will span the chasm between generalized research design textbooks and the statistical and computational methods textbooks.

When learning about applied statistics, our motivations are different than for those who aim to become analysts or statisticians. Healthcare providers, and students who aim to work in the field of healthcare, most often learn about statistics with a goal of evidence-based practice in mind. The clinical practice and the well-being of patients and families are at the heart of our motivation to learn how to analyze data. A perfect time to gain an understanding of and facility with statistics is when you have data that are deeply meaningful to your professional interests and goals. This is because the heart of the matter is not the formula or the software coding language. Those are only important as tools to reach your goal of better understanding a patient population or important health system issue.

When the numbers you are faced with are tied to concepts such as patient data, you can use your professional expertise to think in a nuanced way about how, for example, the treatment might influence a given outcome, but only if you also account for patient demographic variables such as compliance to the treatment and previous surgeries. You are no longer just mapping analysis to a grid of abstract numbers. You are thinking about and using your expertise to identify patterns, relationships, differences, and probabilities that make sense clinically and that can inform your practice.

This goal of evidence-based practice requires building a skill set that involves data and statistical literacy, and in many cases becoming part of a research team and doing our own studies and evaluations of applied practices. It is empowering to be able to utilize the institutional data we have available to us to better understand what is happening at an aggregate level. As clinicians and healthcare staff, we have a great deal of expertise that has been accumulated because of our experience through interaction with specific cases, some of which represent the standard and some of which represent the exceptional. This professional expertise, that is developed over time, is one important source of evidence in our work. We become prepared with greater insight and understanding of the complexity of our workplaces when we can complement that clinical expertise with evidence that comes from our research and evaluation efforts.

Preface

This textbook was designed to introduce individuals to the concepts and strategies of statistical analyses for problems in health-related disciplines. The primary learning objective of this textbook is to introduce the reader to a variety of statistical methods and basic analytical procedures associated with processing data in regard to healthcare research. It is intended that by working through the applications and practice problems, readers should be able to understand and apply some of the methods for developing, implementing and critically evaluating data within the various disciplines of health, health sciences, healthcare, and health services delivery. Secondary objectives of this textbook include the development of an understanding of the theoretical concepts of statistical applications, different strategies for evaluating research questions using statistical methods, and an ability to interpret and critically evaluate statistical analyses, which can be used in measurement and evaluation.

Primary Outcome Linked to Competencies

Working through the material in this textbook, the reader should be able to apply basic statistical methodologies to support decision making within the various disciplines of applied health, including but not limited to nursing, health sciences, healthcare, and health services delivery. Specifically, in this textbook, the reader will be introduced to examples that include the methods by which to:

- Distinguish between descriptive and inferential statistics
- Classify levels of measurement
- Develop and interpret data using frequency distributions
- Apply and interpret various graphing techniques – present data using graphing methods that include, but are not limited to line, bar, bubble and pie charts
- Generate measures of location and measures of dispersion
- Calculate and interpret percentiles
- Apply binomial probability distribution methods
- Calculate normal probabilities using the z-test
- Calculate normal probabilities using the t-test
- Describe methods to select a sample
- Distinguish between measurements for a sample and for a population
- Determine sample size under various scenarios
- Differentiate between a population parameter and a sample statistic
- Compute point estimates and confidence intervals
- Apply hypothesis-testing methodologies
- Apply the computation of confidence intervals to decision making
- Apply tools of non-parametric analyses to tests of hypotheses
- Evaluate the goodness of fit using the chi-squared test
- Evaluate a contingency table using the chi-squared test

A Note about the use of SAS

Throughout this textbook, SAS University Edition is used as the platform upon which to compute statistics, and as the

environment in which students can actively engage and problem solve. The textbook begins with an introduction to SAS University Edition so that the very novice user of statistics, programming languages, and SAS will feel comfortable with the theory and the examples, as presented. Likewise, throughout this textbook, we will use the web as a destination to locate information and gain direct assistance in problem-solving and task resolution.

PART I

BASIC PRINCIPLES

Chapters in This Section

In this first section, information will be presented that introduces the basic principles of applied statistics in healthcare research and includes the following topics

- Introductory and Essential Concepts
- Levels of measurement,
- Defining and describing variables
- Thinking statistically about your research – Essential design considerations
- The hierarchy of evidence
- Types of research designs
- John Snow and the Natural Experiment

1. Introduction

Basic Principles for Applied Statistics in Healthcare

The primary goal of this textbook is to provide the reader with the opportunity to learn fundamental statistical concepts while introducing the reader to data analysis skills that will enable them to become critical consumers of research and enhance their confidence in asking and answering questions about health-related issues.

The main premise of this book is to present the basic concepts of statistical methods while using SAS coding methods to evaluate data and to develop a conceptual understanding of what the results are telling us about the data. While each chapter introduces the essential theoretical foundation of statistical concepts, each concept is presented through the unpacking of relevant examples using SAS programming code.

From a pedagogical perspective, this textbook will introduce the essential elements of applied health using statistical applications at an intermediate level while providing examples for the reader to relate applications of these methods to health data. The methods include but are not limited to examples from health, with a view to understanding and implementing research design and statistical applications that researchers may use as a basis for the development of research hypotheses and a theoretical foundation for program planning, policy changes, and program modifications.

We will begin this textbook with a few "trade secrets" of researchers that are at the cusp of where research methods bridges with applied statistical analysis.

You got your numbers where

It is critical to intimately know your sources of data. Often in healthcare, we can access secondary sources of data from large databases. Other times we are fortunate to have had a summer student or a recent team evaluation effort that involved using some surveys to document the patient experience. The very first step in any statistical analysis is to understand the numbers you are about to analyze. To decide which analysis to use, we need to understand the methodological and measurement properties of our variables. We have parametric options for analyzing continuous data and non-parametric for categorical or discrete data. The understanding you need to have of the numbers you are about to analyze goes deeper than this. You need to understand exactly how data were collected and entered to determine if they are ready for analysis. The key questions you need to be able to answer before trusting the data are worth your valuable analysis time focus on issues of measurement, risks of bias, and accounts of missing data.

Observations

The things we study in applied social and health sciences are very rarely directly observable. Things that are directly observable include the number of days someone stayed in a hospital or waited for a referral to see a specialist. We can directly observe prescriptions filled, but we cannot directly observe the number of times someone took the drug as

prescribed. We can ask people to self-report their compliance and hope that they are being fully honest with perfectly accurate memories. If asking a question about the effects of a drug on given decreased symptom reporting or recovery, this self-report data becomes extremely valuable. What is important as the person analyzing the data is to recognize that this is not a perfect picture of exactly what happened for every person in your sample. There is a chance that some people misremembered and the others are trying to please you and reporting better behavior than what is happening at home. These are, hopefully, small deviations from the true patterns of taking the drug that happened in reality- the events you want to associate statistically with decreased reported symptoms and increased recovery. There are two ways we handle the fact that our data on difficult to observe variables are never perfect. First, we try to be strategic about how the data are collected. If you have a patient report monthly about how they are taking their medication, then there are greater risks of generalizing behaviors and misremembering exactly how many times medication was taken late or not at all. We can design data collection strategies that are closer to the moment of taking the medication – asking the participants to journal their behaviors on a chart beside their meds, or use an electronic application to signal when it is time to take the medication and to click a checkmark at the time when they did, in fact, take the medication. These data can be directly fed into a database to remove data entry errors that are possible with chart reviews. Taking on such a high-tech strategy requires the resources, technological and financial, to create the application. It also requires the participants to have the technology and understanding of the technology to use the application. This in turn can create a selection bias that systemically excludes people who cannot afford a mobile device or who are not tech-savvy enough to use the application. The point of this hypothetical scenario is not to discourage you from collecting data! The lesson to be learned is that researchers and healthcare providers make many nuanced decisions that go into exactly how data are collected. If you are going to analyze data, you need to have the full story (methods of data collection) behind exactly where your numbers came from. With this information, you will better understand and more accurately interpret the results you see when you analyze the data. We cannot strive to capture perfectly things that are not directly observable. We can use principles of good measurement and research methods to come as close to valid (aka accurate) and reliable (aka consistent) data as possible.

As someone who might be about to analyze data and form conclusions from analyses that impact patient care or health system decision making, it is imperative that you take a critical perspective to assess the quality of the data in front of you. When you report your results, it is important that you make transparent the limitations of your measures so that others can draw their own conclusions as relevant to their contexts. It is also important that you make the decision before you start to analyze data, that the data are worth analyzing. It is tempting when one can open a data file and start asking and answering questions to take at face value that the numbers accurately and consistently represent the variables we are interested in. If we do that we run the danger of perpetuating misinformation, which can have negative consequences for the subsequent healthcare decisions made.

Surveys and Data Collection Forms

Sometimes there are questionnaires, surveys, or data collection sheets available to us in our places of work. Forms that people have used before and we would be expected to use moving forward. Sometimes we do not have a tool for data collection waiting for us and we must do a literature search to look for a survey that we can give to patients to collect data. This is not an easy task, certainly not for a beginning researcher, or even for someone who has been doing research for a long time but is switching topics and entering a new field (e.g., a cardiologist and a surgical nurse who have done extensive epidemiological work using secondary data from large databases decide to take on a study about patient experience and quality of life requiring primary data collection tools and sampling strategies). It is possible that you are joining a team that has already selected the data collection tools. At whatever stage of the process you find yourself in, never trust a snappy survey title! If looking for measures, for example, on shared decision making, there are a few

options out there. Each one was constructed in a different way, was used initially with a different patient population and research question in mind by the authors. There are many steps that go into creating consistent and accurate measures. We will not go into the details of these steps here. We will, however, try to convince you to do three things – read the entire measurement tool before distributing it, think about how similar or different the context you plan to use it in is in comparison to the context in which it was developed, and do not change the survey in any way.

Look at the actual measurement tool – read each item. If you understand what you mean by your construct, for example, 'shared decision making' in the study you are about to conduct, read the items of the survey you have found and ask yourself if what is being asked represents what you want to know. This is the first step, and one that anyone can do, with or without measurement expertise.

It will also benefit you to have a sense or consult with a member of your team, to assess the extent to which the survey you have selected has demonstrated validity and reliability in contexts like the context where you plan to collect data. For example, a survey with demonstrated reliability and validity, created for adults in rural South American villages may not be generalizable to your intended sample of adolescents in a European urban setting.

The third and final word of advice on standardized survey tools is to NOT change the wording of any of the items. Standardized surveys are nice when they have demonstrated validity and reliability in contexts like our own because we can trust that the measure will capture the thing we want to know about. There is a science to this process of standardization. Order of items, the wording of items, and the associated response scale are all things that have been carefully thought about. If you change them, then you no longer can report that the tool you are using has been previously demonstrated to be valid and reliable.

Risks of Bias

The risk of bias is a technical term that simply means ways that we can blur the picture of the variables we are intending to represent numerically and in turn, sway the results of our statistical analyses. Data are simply numbers that represent variables or concepts that are important to us in the real world. It is easy to see that a number cannot ever perfectly represent "happiness", "shared decision-making" or "depression". When we have good measurement tools though, we can generate a score that is a good representation of these and so many other complicated constructs. At least good enough to be able to see patterns and relationships to other constructs that can contribute to our knowledge and inform the decisions we make. The goal is to paint the most realistic picture possible of the construct we aim to capture. Knowing the painting will never literally be the exact same as the thing being painted, we can accept that there will be some degree of error in each person's or each observation's score. What we do not want is something so abstract that we do not know if we are looking at a duck or a truck. There are things that we do as researchers and things that happen during study conduct that create risks of bias. Biases can be little, and they can be forgivable. Though blurry, we still see the painting is of a duck. We can also bias our data in extreme ways – ways so extreme that for either our entire sample or for subsets of it, we can no longer decipher what we are measuring at all. We encourage you to learn more about the risks of bias relevant to your area of research and to consider these carefully in relation to how they might impact the conclusions you draw from your research findings. Two examples of the many sources of bias a study can suffer from are performance bias and attrition bias.

Performance biases are the systematic differences experienced by participants in the study that are not relevant to the study. For example, if doing an experiment where you are assessing the impact of online patient education to inform healthcare decision making, you want to measure the effect that the intervention (new online source of useful information) has, and not anything else. If the participants who are randomly assigned to receive the intervention also get 45 minutes of healthcare provider time to talk about the site and have a personal conversation with a healthcare provider that otherwise (and for the control condition) would not have been experienced, then you are at great risk of perfor-

mance bias. You no longer know if any observed differences between your control and intervention are because of the online learning opportunity, the interaction with the healthcare provider, or most likely the combination of the two. The question you are asking yourself here, in the case of an experimental design is "Did people in both groups have exactly the same experiences, aside from the experimental intervention?" This goes the same for descriptive-comparative studies. It is important to engage participants and give experiences of study participants that are the same for the groups you are collecting data from. For example, if comparing prescription compliance behaviors of 30 – 40-year-olds and 70 – 80-year-olds, then you want to have all participants have the same study experience so that you do not add a performance bias from, perhaps, spending more time with the older group and giving them more attention and support to complete daily journals than the younger group.

Attrition bias is a perfect prelude to the next section on the need to account for missing data. Attrition bias addresses risk to your sampling strategy. If you could generate a careful random sample (or stratified random sample) from a population of interest, you will be highly motivated to maintain that random sample for the duration of your study. People might cease participation in your study, and these reasons will be systematic- for example, people with more stress in their lives or fewer resources might not be as able to keep coming to a laboratory to participate or might need to move and you might lose track of them. If those who stop participating are in any way systematically different than those who stay in the study, you have suffered from attrition bias. The representative sample you started with is no longer representative of the population. It is now representative of those in the population with the resources to participate in the study. In healthcare, particularly if there is an intervention being tested, we need to keep track of when and why people stop participating. If doing a drug study, people who have adverse events from the drug might be the ones to stop participating, and those who benefit from the drug are more likely to be the people who remain in the study over time. This is a fatal flaw for a study assessing the risks and benefits of a drug. The same goes for a patient-education study. If those who are just not that interested in learning more about their health stop participating and those who have a high degree of interest stay in, then the learning gains you might see through statistical analysis might have more to do with interest than your educational material. It is important to track the reasons people stop participating, and to be transparent in your reports about all of this information. A question you might ask yourself to assess attrition bias in a study where you are comparing groups (randomly assigned or naturally occurring) is: Did the same proportion of people stop participating across the groups, and for the same reasons?

Accounting for missing data

The final methodological bridge we need to cross to prepare you for understanding the nature of the data you are about to analyze is the need to understand exactly why there are blanks, or empty cells, in your data set. In an ideal world, if you have five variables in your study and 300 participants, you have data on all five variables for all 300 people. This is a complete data set. A complete dataset is a rare occurrence. There are two categories of reasons for missing data that are important for you to keep track of, systematic and random. Systematically missing data means there is a reason the data are missing. When this happens, you need to be able to report why the data are missing. An important step when creating a database is to have pre-set missing data codes so that you do not have any empty cells.

What is a missing data code you might ask? Great question! Before you analyze your data, there should not be any empty cells in your database. All data that are missing should be coded so that you can know why the data are missing, and report those reasons in your study reports. When reporting results of study findings, you must always pair your statistics (e.g. t-statistics and p-values) with sample size. Often we see that the sample size varies from one analysis to another within a study. This is because we rarely have a perfectly complete dataset. Explaining why the data are missing helps us understand the potential risks of bias, and the statistical power for each analysis reported. When deciding what your codes will be, choose numbers that are impossible for your dataset. For example, if -1 is an impossible num-

ber for all the data you will collect, then it is a good code. Disciplines, or even local research groups, create codes that they use consistently to represent missing data. For example, one of the authors of this textbook tends to collect survey and demographic data that would typically range in scores from 0 to 500. If this is the full range of possible values in the dataset, then the research group can consistently use the code 999 for randomly missing data, 888 for attrition due to moving away, 777 for attrition due to participant choice to stop participation for lack of interest. Death would be an extremely rare occurrence for this research group so they do not have a pre-set code for this reason for missing data. If it were to happen in a study, a code would be created.

We will provide two examples of systematic missing data. One reason for systematic missing data is that the variable was not applicable for some of the people in your study. For example, if you are looking at the length of hospital stay as a variable, that will only be relevant for patients who needed to be admitted to the hospital. If your dataset includes patients who may or may not have been admitted to hospital, then the sample for any analysis answering questions about the length of hospital stay is constricted, appropriately so, to those who stayed in the hospital. Another common example in healthcare where non-applicable systematic missing data are adverse events. If adverse events are part of your dataset, then of course you will only have data on adverse events for the few patients who experienced an adverse event. You need to provide a code for those who did not experience an adverse event to indicate that the data are not available because they do not exist. Another systematic missing data, which is quite different in its implications for your study is attrition. When a person starts a study and does not complete it, you might choose (with their informed consent) to keep the data that has been collected to date. When you do this, you need to explain why you no longer have data for them. For each reason a person stopped participating, you need to have a code to indicate why they stopped participating. Three common reasons in health research are death, moving away, and simply choosing to not continue participation (possibly the study was logistically inconvenient, taking too much time they did not want to spend on it, or travel to the lab was inconvenient).

The other kind of missing data will become apparent once you have coded all the data that you have an explanation for. This remaining missing data is random. Random missing data have no explanation for why they are missing, and it does not appear that there are systematic differences for specific groups in your sample. These missing data, if you only have a very small percentage, might be able to be statistically recovered. For small amounts of missing data, we can impute scores that are best guesses as to what those scores might have been, based on mean of nearby points, or multiple imputation based on variance and co-variances of nearby variables. Randomly missing data also need their own unique code so that you can tabulate how much you have, and perhaps, tell the program how to impute scores to replace these codes.

We hope that this section has highlighted for you some measurement and methodological issues that must be addressed to trust and use the data you analyze. If you are still not quite convinced, let us link these issues directly to what you are about to learn about by introducing the concept of the error term. For example, in statistical analyses the bigger the error term, the less likely an effect will be seen even though an effect might actually exist.

A Guiding Principle

An essential guiding principle is an acronym GIGO – garbage in, garbage out. This is especially important when selecting measurement tools, inheriting a dataset with variable names you want to trust at face value and wishing you could ignore how many unexplained empty cells you see in a dataset. If you do not have quality data, there is no way to end with quality answers to your important research questions.

2. Measurement Concepts

Measurement

Measurement, or the ability to assign value to pieces of information and allow comparisons between groups, is an essential aspect of applied health research and clinical practice. Although we might not always think deeply about what we are measuring or how we are measuring it in our daily lives, we are always collecting data for comparisons and decisions that inform our activities of daily living. For example, at the grocery store, your decision to buy one type of apple over another may be informed by the price of each (a measure of their economic value in the market); similarly, if you drove to the market you most likely monitored your speed (a measure of velocity = distance/time) to ensure your safety and the safety of others. In this chapter, we are going to delve more deeply into different types of quantitative data measurements. As applied health researchers we may use measurement to inform clinical treatment decisions or more broadly to make health policy decisions. Therefore, it is crucial that we understand exactly what we are measuring and how to handle different types of data collected from those measurements appropriately. I think we can all agree that these decisions are much more important than deciding between two apples at the store, or how fast we should drive to avoid a speeding ticket and thus, deserve our thoughtful attention. Let's begin.

The term measure can be described as a verb (an action word) or a noun (a naming word).

As a verb, the term *measure* refers to the action of evaluating an entity and in quantitative research assigning it a value; as in, "I am going to *measure* the number of individuals who use aspirin daily". In this example, the term measure is used to describe the action of quantifying the concept of interest – the number of individuals who use aspirin daily. In this context, we count the number of individuals who use aspirin daily within a given sample of participants. In clinical practice, health care professionals assess and measure information about patients frequently, including age, height, weight, heart rate, blood pressure, lung volume, as well as depressive symptoms, or anxiety levels.

Measurement can also involve asking participants to answer questions about themselves (e.g., hair colour, postal code, income) or completing a standardized assessment tool. For example, we might want to see how exercise self-efficacy, or the belief that you can complete an activity, influences junior high students' participation in gym class. Since exercise self-efficacy is a complex concept, instead of counting, we would measure this concept by asking participants to rate their agreement with a series of pre-planned statements and then use this information create a numeric score that represents their perceived level of self-efficacy. Examples of other standardized assessment tools include the Morse Falls Scale which is commonly used to measure a persons' risk of falling, and the Braden Scale for Predicting Pressure Ulcer Risk, which – you guessed it – helps measure a persons' risk of developing a pressure ulcer.

As a noun, the term *measure* can also refer to the instrument or method used to assess the quantity of an attribute of something or someone. For example, the count of the number of individuals who use aspirin daily is the measure or method used to quantify the number of individuals who use this medication every day. Likewise, the General Health Questionnaire is a measure of participants' self-reported overall mental and physical health.

In research, the information obtained during the measurement process is referred to as data. The word data is plural for the term datum, whereby datum refers to a single value. Although data are often presented as numbers, data are not limited to numerical values. For example, data can be the verbal responses to an interview. Similarly, data can be raw materials, artifacts, diagrams, or specimens. However, in this book, we will focus on numerical data which can be analyzed statistically.

When we begin our research we need to consider the type of data that we will be collecting because this information will inform our selection of the appropriate statistical tests for data analysis. Numerical data are often classified into four possible categories: nominal, ordinal, interval, or ratio. In the following sections, each type of data will be described in detail.

Nominal Data

The term *nominal* refers to the first level of measurement. This term is used to describe data that have no intrinsic hierarchy, meaning that there is no underlying number line upon which the measurements are based. Nominal data include those variables that have numeric values but the numbers are not in a logical or meaningful order, as well as those variables that are simply assigned labels as part of the strategy to analyze data. Nominal variables are categorical and must not be used as if they are considered to be on a continuum (see ordinal data). Note that you cannot do math on nominal variables because it doesn't make sense.

Examples of nominal measurements that are numeric include telephone numbers, or license plate numbers. Telephone numbers are randomly assigned to regions within a geographical area. A comparison of the numbers 688-5550 and 978-2345 does not provide us with any information about our participants except that they come from two different telephone regions.

Examples of nominal measurements that are assigned labels include sex, hair colour, or eye colour. For example, when we write a coding strategy in a computer program to analyze data, we often re-code the measure of sex as 1= male and 2= female, 3=other. In this example sex is considered as a nominal variable because the values held by this variable (i.e. 1, 2, and 3) do not have an intrinsic hierarchy with respect to the entities they represent (i.e. males, females, other). In this example, even though other= 3, female = 2 and male = 1, it doesn't make sense to say that the category of other are 3 times as much as males. The labels are entirely arbitrary so you could have coded them the other way around (other= 1 male =2 and female =3) or using other numbers.

Ordinal Data

The term *ordinal* refers to the second level of measurement. This term is used to describe measurements that do not have an intrinsic numerical hierarchy (as defined previously) but do have a distinct order. The importance of the order of ordinal measures is based on a hierarchy established by the researcher. In ordinal data, values represent agreement with subjective anchors. For example, Likert scale response options on questionnaires are often used to measure variables that provide perceptions of constructs such as health, burnout, stress, coping, anxiety, or ratings of emotions. In survey responses, the researcher sets the polarity and order. The values are often discrete, as shown in the following example:

Strongly Disagree 1 –– 2 –– 3 –– 4 –– 5 Strongly Agree

The polarity of these responses (negative to positive) can be reversed (positive to negative) and so respondents need to be vigilant to the meaning of the poles selected for the scale.

Strongly Agree 1 –– 2 –– 3 –– 4 –– 5 Strongly Disagree

One example that illustrates the subjective nature of ordinal scales is the Rating of Perceived Exertion (RPE) scale (Borg, 1982). This scale is commonly used during exercise to determine participants' feelings of exercise intensity and was developed by matching heart rate to perceptions of exercise intensity. Originally the scale ranged from 6 (~ heart rate of 60 beats/min) to 20 (~ heart rate of 200 beats/min) with scores of 6 to 8 corresponding with feelings of "very, very light" intensity and scores of 19 to 20 corresponding with feelings of "very, very hard" intensity (Table X).

Table 2.1. The original rating of perceived exertion (PER) scale (Borg, 1982)

Score	Perceived Exertion
6	
7	Very, very light
8	
9	Very light
10	
11	Fairly light
12	
13	Somewhat hard
14	
15	Hard
16	
17	Very hard
18	
19	Very, very hard
20	

Unfortunately, a scale from 6-20 is not intuitive to most people, making it difficult for people to interpret, especially if they did not know the typical range for an individual's heart rate. Eventually, the RPE scale was changed to 0-10 with 0 being "nothing at all" to 10 being "very, very strong" because it was more meaningful to people and easier to use (Table X). It is noteworthy to mention that the inclusion of the rating estimate: 0.5, was added to convert this ordinal scale to a ratio scale, discussed below.

Table 2.2: The modified RPE scale (Borg, 1982)

Score	Perceived Exertion
0	Nothing at all
0.5	Very, very weak
1	Very weak
2	Weak
3	Moderate
4	Somewhat strong
5	Strong
6	
7	Very strong
8	
9	
10	Very, very strong
.	Maximal

In some applications, a dot is added to the end of the scale so that an individual can provide a rating of their perceived sensation to be over 10. For example, when running on a treadmill to exhaustion, some exercise physiology labs will include the phrase "Saw God!" Regardless, although including an extra indicator at the extreme margin of a scale may seem a bit odd, because it is an ordinal scale the authors are free to assign values to subjective ratings however they wish.

Ordinal-scaled scores can also provide data for ranking, in which the researcher establishes the order of the ranking pattern, usually from highest to lowest or vice versa. An example of ranked ordinal scores is shown in the following table (Table X). Notice that in this example the identification label (ID) assigned to each score is maintained after the original series of scores is ordered by rank (also known as rank-ordered). In this example, notice that the researcher assigned the lowest score to a rank of "1", indicating to the reader that the lower score is perceived to be better (a decision of the researcher). It is important to recognize that while these data represent an arbitrary set of scores, data that are at the level of ordinal, interval and ratio data can be ranked.

Original Scores		**Scores After Ranking**		
ID	Score	ID	Score	Rank
Respondent A	77	Respondent C	68	1
Respondent B	76	Respondent D	74	2
Respondent C	68	Respondent B	76	3
Respondent D	74	Respondent A	77	4
Respondent E	78	Respondent E	78	5

In reviewing these data by rank alone we do not get a sense of the difference in scores between ranks. The assignment of the rank is not in and of itself at an interval level data point because the gap between the ranks is not necessarily consistent or equal.

We often see the ranking of ordinal measures in sporting events where teams or players are evaluated based on their win to loss record or based on other criteria such as percentage goals as performance indicators. This way of presenting the information allows an observer to quickly determine an order to the ranking of the participants (teams or individuals). For example, below in Table X presents the top 10 men's basketball teams from the 2017 NCAA post-season rankings (ESPN, 2017). Notice that teams are ranked by total points and the difference in points between ranks is not the same. For example, North Carolina is ranked #1 and has 31 points more than Gonzaga (rank #2), but Gonzaga has 49 points more than the #3 ranked team, Oregon. This is why you cannot treat ranked data the same way as interval or ratio data – ranks simply tell us the relative position of each value within the data set. Ranks do not provide the absolute scores, as in this case where the total number of points, represents ratio level data.

Table 2.4: The Top 10 men's basketball teams in the NCAA in 2017 according to post-season rankings

Rank	Team	Record	Points
1	North Carolina	33-7	775
2	Gonzaga	37-2	744
3	Oregon	33-6	695
4	Kansas	31-5	653
5	Kentucky	32-6	627
6	South Carolina	26-11	561
7	Arizona	32-5	548
8	Villanova	32-4	498
9	UCLA	31-5	492
10	Florida	27-9	468

Interval Data

The term *interval* refers to the third level of measurement. Data at the interval level of measurement use a constant unit of measurement (i.e., the distance between numbers on the scale represent equal changes in the item being measured) on an underlying real number line. Therefore, measurements made on an interval scale have a distinct order in which the importance of the direction or the polarity of the order is established, previously. However, the "0" value of interval data is subjective and is set by the users of the measurement tool and does not reflect the absence of the characteristic being measured. For example, when measuring temperature in degrees Celsius, 0 represents the freezing point of water. This does not represent the complete absence of heat altogether (absolute zero). Similarly, 100 degrees is not twice as hot as 50 degrees. Another example of an interval scale would be time of day on a clock. Although this measurement of time is meaningful and helps us stay organized, intervals of time only have meaning relative to other times on that scale. It is illogical to say that 12 o'clock is twice as much as 6 o'clock, though we could say 60 minutes is double 30. Again on this time scale, zero o'clock signifies midnight, not the absence of time.

Ratio Data

The term *ratio* refers to the fourth level of measurement. The ratio level of measurement also has an intrinsic hierarchy and is based on the real number line. Ratio measures have a distinct order, a distinct direction, and a distinct polarity. Yet, the distinguishing characteristic of the ratio level of measurement is the presence of a real 0 which indicates an absolute absence of the item being measured. In health research you will find many examples of ratio level measures. often collecting information on height, weight, blood pressure, heart rate, and age, or you may record income, the number of days employed, or the number of times a times a participant experienced an event. For example, when interpreting age as a ratio measure, 0 is meaningful and may indicate that an individual had just been born (current date

subtracting the birth date, on the day an infant was born) and someone who is 30 years old is twice as old as someone who is 15 years old. Similarly, if you were studying the number of ear infections children experience between birth and two years of age, if a study participant recorded 0 ear infections during the study period, this would be meaningful data. When you were interpreting your data, the participant who experienced 2 ear infections would have twice as many as someone who only recorded 1 and so forth.

If you are interested in reading more about levels of measurement the following book is an excellent resource: Pedhazur, E. J., & Pedhazur Schmelkin, L. (1991). Measurement and Scientific Inquiry. In Pedhazur & Pedhazur Schmelkin (Eds.), Measurement, Design, and Analysis: An integrated approach, pp. 15-29. New York: Psychology Press.

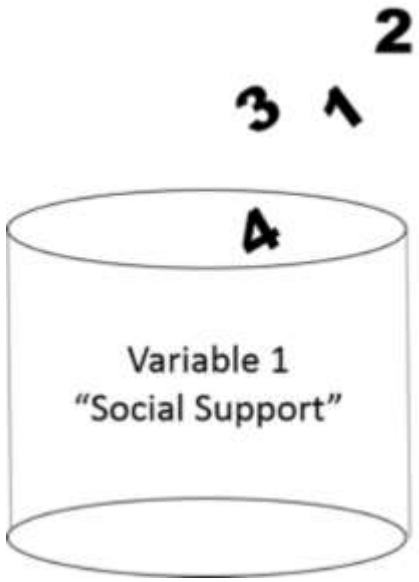
---

[1] National Academies of Sciences, Engineering, and Medicine. 2019. *Implementing strategies to enhance public health surveillance of physical activity in the United States*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/25444.

3. All About Variables

Defining a Variable

As noted previously, variables are central to the application of statistics in research. That is, variables hold the data to which the statistical tests and methods are applied. **Variables** are the containers that we use to collect, organize and process data.



Representation of the variable "Social Support" as a container that holds values.

The values held by the container (variable) represent the data that was reported for the measure of interest in the sample. Before we move on it is essential to understand both the conceptual definition and the operational definition of each variable in your dataset (or planned study if you are collecting your own data).

A **conceptual definition** refers to the abstract **idea or theoretical meaning** of interest (Polit & Beck, 2012). Many concepts are defined in multiple ways so it is important to clearly identify which conceptual definition you are using in your research. For example, the concept of social support can be defined as perceived support, support actually received, emotional support, informational support, and social connections. Though similar, each of these conceptual definitions are distinct. Be careful to ensure that you know exactly the concepts of interest within your research.

In contrast, an **operational definition** refers to **how** you are going to measure the concept of interest (Polit & Beck, 2012). Using our example above, if we decide that we are interested in perceived social support, then we would want to use an instrument that measures perceived social support specifically. We would not want to use an instrument that measures social connections or received support, or something else.

Another example that illustrates the difference and links between a conceptual and operational definition is applying them to the concept of body height. The conceptual definition of height is how tall your body is. The operational definition of height is how long your body is from your feet to the top of your head in cm or inches. Measuring the length of a person's body from head to toe creates a value that represents how tall that person is. That value can then be put into your height container or variable in your dataset for analysis.

Types of Variables

The values held by variable containers can be either fixed or they can change. That is, they can vary (that's right, variables (noun) can be variable (adjective)!). **A random variable** is a variable that can take on any value.

There are two types of random variables:

Discrete random variables – these have finite values, typically based on the whole number line (1, 2, 3…etc.).

Discrete random variables can represent counts or categories. Some examples of discrete random variables include classification of sex as male, female, or other; the number of symptoms reported; the number of guests at a resort; the number of airlines using a specific terminal; the number of cavities for a patient, an individual's socio-economic status (SES); disease state – classification as a case or non-case; and age category in years.

Continuous random variables – these can have an infinite number of values since continuous random variables use the continuous measurement scale based on the REAL number line to record outcomes (0, 1, 1.1, 1.2, etc.).

Examples of continuous random variables include time, distance, velocity (which is computed from distance : time and reported in km/hr), body weight (reported in kg), oxygen consumption (reported as ml/kg•min-1), and pharmaceutical prescription (sometimes denoted as mg per 100 grams body weight). Notice in the description of discrete variables we refer to the measurement unit as a whole number or as an integer value (no decimals) while the reference to continuous variables often describe rates or ratios where the measurement can include a decimal value and is often described as a fraction.

Creating New Categorical Variables

After determining whether the measurement is either a discrete or continuous random variable we can group our data to produce new variables that can also take on specific or discrete boundaries. The result is a set of variables, which produce categories of items. The term for a variable, which is used to organize data, is a "categorical variable". The categorical variables are used to group results or measurements into sub-classifications so that specific statistical analyses can be performed on each of the sub-classifications within the categorical variable.

For example, we could use age measured as a continuous variable to create age categories. In Figure 1.6 you can see how this might be done. Note that you always have less information in the categorical variable than the continuous variable. Keep this in mind. If you ask participants to select their age category that is the maximum amount of information you will have about their age. If instead, you ask them for their exact age you can still create categories later if desired.

Age	Age Group
25	1 (age 20-29)
36	2 (age 30-39)
45	3 (age 40-49)
32	3 (age 30-39)
26	2 (age 20-29)

Example of creating a categorical variable from a continuous variable

Independent versus Dependent Variables

Research studies can take on several different design types. Some are purely descriptive, some are correlational and some are comparative. However, many research studies are designed to prove cause and effect relationships between variables.

Cause (Input) -> Effect (Outcome)

Input variables are often referred to as *independent* variables. In experimental research, these can be controlled and are typically categorical or organizing variables. Often the independent variables are called the intervention or predictor variables. Outcome variables are often called *dependent* variables because they depend on the influence of the independent variables. Throughout this text we will explore the application of statistical analyses to variables of different types within research designs that collect measures with different considerations for the role of variables.

Here is what we covered in this section

Here you were introduced to:

- The different types of research variables that are essential to creating a statistical analysis.
- The concepts of thinking statistically and creating a process before embarking on the research.

In several examples, we will generate data for our examples using random number generators. The SAS program provides a powerful platform for creating computer simulations that can demonstrate potential outcomes.

4. Thinking Statistically About Your Research

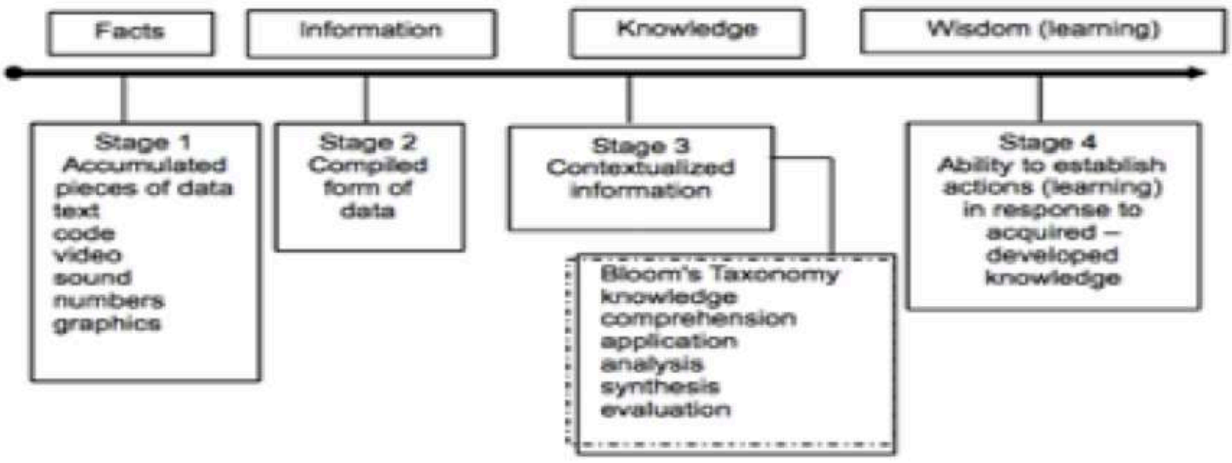
Essential Design Considerations

Thinking statistically means being able to identify real-life examples of data and consider the following questions: Data are everywhere, but do we recognize data? How do we give meaning to the data around us? How do we provide context for data? How do we make data relevant to us or to others?

While the focus of this book is on analyzing quantitative data, we also need to consider other forms of data that help us to recognize behaviours, attitudes, outcomes, and patterns that can lead to predictions of future outcomes or behaviours. Recognizing data, thinking about data patterns and the connectivity of information, and providing structure to data sources are important exercises. These are the kinds of exercises that underlie the questions that give value to statistical enquiry. These are the thought processes that can help us generate new ways of thinking and new applications from existing data.

In an earlier study**[1]**, we considered a famous quote by T.S. Eliot from 1962: "*Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?*" This is the fundamental construct of a constructivist paradigm approach to developing knowledge and action. In our mapping exercise, we explained that using a constructivist paradigm we begin with raw data, contextualize the data, apply the data to a set of rules, a rubric, a plan, an algorithm, a taxonomy, and eventually create a product, a process, a policy.

Figure 4.1 of the Constructivist Perspective for the Flow of Facts to Learning



Schema of the Constructivist Pedagogy

A constructivist paradigm approach was used to develop the material presented in the following chapters. Data are contextualized – referenced to various scenarios and processed using SAS programming statements. This process enables us as researchers to make sense of the data, thereby forming the evidence for knowledge synthesis. This new knowledge can then be used in specific actions like policy development or policy evaluation.

The Research Process

We conduct research for a variety of reasons such as:

- to describe a phenomenon
- to evaluate outcomes
- to evaluate diagnostic testing
- to evaluate screening procedures
- to determine cause-effect relationships
- to describe a disease, prevalence and natural history
- to determine a prognosis
- to review existing practice
- to evaluate the effectiveness of interventions.

Phase 1: Conceptual Work

First, we do the conceptual work needed to develop a research problem and identify potential approaches to addressing it or understanding it better. We review the related literature to support our rationale for conducting the study, establish a clear statement of the problem, and identify a target population from which we will select an accessible sample.

In this phase we also identify and clarify each variable in our study, paying close attention to their conceptual definitions. This helps ensure that we select appropriate measures that actually measure the concepts we want to measure (you'd be surprised how often there is a mismatch between what people say they are measuring and what they *actually* measure!).

Last, but certainly not least, in the conceptual work phase we develop a logical, theory-informed model of how we think variables are related to one another. The relationships that we propose between variables are called hypotheses and will be tested using statistics after we collect our data. It is important to note that there is not just one way to approach a research problem and sometimes several possible theories could be used. It is up to the researcher to think critically, evaluate past evidence, and decide what makes the most sense.

For example, we might be interested in helping people stop smoking.

To do this we could develop an intervention based on Bandura's (YEAR) self-efficacy theory. According to this theory, people's behavior is influenced by their confidence that they will successfully be able to engage in that behavior and their outcome expectations (i.e., what they think will happen if they engage in that behaviour). Therefore we would plan an intervention that will improve self-efficacy to quit smoking and provide information about the benefits of quitting smoking. We would later use statistics to test whether our intervention improved participant's self-efficacy to quit smoking, and perhaps, whether increased self-efficacy, in turn, led to actual changes in behavior (i.e., did they stop smoking?).

Alternatively, we could evaluate the effectiveness of a drug or health product such as nicotine gum on smoking cessation. In this case, the theory to inform the study might be more physiological, centring on the effects of nicotine on the body as well as the benefits of not inhaling cigarette smoke.

These are just two ways that you could address the research problem (how do we help people stop smoking?) and there are many more! Both are equally valid but they take different angles. The thing you want avoid though it collecting data BEFORE you do this crucial conceptual work. Even if you are working with secondary data (i.e., data that someone else collected), the last thing you want to do is "go fishing" and simply run statistics until you find something significant.

Phase 2: Study Design & Planning

After you've decided what the research problem is, the variables and relationships of interest, and your hypotheses, you design the study and create a plan to conduct the research. This involves selecting the appropriate methods needed to answer our research questions outlined in Phase 1. We decide which study design makes the most sense to test our hypotheses of interest. For example, should the study be cross-sectional or longitudinal? Are you going to compare groups? Are you doing an observational study or an intervention? There are many choices here and they will influence the statistics that you are able to do so choose wisely. For example, if you want to compare groups using a t-test, you need to have data for the same variables from the two different groups you want to compare. Seems obvious, right?

The point is that your statistical analysis actually begins well before you are anywhere close to writing your first SAS command. You need to know what hypotheses you are going to be testing in order to know what statistical tests you will need to run. And, you need to know what statistical tests you plan to run in order to calculate your anticipated effect size (which is always specific to a particular statistical test) and sample size which is a vital step for planning your recruitment and data collection plan.

Creating a data analysis plan is an integral part of creating a well-designed research study. It should describe:

- Your plan for assessing and dealing with missing data
- What statistical tests you will be using to evaluate the reliability and validity of your measures
- What statistical tests you will be using to test your hypotheses (including post-hoc tests, if applicable) and how you will examine the underlying assumptions of those tests
- What software program you are using (SAS)
- The significance (alpha) level used (or model fit statistics if appropriate)

Phase 3: Research Implementation

After you have developed your research plan and obtain ethics approval, it is finally time to do the study! This is one of the most exciting parts of research and where all of your hard work and energy spent in the conceptual and planning phases pay off.

If you have time and the resources to do so, it is usually a good idea to pilot test your research protocol on a small group of people before you roll it out on a large scale. This allows you to work out any potential kinks, catch typos, and assess the feasibility of what you are planning to do. In the long run, this can save you a lot of time and energy and will help ensure that you get the data that you want.

Phase 4: Analysis & Interpretation

Once you have collected your data, you can move on to the data analysis phase. This is the central focus of this book but hopefully, you can see why data analysis is not a stand-alone part of doing research. All phases of the research process work together and quite frankly, just because you find significant results using a statistical test doesn't mean that they are important or meaningful.

In the data analysis phase, you analyze your data using the plan that you developed in phase 2. However, it is often the case that you need to make decisions to deal with the imperfections of real data. More often than not you will have data with missing values or variables that don't meet the underlying assumptions of the test you planned to do. For this reason, it can be helpful to keep a research journal or log outlining all of your data analysis and findings and the decisions that you make along the way so that you remember why you changed your plan and have evidence to support

your choices. If you are a novice researcher it is also a good idea to consult with your supervisor or a more experienced researcher with statistical expertise. Being well-organized and being able to show them exactly what you did and why will make it easier for them to help you.

After you finalize your results, you need to report them systematically and can include graphs and tables to summarize essential information. Interpretation of the results involves deciding whether or not they support your hypotheses and situating your findings within the current evidence relevant to your research problem. This includes discussing your results in light of previous research and with regard to similarities or differences to previous research and then conclusions are stated about the knowledge gained by conducting this research.

Phase 5: Dissemination

The final phase of any research project is the dissemination of the findings to stakeholders, which simply means people who have a "stake" or interest in the results. For example, your research on smoking cessation would affect people who smoke cigarettes, their loved ones, the health care system, store owners of stores that sell cigarettes, tobacco companies, and tobacco farmers, to name a few. Strategies and media used to disseminate new knowledge are diverse and can range from more traditional approaches such as conference presentations and peer-reviewed journal articles to more modern ones like infographics tailored to social media platforms or webinars with stakeholders.

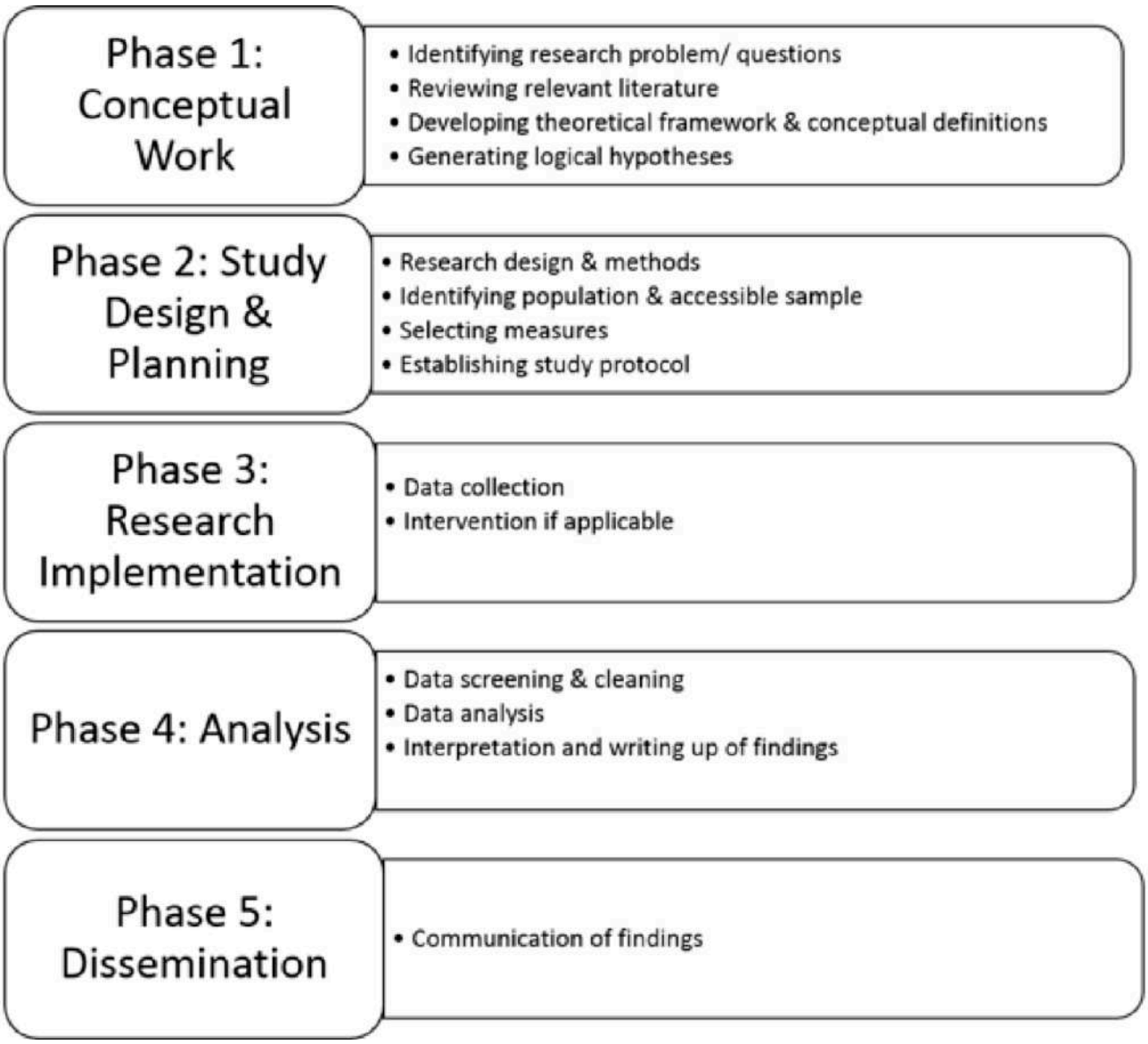


Figure of The Research Process for a Quantitative Research Project

[1] *Montelpare, W.J., and Williams, A.M., (2000) Web-Based Learning: Challenges in using the internet in the undergraduate curriculum. Education Information Technology, Volume 5(2), pp. 85-101.*

5. The Hierarchy of Evidence

The hierarchy of evidence provides a useful framework for understanding different kinds of quantitative research designs. As shown in Figure 2.1, studies at the base of the pyramid involving laboratory and animal research are at the lowest level of evidence because they tend to be focused on understanding how things work at the cellular level and it is difficult to establish a direct link between the research findings and implications for practice.

This type of research is still valuable because it provides the researcher with a very high level of control which allows them to study things that they can't do in humans. For example, you could breed genetically modified mice and compare them to regular mice in order to examine the influence of specific genes on behavior. Obviously this type of experiment would be unethical to do with humans but it can provide initial evidence to help us better understand phenomena (in this case, the influence of genes on behavior), intervention, or drug.

The next level includes research with no design and include case reports or case series reports that are commonly used in novel or rare situations (for example, a patient with a rare disease). Expert opinions, narratives, and editorials also fall into this category because they rely on an individual's expertise, knowledge, and experience which is not necessarily objective.

Above this are retrospective observational studies such as case-control studies or chart reviews that seek to find patterns in data that has already been collected. One downside of this type of research is that the researcher has no control over the variables that were collected or the information that is available.

Next in the hierarchy are prospective observational studies which include cohort studies as well as non-experimental research designs such as surveys. Here the researcher does have control over what variables are measured as well as how and when they are measured. If done well, this approach can strengthen the findings because it provides the researcher with the opportunity to control for confounding variables and bias, take measures to improve response rates, and select their sample.

Randomized controlled trials (RCTs) are often hailed as "the gold standard" for quantitative research studies in health care because they allow the researcher to control the experiment and isolate the effect of an intervention by comparing it to a control group. However, the inclusion and exclusion criteria for participation can be quite strict and the high level of control is not consistent with real-world conditions, which can reduce the generalizability of findings to the population of interest. Pragmatic RCTs (PRCTs) have begun to gain more popularity for this reason. The goal of a PRCT is to keep the treatment that the control group receives consistent with usual care and the treatment that the intervention group receives consistent with what is practical in the context of real life. While PRCTs don't provide the same degree of control and standardization as an RCT, the idea is that they provide more realistic evidence about how effective an intervention will be in real life.

Meta-analysis and systematic reviews come next on the hierarchy. The main benefit of systematic reviews and meta-analyses is that they include findings from a number of different studies, and thus, provide more robust evidence about the phenomenon of interest. Whenever possible, this type of evidence should be used to inform decisions about health policy and practice (rather than that from a single study).

It is important to note here that meta-analysis generally occurs as part of a systematic review and combining the data from several studies is not always appropriate. If the designs, methods, and/or measures used in different studies vary considerably then the researcher should not combine the data and analyze it as a group. It is also important not to include multiple studies that use the same dataset because the same sample gets used more than once which will skew the results.

Lastly, at the top of the hierarchy of evidence are clinical practice guidelines. These are at the very top because they are created by a team or panel of experts using a very rigorous process and include a variety of evidence ranging from quantitative and qualitative research studies, white papers and grey literature. Clinical practice guidelines also examine the quality of the evidence and interpret it in order to provide clear recommendations for practice (and often, research and policy as well).

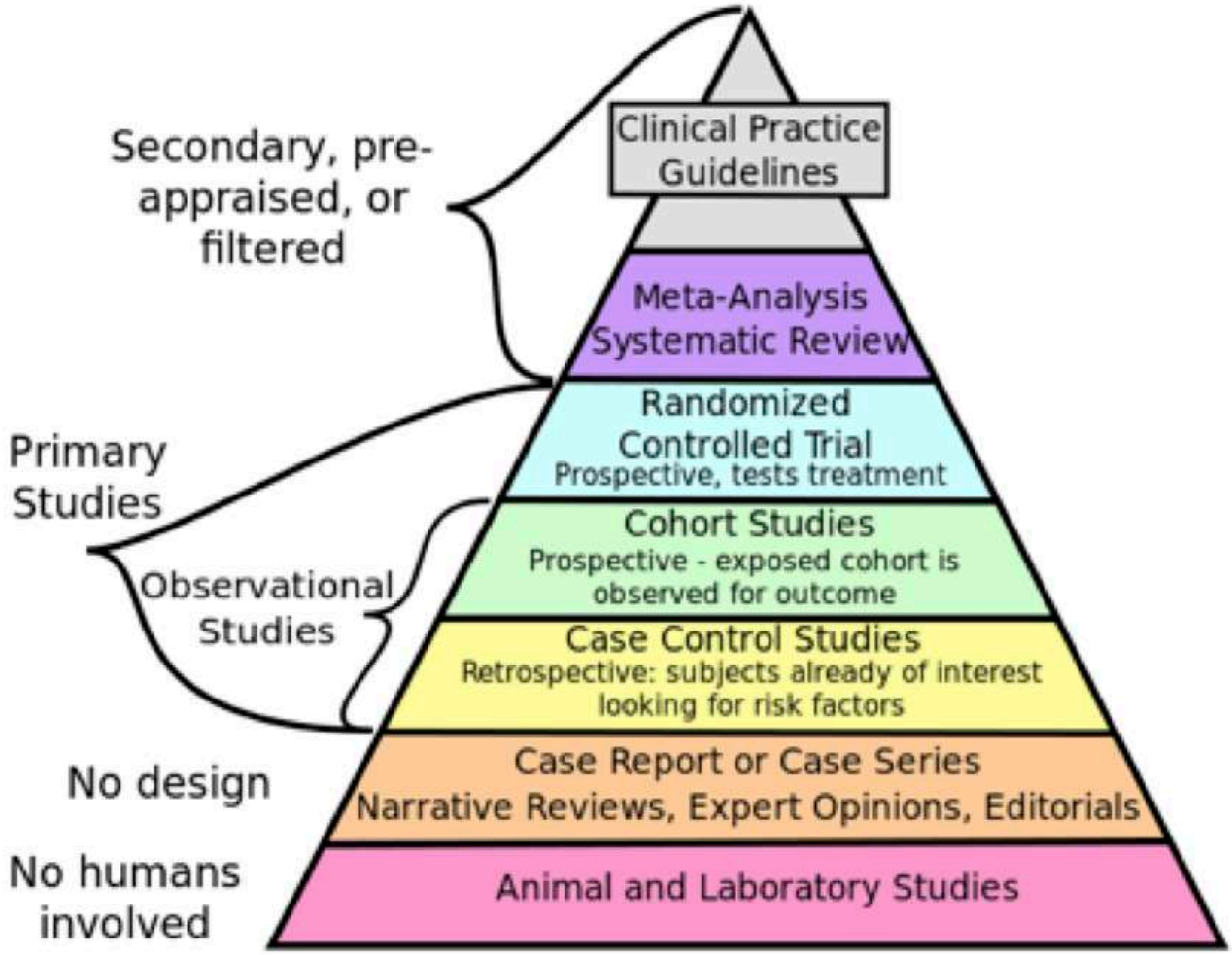


Figure of The hierarchy of evidence (image available to use as per Creative Commons license – https://commons.wikimedia.org/wiki/File:Research_design_and_evidence.svg)

6. Types of Research Designs

Research Designs

While we discussed the hierarchy of evidence previously, it is important to note that different research designs are appropriate for different situations and research questions. In other words, each design type is fit for a specific purpose. For example, a case report might be the best choice for examining a new condition that has not been seen before, while a randomized controlled trial would not be appropriate. In this chapter we will review a number of research designs that fall into two broad categories: 1) Observational Research Designs, and 2) Experimental Research Designs.

Common research designs include:

- Case report or case series
- Ecologic
- Cross-sectional
- Case-control
- Longitudinal
- Cohort
- Randomized controlled trials

Observational Research Designs

Observational research designs are used by researchers to draw inferences about a sample or population without intervention by the researcher or research team- exposure to the independent variable, such as a treatment for a disease, is not influenced by the study. Observational studies provide data that allow the researcher to conduct descriptive and analytic research. *Descriptive observational* studies typically focus on the person, place or time of an event of interest and are useful to help identify patterns and generate hypotheses, but cannot test them. For example, you may be interested in describing the incidence of colon cancer in men and women living in Prince Edward Island, Canada between 2000 and 2020. In an *analytic observational study*, the researcher goes a step further to include comparisons groups and test hypotheses. For example, in your previous study you may have noted the pattern of colon cancer diagnoses appears different between men and women. In your analytic study, you may then compare the risk of colon cancer between the two groups. Observational studies allow you can find out a lot about the current situation and examine relationships between variables but can be limited in their understanding of *causality*, or attributing cause and effect between two variables.

Case Reports and Case Series

Case reports in health research are generally defined as published reports of an individual patient or event or small group of patients or events that demonstrates unique characteristics of interest. Typically, the reporter uses a comprehensive and detailed assessment of the individual to describe features that are atypical of normal observations, such as combinations of presenting signs and symptoms, disease sequelae or trajectory, or unexpected outcomes. For example, a case report in the February 27 2014 issue of the *New England Journal of Medicine* described the a four month old infant who required surgery for a brain tumour, which contained multiple fully formed teeth.

Case series can be described simply as a collection of individuals or events that are used to describe aspects of a disease, treatment or diagnostic procedure, also known as a series of case reports. Typically a practitioner, such as a clinician, researcher, healthcare provider, and social worker, among others, using a standardized format, draws the cases from an accumulation of documented case reports. A major condition in forming a case series is that although collected as independent units, to be eligible as a case within the series, each individual selected for the case series demonstrates common characteristics of interest. For example, a seminal case series published in the *Lancet* in September 1981 described eight men with Kaposi's sarcoma and hypothesized a link between sexually transmitted infections and the development of these lesions. It has since been established that Kaposi's sarcoma, a cancer, is caused by the human herpesvirus 8 (HHV-8) and is thought to spread through blood and saliva during sex or from maternal-to-infant transmission during birth.

Case reports and case series can identify new trends or diseases, detect drug side effects or new uses, share experiences with rare events, or describe uncommon disease phenomenon. These studies are important as they can be where new issues or ideas emerge and generate hypotheses for further exploration. However, because they are single events or a small series of events, they may not be generalizable, may incorrectly attribute cause and effect and as a result be misleading, and are not based on rigorous research methods. Rather, the case report, as a unit, provides a reference to a specific observation that alerts a community of practice and which can then be documented for future study and comparison. The information from the case series can be used to build a knowledge base and develop hypotheses tested in more rigorous study designs.

Cross-sectional Research Studies

Before we begin discussing cross-sectional studies, let's define the term prevalence. Prevalence is described as the frequency of a condition of interest measured at an instantaneous point in time. The prevalence score, or prevalence estimate, is sometimes referred to as the *point-prevalence rate* because it provides an estimate of the number of cases within a sample taken at a *point in time*. The prevalence estimate takes on a rate value when compared to the total number of individuals at risk within the sample.

Given this definition of prevalence, we can say that a cross-sectional research design enables the study of prevalence. The prevalence rate is typically used in cross-sectional studies where the measurements are made only once across a selected sample. Prevalence rates can be considered generalizable to a larger population if the sample upon which the estimate is made is truly representative of the population from which it was drawn.

Figure 2.2 below depicts the application of a cross sectional research design. Notice a sample is first identified from an existing population and then from the sample (k) groups are drawn. In this image 2 groups were drawn at random from the sample, each group having the same number of individuals, with similar characteristics.

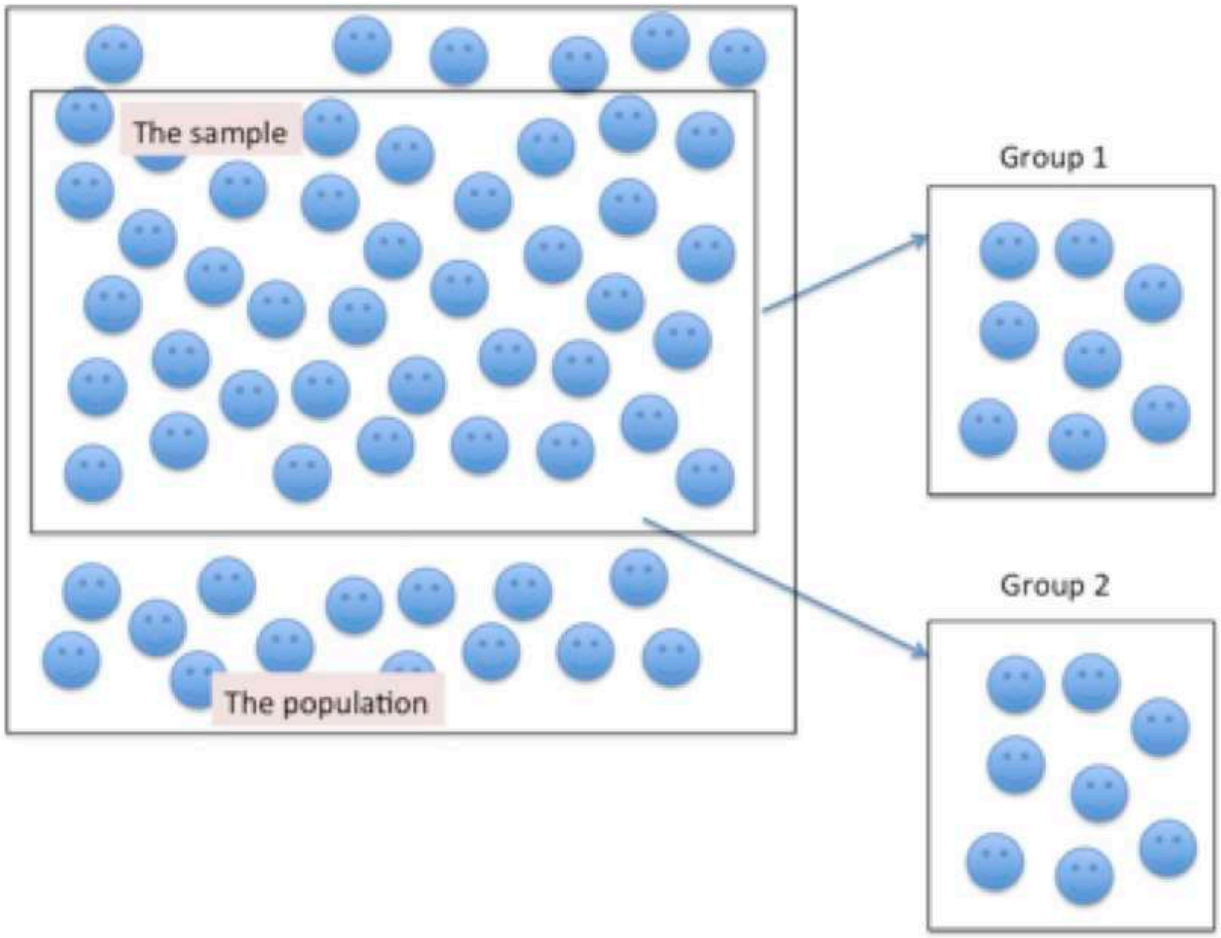


Figure 6.1. Sampling Participants for a Cross-sectional Research Design

Characteristics of the cross sectional research design

A positive characteristic of the cross sectional study design is that because data collection occurs only during a single bout (one time sampling), there are fewer costs than in a longitudinal research study. However, a negative consequence of one time sampling is that some individuals that may influence the outcome for the study can be missed during the sampling phase. Likewise, because there is no follow-up of the sample, and often no historical information preceding the sample, very little can be inferred about causal relationships from the sample. The cross sectional study design provides a snapshot of sample characteristics at a specific **point in time** – when the data were collected.

Case-Control Research Studies

Case-control research studies are typically retrospective studies as they are used to compare groups of individuals that were identified as having a condition of interest – the cases – against a similarly matched group of individuals that do not have the condition of interest – the controls. Cases are identified from the sampling frame, by the investigator, for individuals that demonstrate the specific characteristics of the condition of interest. Controls – individuals that do not demonstrate the characteristics of the condition of interest– are matched to the case subjects on the basis of relevant

measures that help to identify (or exclude) causes of the condition of interest. In conducting the case-control research study, the investigator identifies a set of suspected causes and then searches backward in time to compare differences in the suspected causal variables between cases and controls to help explain why cases differ from matched controls.

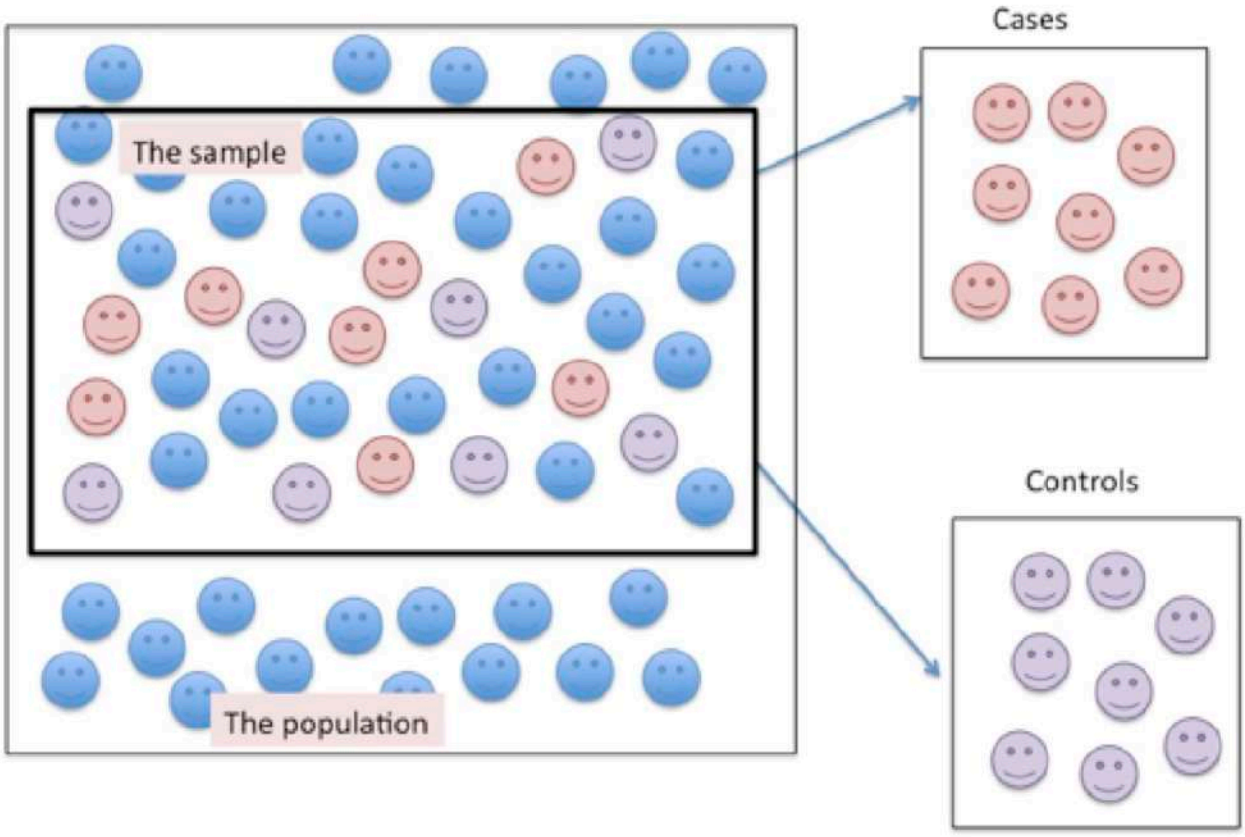


Figure 6.2. Sampling Participants for a Case-control Research Design

As shown in Figure 6.1 above, the cases are similar to the controls but differ only on the condition of interest. In both the case and the control groups the researcher is not trying to make an inference to the population in general, from which the sample was drawn, but is demonstrating the likelihood that an event can be attributed to a specific causal mechanism.

Characteristics of the case-control research design:

In the case-control research design, the researcher is attempting to identify causal mechanisms for a condition of interest by comparing an outcome in known cases against the lack of an outcome in similarly matched controls. Since the sample is not representative of the larger population, and because the cases are selected by the investigator, the influence of selection bias is high, and the generalizability to a population from which the cases were identified is low. Likewise, because the total number of at-risk individuals in the population is not considered in this design, the incidence of the outcome cannot be estimated and relative risk can only be estimated in specific scenarios. However, the case-control research design enables the researcher to establish a proportional representation of cases drawn from a sample against the number of individuals not demonstrating the condition of interest and thereby present estimates of likely

causes. Using the odds ratio (OR) estimator, the researcher is able to compute the odds that a case is (OR) times more likely to be affected by a suspected causal mechanism (exposed) than a control (unexposed).

In the case-control design, the stimulus (or exposure condition) suspected to cause the condition of interest is not manipulated by the researcher. However, because of the comparison between observed cases and matched controls, researchers can evaluate risk factors by comparing the proportion of individuals exposed to a suspected causal agent that demonstrate a particular outcome versus the proportion of individuals not exposed to a suspected causal agent that do not demonstrate a particular outcome, while controlling for the proportion of individuals that demonstrate the outcome of interest but were not exposed versus the proportion of individuals that were exposed and did not demonstrate the outcome of interest.

Case-control studies are used extensively in epidemiological investigations to assess risk and help to determine the cause of outcomes. For example, in outbreak investigations, the case-control design is used to determine the risk associated with exposure to the suspected causal agent; in lung cancer studies the case-control design is used to establish the link between and cancer outcomes; in maternal health, the case-control design is used to show the benefits of breastfeeding duration and on the development of asthma and wheeze.

One of the difficulties in conducting the case-control design is that the researcher needs to establish the case definition precisely, and maintain consistency in selecting participants to the appropriate cells of the 2×2 matrix when analyzing the data with a simple odds ratio estimate, as shown in Table 6.1.

Table 6.1. Structure of the 2 x 2 table to calculate the Odds Ratio

		Outcome of Interest	
		Present (yes)	**Absent (no)**
Suspected Mechanism (Predictor)	**Exposed (yes)**	Cell A (+ve case, exposed)	Cell B (-ve case, exposed)
	Not Exposed (No)	Cell C (+ve case, **un**exposed)	Cell D (-ve case, **un**exposed)

In computing the odds ratio for the case-control study, the researcher is comparing the ratio of exposure within the sample of cases (a/c) to the ratio of the exposure within the control group (b/d). The odds ratio is then the ratio of the two ratios: (a/c) : (b/d).

Longitudinal Studies

Longitudinal studies are a form of observational study. However, unlike the cross-sectional study where variables in a sample of participants are measured at a single point in time, and the case-control study design where the outcome is measured first and the risk factors after the outcome has occurred, the longitudinal study measures variables in a sample of participants over a period of time. Further, because the longitudinal study design is observational, the study design does not allow the researchers to intervene or impose any stimuli on the participants. Yet, because the study

design is longitudinal, the researcher can collect data throughout the time period according to an apriori measurement schedule and evaluate outcomes at the level of the individual. Longitudinal study designs are thus often referred to as time-lagged, pre-post, or repeated measures designs. Note that for a study to be truly longitudinal, at least three time-points of data must be collected.

An important estimate in longitudinal studies is often the incidence rate (also referred to as the incidence density rate). The incidence rate is defined as the number of new cases observed within a given time period divided by the population at risk during the specified time period. Computing the incidence rate enables the researcher to determine the effects of a specific treatment regimen, monitor adherence to a prescription, assess the effects of ageing, or determine the compliance to policies.

Cohort Studies

One of the most common types of longitudinal study designs is the cohort study. The cohort study is an observational study in which the researcher simply observes an outcome without intervening. Cohort studies follow a group of individuals with similar characteristics either forward in time (prospectively) or backward in time (retrospectively). Cohort study designs are used to study incidence rates. In the cohort study, the group demonstrating the characteristic(s) of interest are followed for a period of time while being compared to a similar group that does not demonstrate the characteristic(s) of interest. Specific measures in the designated group are compared to those reported for the comparison cohort. Throughout the monitoring stage, measures are taken typically at the onset of monitoring, at pre-designated points throughout, and then again at the completion of the study.

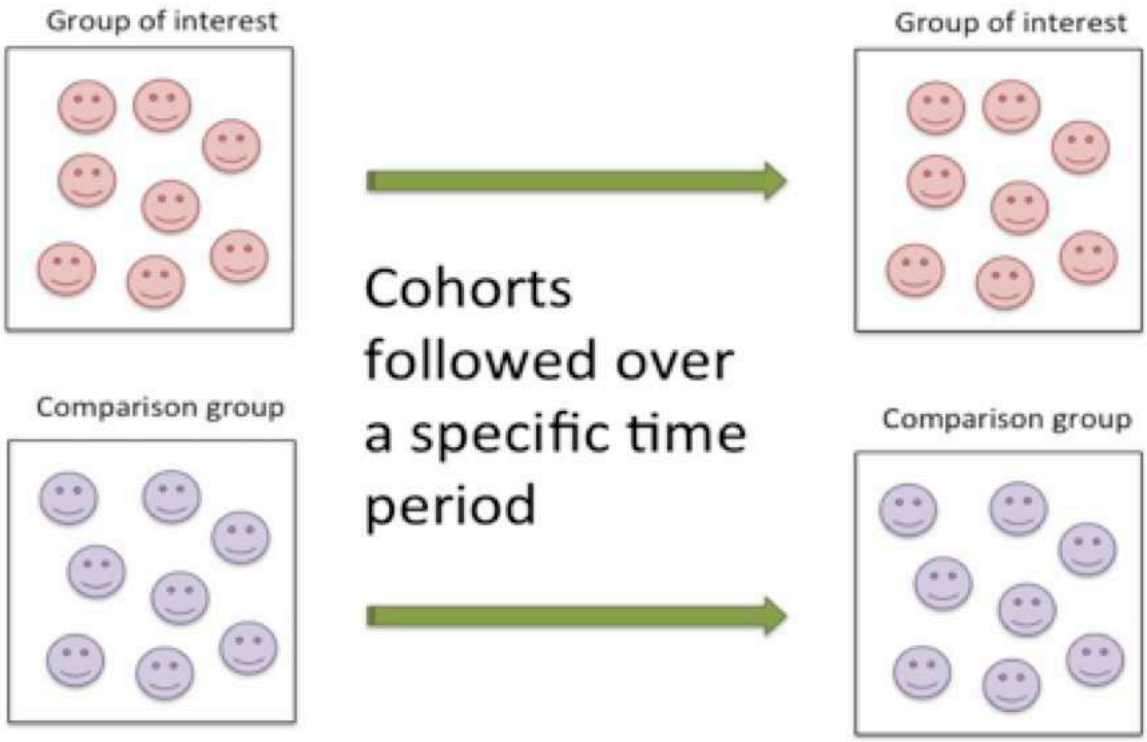


Figure 6.3 Monitoring Characteristics of the Cohort Research Design

Characteristics of the cohort research design:

Strengths of the cohort research design include the capability to monitor individuals with rare exposures and study the multiple effects of a single exposure. Cohort studies are also useful when evaluating a temporal relationship (time-based) between exposure and outcomes because the individuals are followed over a specific period of time and are compared to non-exposed individuals drawn from the same population during the same time period. Likewise, given that cohorts are homogenous groups of individuals demonstrating like characteristics and only differing on the measure of interest, concurrent cohort research designs can be used to minimize the bias due to sampling as in selection bias and over-estimation in the treatment group.

Since cohort studies require that the researchers follow the designated groups for prolonged time periods, cohort studies tend to be expensive. Often as a result of the prolonged-time period required to ensure complete observations, the cohort study suffers from loss to follow-up, which can essentially invalidate the study. In order to ensure the precise estimation of incidence, cohort studies require attention to detail in the follow-up stages.

In cohort studies, the estimate of relative risk is used to show the ratio of the probability of those exposed versus the probability of those not exposed. In table 6.2, the structure of a 2×2 computational arrangement to calculate relative risk is shown.

The formula for relative risk is given as the ratio of – the proportion of individuals within an exposed group showing a condition versus the proportion of individuals within a non-exposed group showing a condition à (cell a / (cell a + cell b)) : (cell c / (cell c + cell d)).

Exposed	Cell A + condition + exposed	Cell B – condition + exposed	The numerator (a/(a+b))
Not exposed	Cell C + condition – exposed	Cell D- condition – exposed	Denominator (c/(c+d))

Relative rRisk = (a/(a+b)) + (c/(c+d))

Table 6.2 Relative Risk derived from a 2 x 2 Computational Arrangement used in the Cohort Research Design

According to Grimes and Schulz (2008), an estimates odds ratio can be used to compute the relative risk score as shown here:

RR = (Odds Ratio) : [(1- P0) + (P0 x OR)] Eq (6.1)

However, in cohort studies of rare diseases the odds ratio factor is negligible to the estimate of relative risk because P0 approaches 0 when the disease is rare. As such the relative risk estimate in a cohort study of rare disease cases is equal to the odds ratio. Computations of relative risk and odds ratios are discussed in more detail later in this text.

Experimental Research Designs

In its simplest form, an experiment is an evaluative procedure in which the researcher controls the conditions that are applied to a selected group of participants and they observe the occurrence of an outcome. For example, consider an experiment to determine the effect of using a drug versus a placebo on changes in resting systolic blood pressure.

In step 1, the researcher randomly selects a group of individuals from the population. The individuals in the group are then randomly allocated to either of two groups – the drug group versus the placebo group. In step 2, the individuals in both groups are measured for their resting systolic blood pressure. In step 3, individuals from the drug group receive a

pill that is purported to alter blood pressure. In the other group-the placebo group, the individuals receive a similarly shaped pill that is a placebo. In step 4, the individuals in both groups wait for one hour and then are again measured for their resting systolic blood pressure. The difference between the systolic blood pressure measurements is compared in two ways.

First, the average change in blood pressure in the drug group is compared to the average change in blood pressure in the placebo group. Next, the average pre blood pressure is compared to the average post blood pressure in the drug group separate from the placebo group. The series of events in this simple experimental design are shown in the following illustrations. In Figure 6.3, participants are randomly selected from the population and randomly allocated to either the drug group or the placebo group.

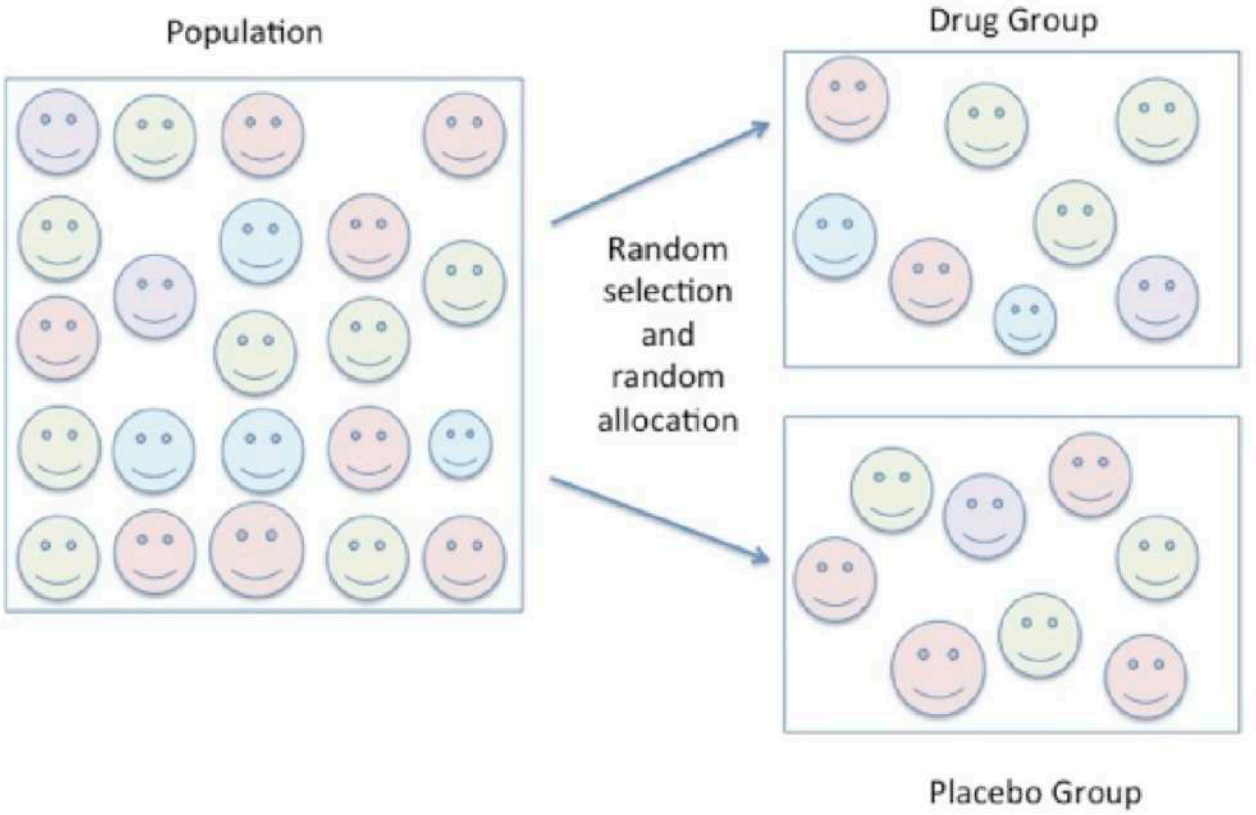


Figure 6.4 Random Selection and Random Allocation

Next, resting systolic blood pressure is measured for each participant in each group after the individual has been allocated to the groups and then again after one hour following the ingestion of either the drug or the placebo.

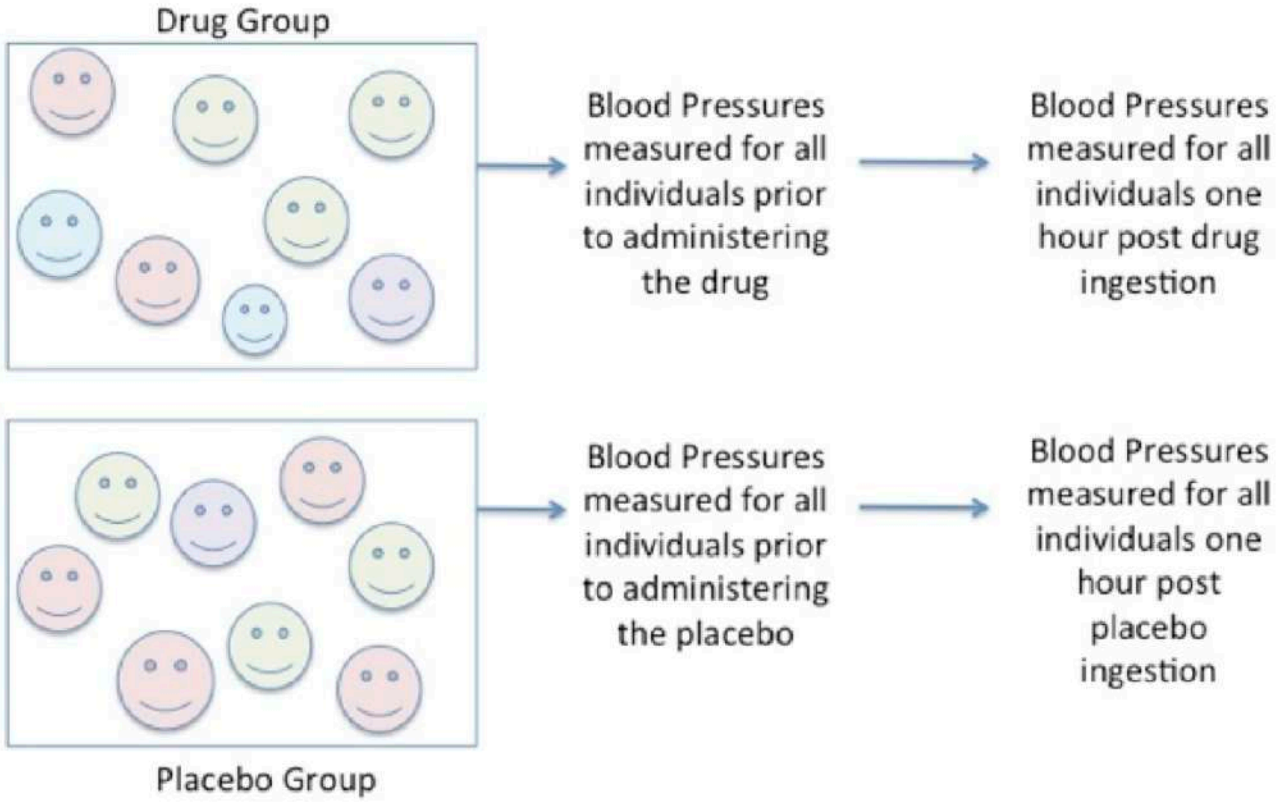


Figure 6.5. Consistent Measures in Each Group

Statistical comparisons can be made using an estimate from the group, like the mean to compare the effect of the drug versus the placebo. Pre to post means are compared within each group as shown in comparison 1(a) and in 1(b). A second comparison is made between the pre to post difference in the drug group versus the pre to post difference in the placebo group.

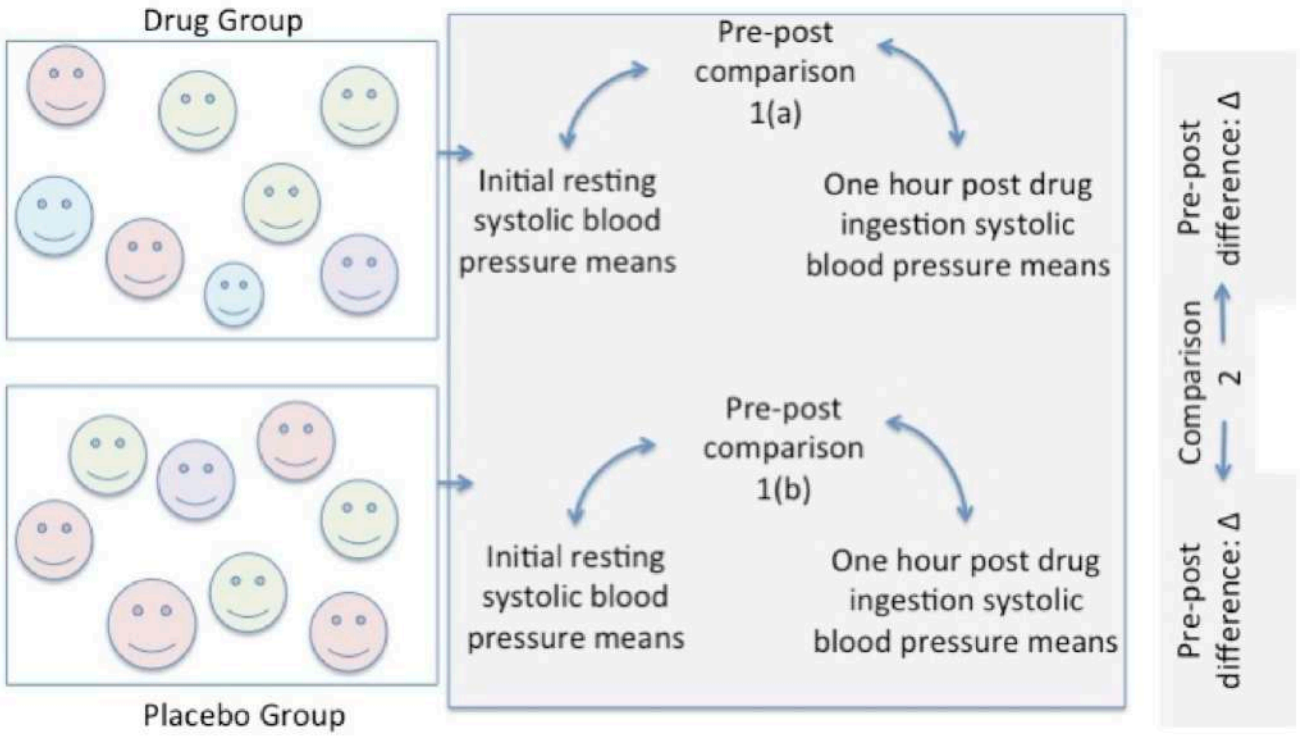


Figure 6.6 Pre-Post Comparison in Each Group

7. John Snow and the Natural Experiment

Historical Background

In a **traditional experiment**, the researcher maintains control over the organizational considerations of the methodology, but, to ensure that the experiment is unbiased, the researcher may randomly select participants from a designated larger population, and randomly allocate the selected participants to the various control and experimental groups. The process is shown in Figure 7.1 below.

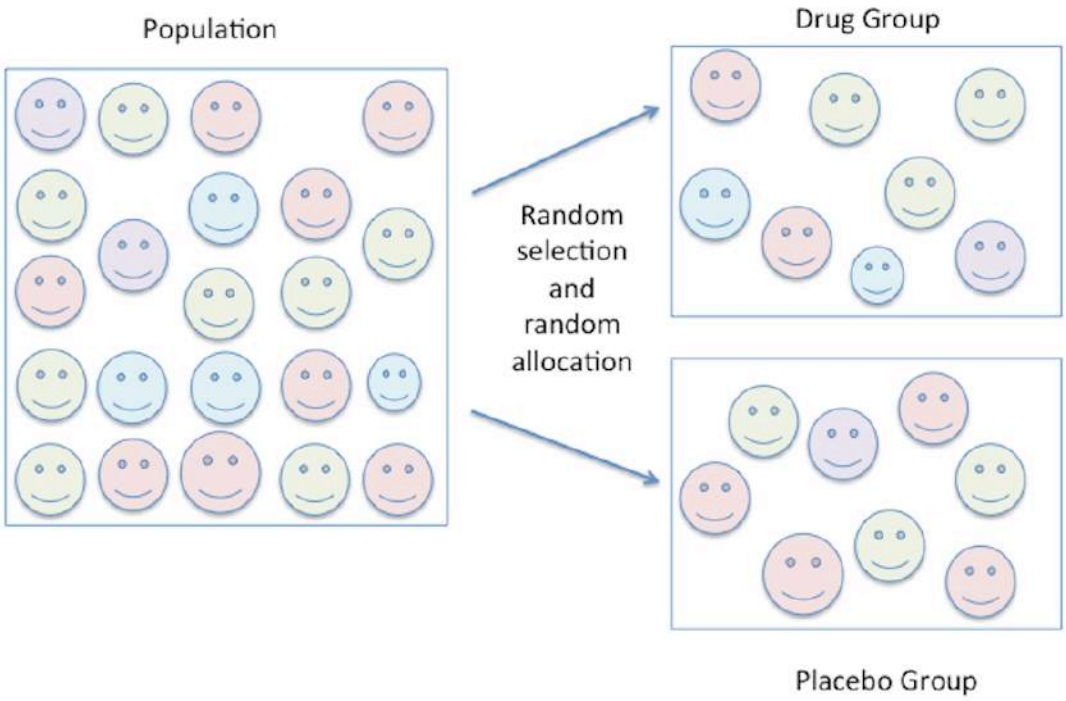


Figure 7.1 Random Selection and Random Allocation in the Typical Experiment

In the **natural experiment**, the researcher is merely the observer. While they may have decision making control with respect to identifying which areas of a population will be investigated, all of the efforts in enumerating the outcomes of the naturally selected and distributed participants, as well as the exposure to the suspected stimulus will occur aposteriori – after the fact.

The birth of epidemiology and public health is often attributed to the **Natural Experiment** described by Dr. John Snow in the mid-1800s when he investigated the relationship between drinking contaminated water and the incidence of cholera.

A natural experiment is defined as one in which the researcher has no manipulation of the stimulus or outcome but rather identifies the cohort of interest and the comparison control group, and then waits to observe outcomes in the identified groups. In the measurement of the cholera outbreak of 1854 in London, England, Snow identified individuals that were patrons of two separate water companies (exposed versus not exposed to the stimulus: *vibrio cholerae*), and simply counted and compared the number of cases of cholera in each group of patrons. In other words, he let the events occur naturally and without his direct influence on the stimulus or outcome.

Figure 7.2 Dr. John Snow

About John Snow

John Snow (1813-1858) was a physician who lived and worked in London, England. In addition to his reputation as an anesthetist for Queen Victoria, he is also known as a pioneer in modern epidemiology. John Snow used the design of a natural experiment to demonstrate the causal association between the waterborne bacterium: Vibrio-Cholerae and the acute intestinal infection that led to widespread death in London during the early 1850s. During the 1830s there were various epidemics, which spread across Europe and although described as cholera, were attributed to social unrest and political upheaval. A later widespread epidemic in 1848 led John Snow to publish a public health pamphlet that described the spread of cholera as being a waterborne infection.

In 1854 there were several misconceptions and misguided beliefs about cholera, especially in regard to the causes of a cholera outbreak. Most commonly, the miasma theory (also called the miasmatic theory) was a commonly accepted explanation for diseases and outbreaks like cholera or the bubonic plague and even sexually transmitted diseases like chlamydia. It was suggested that these illnesses were caused by a noxious form of "bad air", also known as the night air.

John Snow theorized that the cause of the most recent cholera epidemic (1853-1854) was due to the presence of human waste which included the Vibrio Cholerae bacterium in the water source from selected water supplying companies – most notably the Southwark and Vauxhall Company.

Snow's theory was based on the location of the water company's intake spout in relation to the effluent discharge in the Thames River. Snow used a natural experiment design to show that households which received their water from the Southwark and Vauxhall Company, which had an intake spout downstream from the effluent discharge were more likely to present with cholera and cholera-like symptoms than persons in households in a comparative neighborhood that received their drinking water from the Lambeth Water Company, which had placed their intake spout upstream in the Thames River and away from the effluent discharge.

The image on the left, below illustrates the comparison of the positions of the water intake valves for the two companies. The image on the right, below illustrates the distribution of water to the different households by the company supplying the water. Here Snow compared the water from the Southwark and Vauxhall Company to that of the Lambeth Company.

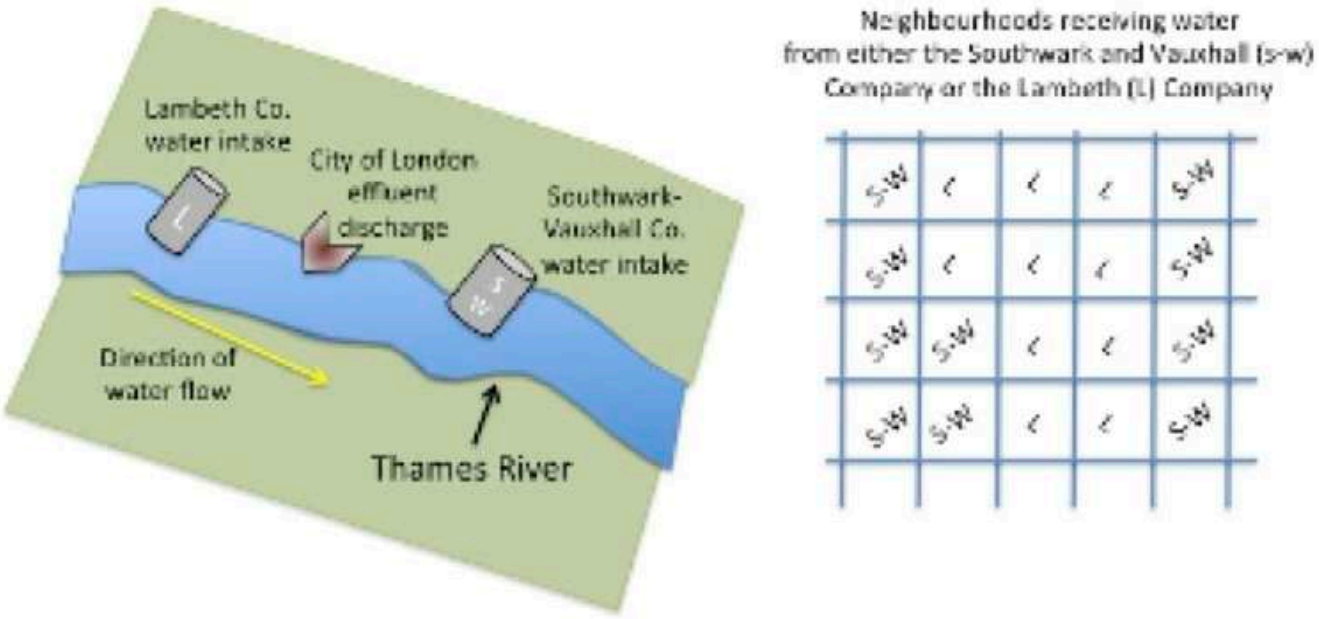


Figure 7.3 Drinking-Water Distribution by the Lambeth Water Company and by the Southwark and Vauxhall Company

The image below is that of the area mapped by John Snow in 1854 to show the location of deaths among households that received their water from the Broad Street Pump. This map presents the area used for patient selection in the natural experiment.

Figure 7.4 Map of deaths among households for the area serviced by the Broad Street Pump

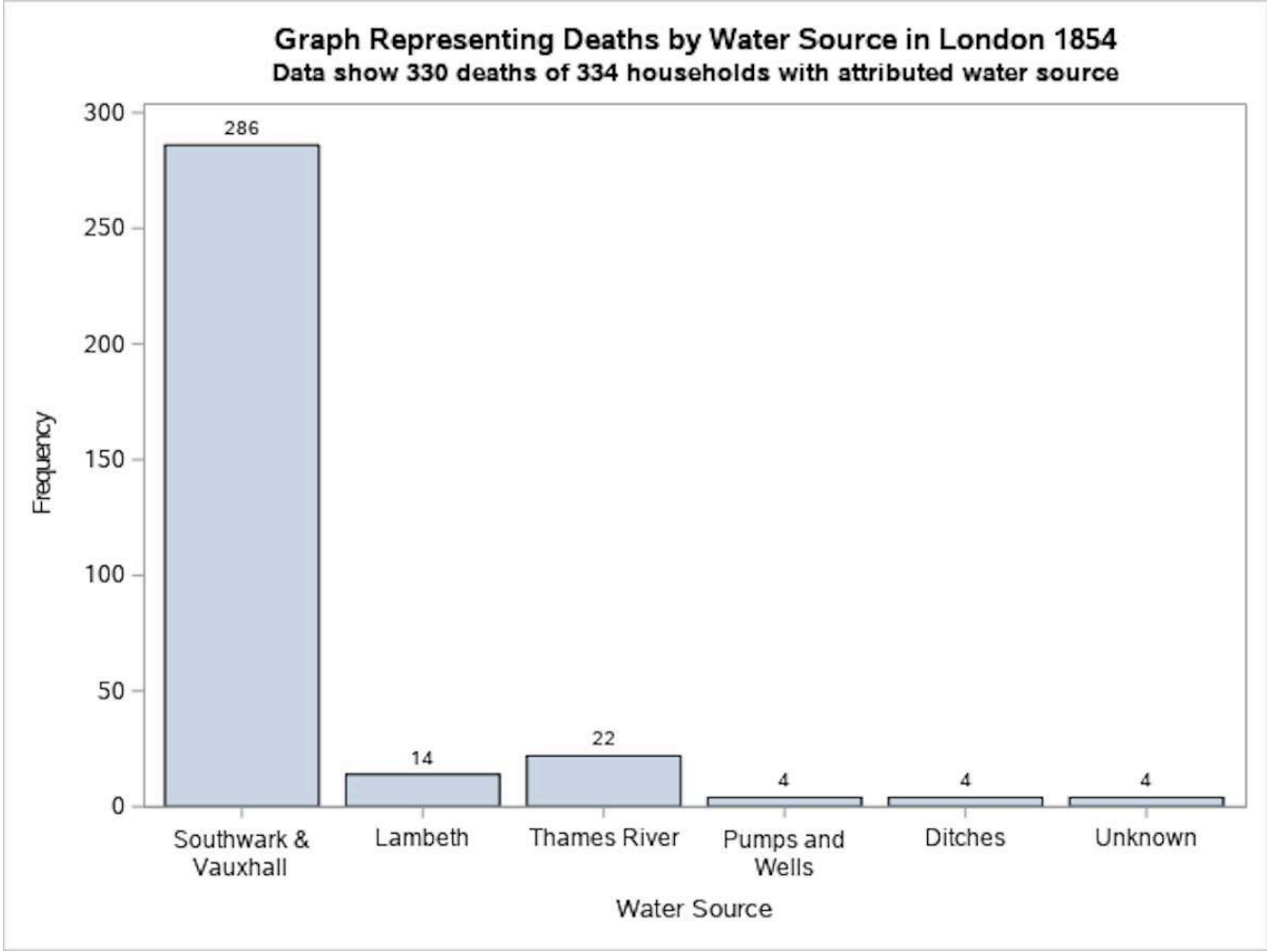
Snow visited households in the district to confirm the number of deaths per household within the given area. The following table shows the number of deaths to households by water source. The data indicate that there were 330 deaths in 334 households that could be attributed to a water source.

Source of water	Number of Deaths
Southwark and Vauxhall company	286
Lambeth company	14
Direct from the river	22
Pump wells	4
Ditches	4
Unknown	4

Table 7.1 Distribution of the Number of Deaths Per Household

A SAS graph to evaluate tabled data – reporting on the London Cholera Epidemic

Imagine if John Snow had access to SAS in the 1800s. How might he have presented his findings with respect to comparing cases between Lambeth water consumers versus households that consumed water from Southwark and Vauxhall? Below is a SAS vertical bar chart that presents the data from Table 7.1 above. The program to generate this graph is annotated later in Chapter 12.



In a follow-up to his original survey of deaths, the following table presents the data for the individuals who died of cholera from 8th July to 26th August 1854. These data indicate the source of water in the households and thereby provide the denominator of all those individuals at risk of death from cholera.

Source of Water	Total Number of Houses Supplied	Cholera Deaths (n)	Proportion per 10000 persons
Southwark -Vauxhall	40,046	1263	315.00/104
Lambeth	26,107	98	37.53/104
Other	256,423	1422	55.46/104

Table 7.2 Summary Of Cholera Related Deaths In The Two Neighbourhoods

The computation of deaths attributed to water sources showed that patrons of the Southwark and Vauxhall Company were more than eight times more likely to die from cholera than patrons of the Lambeth Company (315_{SV} / 37.53_{L}= 8.39).

Many years later, in 2017, the World Health Organization described cholera as an acute intestinal infection that is caused by the Vibrio Cholerae bacterium and that causes individuals to experience severe diarrhea leading to dehydration and eventually death. One of the ways the infection is treated is by rehydration through copious consumption of clean water. Unfortunately for those villages that continue to overcome this epidemic, the difficulty is in the ability to locate clean water.

PART II

SAS PROGRAMMING

Learning Objectives

After reading this section you should be able to:

- Create and process SAS code using the SAS Studio feature of the SAS University Edition
- Open the SAS editor and create SAS files
- Run a SAS program and retrieve the output from the processing of a dataset
- Create directories and store/retrieve files for processing with SAS
- Create SAS commands to read embedded data as well as external data files

In this text, we will use SAS University Edition, also known as SAS Studio. Throughout the textbook we will simply refer to SAS. This program provides a powerful platform for problem-solving and conducting quantitative research analyses.

What is SAS?

SAS is an acronym for "Statistical Analysis System". It is a statistical software program that you can use to evaluate quantitative research questions in a variety of fields. SAS uses standard computer logic common to many languages combined with a unique programming-style syntax that gorms the framework for SAS procedures, or procs, pronounced "prock". You use PROC statements to construct evaluative approaches for analysis of any kind of data set you may encounter. Moreover, because you can write SAS programs using standard syntax resembling traditional programming language, SAS is a versatile tool with a wide range of applications. For example, you can apply standard logic statements within a SAS program to run computer simulations and test all kinds of emerging ideas.

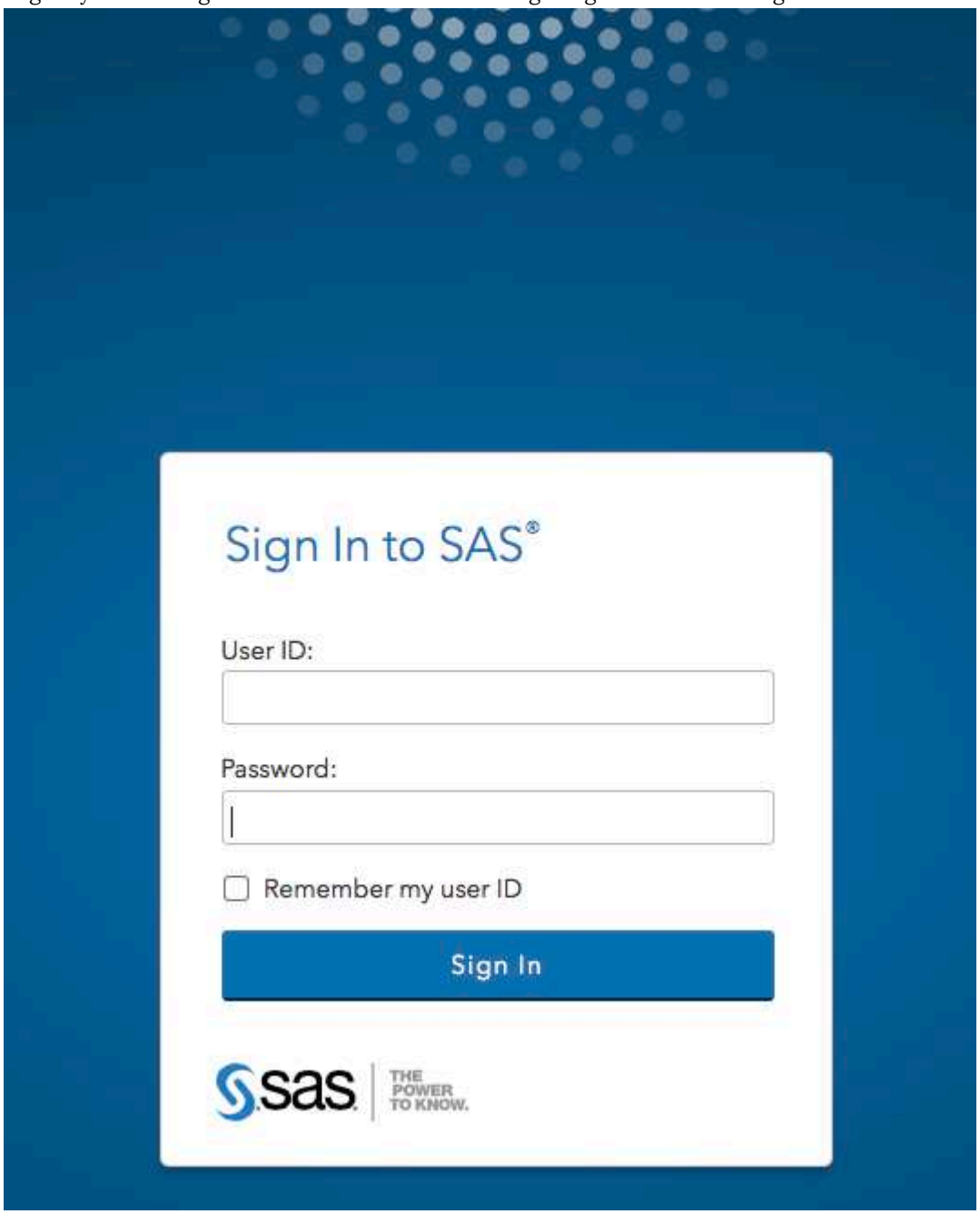
Connecting to the SAS Program

In order to use SAS you must first connect to the SAS Studio program (SAS University Edition). SAS Studio runs on multiple operating systems and can be installed as either a network operation or as a single user system. What this means is that you can access the SAS system by connecting to a Cloud-based system using your web browser, or you could have the entire SAS University Edition running on your personal computer. See your IT expert at your institution to assist

you in setting up SAS Studio with the SAS University Edition. You can also download the SAS Education Analytic Suite directly from SAS for free:

http://www.sas.com/en_ca/software/university-edition.html

Begin by connecting to the SAS Studio. The following Image is that of the Log-in Screen.



The nexr image may be available when using a stand alone version of SAS Studio:

sas
THE POWER TO KNOW.
Start SAS Studio >

8. Components of a SAS Session

Every SAS session produces three files:

- **The SAS code file** (.sas): This is the set of instructions that are submitted to the central processing unit of the computer through the SAS engine to generate output (results). You can use the program file editor in SAS Studio to create the .sas file. Your SAS instructions are included in the filename.SAS program file as it is the file that you submit to be processed by the SAS engine which in turn creates output files and log files.

After you have successfully started the SAS Studio you will see the image below on your screen. Be sure to select the **SAS Programmer work mode** as shown in the image. This is the SAS Studio operation in which you can function as a SAS programmer. In this text, we will use the SAS Studio software to practice statistical analyses for a variety of questions that you may generate in health-related research.

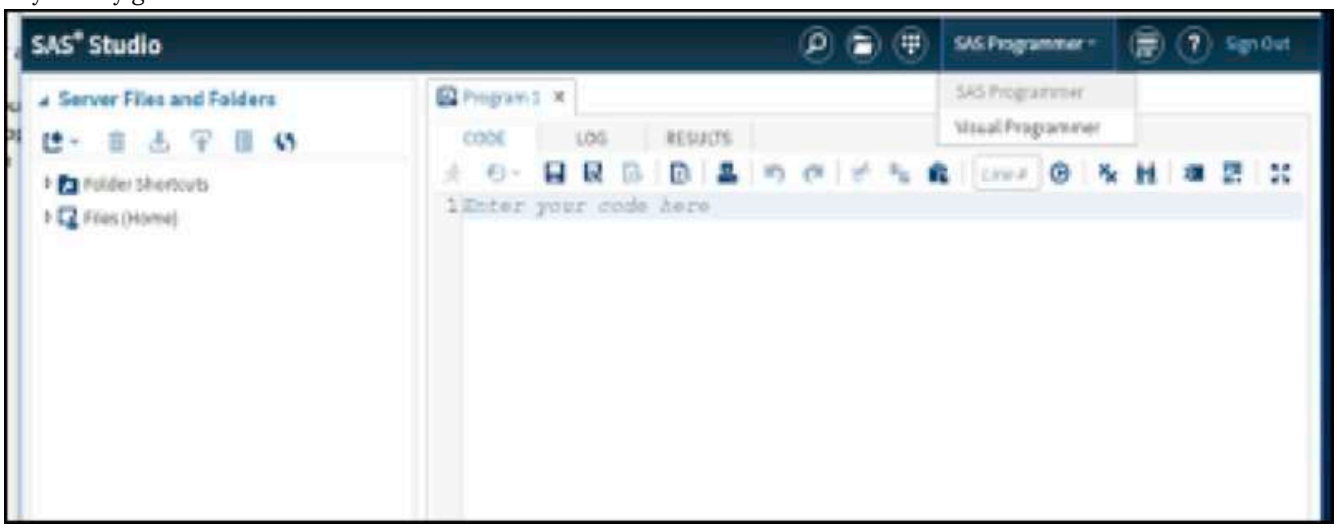


Figure 8.1 The landing page for SAS Studio in SAS University Edition

Note that there are two work modes in the SAS Studio: 1) SAS Programmer and 2) Visual Programmer. In the examples presented in this text, we selected the SAS Programmer mode (Figure 8.1).

SAS programs are written to the code page provided within the SAS Studio. The editor function enables you to create the program, submit the program and evaluate the output.

- **The output file.** In SAS Studio, the output file is accessed from the RESULTS tab of the program editor page, after your SAS code file has been submitted for processing. The RESULTS can be downloaded in the form of either an HTML file – which you can access through a browser, as a PDF file, or as a word document file – in RTF format. The output file is what you are trying to produce with your SAS code as it is the file that is generated from the statistical processing of your data. If you should end the session without retrieving the results in one of these formats then it will not be stored, and you will need to resubmit the program in order to view the output.

After you have successfully submitted the SAS program you can click on the RESULTS tab to view the output that you have generated with your SAS code.

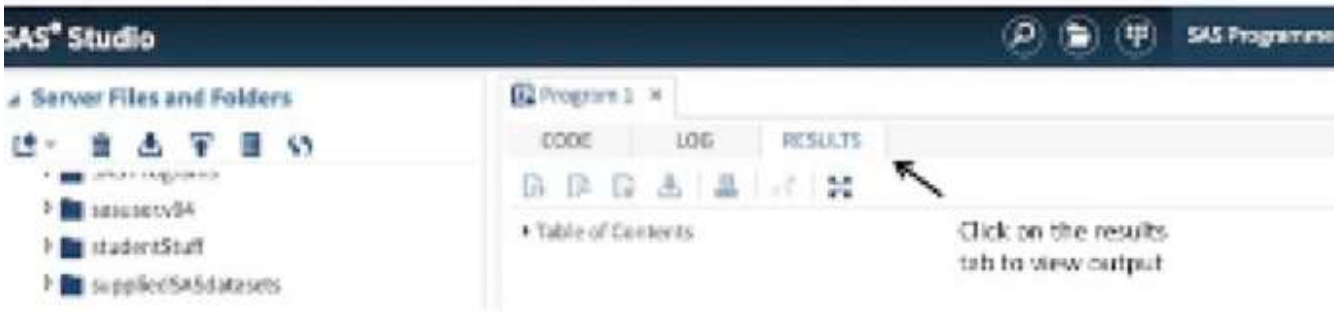


Figure 8.2 Where to find the RESULTS screen for the SAS Studio in SAS University Edition

**** SAS code that runs without errors will generate output in the window shown in Figure 8.1 above.***

- **The log file:** The LOG file keeps track of what you do in SAS. It is accessed from the LOG tab of the program editor after your SAS code has been processed. The log file is extremely valuable as it details the steps that the SAS *engine* used to generate the results from the series of commands that you entered in your SAS code file. The log file also includes any error messages and warnings that result from the submitted SAS code file. The LOG file presents the specific processing activities of the SAS engine which thereby show you the incorrect syntax or inappropriate command choices and sequences that you may have included in your SAS code file. The SAS LOG file is extremely valuable to you as a programmer as it can show you exactly what you did in the code file that the SAS processing engine didn't like.

Similar to the generation of the RESULTS file, after you have successfully submitted the SAS program you can click on the LOG tab to view the processing sequence of the commands that you submitted in your SAS code. All SAS code submissions will generate a LOG file as shown in the image below.

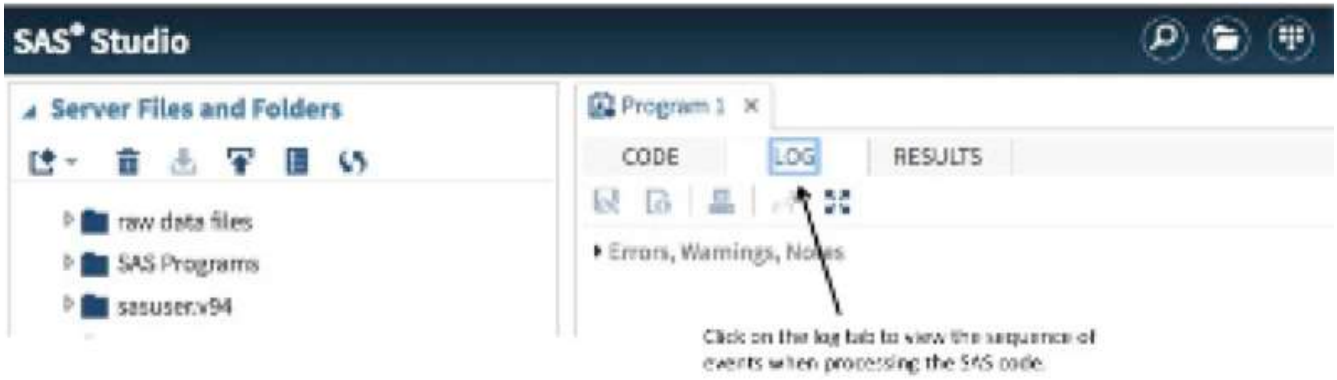


Figure 8.3. Log screen for the SAS Studio in SAS University Edition

The log file includes any error messages and warnings that result from the submitted SAS code file, allowing you to identify and resolve problems.

Entering Data and Writing a SAS Program

In the following section, we describe the process of creating a SAS program to analyze data that we have collected.

Later we will learn how to import data from several sources as external data sets but for now, we will enter the data by hand.

Syntax and Variable Type are Important

- Syntax refers to the structure of the language. All computer languages have a specific syntax with distinct rules related to the composition, arrangement and phrasing of commands.
- In SAS programming there are distinct composition and structural arrangement requirements in order for the SAS Processing Engine to understand the code sequence and perform the anticipated analysis.
- In SAS the Statistical Procedures are referred to as Procs – pronounced "PROCK"
- In SAS all command paragraphs end with a semi-colon
- In SAS we **must** define all alphanumeric variables by including a $ after the variable name, but we **can** include decimal length indicators if we choose when using continuous measures.

An annotated practice example

Let's write our first SAS program using data for a simple reaction time experiment.

In our experiment, we will use a sample of 5 males and 5 females (total n = 10). We record each participant's age and their score on a simple reaction time test.

To measure reaction time, we drop a meter stick from 1.5m off the floor and have each participant to catch it between their fingers.

This simple test has been modified for use in the assessment of reaction time testing for concussed patients and is commonly referred to as the Sideline-Drop stick test[1]. The score is a measurement in centimetres, of the distance that the metre stick travels between the participant's fingers from the start of the test to when the participant secured the stick. We can use the distance and the speed of gravity to calculate how quickly each participant grabbed the stick. We can also use distance scores as a proxy measure as we will do in this example, for simplicity.

The data we collected for this experiment is shown in the following table:

Participant ID	Age in Years	Sex	Reaction time Score
1	21	M	2.3
2	21	F	3.2
3	22	M	4.2
4	21	F	2.4
5	23	M	5.8
6	20	F	4.3
7	21	M	3.6
8	21	F	5.4
9	21	M	7.5
10	21	F	1.2

Table 8.1 Sideline Drop stick Test data for Annotated Practice Example 1

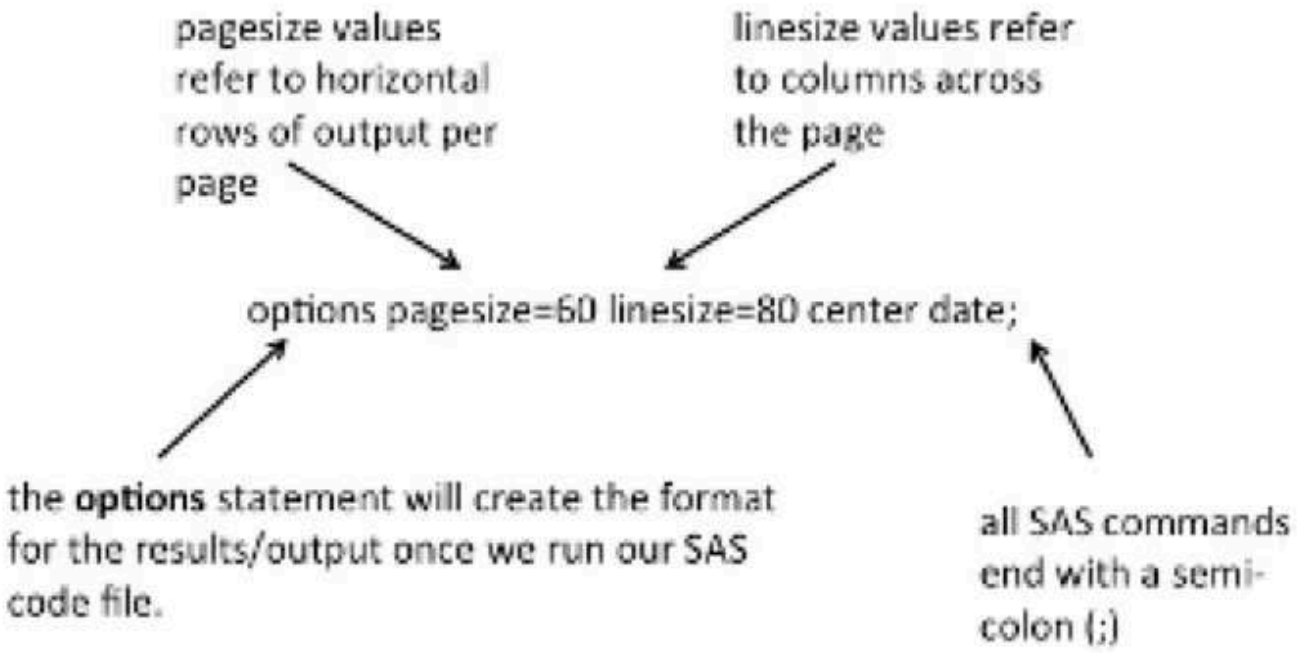
Figure 8.4 SAS Studio editor space in SAS University Edition

How to Write the SAS Program

- The first line we enter into our SAS program is the **options** statement. The options statement tells SAS how to report the results in the output file. Notice that every SAS command statement ends with a semi-colon (;). This is required to tell SAS that the command is complete – remember, SAS is not magic! You are the programmer (aka coder) and so you need to tell the program what to do.

```
options pagesize=60 linesize=80 center date;
```

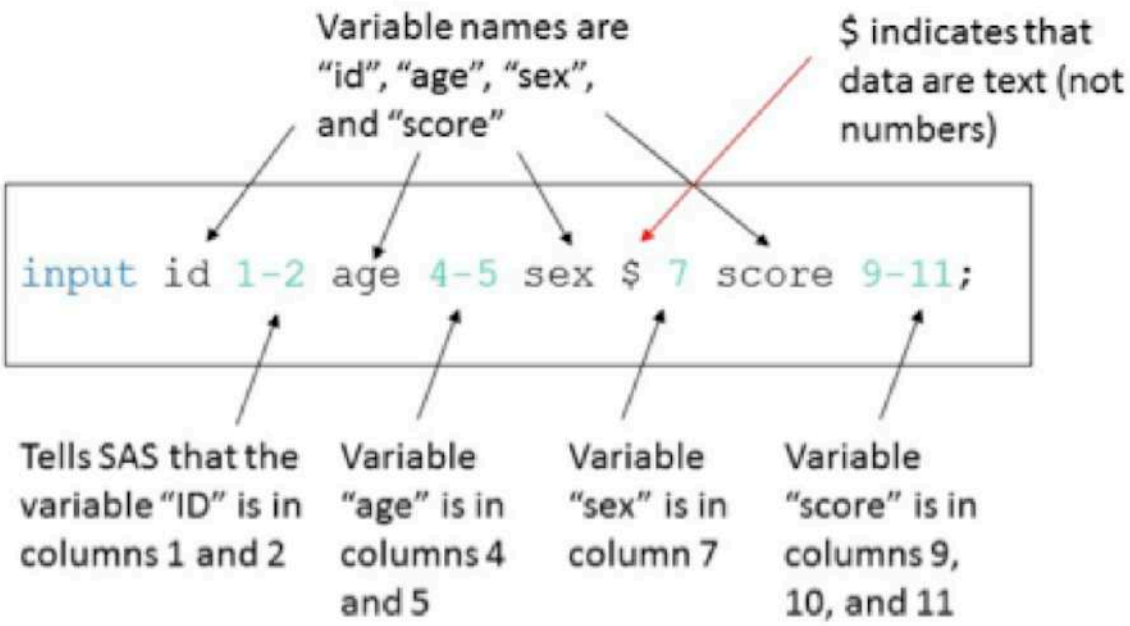
The SAS CODE explained.



- In line 2 of the program, we name the workspace. Naming the workspace is important so that we can recycle our code and reuse features of programs that we have already written. In other words, you don't have to start from scratch every time you use SAS. How awesome is that!? To name the workspace use the **data** command shown here as:
- Next, we use the **input** command to tell SAS where each variable is located in our code file and whether the data are numeric or include text characters (i.e., letters or words). First we type the word INPUT. In SAS we have an entire lexicon of key words that invoke specific functions. INPUT is a KEYWORD that lets the SAS engine know that the text that follows identifies the column headings (aka variable names) the variable types and the width of the column that holds the data for the variable

Recall that our data set for the **reaction time test** included the following variables: **participant's id**, **age**, **sex**, and **reaction time test score**. After each variable name, we provide the column numbers where the values for that variable are located. To indicate that a variable includes text characters add a dollar sign $ after the variable name. Don't forget to include a semi-colon at the end!

```
INPUT ID 1-2 AGE 4-5 SEX $ 7 SCORE 9-11;
```



1. Next we can add a **label** statement that will help SAS understand the names of the variables that we used in the SAS program.

```
LABEL ID='PARTICIPANT ID'
  SCORE='REACTION TEST SCORE';
```

These first few lines of SAS code set up the input environment. There are more commands that we can add in this section, but for now, these are sufficient to enable us to conduct a simple analysis of our data.

2. Our next step is to provide the data that SAS will analyze. For this practice example, we will type the data into our SAS code file. In later exercises, we will use external datasets that are saved as separate files and tell SAS where to find it.

The data entry paragraph begins with the command data lines;[1].

This command is followed by the data set organized according to the column format that you specified in the input statement above. A semi-colon closes the data paragraph, and the data paragraph is followed by the run; statement.

```
DATALINES;
01 21 M 2.3
02 21 F 3.2
03 22 F 4.2
04 21 F 2.4
05 23 F 5.8
06 20 F 4.3
07 21 F 3.6
08 21 F 5.4
09 21 M 7.5
10 21 F 1.2
```

;

The image below illustrates the program up to this point. So far in this program, we have used the following SAS KEY-WORDS: DATA, INPUT, LABEL, and DATALINES. Notice also that each SAS COMMAND ends with a semi-colon (;).

react1.sas

CODE | LOG | RESULTS | OUTPUT DATA

```
1  options pagesize=60 linesize=80 center date;
2
3  data react;
4  input id 1-2 age 4-5 sex $ 7 score 9-11;
5  label id='Participant ID'
6        score='Reaction Test Score';
7  datalines;
8  01 21 M 2.3
9  02 21 F 3.2
10 03 22 M 4.2
11 04 21 F 2.4
12 05 23 M 5.8
13 06 20 F 4.3
14 07 21 M 3.6
15 08 21 F 5.4
16 09 21 M 7.5
17 10 21 F 1.2
18 ;
```

The datalines statement begins the data entry step. Data are entered in rows, representing cases, and columns representing variables.

Figure 8.5 How your SAS code file should look

Notice the structure of the program. The code itself does not necessarily need to begin in column 1. This is because SAS begins reading at the start of a command line and ends reading when it reaches a semi-colon. However, **the data are intentionally lined up in the left-hand margin**, as the column position of the data is important. Correct arrangement and location of the data is essential in order for your program to work properly.

3. **Be sure to save the code by clicking on the save as icon (it looks like an old-school floppy disk with a pen on it).** In SAS there are different ways to store files but **there is no auto-save** so if you exit the program without saving you will lose your work. For the most part, the storage/saving of a file is similar to that which you would use in any end-user application (like a word processor or spreadsheet application).

If this is the first time saving this SAS file, then click the SAVE_AS icon shown here:
If you are saving an updated version of the file instead, click the SAVE icon:
For this practice example, save the file as **react1.sas** in your folder space.

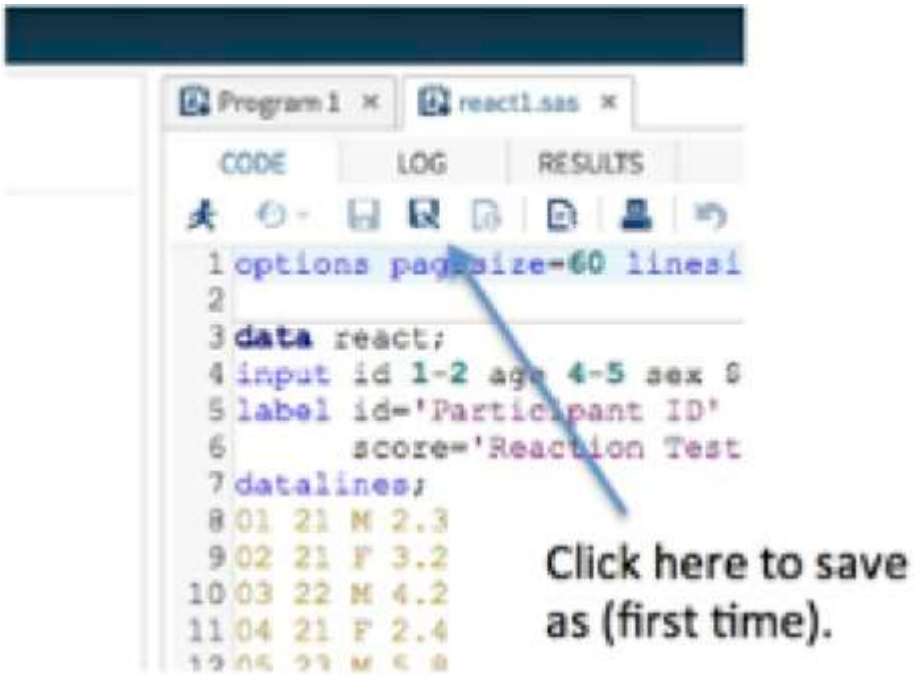


Figure 8.6 Using the SAVE-AS feature in the SAS program editor

Later, you can retrieve the saved file by locating the file name in your SAS folders. Just select the file and then double click to open the file in your SAS editor.

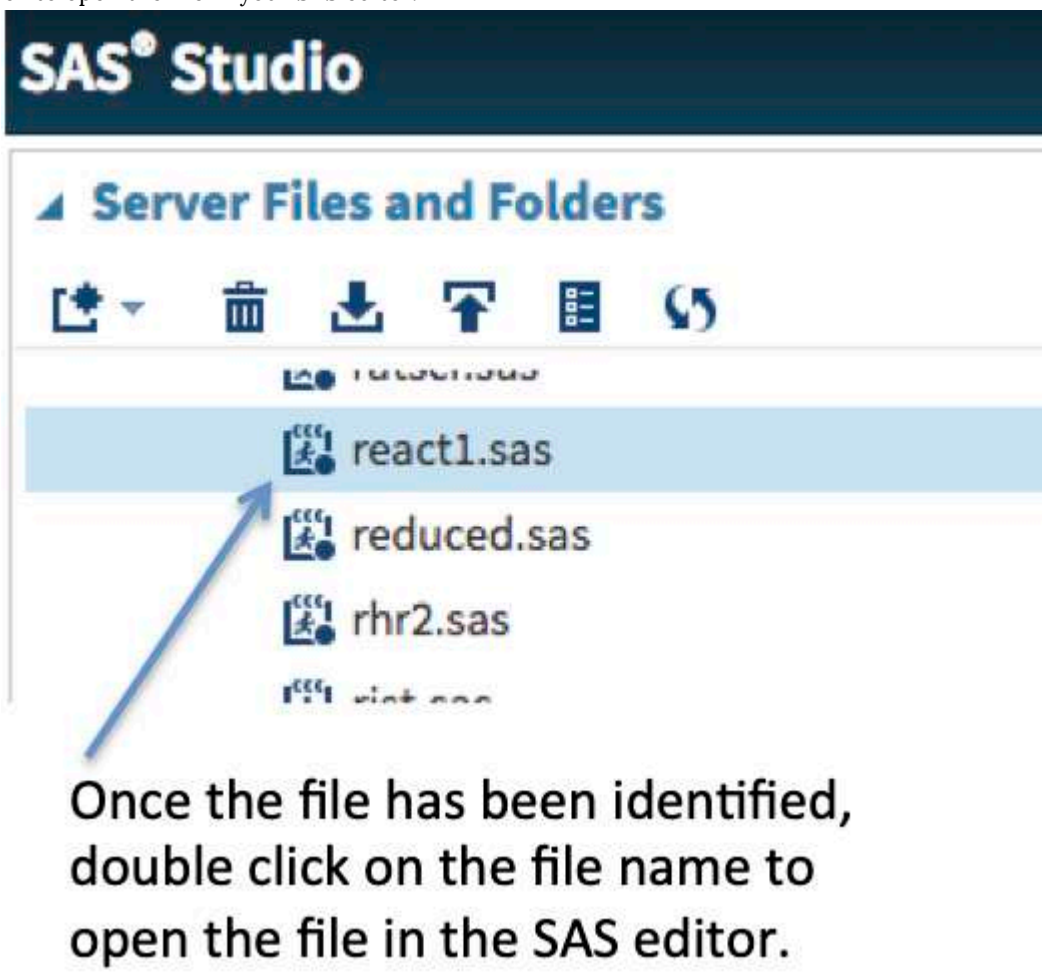


Figure 8.7 Retrieving a saved file for use in the SAS program editor

Now that the data is entered and your file is saved we can tell SAS to analyze the data. To do this we write SAS commands that tell the program what to do with the data. These are procedural statements so they each begin with the word **proc** (pronounced as "prock").

4. The first thing we want to do is to sort the data so we use the following command: **proc sort**.

So, what are we sorting? Well, we are sorting the data we entered by hand from our reaction test experiment. Recall that the name we gave to the working file was **react**, because it represented the reaction test scores for a group of participants.

Let's begin by sorting the data by sex and produce a printout of the data we entered. First we write proc sort which tells SAS to sort the data. Then we write **data=react;** to tell SAS which data we want it to sort. Finally, we tell it how to sort the data by using the word **by** + *the name of the variable to sort on* (in this case, **sex**).

```
PROC SORT DATA=REACT; BY SEX;
  PROC PRINT; VAR ID AGE SEX SCORE;
```

5. Let's also compute some basic descriptive statistics on this data using the **proq freq** command and the **tables** command to produce a frequency table to count the number of females and the number of males in our data set. Shown here:

```
PROC FREQ; TABLES SEX;
```

6. We also want to look at the average scores for the reaction test in the total group. To do this we first use the **proc univariate;** command which will tell SAS to provide descriptive statistics. We are only interested in looking at the reaction test score so we indicate this by writing **var** + *the name of the variable of interest* (in this case, **score**).

```
PROC UNIVARIATE; VAR SCORE;
```

7. Finally, let's compare the average reaction test scores between male and female participants. To do this we repeat the line of code for the total group but we add a **by** statement which tells SAS to apply the procedure based on groupings identified by that variable (in this case, **sex**).

```
PROC UNIVARIATE; VAR SCORE; BY SEX;
```

Congratulations! Your first SAS program is complete and ready to run.

react1.sas

CODE | LOG | RESULTS | OUTPUT DATA

```
1 options pagesize=60 linesize=80 center date;
2
3 data react;
4 input id 1-2 age 4-5 sex $ 7 score 9-11;
5 label id='Participant ID'
6       score='Reaction Test Score';
7 datalines;
8 01 21 M 2.3
9 02 21 F 3.2
10 03 22 M 4.2
11 04 21 F 2.4
12 05 23 M 5.8
13 06 20 F 4.3
14 07 21 M 3.6
15 08 21 F 5.4
16 09 21 M 7.5
17 10 21 F 1.2
18 ;
19 proc sort data=react; by sex;
20 proc print; var id age sex score;
21 proc freq; tables sex;
22 proc univariate; var score;
23 proc univariate; var score; by sex;
24 run;
```

Figure 8.8. The complete program in the SAS program editor

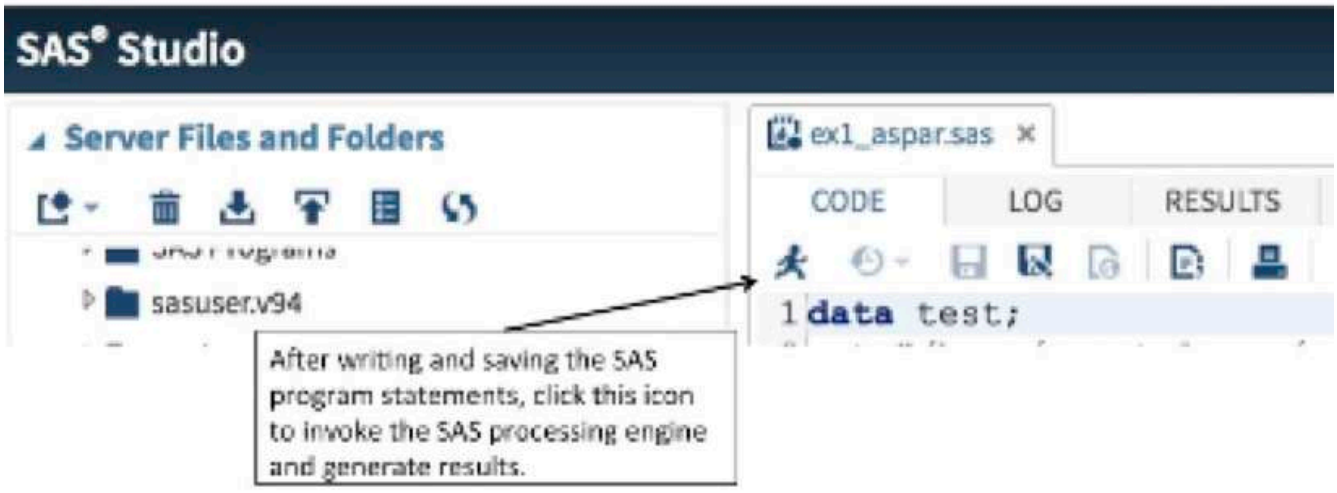
Save the program again before you run it. Saving continuously reduces the risk of losing work as you write program code.

[1] In older versions of SAS the word `cards` was used instead of `datalines`. Notice that the command ends with a semi-colon.

9. Running a SAS Program

In case you are new to programming, a SAS program simply refers to several lines of code that you wrote (like in our practice example about reaction time test scores). We are not referring to the entire SAS software program!

To run your program and generate results, click the running person icon. located at the top of the editor page.



Click the running person icon to invoke the SAS processing engine

In our first submission, we asked to sort the data and then create a printout of the data. The specific program statements are:

```
PROC SORT DATA = SAMPLE.REACT; BY SEX;
  PROC PRINT; VAR ID AGE SEX SCORE;
```

Obs	id	age	sex	score
1	2	21	F	3.2
2	4	21	F	2.4
3	6	20	F	4.3
4	8	21	F	5.4
5	10	21	F	1.2
6	1	21	M	2.3
7	3	22	M	4.2
8	5	23	M	5.8
9	7	21	M	3.6
10	9	21	M	7.5

Notice in our RESULTS that the data have been re-organized (sorted, if you will) from what was entered initially. As a result of the PROC SORT command. The data are now organized by sex, whereby the data for the females is entered first and the data for the males follows (because F precedes M in the alphabet).

The next SAS command asked the SAS engine to calculate the number of males and the number of females in the data set. The PROC FREQ command used here produced a frequency distribution table for the data organized BY SEX;

The result of the **PROC FREQ; TABLES SEX; c**ommand is shown below.

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	5	50.00	5	50.00
M	5	50.00	10	100.00

Our next SAS commands will produce descriptive statistics for the continuous variable reaction test score. We can process this variable first for the whole group and then for the separate subgroups of males and females. Notice that the SAS commands are all written in upper case. Using uppercase text is not required but is considered good SAS programming style.

PROC UNIVARIATE; VAR SCORE;

PROC UNIVARIATE; VAR SCORE; BY SEX;

Several important measures can be generated from the PROC UNIVARIATE procedure as we will see later in the text. Below is a simple summary from the descriptive statistics for the total group

The UNIVARIATE Procedure

Variable: score (Reaction Test Score)

Our next SAS commands will produce descriptive statistics for the continuous variable reaction test score. We can process this variable first for the whole group and then for the separate subgroups of males and females. Notice that the SAS commands are all written in upper case. Using uppercase text is not required but is considered good SAS programming style.

PROC UNIVARIATE; VAR SCORE;

PROC UNIVARIATE; VAR SCORE; BY SEX;

Several important measures can be generated from the PROC UNIVARIATE procedure as we will see later in the text. Below is a simple summary from the descriptive statistics for the total group

The UNIVARIATE Procedure –> Variable: score (Reaction Test Score)

Moments			
N	10	**Sum Weights**	10
Mean	3.99	**Sum Observations**	39.9
Std Deviation	1.87584055	**Variance**	3.51877778
Skewness	0.43975108	**Kurtosis**	-0.0729366
Uncorrected SS	190.87	**Corrected SS**	31.669
Coeff Variation	47.0135477	**Std Error Mean**	0.59319287

The following tables show a simple summary from the descriptive statistics for the separate subgroups organized BY SEX ? results for females followed by results for males.

sex=F

Moments			
N	5	**Sum Weights**	5
Mean	3.3	**Sum Observations**	16.5
Std Deviation	1.63095064	**Variance**	2.66
Skewness	0.02593164	**Kurtosis**	-0.8358726
Uncorrected SS	65.09	**Corrected SS**	10.64
Coeff Variation	49.4227468	**Std Error Mean**	0.7293833

sex=M

Moments			
N	5	**Sum Weights**	5
Mean	4.68	**Sum Observations**	23.4
Std Deviation	2.01668044	**Variance**	4.067
Skewness	0.45615378	**Kurtosis**	-0.5702631
Uncorrected SS	125.78	**Corrected SS**	16.268
Coeff Variation	43.0914624	**Std Error Mean**	0.90188691

Recall that by clicking the SAS LOG tab in the SAS program editor, we can review the sequence of the steps SAS used to process the program file. The LOG file lists any notes, warnings or errors that you may have generated as a result of your program entry. If the LOG file is clean, that is without errors, then your listing file will show the output relative to the commands that you submitted. A portion of the output in the file for the reaction time test is shown below.

```
OPTIONS PAGESIZE=60 LINESIZE=80 CENTER DATE;
   DATA REACT;
   INPUT ID 1-2 AGE 4-5 SEX $ 7 SCORE 9-11;
   LABEL ID='PARTICIPANT ID'
   SCORE='REACTION TEST SCORE';
   DATALINES;

NOTE: The data set WORK.REACT has 10 observations and 4 variables.
   NOTE: DATA statement used (Total process time):
   real time          0.00 seconds
   cpu time           0.00 seconds

PROC UNIVARIATE; VAR SCORE;

NOTE: PROCEDURE UNIVARIATE used (Total process time):
   real time          0.13 seconds
   cpu time           0.08 seconds
```

Observe the highlights

- In our data set we have four variables: The participant's ID, their AGE, SEX, and their reaction test score (cm)
- There were 10 participants
- Three of the variables were recorded as numerals, 1 variable was recorded as an alpha-numeric value and denoted by the symbol $.
- The variable set can be summarized as: ID AGE SEX $ SCORE
- The measure of interest is SCORE

10. Data Screening and Cleaning

This textbook was developed to demonstrate biostatistical research applications that use SAS coding and the Webulators to resolve questions that arise in healthcare research. In the following sections, we will work through the concepts of biostatistical applications using specific examples that are demonstrated with SAS coding or the application of the Webulators. And while it may be tempting to dive straight into your main analysis once you have your data but in most situations you need to do some work to first prepare your data. Before you start testing your main research hypotheses it is vital to get to know your data, screen it for errors, and make informed decisions about how you deal with missing values, outliers, and violations of underlying assumptions of the statistical applications you plan to use.

It may sound like a lot of work, but in the long run, taking the time to get to know your data and address these issues will ensure that you are confident about your results and you won't have to re-do your analysis because you later find a mistake or erroneous scores in the dataset. It is rare that you have a dataset that is 100% perfect as you begin your analysis so as a researcher you need to make some informed decisions about how to deal with the limitations of real-world data.

The process of screening and cleaning quantitative data generally involves the following components:

- Checking data accuracy (Is the data entered correctly?)
- Checking data completeness (How much data is missing? Are there patterns of missingness within the set of responses, or recorded values in the dataset?)
- Assessing the distribution of the data (How are values spread out in your sample?)
- Assessing the validity & reliability of measures (Are you measuring what you want to measure? Are your results repeatable?)

---

11. Working with Missing Data

In this section, we will work through the concepts of dealing with missing data using specific examples that are demonstrated with SAS coding, and which are based on the SAS Studio Education Analytic Suite.

Missing Values

Missing data are observations that we intended to record but did not. Values can be missing for different reasons and most of the time we don't know the exact reason why people didn't answer certain questions. However, we can look at how much data is missing as well as the patterns of missing values and determine whether missingness is related to the variable itself, other variables in the dataset, or has no apparent pattern. In the following sections, we will go through three categories of missing data that are commonly used in research to explain why data is missing.

How much data is missing?

The overall percentage of data that is missing is important. Generally, if less than 5% of values are missing then it is acceptable to ignore them (REF). However, the overall percentage missing alone is not enough; you also need to pay attention to *which* data is missing. Often you may need to consider deleting cases (participants) or individual variables that are missing a ton of values. This step alone can drastically improve the integrity of your data and reduce the overall percentage of missing values in your dataset.

Types of Missing Data

There are several types of missing data, as we will discuss here. Some types are easy to consider and account for, while others are confusing and may be less obvious to the novice researcher.

Data Missing at Random

In this situation, data is not actually missing at random which makes the name of the category very confusing! MAR data happen when missing values are related to another variable in the data set. That is, the missing value (y) depends on x, but not y (itself). Here are some examples:

Example 1:

In a survey of health care professionals, **nurses** do not report their **age**. In this case, being a nurse (x) predicts the missing data for age (y).

Example 2:

In a family survey, **single parents** do not report their income. In this case, being a single parent (x) predicts the missing data for income (y)

Example 3:

Employees who fear their manager do not report their job satisfaction. In this case, employees might be afraid to report their job satisfaction for fear of reprisal.

Note that in real life you might find MAR patterns in your data but the rationale behind them is still speculative. Unless you go back and check with the participants it is impossible to prove.

Missing Completely at Random (MCAR)

Missing data that doesn't have a pattern of missingness is referred to as data missing completely at random (MCAR). This is the ideal situation when you have missing data because missing values are random so any influence they have on your analysis is also random. Here are some examples of MCAR situations:

Example 1:

You conduct a study on heart transplant patients and discover that 10 patients did not answer 2-3 questions on your survey. The questions that are missing are different for each person and there is no pattern.

Example 2:

You conduct an RCT comparing the effects of fish oil supplementation versus placebo on anxiety levels in nursing students. 1 patient in the control group forgot to take their supplement on Sept 10th because they were busy. 2 patients in the experimental group missed their dose on October 3rd and Nov 19th, respectively. One woke up late. The other one burnt their breakfast and got distracted.

There is NO pattern causing the missing data.

Not Missing at Random (NMAR)

The last category is data not missing at random. In this situation, missingness is because of the variable itself. In other words, there is a reason why people don't want to answer that particular question. Usually this happens with sensitive questions.

Example 1:

People who are overweight do not report their weight. In this case, being overweight (x) predicts the missing data for weight (x).

Example 2:

Single parents do not report their marital status. In this case, being a single parent (x) predicts the missing data for marital status (x).

Analyzing Missing Values

The default in SAS is to delete missing values from your analysis. The effect this has on your results depends on how much of your data is missing. SAS offers a number of robust options for dealing with missing data but the focus of this section is on being able to see how much of your data is missing and examine patterns.

One of the easiest ways to examine missing data patterns is to use the PROC MI command which is the multiple imputation (MI) procedure in SAS.

The following code uses the NIMPUTE=0 option to create the "Missing Data Patterns" table for the specified variables.

```
  ODS SELECT MISSPATTERN;
PROC MI DATA = NAMEOFDATASET NIMPUTE=0;
VAR VAR_1 VAR_2 VAR_3;
RUN;
```

Let's do an example together:

In this example you are interested in knowing more about stress levels of caregivers of older adults with dementia. You send out a pilot survey and get an initial sample of 16 people to answer it. The questionnaire includes demographic variables and a five-item questionnaire to measure stress rated on a 5-point Likert scale from 1 = strongly disagree to 5 = strongly agree. Higher scores indicate higher levels of stress.

The first step of course is to set up your data file in SAS:

```
  OPTIONS PAGESIZE=60 LINESIZE=80 CENTER DATE;
DATA CAREGIVER;
LABEL
ID = 'PARTICIPANT ID'
STRESS1 = 'CAREGIVER STRESS QUESTIONNAIRE ITEM 1'
STRESS2 ='CAREGIVER STRESS QUESTIONNAIRE ITEM 2'
STRESS3 = 'CAREGIVER STRESS QUESTIONNAIRE ITEM3'
STRESS4 = 'CAREGIVER STRESS QUESTIONNAIRE ITEM4'
STRESS5 = 'CAREGIVER STRESS QUESTIONNAIRE ITEM5'
SEX = 'SEX';
INPUT ID 1-2 SEX 4 STRESS1 6 STRESS2 8 STRESS3 10 STRESS4 12 STRESS5 14;
DATALINES;
01 0 4 3 4 5 4
02 1 3 2 3 4 4
03 1   3 3   3
04 0 1 2 1 2 3
05 1 4 4   5 3
06 1 2 3 3 4 4
07 0 3 3 5   5
08 1 3 5 4 5 3
```

```
09 1 4 4 5 4 4
10 1   2 4 4 4
11 0 4 3 4 5 5
12 1 2 1 2 3 4
13 0 1 2 4 4 2
14 1 3 4 4 5 4
15 1 3 4 5 4 3
16 1   4 3 2 1
;
RUN;
```

Next we use the MI code template but we replace NAMEOFDATASET with the actual name of our dataset (CAREGIVER) and replace the VARIABLE NAMES with the actual names of the variables in our dataset:

```
  ODS SELECT MISSPATTERN;
PROC MI DATA = CAREGIVER NIMPUTE=0;
VAR ID SEX STRESS1 STRESS2 STRESS3 STRESS4 STRESS5;
RUN;
```

As you can see in the figure below, SAS uses this code to produce a table showing the number of cases with each pattern of missing data. First we look at the lefthand side of the table to examine the patterns of missingness in our dataset. In this example, you can see that Group 1 has no missing data because there is an "X" in each of the variables in the dataset. This means that there is data for each variable for participants in this group. The frequency of participants in this group is in the column labelled "freq" and you can see that there are 11 people with no missing values. By looking at the next column over which is labelled "percent", we can see that this represents 68.75% of the sample.

The next pattern of missing data is Group 2. Looking across the columns, we can see that there is no "X" for **stress4**. This means that participants in Group 2 answered all the questions except that one. There is only 1 person in this group and they represent 6.25% of the data. We can continue doing the same interpretation for Groups 4-6 in the table.

This table also provides the means for each of the variables. Again, don't forget that "means" for some variables are not meaningful. For example, the mean values provided for Participant ID and sex should be ignored here. What is valuable though is that for continuous variables you can compare their means for participants with different patterns of missing data. For example, for the variable Stress2, you can see that the mean is the same for Groups 1, 2, 4, and 5 but it is higher for Group 3. Although this is a small sample for illustrative purposes, you can hopefully see how the information in this table can help you understand the patterns of missing values in your data better.

Output table showing patterns of missing values

The MI Procedure

Missing Data Patterns

Group	id	sex	stress1	stress2	stress3	stress4	stress5	Freq	Percent	Group Means						
										id	sex	stress1	stress2	stress3	stress4	stress5
1	X	X	X	X	X	X	X	11	68.75	8.636364	0.636364	2.727273	3.000000	3.545455	4.090909	3.636364
2	X	X	X	X	X	.	X	1	6.25	7.000000	0	3.000000	3.000000	5.000000	.	6.000000
3	X	X	X	X	.	X	X	1	6.25	5.000000	1.000000	4.000000	4.000000	.	5.000000	3.000000
4	X	X	.	X	X	X	X	2	12.50	13.000000	1.000000	.	3.000000	3.500000	3.000000	2.500000
5	X	X	.	X	X	.	X	1	6.25	5.000000	1.000000	.	3.000000	3.000000	.	3.000000

Finally, consider that in research we can always expect to have missing data for a variety of reasons. The SAS program provides a powerful platform for calculating data while recognizing strategies to handle missing data.

12. Graphing Data for Effective Presentations

Learner Outcomes

After reading this chapter you should be able to:

- Organize data from various sources, and especially identify the specific variables that comprise a data set and separate data into independent and dependent measures.
- Present your data in a graphical format to convey your message to the reader by identifying the level of measurement for each of the variables within the data set and organize the data appropriately for the type of graph or table selected
- Select from a variety of graphing and charting features in SAS to create an appropriate visual presentation of variables in your dataset and thereby demonstrate when a graph of a particular type is most appropriate
- Prepare your data set for graphing and charting by transposing from wide to narrow or narrow to wide
- Create a simple SAS program to produce different types of graphs and tables
- Add specific features such as axis labels, legends, and colors to enhance your graphical presentation of variables in the data set

Preamble

Producing graphical images is one of the most important features of conveying the statistical message. Face it, statistics is a hard area to wrap your brain around because it is based on the complexity of mathematically derived outcomes. What is the chance of picking the correct lottery number? How many people do I need in my study to know that I can represent the population? What is the best outcome from the randomized clinical trial that I should achieve before I decide that the vaccine is effective? Chance, probability, estimation, hypotheses, confidence, prediction, these are all complex concepts in which we only estimate the likelihood of our accuracy.

Statistics is the science that helps us to create knowledge based on the information attributed to the facts that make up our datasets.

Graphing is an approach that enables us to bring the data from the abstractness of a fact to the reality of a contextualized image. With a graph, we view the image and then interpret the meaning of the image from our understanding of the mathematical system that the image represents.

In applying statistics, and more specifically learning statistics, graphs are essential.

Creating a Visual Presentation of Your Data

In this section, we will organize data from various sources, and especially identify the specific variables that comprise

a data set and separate data into independent and dependent measures. The examples here will enable us to present data in a graphical format to convey our message to the reader by identifying the level of measurement for each of the variables within the data set and organize the data appropriately for the type of graph or table selected.

Graphing is a useful technique to illustrate:

- the shape of data sets relative to how scores are distributed
- relationships or associations between variables within or between data sets
- the magnitude of differences for numbers within and between datasets

In this section, we will use several different examples of graphing and charting features in SAS to create the appropriate visual presentation of variables in our dataset and thereby demonstrate when a graph of a particular type is most appropriate. In addition, we will work through examples that prepare data sets for graphing and charting by transposing from wide to narrow or narrow to wide, as well as adding specific features such as axis labels, legends, and colors to enhance your graphical presentation of variables in the data set.

Creating a Vertical Bar Chart to Represent John Snow's Natural Experiment

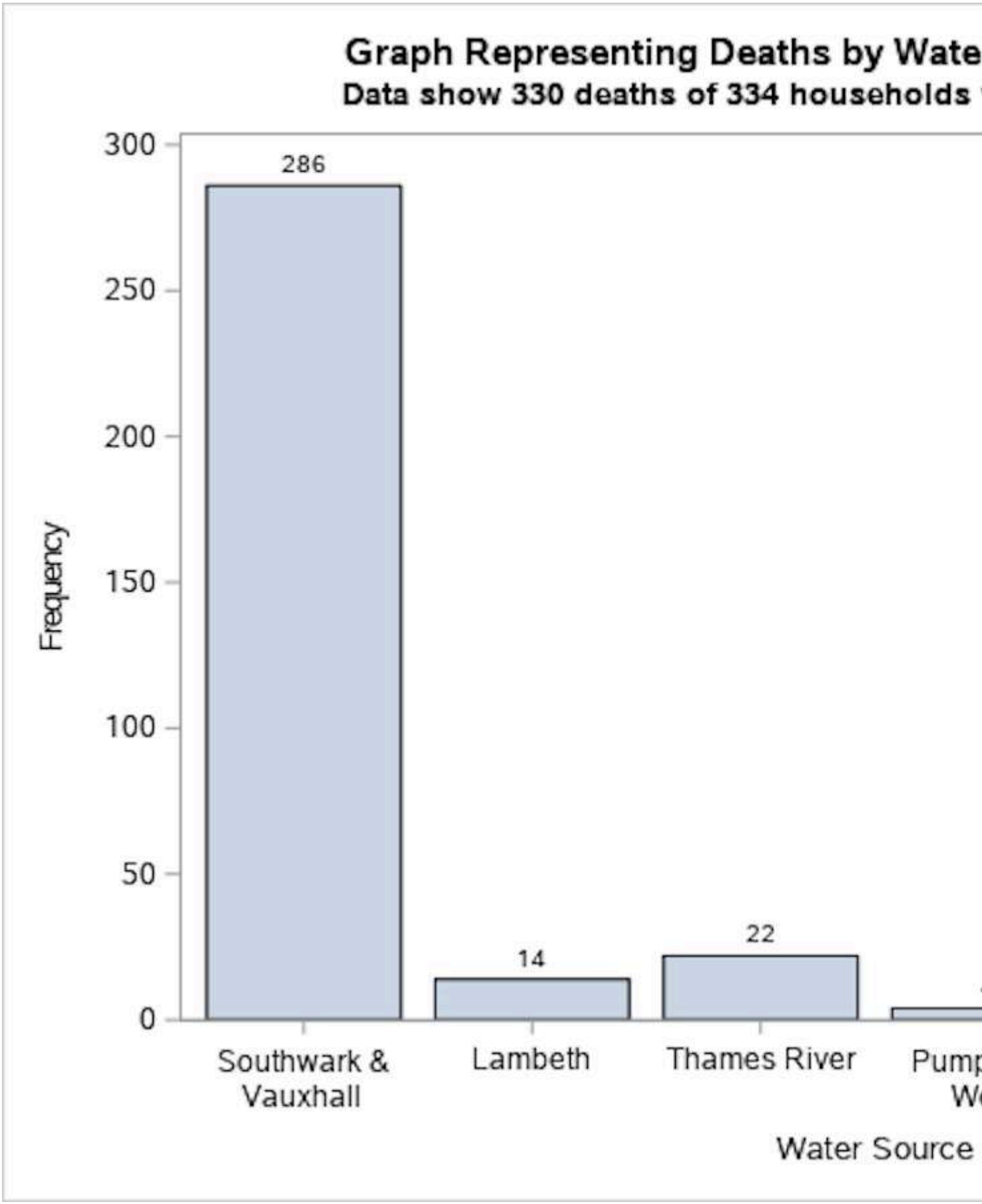
In this first example, we will use the PROC SGPLOT command to create a vertical bar graph to represent the data that John Snow reported for the water source by household in his 1854 surveillance during the London Cholera epidemic. The data are discrete frequencies of households which are then plotted against the source of drinking water for the household. The SAS code used to generate this vertical bar chart is presented below the image.

In this SAS graphing program we create a vertical bar graph to represent the data that John Snow reported for the water source by household in his 1854 surveillance during the London Cholera epidemic. The data are discrete frequencies of households and these are plotted against the source of drinking water for the household.

The SAS Code to generate the Vertical Bar Chart above.

```
options pagesize=55 linesize=120 center date;
PROC FORMAT;
VALUE SLICE 1 = 'Southwark &amp;amp; Vauxhall'
2 = 'Lambeth'
3 = 'Thames River'
4 = 'Pumps and Wells'
5 = 'Ditches'
6 = 'Unknown';
data snow1;
input source deaths ;
label Source = 'Water Source';
datalines;
1 286
2 14
3 22
4 04
5 04
6 04
;
run;
proc sgplot data=snow1; vbar source / freq=deaths datalabel;
```

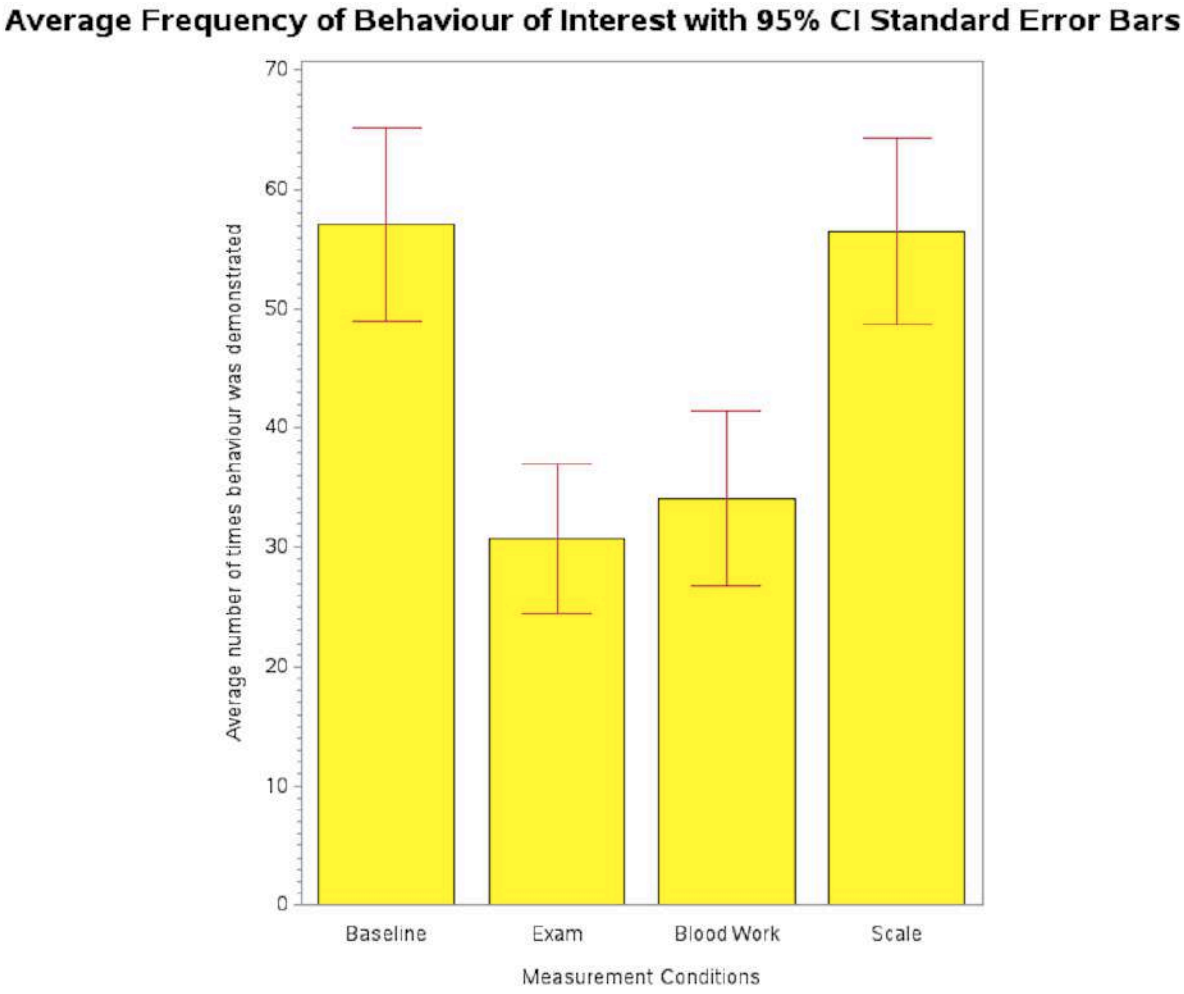
```
FORMAT source SLICE. ;
run;
```


Graph Representing Deaths by Wate
Data show 330 deaths of 334 households
300
250
200
150
100
50
0
Frequency
286
14
22
Southwark & Vauxhall
Lambeth
Thames River
Pump
W
Water Source

In the following program we added error bars based on 95% confidence interval calculations to each vertical bar. The SAS code is annotated with comments throughout.

SAS Program to Create Vertical Bar Chart with error bars

```
   proc format;
value cndfmt 1 = 'Baseline' 2 = 'Exam'  3 = 'Blood Work' 4 = 'Scale';
data behave;
input Code Cond bhvscr @@;
xvar=cond; yvar=bhvscr;
label xvar='Measurement Conditions';
label yvar='Average number of times behaviour was demonstrated';
datalines;
01 1 79 01 2 54 01 3 51 01 4 85 02 1 21 02 2 15 02 3 23 02 4 80
03 1 37 03 2 14 03 3 18 03 4 38 04 1 61 04 2 21 04 3 13 04 4 79
05 1 32 05 2 30 05 3 34 05 4 58 06 1 60 06 2 30 06 3 15 06 4 50
07 1 78 07 2 53 07 3 67 07 4 53 08 1 67  08 2 42 08 3 47 08 4 48 09 1 41
09 2 10 09 3 28 09 4 28 10 1 72 10 2 52 10 3 33 10 4 24 11 1 62
11 2 21 11 3 60 11 4 47 12 1 44 12 2 46 12 3 54 12 4 52 13 1 32
13 2 11 13 3 25 13 4 32 14 1 39 14 2 37 14 3 12 14 4 36 15 1 55 15 2 20
15 3 23 15 4 49 16 1 62 16 2 22  16 3 28 16 4 49 17 1 83 17 2 34
17 3 22 17 4 43 18 1 86 18 2 14 18 3 47 18 4 80 19 1 54 19 2 47 19 3 77
19 4 44 20 1 76 20 2 57 20 3 24 20 4 88 21  1 56 21  2 14 21 3 18 21 4 59
22 1 37 22 2 43 22 3 39 22 4 75 30 1 28 30 2 13 30 3 15 30 4 81 31 1 94
31 2 31 31 3 69 31 4 90 52 1 90 51 2 55 52 3 26 52 4 70 53 1 48 53 2 14
53 3 29 53 4 53 54 1 47 54 2 29 54 3 24 54 4 34
;
/* Define the axis characteristics */
axis1 offset=(0,70) minor=none;  axis2 label=(angle=90);
pattern1 color = yellow;
   /* The term pattern1 refers to the first item to be graphed. If there were two variables being graphed then
we we use pattern1 and pattern3.  */
proc sort; by xvar;
proc gchart data = behave;
vbar xvar/ type=MEAN errorbar=BOTH clm=95
sumvar=yvar discrete raxis=axis2 cerror=crimson cr=biv;
format xvar cndfmt.;
/* Define the title */
title1 'Average Frequency of Behaviour of Interest with 95% CI Standard Error Bars ';
run;
```

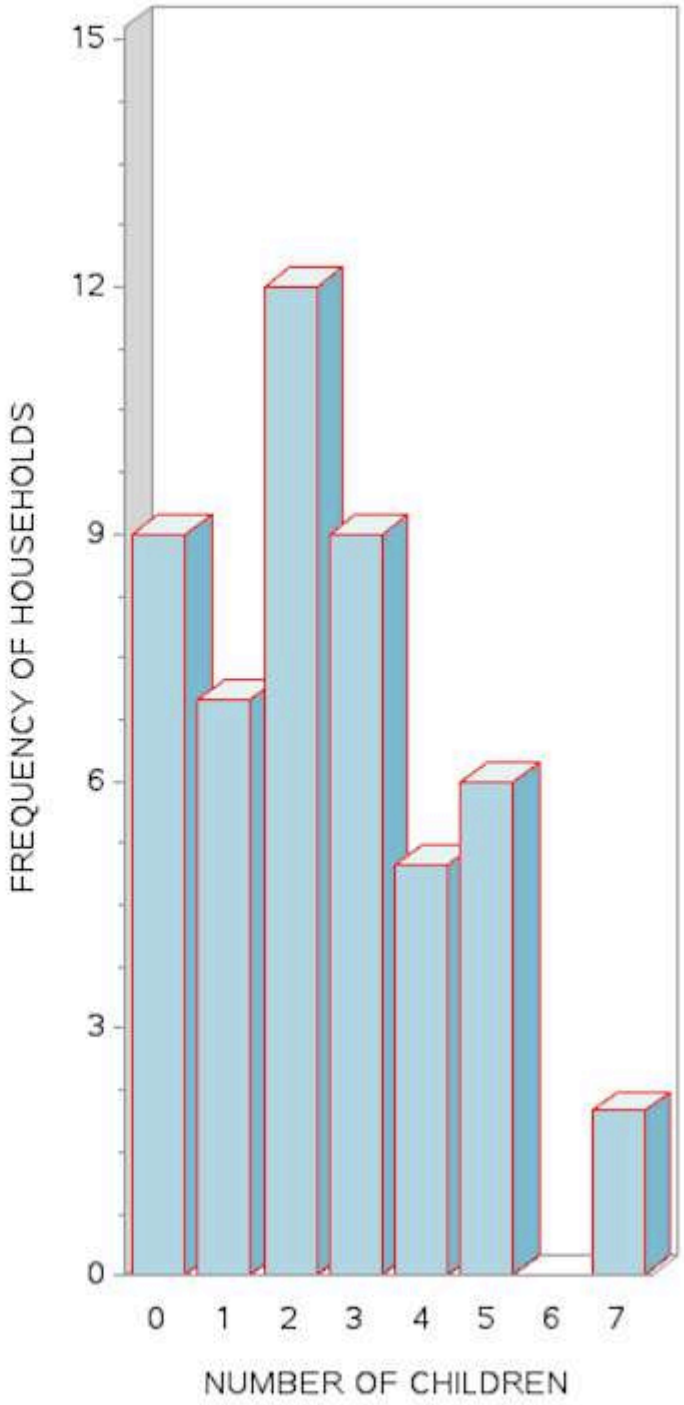


More Simple Barcharts — Graphing data as a Frequency Distribution Bar Chart

In the following examples, we use SAS commands to create a three-dimensional vertical bar chart and a horizontal bar chart with a frequency table of the data. In this SAS code, we include formatting commands for the graphical output – defining the characteristics of each axis – prior to having the SAS program read the data set.

Using GCHART to Produce A Vertical Bar Chart

VERTICAL BAR CHART NUMBER OF CHILDREN IN EACH



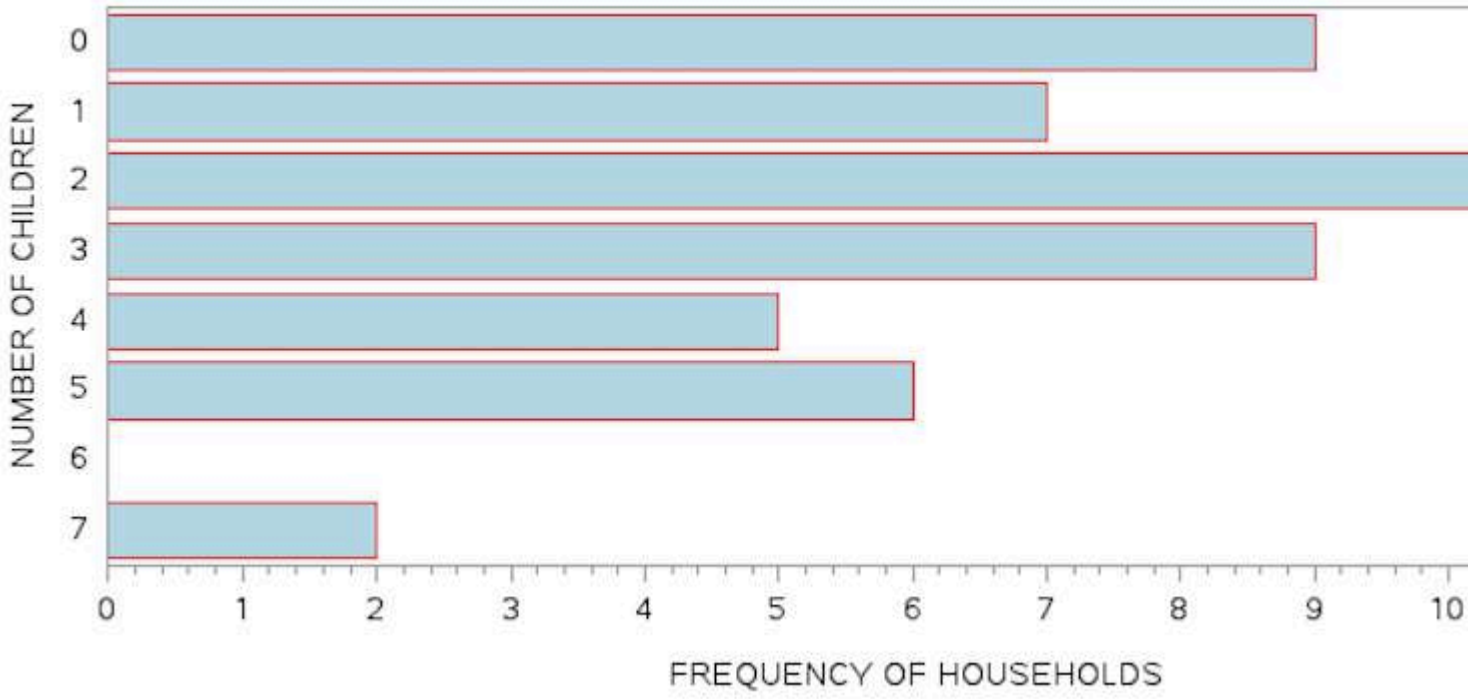
```
DATA FAMILIES;
INPUT NKIDS HSEHLD;
/* DEFINE THE AXIS CHARACTERISTICS */
AXIS1 LABEL=("NUMBER OF CHILDREN")
VALUE=(JUSTIFY=CENTER);
AXIS2 LABEL=(ANGLE=90 "FREQUENCY OF HOUSEHOLDS")
ORDER=(0 TO 15 BY 3)
MINOR=(N=3);
AXIS3 LABEL=(ANGLE=90 "NUMBER OF CHILDREN");
AXIS4 LABEL=("FREQUENCY OF HOUSEHOLDS") ;

DATALINES;
00 9
01 7
02 12
03 9
04 5
05 6
06 0
07 2
;
PROC GCHART DATA=FAMILIES;
VBAR3D NKIDS/SUMVAR=HSEHLD TYPE=SUM DISCRETE
COUTLINE=RED WOUTLINE=1 WIDTH=3 MAXIS=AXIS1 RAXIS=AXIS2;
TITLE1 'VERTICAL BAR CHART NUMBER OF CHILDREN IN EACH HOUSEHOLD';
PATTERN1 COLOR = LIGHTBLUE;
RUN;
   PROC GCHART DATA=FREQ4_3;
HBAR NKIDS/DISCRETE SUMVAR= HSEHLD
TYPE=SUM DISCRETE COUTLINE=RED WOUTLINE=1 WIDTH=2
MAXIS=AXIS3 RAXIS=AXIS4;
Title1 'HORIZONTAL BAR CHART NUMBER OF CHILDREN IN EACH HOUSEHOLD';
TITLE2 'INCLUDES FREQUENCY VALUES AT END OF EACH BAR';
PATTERN1 COLOR = LIGHTBLUE;
RUN;
```

The essential SAS processing command to produce the vertical bar chart is PROC GCHART. However, the important commands to discriminate the independent and dependent variables are given in the command line: VBAR3D NKIDS/ SUMVAR=HSEHLD TYPE=SUM DISCRETE. Here we tell SAS to read the variable NKIDS as the categorical independent variable, while HSEHLD is the dependent variable and is read by the option SUMVAR= HSEHLD. We include the second option TYPE=SUM to indicate that the values entered are actually the sum scores for each category of the independent variable.

Using the same data from the SAS program above and adding two new AXIS labels we can generate a horizontal bar chart with the frequency values included at the end of each horizontal bar. Notice in both the vertical and horizontal bar charts, the length of the bar is proportional to the value of the frequency.

Horizontal Bar Chart with Frequency Values Included




```
DATA FAMILIES;
INPUT NKIDS HSEHLD @@;
/* DEFINE THE AXIS CHARACTERISTICS */

AXIS3 LABEL=(ANGLE=90 "NUMBER OF CHILDREN");
AXIS4 LABEL=("FREQUENCY OF HOUSEHOLDS") ;
   DATALINES;
00 9 01 7 02 12 03 9 04 5 05 6 06 0 07 2
;
PROC GCHART DATA=FAMILIES;
HBAR NKIDS/DISCRETE SUMVAR= HSEHLD
TYPE=SUM DISCRETE COUTLINE=RED WOUTLINE=1 WIDTH=2
MAXIS=AXIS3 RAXIS=AXIS4;
Title1 'HORIZONTAL BAR CHART NUMBER OF CHILDREN IN EACH HOUSEHOLD';
TITLE2 'INCLUDES FREQUENCY VALUES AT END OF EACH BAR';
PATTERN1 COLOR = LIGHTBLUE;
RUN;
```

Notice in the input statement, the variables are defined and two @ symbols are used to hold the cursor at the line until all values are entered in sequence.

INPUT NKIDS HSEHLD @@;

This style for data entry economizes space in programming.

In the following example, we return to the HDX dataset to observe the total number of cases of the Ebola virus across selected countries. These data are based on the actual reports of cases and deaths related to the 2014 West Africa Ebola Outbreak.

Sample Data Of Total Cases Of Ebola Virus Across Selected Countries

Country	Case definition	Total cases	Total deaths	Country report date
Guinea	Confirmed	2384	1422	2014-12-27
	Probable	275	275	
	Suspected	36	0	
	All	2695	1697	
Liberia	Confirmed	3108	..	2014-12-24
	Probable	1773	..	
	Suspected	3096	..	
	All	7977	3413	
Sierra Leone	Confirmed	7326	2366	2014-12-27
	Probable	287	208	
	Suspected	1796	158	
	All	9409	2732	

The SAS program and corresponding output from the analysis is presented below. Notice that only the data for *confirmed*, *probable* and *suspected* cases are being used in the dataset. These data represent the summary of counts whereby the units of measurement are the total number of cases and the total number of deaths.

SAS Program to Create a Frequency Distribution for Ebola Outbreak

```
   OPTIONS PAGESIZE=60 LINESIZE=80;
DATA GRAPH1;
INPUT COUNTRY $ 1-12 DEF $ 15-23 CASES 26-29;
LABEL DEF='DEFINITION OF CASES';
DATALINES;
GUINEA          CONFIRMED  2384
GUINEA          PROBABLE    275
GUINEA          SUSPECTED    36
LIBERIA          CONFIRMED  3108
LIBERIA          PROBABLE   1773
LIBERIA          SUSPECTED  3096
SIERRA LEONE  CONFIRMED  7326
SIERRA LEONE  PROBABLE    287
SIERRA LEONE  SUSPECTED  1796
;
PROC SORT; BY COUNTRY;
PROC FREQ; TABLES COUNTRY/OUT=CASEPCT; WEIGHT CASES;
RUN;
```

The output from the PROC FREQ procedure is shown here.

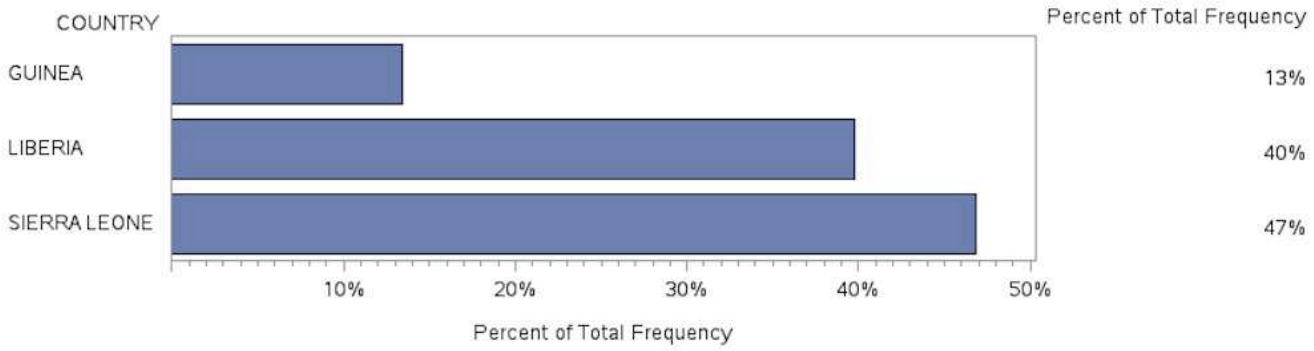
COUNTRY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
GUINEA	2695	13.42	2695	13.42
LIBERIA	7977	39.72	10672	53.14
SIERRA LEONE	9409	46.86	20081	100.00

In the code above we produce an output file that represents the percent value of the cases based on the sum of cases in each country. For example, all of the cases for GUINEA regardless of whether the cases were PROBABLE, SUSPECTED, or CONFIRMED, equal 2695 which represents 13.42 percent of the total number of cases. The total number of cases across all countries is reported in the last row of the **Cumulative Frequency** column and is 20081.

Below, the output data file (DATA=CASEPCT) is used in a PROC GCHART procedure to produce a horizontal bar chart using the SUMVAR option with the data from the PERCENT column.

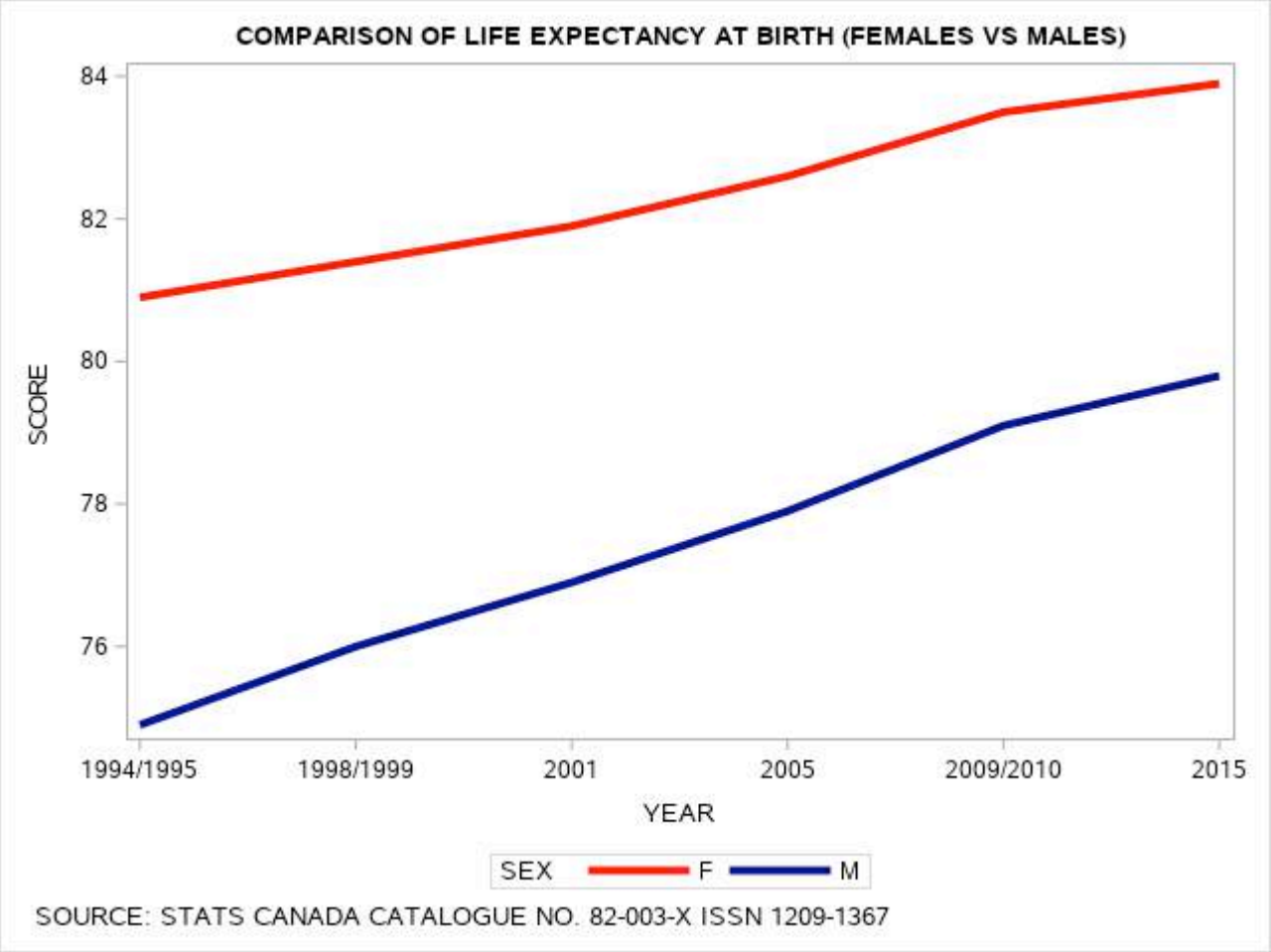
```
PROC FORMAT; PICTURE PCTFMT (ROUND) 0-HIGH='000%';

PROC GCHART DATA=CASEPCT; HBAR COUNTRY/ SUMVAR = PERCENT;
   FORMAT PERCENT PCTFMT.;
   RUN;
```



Creating a Line Graph to Summarize Data

In this program, we will use the SAS PROC GPLOT functions to observe "at a glance" a comparison of the unadjusted differences in life expectancy at birth for males versus females in Canada, based on data reported since 1994. The data used in this example range from an initial value of 74.9 years for males and 80.9 years for females in 1994, to life expectancy scores of 79.8 years for males and 83.9 years for females, in 2015. Here we see that at each year of reporting life expectancy estimates, the females are predicted to live longer than males, on average.



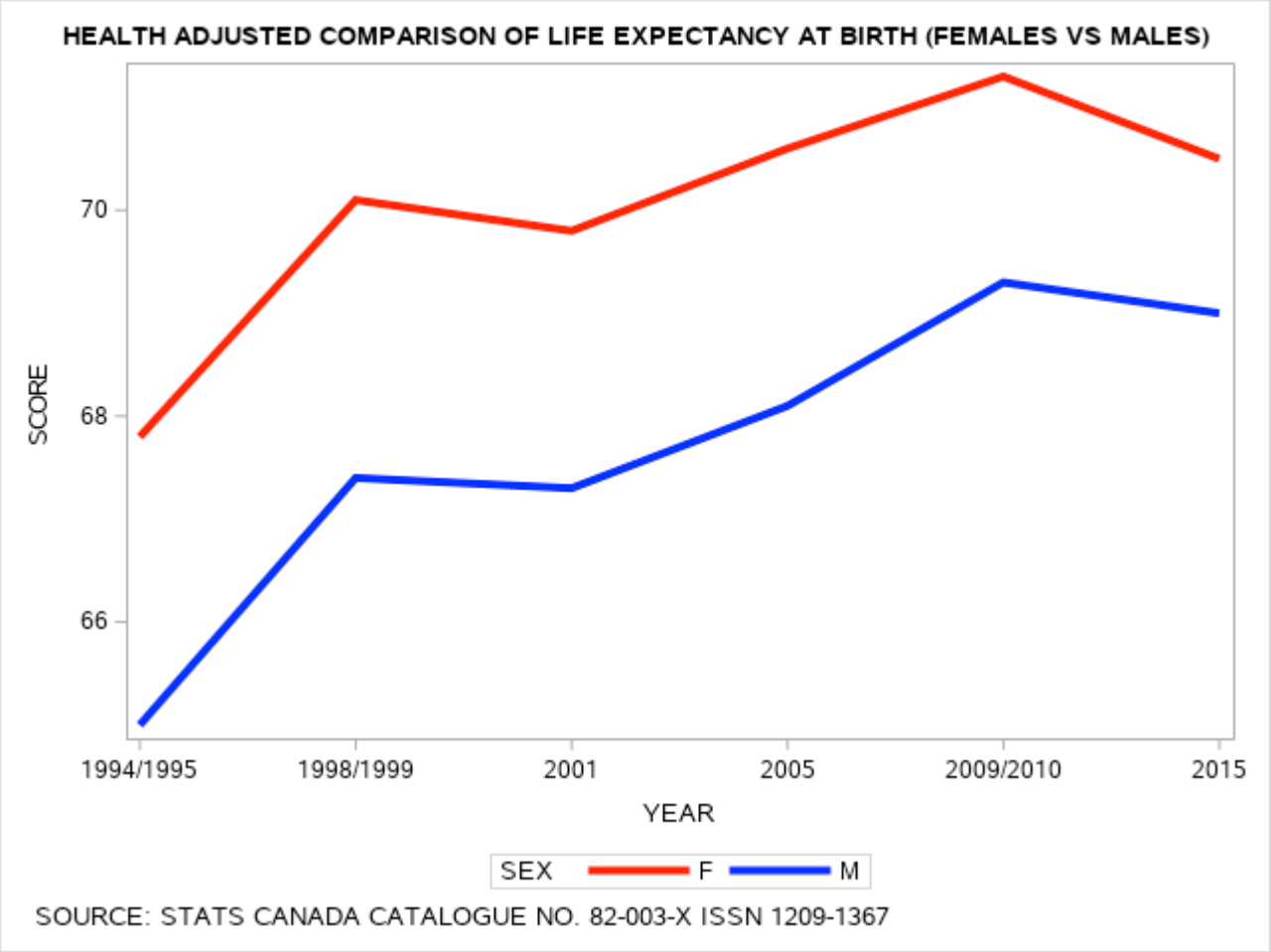
SAS SGPLOT to Produce Comparison of Life Expectancy Scores

```
   PROC FORMAT;
VALUE $SXFMT 'F'='FEMALE' 'M'='MALE';
VALUE YRFMT 1='1994/1995' 2='1998/1999' 3='2001' 4='2005' 5='2009/2010' 6='2015';
VALUE SRCFMT 1='UNADJUSTED LIFE EXPECTANCY' 2='HEALTH ADJUSTED LIFE EXPECTANCY';
   LABEL SCORE= 'LIFE EXPECTANCY AT BIRTH';
LABEL YEAR= 'YEAR OF REPORTING';
   DATA CH7FIG1;
INPUT ID SEX $ YEAR SOURCE SCORE @@;
DATALINES;
01 M 1 1 74.9 02 M 2 1 76.0 03 M 3 1 76.9 04 M 4 1 77.9 05 M 5 1 79.1  M 06 6 1 79.8 07 M 1 2 65.0 08 M 2 2 67.4 09 M
3 2 67.3 10 M 4 2 68.1
```

```
11 M 5 2 69.3 12 M 6 2 69.0 13 F 1 1 80.9 14 F 2 1 81.4 15 F 3 1 81.9
16 F 4 1 82.6 17 F 5 1 83.5 18 F 6 1 83.9 19 F 1 2 67.8 20 F 2 2 70.1
21 F 3 2 69.8 22 F 4 2 70.6 23 F 5 2 71.3 24 F 6 2 70.5
;
TITLE2 'COMPARISON OF LIFE EXPECTANCY AT BIRTH (FEMALES VS MALES)';
   FOOTNOTE1 J=L " SOURCE: STATS CANADA CATALOGUE NO. 82-003-X ISSN 1209-1367";
   AXIS1 ORDER=(1990 TO 2015 BY 5) OFFSET=(2,2) LABEL=NONE
MAJOR=(HEIGHT=2) MINOR=(HEIGHT=1) ;
   AXIS2 ORDER=(50 TO 100 BY 5) OFFSET=(0,0) LABEL=NONE
MAJOR=(HEIGHT=2) MINOR=(HEIGHT=1);
   LEGEND1 LABEL=NONE POSITION=(TOP CENTER INSIDE)
MODE=SHARE;
   RUN;
PROC SORT; BY SEX;
PROC SGPLOT;
   SERIES X = YEAR Y = SCORE/GROUP=SEX lineattrs=(thickness=4);
   XAXIS TYPE = DISCRETE;
styleattrs datacontrastcolors=(RED NAVY)
datalinepatterns=(SOLID);
   WHERE SOURCE=1 ;
   FORMAT YEAR YRFMT. SOURCE SRCFMT. ;
   RUN;
```

Adding the WHERE command to restrict output

In the following line graph, we observe the Health Adjusted Life Expectancy, also referred to as HALE data comparison between males and females changes the contours of the lines for the predicted values of the female and male response data. Again these data range from first reports in 1994 taken from the National Population Health Survey and the Canadian Census in 1993 to 1995 to data from the Canadian Community Health Survey, as well as the NPHS and Census up to and including 2015 (Bushnik, Tjepkema, Martel, 2018)[1].



The SAS program to produce the line graph above includes the **PROC SGPLOT** statement and the command **WHERE SOURCE=1**; This restricts the processing of the graphing procedure to only select the unadjusted life expectancy values from the dependent variable **SCORE**. When we change the command **WHERE SOURCE=2**; then we change the output to only consider health adjusted values from the dependent variable SCORE.

Here we use the PROC FORMAT feature to ensure that the data are converted to explanatory labels and these labels are included in the graphs.

SAS SGPLOT to produce Health Adjusted Comparison of Life Expectancy Scores

This program is incorporating the WHERE command to restrict output to a subgroup.

```
PROC SORT; BY SEX;
PROC SGPLOT; SERIES X = YEAR Y = SCORE/GROUP=SEX
lineattrs=(thickness=4);XAXIS TYPE = DISCRETE;
styleattrs datacontrastcolors=(RED BLUE)
```

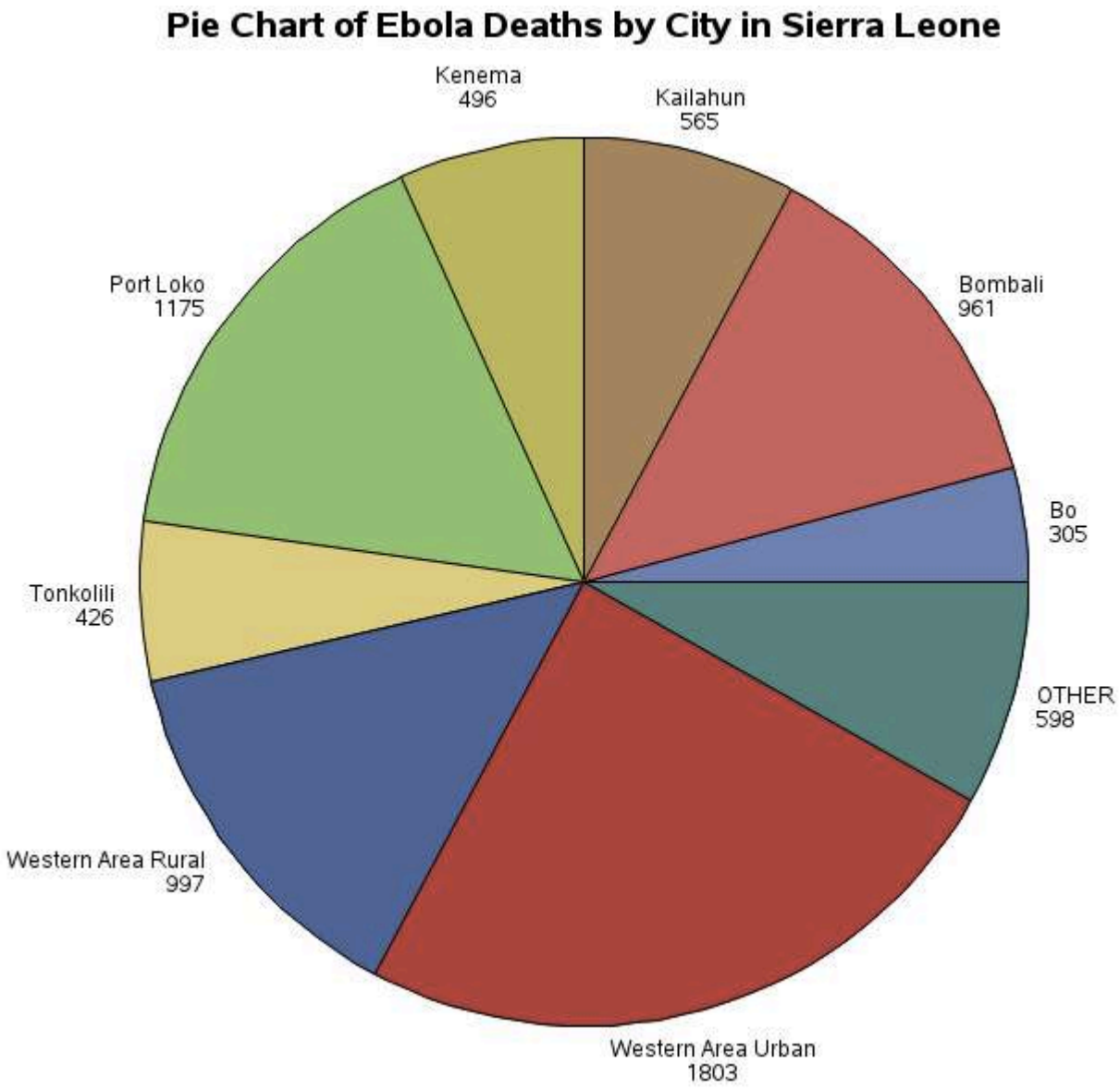
```
datalinepatterns=(SOLID);
WHERE SOURCE=2 ;
   FORMAT YEAR YRFMT. SOURCE SRCFMT. ;
   RUN;
```

[1]Bushnik, T., Tjepkema, M., & Martel, L., Health-adjusted life Expectancy in Canada, Statistics Canada. Catalogue no. 82-003-X ISSN 1209-1367

Creating a Pie Chart to Represent Summary Data

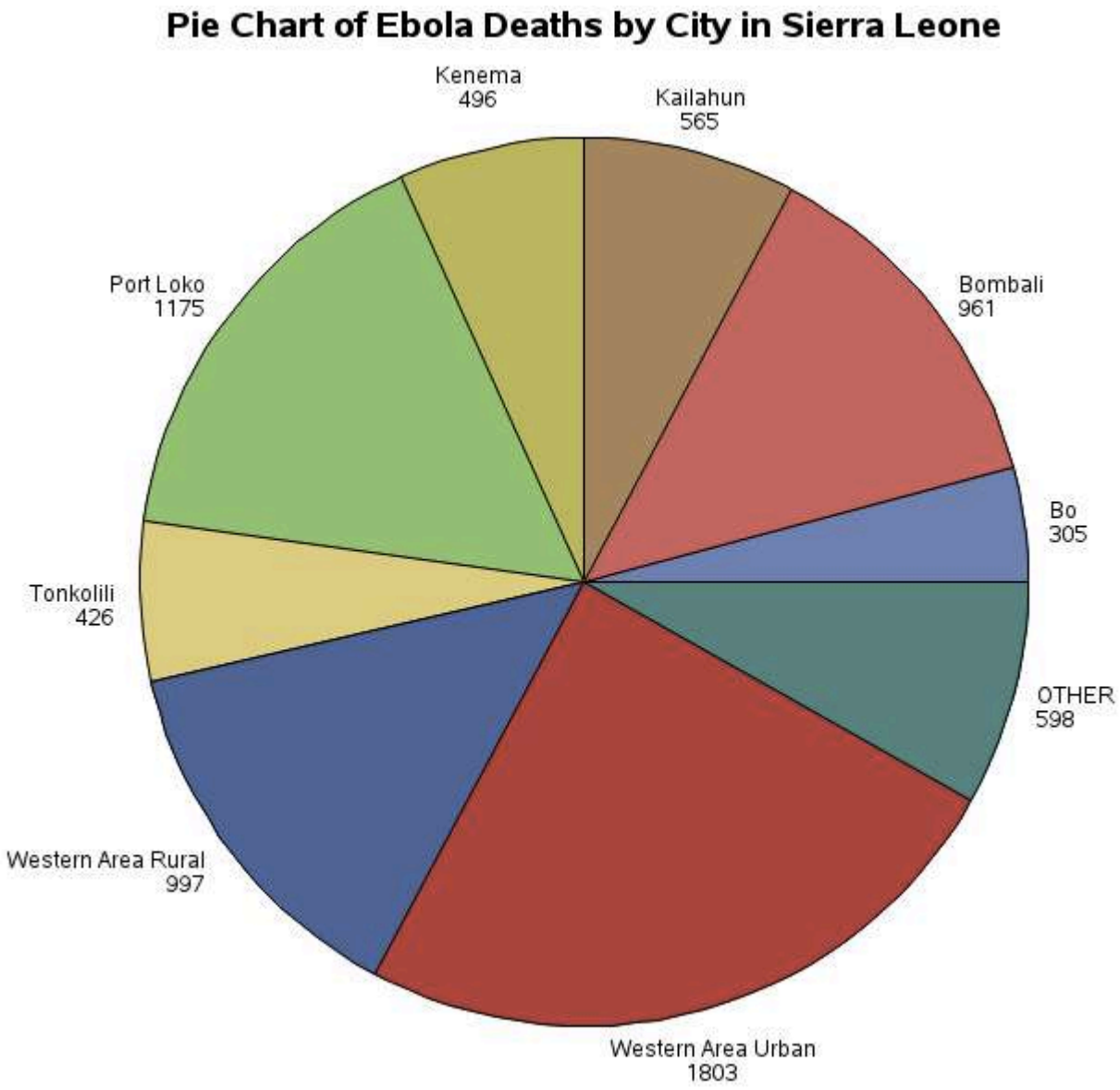
In the following example, we present a pie chart of the data from The Humanitarian Data Exchange (url: https://data.hdx.rwlabs.org/) a project from the United Nations Office for the Coordination of Humanitarian Aid (url: http://www.unocha.org/).

On January 15th, 2016 the World Health Organization declared the country of Sierra Leone as Ebola-free. However, by that time Sierra Leone had recorded approximately 4000 deaths from the Ebola Virus. In this second example, we will generate a pie chart. The data are based on confirmed cases of Ebola for Sierra Leone by region from 2014 to December 28, 2014. The data represent the cumulative deaths since the recognized beginning of the Ebola virus outbreak in April 2014.




```
/* ***************************************************************
* SAS CODE TO PRODUCE PIE CHART FOR EBOLA RELATED
* DEATHS IN CITIES OF SIERRA LEONE
* BE SURE TO CHECK COLUMN ALIGNMENT
*************************************************************** */.

OPTIONS PAGESIZE=55 LINESIZE=120 CENTER DATE;
LIBNAME SAMPLE '/HOME/WMONTELPARE/MN636_EXAMPLES/';
DATA SAMPLE.PIE1;
INPUT CITY $ 1-19 @23 CONFIRM D21DAYS 29-31;
DATALINES;
```




```
WESTERN AREA URBAN 1803 400
WESTERN AREA RURAL 997 112
KAMBIA 108 22
PORT LOKO 1175. 219
TONKOLILI 426 41
KONO 176 70
KAILAHUN 565 3
KENEMA 496 2
PUJEHUN 31 0
KOINADUGU 97 11
BO 305 36
BONTHE 5 1
BOMBALI 961 81
MOYAMBA 181 11

;
RUN;

TITLE1 'PIE CHART OF EBOLA DEATHS BY CITY IN SIERRA LEONE';
PROC GCHART DATA=SAMPLE.PIE1;
PIE CITY / SUMVAR=CONFIRM NOHEADING;
RUN;
```

The data for this example are taken from the HDX: The Humanitarian Data Exchange to represent deaths from the Ebola outbreak in Sierra Leone in 2014.

Producing Bubble Plots

SOURCE: HDX: The Humanitarian Data Exchange [1]

In the following example data set the cumulative number of health-care workers deaths by Ebola Disease Virus are reported. These data were extracted from WHO: Ebola Response Roadmap Situation Reports, the data are based on extraction from data reported on 24 December 2014. Here we can plot the total deaths from these data by country, and within each country by month and use appropriate axes titles and legend. The data are presented first in the table below and then as two separate bubble plots. The size of the bubbles represents the frequency value for the total number of deaths reported.

Number Of Health-Care Workers Deaths By Ebola Disease Virus (Sept 2014 - Dec 2014)

Country	Total deaths	Month	Country	Total deaths	Month	Country	Total deaths	Month
Guinea	27	Sept	Sierra Leone	81	Sept	Liberia	103	Oct
Liberia	81	Sept	Guinea	35	Sept	Nigeria	5	Oct
Sierra Leone	31	Sept	Liberia	95	Sept	Sierra Leone	95	Oct
Guinea	30	Sept	Nigeria	5	Sept	Guinea	43	Oct
Liberia	85	Sept	Sierra Leone	81	Sept	Liberia	123	Oct
Nigeria	5	Sept	Guinea	40	Oct	Nigeria	5	Oct
Sierra Leone	31	Sept	Liberia	96	Oct	Sierra Leone	101	Oct
Guinea	35	Sept	Nigeria	5	Oct	Guinea	46	Nov
Liberia	87	Sept	Sierra Leone	95	Oct	Liberia	157	Nov
Nigeria	5	Sept	Guinea	41	Oct	Nigeria	5	Nov
Guinea	56	Nov	Mali	2	Nov	Sierra Leone	102	Nov
Liberia	172	Nov	Guinea	59	Nov	Guinea	55	Nov
Sierra Leone	105	Nov	Liberia	174	Nov	Liberia	170	Nov
Nigeria	5	Nov	Sierra Leone	106	Nov	Nigeria	5	Nov
Guinea	51	Nov	Nigeria	5	Nov	Sierra Leone	104	Nov
Liberia	162	Nov	Guinea	62	Dec	Guinea	62	Dec
Nigeria	5	Nov	Liberia	174	Dec	Liberia	174	Dec
Sierra Leone	102	Nov	Sierra Leone	106	Dec	Sierra Leone	106	Dec
Nigeria	5	Dec	Liberia	177	Dec	Nigeria	5	Dec
Mali	2	Dec	Sierra Leone	110	Dec	Mali	2	Dec
Guinea	72	Dec	Nigeria	5	Dec	Guinea	72	Dec
Liberia	177	Dec	Mali	2	Dec	Liberia	177	Dec
Sierra Leone	109	Dec	Sierra Leone	110	Dec	Sierra Leone	109	Dec
Nigeria	5	Dec	Nigeria	5	Dec	Nigeria	5	Dec
Mali	2	Dec	Liberia	177	Dec	Mali	2	Dec
Guinea	72	Dec	Mali	2	Dec	Guinea	72	Dec

Data Source: (url: https://data.hdx.rwlabs.org/) a project from the United Nations Office for the Coordination of Humanitarian Aid (url: http://www.unocha.org/)

The SAS program to analyze these data is presented below. Notice that the dataset presented above used three columns: Country, Total Deaths, and Months, which are repeated three times, using the following input statement. The double trailing @@ symbols hold the pointer at the end of the line to ensure that the data read as three variables repeated three times.

Sample Code:

```
INPUT COUNTRY $ TOTDTH MONTH $ @@;
```

In this way, the SAS program reads the data and produces the output for the entire dataset. Notice that we precede the input statement by declaring the length of the contents of the variable COUNTRY to be **more than 12 characters** in length.

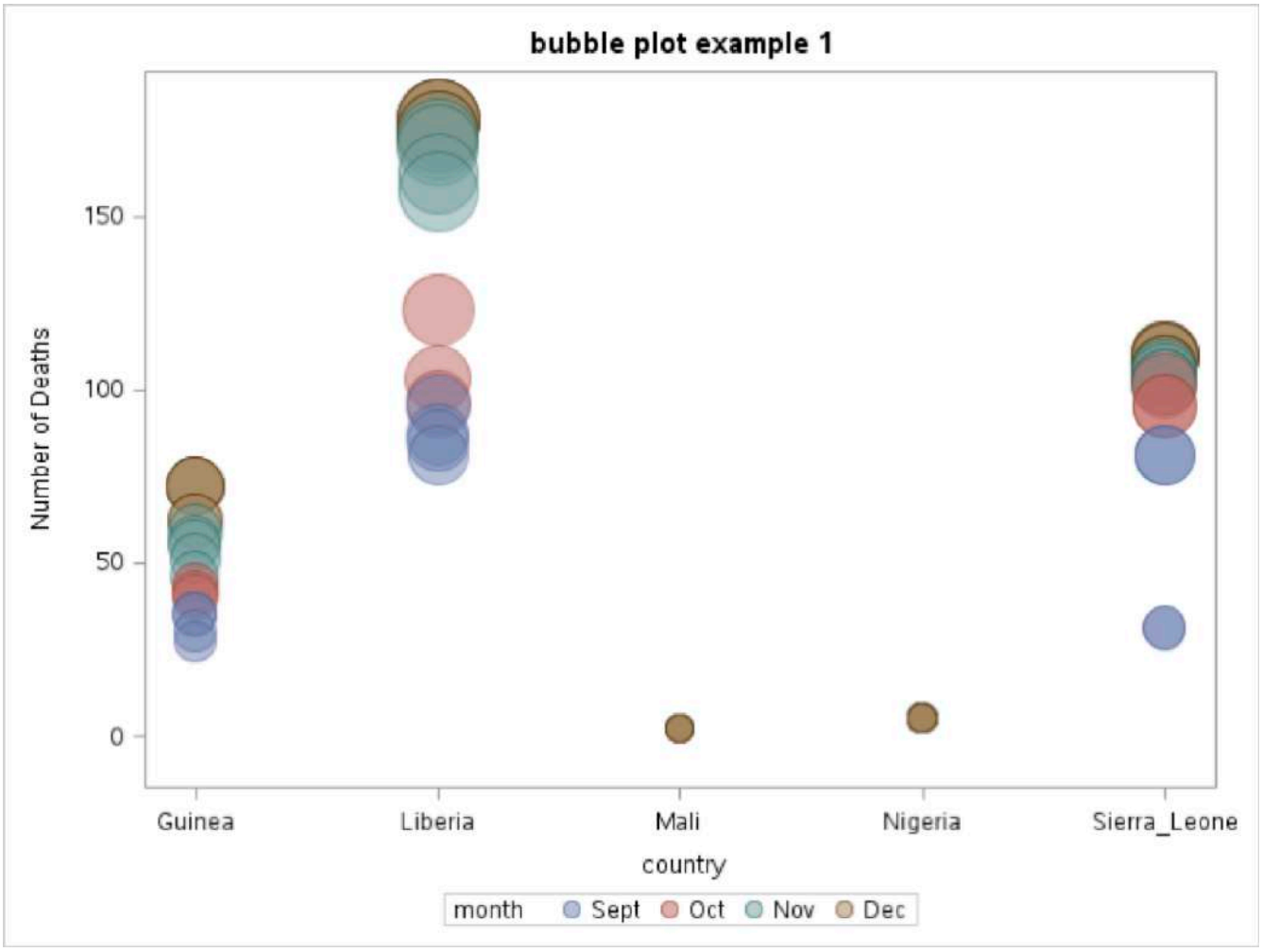
SAS code to Produce Bubble Chart

```
   OPTIONS PAGESIZE=55 LINESIZE=120 CENTER DATE;
DATA GRAPH2;
LENGTH COUNTRY $12.;
INPUT COUNTRY $ TOTDTH MONTH $ @@;
LABEL TOTDTH ='NUMBER OF DEATHS';DATALINES;
<DATA GOES HERE>
   Sample of the raw data:
   Guinea 27 Sept  Sierra_Leone 81 Sept Liberia  103 Oct  Liberia 81 Sept  Guinea  35 Sept Nigeria 5 Oct
Sierra_Leone 31 Sept  Liberia 95 Sept Sierra_Leone  95 Oct
   …
   ;
RUN;
PROC SORT; BY COUNTRY;
PROC FREQ DATA=GRAPH2;WEIGHT TOTDTH; TABLES MONTH*COUNTRY;
RUN;
* NOTICE THE WEIGHT STATEMENT IS USED WHEN THE RAW DATA ARE SUMS;
PROC SGPLOT DATA=GRAPH2;
TITLE1 'BUBBLE PLOT';
TITLE2 'EXAMPLE 1: TOTAL DEATHS BY COUNTRY';
BUBBLE X = COUNTRY Y = TOTDTH SIZE = TOTDTH / GROUP = MONTH TRANSPARENCY = 0.5;
FOOTNOTE1 J=L "SOURCE: HTTPS://DATA.HUMDATA.ORG/DATASET/NUMBER-OF-HEALTH-CARE-WORK-
ERS-DEATHS-BY-EDV";PROC SGPLOT DATA=GRAPH2;
TITLE1 'BUBBLE PLOT';
TITLE2 'EXAMPLE 2: TOTAL DEATHS BY MONTH';
BUBBLE X = MONTH Y = TOTDTH SIZE = TOTDTH / GROUP = COUNTRY TRANSPARENCY = 0.5; YAXIS
GRID    ;
RUN;
```

The summary frequency distribution is presented here first.

Month	Guinea	Liberia	Mali	Nigeria	Sierra_Leone	Total
Sept	f= 127 % total = 2.41 row % = 17.79 col % = 13.66	f= 1056 % total =20.02 row % = 48.89 col % = 41.23	f= 12 % total =0.23 row % = 0.56 col % = 85.71	f= 30 % total =0.57 row % = 1.39 col % = 35.29	f= 650 % total =12.32 row % = 30.09 col % = 38.60	f= 2160 % total =40.96
Oct	f= 124 % total = 2.35 row % = 16.49 col % = 13.33	f= 322 % total = 6.11 row % = 42.82 col % = 12.57	f= 0 % total = 0.00 row % = 0.00 col % = 0.00	f= 15 % total = 0.28 row % = 1.99 col % = 17.65	f= 291 % total = 5.52 row % = 38.70 col % = 17.28	f= 752 row % = 14.26
Nov	f= 267 % total = 5.06 row % = 16.20 col % = 28.71	f= 835 % total = 15.83 row % = 50.67 col % = 32.60	f= 2 % total = 0.04 row % = 0.12 col % = 14.29	f= 25 % total = 0.47 row % = 1.52 col % = 29.41	f= 519 % total = 9.84 row % = 31.49 col % = 30.82	f= 1648 row % =31.25
Dec	f= 412 % total = 7.81 row % = 19.07 col % = 44.30	f= 1056 % total = 20.02 row % = 48.89 col % = 41.23	f= 12 % total = 0.23 row % = 0.56 col % = 85.71	f= 30 % total = 0.57 row % = 1.39 col % = 35.29	f= 650 % total = 12.32 row % = 30.09 col % = 38.60	f= 2160 row % = 40.96
Total	f= 930 col % = 17.63	f= 2561 col % = 48.56	f= 14 col % = 0.27	f= 85 col % = 1.61	f= 1684 col % = 31.93	f= 5274 100.00

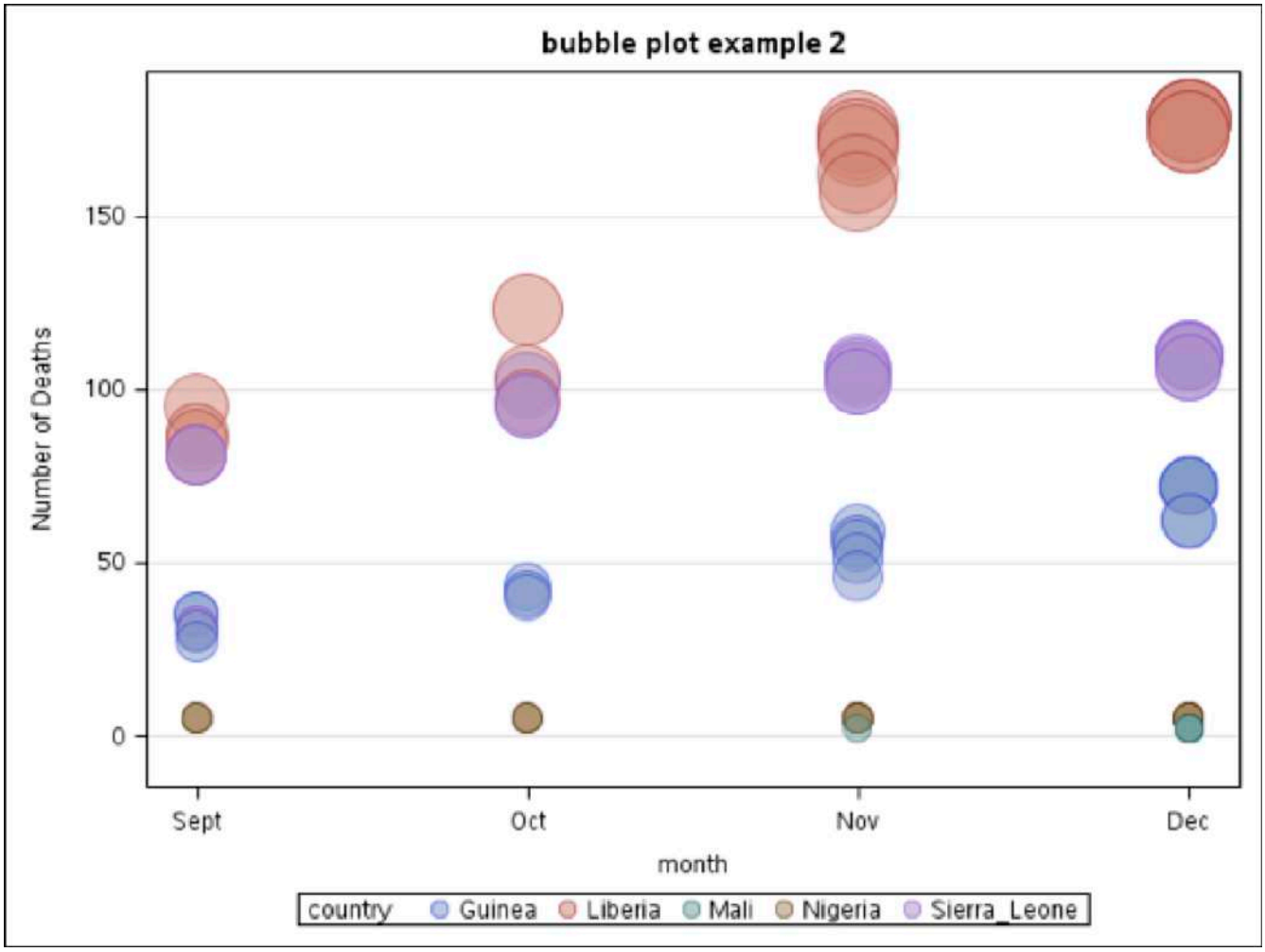
Bubble plots can be used to illustrate the distribution of outcomes within specific groups. In the following two graphs the data from the summary frequency table of *month by country* above, which showed deaths within the countries monitored across months are presented using two different grouping strategies. In the first example (bubble plot example 1) the data showing the number of deaths (Y-axis) are separated using countries as the main X-Axis variable and months as the grouping variable.



The specific SAS code is:

```
   PROC SGPLOT DATA=GRAPH2;
TITLE1 'BUBBLE PLOT';
TITLE2 'EXAMPLE 1: TOTAL DEATHS BY COUNTRY';
BUBBLE X = COUNTRY Y = TOTDTH SIZE = TOTDTH / GROUP = MONTH TRANSPARENCY = 0.5;FOOT-
NOTE1 J=L "SOURCE: HTTPS://DATA.HUMDATA.ORG/DATASET/NUMBER-OF-HEALTH-CARE-WORKERS-
DEATHS-BY-EDV";
```

In the second example (bubble plot example 2) the data showing the number of deaths (Y-axis) are separated using months as the main X-axis variable and country in which the deaths occurred is the grouping variable.



The specific SAS code is presented here. Notice we did not need to repeat the footnote statement from Bubble Plot 1 for it to be included in Bubble Plot 2 because the RUN; statement was held until the end of the program.

```
   PROC SGPLOT DATA=GRAPH2;
TITLE1 'BUBBLE PLOT';
TITLE2 'EXAMPLE 2: TOTAL DEATHS BY MONTH';
BUBBLE X = MONTH Y = TOTDTH SIZE = TOTDTH / GROUP = COUNTRY TRANSPARENCY = 0.5; YAXIS
GRID    ;
RUN;
```

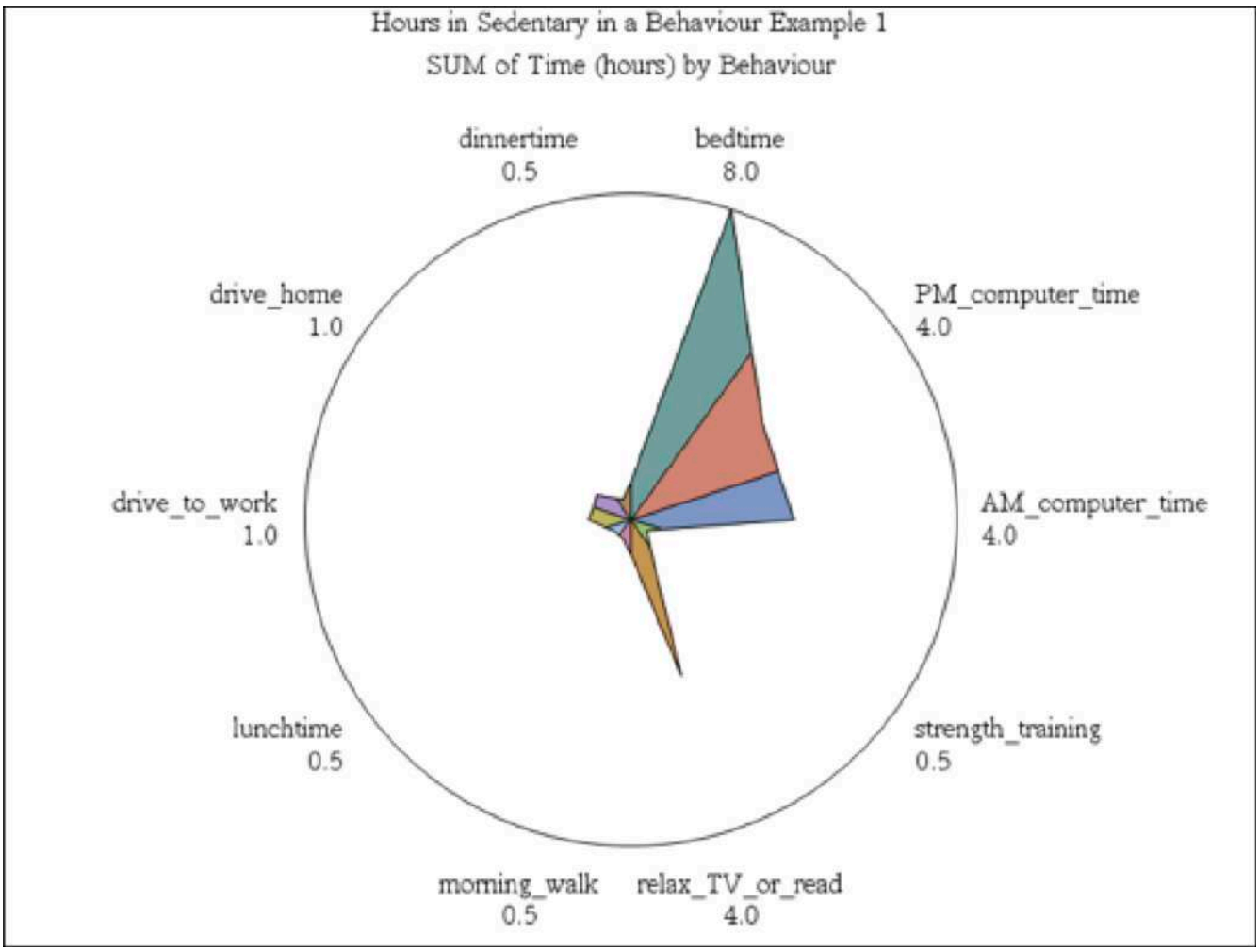
Producing Star Charts

In this SAS graphing procedure we show how out of balance sedentary behaviour can be in comparison to other activities of daily living.

Primary healthcare has continued to support the notion that sedentary behaviours are major risk factors for most chronic diseases. In particular, there has been a growing awareness of the relationship between sitting for prolonged periods during the day as a risk factor for chronic diseases such as CVD/CHD, type II diabetes, and hypertension. The data reported here is the estimated time in non-standing related activities. We can use a star chart to demonstrate an effective approach to representing unbalanced data for a given outcome. These data are from the American Heart Foundation (2015).

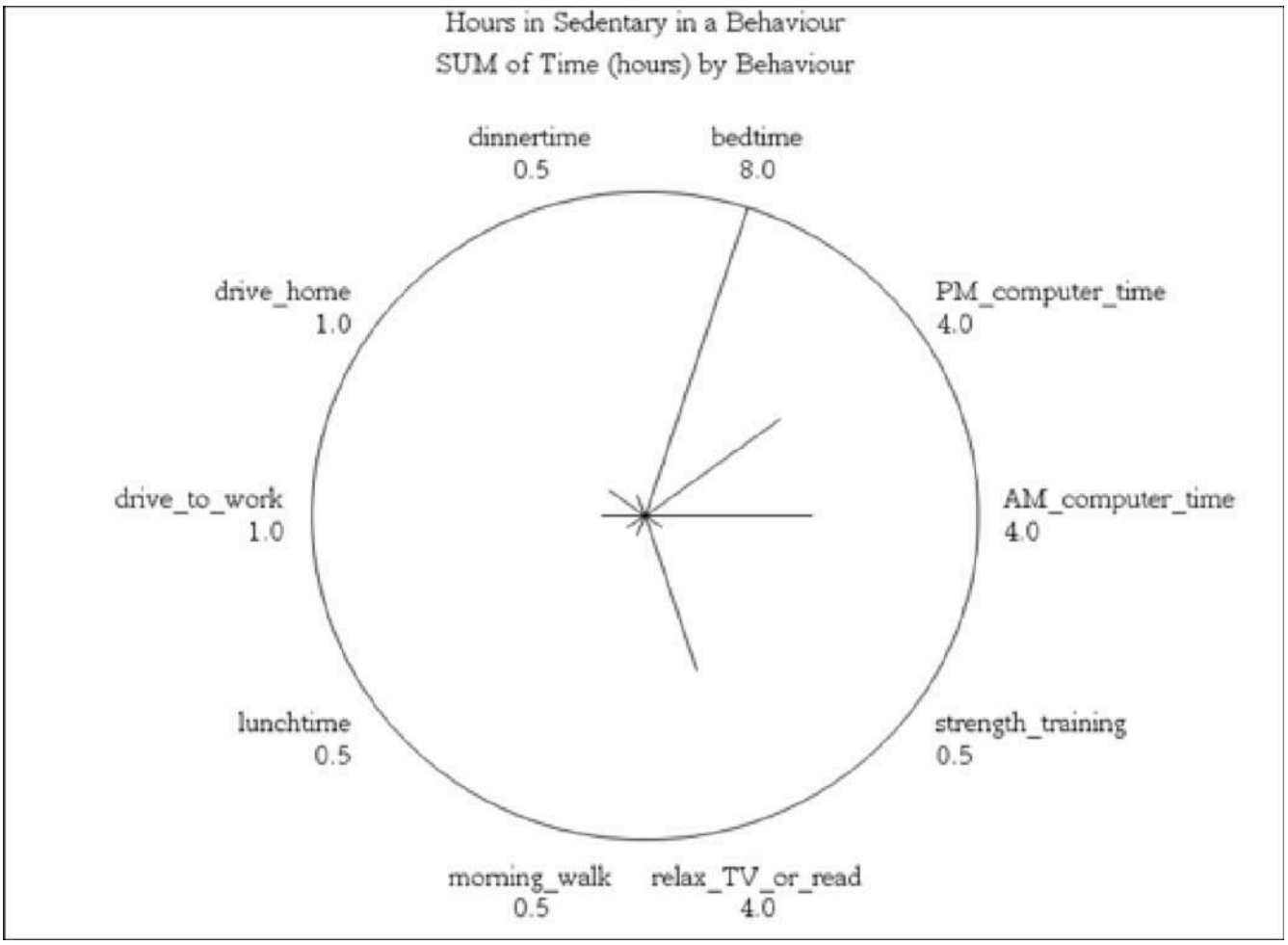
The SAS code to generate a horizontal bar chart with a corresponding frequency distribution table and two different star graphs are shown below. Notice in this SAS code we predefine the length of the input data for the variable BEHAVIOR and we use a fixed input format to enter the data values for the variables TIME in columns 22 to 24 and the variable GROUP in columns 27 to 28.

```
   DATA STARS;
LENGTH BEHAV $20.;
INPUT BEHAV $ 1-20 TIME 22-24 GRP 27-28 ;
LABEL TIME='TIME (HOURS)';  LABEL BEHAV='BEHAVIOUR';
DATALINES;
MORNING_WALK        0.5  1
DRIVE_TO_WORK       1.0  1
AM_COMPUTER_TIME    4.0  1
LUNCHTIME           0.5  1
PM_COMPUTER_TIME    4.0  1
DRIVE_HOME          1.0  2
STRENGTH_TRAINING   0.5  2
DINNERTIME          0.5  2
RELAX_TV_OR_READ    4.0  2
BEDTIME             8.0  2
;
PROC GCHART;
HBAR BEHAV/SUMVAR=TIME;
TITLE1 "HOURS SPENT IN SEDENTARY BEHAVIOUR-HORIZONTAL BAR CHART";
RUN;
   PROC GCHART ;
TITLE1 "EXAMPLE STAR GRAPH 1";
TITLE2 "HOURS SPENT IN SEDENTARY BEHAVIOUR";
STAR BEHAV / DISCRETE SUMVAR=TIME FILL=S;
RUN;
   PROC GCHART ;
STAR BEHAV / DISCRETE SUMVAR=TIME NOCONNECT;
TITLE1 "EXAMPLE STAR GRAPH 2";
TITLE2 "HOURS SPENT IN SEDENTARY BEHAVIOUR";
RUN;
```



In the image above we include the FILL=S; option in the SAS code

```
   PROC GCHART ;
TITLE1 “EXAMPLE STAR GRAPH 1”;
TITLE2 “HOURS SPENT IN SEDENTARY BEHAVIOUR”;
STAR BEHAV / DISCRETE SUMVAR=TIME FILL=S;
```



In the image above we we ***remove*** *the FILL=S option and* ***include*** *the NOCONNECT option in the SAS code*

```
   PROC GCHART ;
STAR BEHAV / DISCRETE SUMVAR=TIME NOCONNECT;
TITLE1 “EXAMPLE STAR GRAPH 2”;
TITLE2 “HOURS SPENT IN SEDENTARY BEHAVIOUR”;
```

Preparing data for graphing by transposing datasets

In this next section, we will rotate the perspective of the data set – a term we refer to as TRANSPOSING. With the PROC TRANSPOSE feature, we can re-orient the data set from written in a wide format to a narrow format.

The wide-format of the dataset is shown in the table below. With the SAS code below we can transpose four variables into one variable. The following table is the original four variables from the raw data.

Obs	ID	var1	var2	var3	var4
1	7	0.350	0.326	0.333	0.333
2	9	0.346	0.328	0.318	0.325
3	10	0.350	0.352	0.345	0.355
4	11	0.345	0.330	0.341	0.321
5	13	0.348	0.342	0.335	0.330
6	14	0.347	0.334	0.342	0.350
7	15	0.349	0.325	0.324	0.327
8	16	0.338	0.322	0.334	0.324
9	18	0.331	0.329	0.314	0.335
10	19	0.342	0.332	0.323	0.328
11	20	0.338	0.318	0.325	0.331

SAS Code to transpose the data from a wide to a narrow format

```
   DATA TRNSPS_W2N;
/* TRANSPOSING WIDE DATA TO NARROW DATA */
INPUT ID VAR1 VAR2 VAR3 VAR4;
DATALINES;
7 0.35 0.326 0.333 0.333
9 0.346 0.328 0.318 0.325
10 0.35 0.352 0.345 0.355
11 0.345 0.33 0.341 0.321
13 0.348 0.342 0.335 0.33
14 0.347 0.334 0.342 0.35
15 0.349 0.325 0.324 0.327
16 0.338 0.322 0.334 0.324
18 0.331 0.329 0.314 0.335
19 0.342 0.332 0.323 0.328
20 0.338 0.318 0.325 0.331
;
   TITLE1 'TRANSPOSING FOUR VARIABLES INTO ONE VARIABLE';
PROC SORT DATA=TRNSPS_W2N; BY ID;
TITLE2 'PRINT OF THE ORIGINAL FOUR VARIABLES FROM THE RAW DATA';
PROC PRINT; VAR ID VAR1 VAR2 VAR3 VAR4 ;
RUN;
PROC TRANSPOSE DATA=TRNSPS_W2N OUT=NARROW;
BY ID;
```

```
   TITLE2 'PRINT OF THE TRANSPOSED DATA TO A SINGLE VARIABLE';
RUN;
PROC PRINT DATA=NARROW; VAR ID    _NAME_       COL1;
RUN;
```

A portion of the narrow format of the data is shown here in this printout of the transposed data. Here we show the four variables as one categorical variable and one outcome variable, which can then be graphed.

Obs	ID	_NAME_	COL1
1	7	var1	0.350
2	7	var2	0.326
3	7	var3	0.333
4	7	var4	0.333
5	9	var1	0.346
6	9	var2	0.328
7	9	var3	0.318
8	9	var4	0.325
9	10	var1	0.350

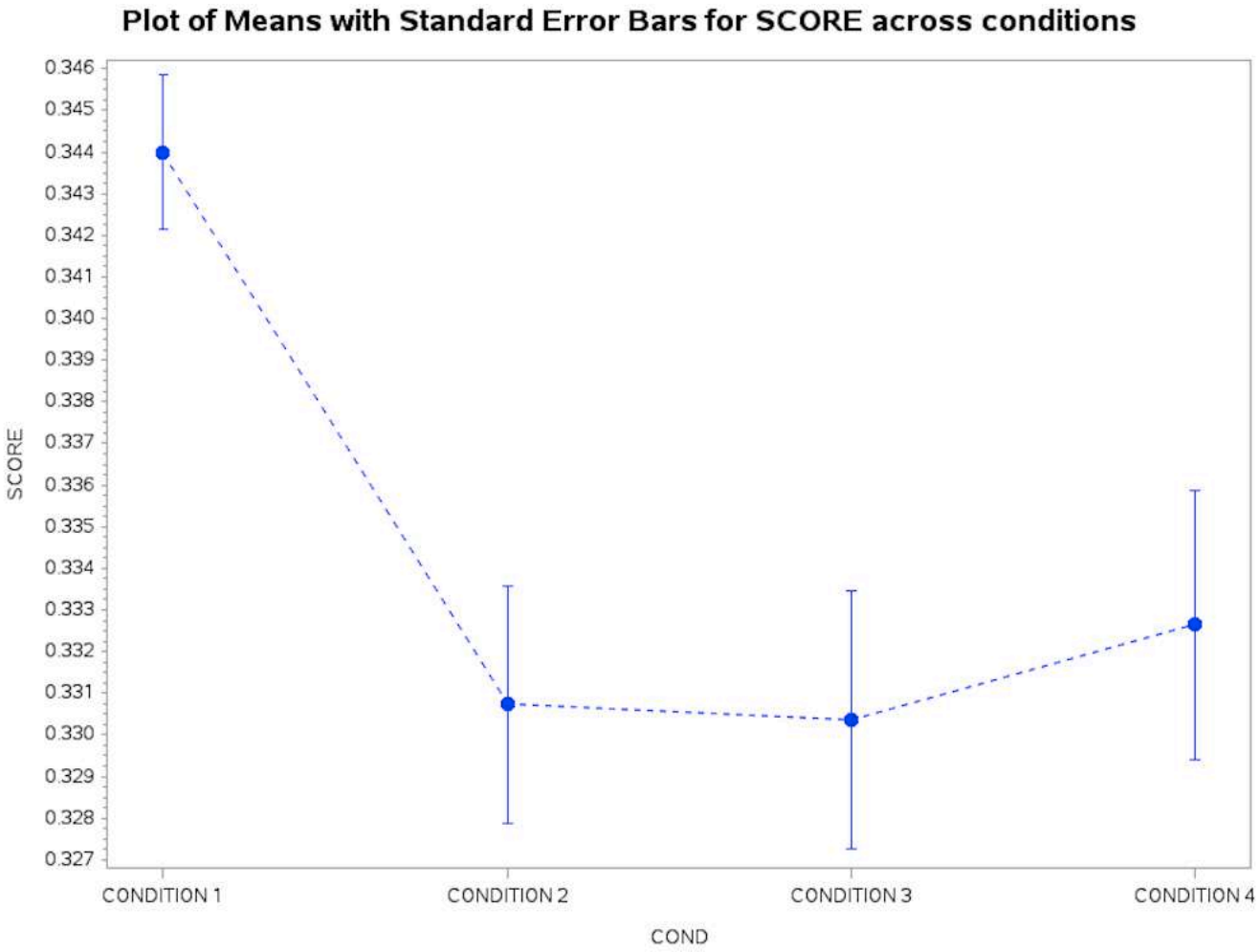
The data in the transposed table above can be used in a graph to show the response of each participant for the single dependent variable, which we called SCORE, across four measures. The SGPLOT procedure was modified from SAS SUPPORT CODE: Sample 50217: Plot means with standard error bars from calculated data for groups with PROC GPLOT[1].

SAS Code for Wide to Narrow

```
   PROC FORMAT;
VALUE CNDFMT 1 = 'CONDITION 1'
2 = 'CONDITION 2'
3 = 'CONDITION 3'
4 = 'CONDITION 4' ;
   /* PLOT OF DEPENDENT VARIABLE AFTER TRANSPOSE TO NARROW DATA */
DATA W2N;
INPUT OBS ID COND SCORE;
/* USE THE TRANSPOSED DATASET IN A LINE GRAPH ACROSS 4 CONDITIONS */
DATALINES;
1 7  1 0.350
2 7  2 0.326
   <MORE DATA HERE >
```

```
   41 20  1 0.338
42 20  2 0.318
43 20  3 0.325
44 20  4 0.331
;
TITLE1 'TRANSPOSING FOUR VARIABLES INTO ONE VARIABLE';
TITLE2 'TRANSPOSED VARIABLE AS A SINGLE RESPONSE ACROSS FOUR TIME POINTS';
AXIS1 ORDER=(1 TO 4 BY 0.55) OFFSET=(2,2)
LABEL=NONE  MAJOR=(HEIGHT=2) MINOR=(HEIGHT=1);
AXIS2 ORDER=(0.3 TO 0.4 BY 0.01) OFFSET=(0,0)
LABEL=NONE   MAJOR=(HEIGHT=2) MINOR=(HEIGHT=1);
RUN;
PROC SORT DATA=W2N; BY COND;
PROC MEANS DATA=W2N  NOPRINT;
BY COND;
VAR SCORE;
OUTPUT OUT=MEANSOUT MEAN=MEAN STDERR=STDERR;
TITLE1 'DESCRIPTIVE STATISTICS FOR SCORE ACROSS 4 CONDITIONS';
RUN;
/* RESHAPE THE DATA TO PRESENT ONE Y VALUE FOR */
/* EACH X FOR USE WITH THE HILOC INTERPOLATION.  */
DATA RESHAPE(KEEP=COND SCORE MEAN);
SET MEANSOUT;
SCORE=MEAN;
OUTPUT;
SCORE=MEAN − STDERR;
OUTPUT;
SCORE=MEAN + STDERR;
OUTPUT;
RUN;
/* DEFINE THE TITLE */
TITLE1 'PLOT OF MEANS WITH STANDARD ERROR BARS FOR SCORE ACROSS CONDITIONS';
/* DEFINE THE AXIS CHARACTERISTICS */
AXIS1 OFFSET=(5,5) MINOR=NONE;
AXIS2 LABEL=(ANGLE=90);
/* DEFINE THE SYMBOL CHARACTERISTICS */
SYMBOL1 INTERPOL=HILOCTJ COLOR=BLUE LINE=2;
SYMBOL2 INTERPOL=NONE COLOR=BLUE VALUE=DOT HEIGHT=1.5;
/* PLOT THE ERROR BARS USING THE HILOCTJ INTERPOLATION */
/* AND OVERLAY SYMBOLS AT THE MEANS. */
PROC GPLOT DATA=RESHAPE;
PLOT SCORE*COND MEAN*COND / OVERLAY HAXIS=AXIS1 VAXIS=AXIS2;
FORMAT COND CNDFMT.;
RUN;
```

This SAS code from the transposed dataset produced the following graph.



Transposing data from narrow to a wide format

Consider now if our data were in a long format, as in a single column with 3 categories but we wanted to reshape the data so that each of the categories became a separate measure of interest. In the following data set consisting of a categorical variable that we called employment status and a dependent variable based on household savings in the bank on January 1. Here, we will transpose the data from a long format to a wide format and convert the initial measure of interest to three variables.

The initial SAS code with data is as follows[2]:

SAS Code for Narrow to Wide

```
PROC FORMAT;
```

```
VALUE EMP 1= 'FULL-TIME' 2 = 'PART-TIME' 3= 'CASUAL';
DATA EMPSTAT;
LABEL       ID = 'PARTICIPANT ID'
EMPSTAT = 'EMPLOYMENT STATUS'
SAVINGS = 'SAVINGS IN BANK';
INPUT ID 1-2 EMPSTAT 4 SAVINGS 6-9;
DATALINES;
01 3 0020
02 1 0120
03 2 0050
04 3 0030
05 3 0000
06 1 4500
07 1 8900
08 2 0540
09 3 0900
10 1 3220
11 2 0240
12 2 0400
;
PROC SORT data=EMPSTAT; BY EMPSTAT;
PROC FREQ; TABLES EMPSTAT;
FORMAT EMPSTAT EMP. ;
PROC FREQ; TABLES EMPSTAT*SAVINGS;
FORMAT EMPSTAT EMP. ;
TITLE1 ' FREQUENCY DISTRIBUTION FOR EMPLOYMENT STATUS';
RUN;
PROC SORT data=EMPSTAT; BY ID;
PROC TRANSPOSE data=EMPSTAT out=NEW_WIDE prefix=GROUP_;
by ID ;
id EMPSTAT;
var SAVINGS;
RUN;
proc print data = NEW_WIDE; VAR ID GROUP_1 GROUP_2 GROUP_3;
TITLE 'OUTPUT FOR WIDE FORMATTED DATA';
RUN;
PROC MEANS MEAN MEDIAN STD STDERR CV; VAR GROUP_1 GROUP_2 GROUP_3;
TITLE 'USING PROC MEANS- DESCRIPTIVE STATISTICS FOR WIDE FORMATTED DATA';
RUN;
PROC TABULATE data = NEW_WIDE;
LABEL GROUP_1 = 'EMPLOYED FULL TIME'
GROUP_2 = 'EMPLOYED PART TIME'
GROUP_3 = 'EMPLOYED CASUALLY';
VAR GROUP_1 GROUP_2 GROUP_3;
TABLE (GROUP_1 GROUP_2 GROUP_3)* (N MEAN STD CV);
```

```
TITLE 'USING PROC TABULATE – DESCRIPTIVE STATISTICS FOR WIDE FORMATTED DATA';
RUN;
```

The SAS code above produced the following output after transposing the data from the dependent variable to produce three measures of interest which we called GROUP_1 GROUP_2 and GROUP_3. Each variable now represents the data within the specific employment category and the PROC TABULATE and PROC MEANS commands were used to produce descriptive statistics for each separate dependent measure.

FREQUENCY DISTRIBUTION FOR EMPLOYMENT STATUS

EMPLOYMENT STATUS	FREQUENCY	PERCENT	CUMULATIVE FREQUENCY	CUMULATIVE PERCENT
FULL TIME	4	33,33	4	33.33
PART-TIME	4	33.33	8	33.33
CASUAL	4	33.33	12	100

USING PROC MEANS TO PRODUCE DESCRIPTIVE STATISTICS FOR WIDE FORMATTED DATA

The MEANS Procedure

Variable	Mean	Median	Std Dev	Std Error	Coeff of Variation
GROUP_1	4185.00	3860.00	3641.70	1820.85	87.01
GROUP_2	307.50	320.00	210.93	105.47	68.60
GROUP_3	237.50	25.00	441.84	220.92	186.04

USING PROC TABULATE –DESCRIPTIVE STATISTICS FOR WIDE FORMATTED DATA

DESCRIPTIVE STATISTICS CALCULATED WITH PROC TABULATE

EMPLOYED FULL TIME			
N	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION
4	4185	3641.7	87.02
EMPLOYED PART TIME			
N	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION
4	307.5	210.93	68.6
EMPLOYED CASUALLY			
N	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION
4	237.5	441.84	186.04

A first look at the features of SAS PROC TABULATE

---

[1] http://support.sas.com/kb/50/217.html

[2] The structure of this code was derived from: Introduction to SAS. UCLA: Statistical Consulting Group.from https://stats.idre.ucla.edu/sas/modules/sas-learning-moduleintroduction-to-the-features-of-sas/ (accessed August 22, 2016).

[1] (URL – https://data.humdata.org/dataset/number-of-health-care-workers-deaths-by-edv) a project from the United Nations Office for the Coordination of Humanitarian Aid (url: http://www.unocha.org/)

PART III

GOODNESS OF FIT AND RELATED CHI-SQUARE TESTS

Learning Objectives

After reading this section you should be able to:

- Create frequency distribution tables
- Compute percentiles for scores within a frequency distribution
- Apply the goodness of fit test – chi-square statistical procedure to evaluate nominal level data for one or more samples
- Compute the contingency table using a chi-square statistical procedure
- Describe and compute the chi-square statistical procedure for a 2 x 2 research design to test for difference in two variables measured at the nominal level
- Describe and compute the phi-coefficient in a 2 x 2 design to evaluate the association between two variables measured at the nominal level
- Describe and compute Fisher's Exact test in a 2 x 2 design to determine the exact probability associated with the computation of the chi-square statistic

This textbook was developed to demonstrate biostatistical research applications that can use either web-based JavaScripted calculators –**aka the Webulators** or the Statistical Analysis System –**aka SAS coding** to resolve questions arising in healthcare research. In the following sections, we will work through the concepts of statistical applications using specific examples that demonstrate the robustness of **the Webulators** and **SAS coding**. The SAS applications are based on the SAS Studio Education Analytic Suite. The Webulators are located at https://health.ahs.upei.ca/webulators to resolve specific questions using online tools. The "webulators–the web-based calculators", provide several useful resources that extend well beyond this textbook.

13. Frequency Distributions

13.1 Analyzing Distributions of Data

Throughout this text, we will focus on using frequency analysis and descriptive statistics. These simple but powerful analyses enable you to examine your data and identify patterns including the shapes and distributions of data, missing values, and outliers. Frequencies and distributions are important concepts in the quantitative analysis of data that underlie the overall statistical approach covered in this book. In fact, this approach is often referred to as "frequentist statistics" because it relies on frequencies to make inferences about the data. An alternative approach is called the Bayesian statistical analysis which relies on probabilities. While we won't go into detail about the differences between frequentist and Bayesian statistical approaches, it is important to recognize that frequencies play a key role in the approach that we are demonstrating here but that a frequentist approach is not the only way to analyze your data.

A *frequency* is simply the number of times something happens. It could be, for example, the number of people with brown hair, the number of children in a family, the number of deaths in a hospital. It could also be the number of times an electrical signal with a given level of energy-intensity is recorded.

A *distribution* shows the relative frequencies of each possible value or category for a variable. Distributions are used to describe the organization or shape of a set of scores or values for a particular variable. If you studied statistics previously you are most likely familiar with the normal distribution or bell curve. What you may not realize is that distributions other than the *normal distribution* are also used in statistic analyses and that datasets can take the shape of these other distributions. For example, datasets that include only discrete scores ranging from 1 to 5 would not be expected to fit a normal distribution curve but would rather be compared to a categorical distribution curve – like the chi-square distribution or the Poisson distribution.

Distributions can be obtained by counting the number of events that occur or how many participants in a sample have a specific score on a questionnaire or measure (i.e., counting frequencies). For example, you might look at the number of patients presenting to the Emergency Department for different reasons: cardiovascular concerns, accidents, infections, reported symptoms. You might also consider responses to an anxiety questionnaire scored on a Likert scale (using discrete scaled scores) ranging from 1 to 5. You may consider reviewing the number of respondents in your sample had each possible score (i.e. 1, 2, 3, 4, or 5) – in other words, how *frequent* each score appeared within the total set of scores.

The following is an example of a table showing a frequency distribution for a set of responses to a categorical variable ranging from 1 to 5, and a graphical representation of the frequency of responses in each category. The Category Label is presented on the x-axis, and the number of responses–frequencies for each response are presented on the y-axis.

Table 13.1 Frequency Distribution For Categorical Responses

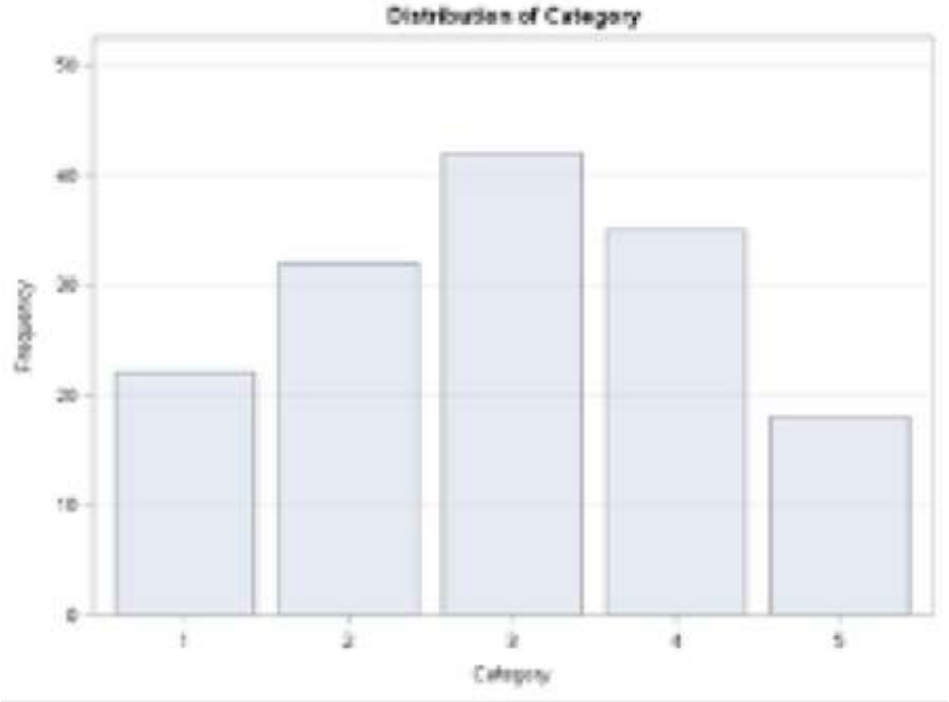
The PROC FREQ Procedure

Category	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	22	14.77	22	14.77
2	32	21.48	54	36.24
3	42	28.19	96	64.43
4	35	23.49	131	87.92
5	18	12.08	149	100.00

SAS Code to produce the Frequency Distribution and Corresponding Figure

```
  OPTIONS PAGESIZE=55 LINESIZE=120 CENTER DATE;
DATA FREQ13_1;
INPUT CATEGORY 1 RESPONSES;
DATALINES;
1 22
2 32
3 42
4 35
5 18
;
TITLE1 'TABLE 13.1. FREQUENCY DISTRIBUTION FOR CATEGORICAL RESPONSES';
PROC FREQ ORDER=FORMATTED;
TABLES CATEGORY/PLOTS=FREQPLOT;
WEIGHT RESPONSES;
RUN;
```

Figure 13.1 Frequency Distribution For a Categorical Response Variable



Frequency distributions are useful in describing variables, helping to identify errors (impossible values) and outliers, assessing how well a continuous variable fits the normal distribution, or to test hypotheses using specific statistical tests such as using a chi-square test to evaluate categorical variables.

Frequency & Distribution of a Count Variable

Count variables refer to those that simply tally the number of items or events that occur. For example, you might want to count the number of adverse events that occur when people take a medication, the number of times nurses wash their hands during their shift or the number of babies born in each month of the year. In health research, there are many items or events that can be counted!

Note that for a count variable, the values are arithmetically meaningful and represent the ***number*** of events or items for a specific variable– the count variable is quite literally storing the count of items of interest. Therefore, values differ by a magnitude and are meaningful. For example, 4 adverse events are twice as many as 2 adverse events.

Count variables are different than categorical variables.

Categorical variables are used when the researcher wishes to use numbers to represent different ***kinds*** of items or events. In the categorical variable the numbers are arbitrary. For example, hair colour could be coded as 1 = blonde, 2 = brown, 3 = gray, 4 = red, and 5 = other but it could also be coded as 11 = blonde, 22 = brown, 33 = gray, 44 = red, and 55 = other. The numbers representing a category label are not mathematically meaningful and do not represent the number of people with a specific hair colour. Of course, you can analyze the frequency of people with each response which we will cover later when we talk about categorical variables in more detail.

Working example to process a "count" variable

Let's say we would like to take a sample of 50 families from a population of 1000 households in a small town and record the number of children in each household. Here we will create two variables, the first we will call "**NKIDS**" and the second we will call "**HOUSEHOLDS**". The variable NKIDS is the categorical variable for the number of children in each household that we sampled, while the variable HOUSEHOLDS represents the number of response houses that report having a given number of children.

The Scenario

We arrive at the small town and knock on the front door of the first house. Below is the dialogue between the researchers and the respondents.

"Good day, we are Biostatisticians and we are conducting a study of the number of children in your family."

"Oh we don't have any children."

"Okay, thank-you."

We note that for Household #1 there are 0 children. We then knock on the front door of the second house.

"Good day, we are Biostatisticians and we are conducting a study of the number of children in your family."

"We have 7 children. Would you like some?"

"No thank-you, but have a nice day."

We note that for Household #2 there are 7 children. We then knock on the front door of the third house, and continue our process for each of 50 houses in the town.

In this example, the categorical variable is NKIDS and is considered the independent variable, while the continuous-discrete variable is HOUSEHOLDS and is considered the dependent variable – aka the measure of interest.

Since there can only be whole numbers for the variable NKIDS (i.e., you can't *actually* have 1.2 children), the variable NKIDS is a discrete categorical variable, and likewise, because we are counting families on a whole number line (i.e. not partial families) then the variable NFAMILIES is a discrete random variable.

The frequency distribution recording sheet for this example is shown below. Notice that as a rule, we want to keep our variable labels at or near 8 characters so that HOUSEHOLDS is shortened to HSEHLD.

Table 13.2 Tally Sheet to produce the Frequency Distribution for Number of Children in Each Household Sampled

Number of Children	Tally of Households	Frequency (f)	Relative frequency (f/n)
0	‖‖‖ ‖‖‖	9	9/50 = 0.18
1	‖‖‖ ‖	7	7/50 = 0.14
2	‖‖‖ ‖‖‖ ‖	12	12/50 = 0.24
3	‖‖‖ ‖‖‖	9	9/50 = 0.18
4	‖‖‖	5	5/50 = 0.10
5	‖‖‖		6	6/50 = 0.12
6	––	0	0/50 = 0
7	‖	2	2/50 = 0.04
		N=50	Proportion = 1.00

Counting events such as the number of children in a family, the number of needles found on the ground near a safe injection site, or the number of patients readmitted to the hospital after discharge, typically follow the whole number line. Frequency tables are often used to show how many times an event has occurred.

In our example, we can say that the variable HOUSEHOLDS is a discrete random variable because in a given sample of 50 families the variable can take on (contain) any value between 0 and 50 (the total sample) on the whole number line.

Table 13.1 shows how we can determine the frequency and relative frequency (percentage out of 100) for the number of children in each of the families in our sample. Of the 50 families in our sample, nine families did not have children, 7 families had 1 child, 12 families had 2 children, 9 families had 3 children, 5 families had 4 children, 6 families had 5 children, no families had 6 children, and 2 families had 7 children. Notice here that the variable of interest is the number of families reporting each of the possible number of children.

Relative frequency refers to the proportion of the entire sample that had a particular value. In this example, the relative frequency tells us what percentage of the sample had a specific number of children. To calculate the relative frequency, simply divide each frequency by the total number of families and then multiply the result by 100 to calculate the percentage value. For example, from the data in Table 4.1 we see that in this sample, 24% of the families had 2 children while only 4% had 7 children.

Creating the SAS Program to compute a frequency distribution for a discrete random variable

Below are the SAS commands to produce the frequency distribution table of the data recorded for the number of children in our sample of 50 families.

SAS Code to produce a Frequency Distribution Table For Number of Children in Each Household sampled

```
  DATA FREQ13_2;
INPUT ID NKIDS HSEHLD;
DATALINES;
01 00 9
02 01 7
03 02 12
04 03 9
05 04 5
06 05 6
07 06 0
08 07 2
;
PROC FREQ DATA=FREQ13_2 ORDER=DATA;
TABLES NKIDS;
WEIGHT HSEHLD;
RUN;
```

In this SAS program, we are using the PROC FREQ statistical processing command with the keyword TABLES to produce a frequency distribution for the data recorded for our sample of 50 households. Notice in the PROC FREQ command sequence we included the statement WEIGHT HSEHLD. In this example, the independent or categorical variable is NKIDS and the dependent discrete random variable is HSEHLD. The WEIGHT command enables us to enter the summary data for the dependent variable HSEHLD as the count related to the categorical variable NKIDS.

Notice the table indicates that 9 households reported no children, while no households reported having 6 children. The table also indicates that most households reported having 2 children.

TABLE 13.3 Frequency distribution for number of children in each household

The FREQ Procedure

NKIDS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	9	18.00	9	18.00
1	7	14.00	16	32.00
2	12	24.00	28	56.00
3	9	18.00	37	74.00
4	5	10.00	42	84.00
5	6	12.00	48	96.00
7	2	4.00	50	100.00

The following is a SAS program to compute elements of PROC FREQ for frequency distributions. The data are fictitious and are used here to enable you to work through the various options and features of the PROC FREQ command with relevant options.

As you work through the SAS program take note of the specific features that are identified, therein. The scenario is based on a public health study in which a group of researchers intended to determine the number of discarded needles left on the ground within a 100-metre radius of safe injection sites. We begin the program first by reading the data set and then using the essential SAS statistical processing commands with relevant *options* for PROC FREQ.

The program begins by labeling the working SAS program as DATA FREQ13_4; – which simply creates a label for the SAS program in the present SAS work session;

The second line is the listing of variables to be read within the sample data set. The SAS command begins with the SAS keyword INPUT which is followed by the names of each variable. Notice that the variable names are kept to eight characters and each variable name begins with an alphabetic character rather than a number or a special character. In this example the variables SITE, NDLCNT and INCREG are used to indicate that we have a variable to list the various sites from which the data were collected (SITE), the number of needles found on the ground within a 100-metre radius of the exit door of the safe injection site (NDLCNT), and the estimated average household income reported in thousands of dollars for the region in which the injection site is located (INCREG).

We also use simple IF-THEN logic commands to create summary groups for both the variable NDLCNT – number of needles recorded at each site, as well as to group the average household income – INCGRP.

SAS Code to Demonstrate Features of PROC FREQ

```
DATA FREQ13_4;
INPUT SITE NDLCNT INCOME;
LABEL SITE = 'INJECTION SITE'
NDLCNT='# NEEDLES FOUND'
INCOME = 'AVE HSHLD INCOME ($000.00)'
INCGRP = 'INCOME GROUPS'
NDLGRP ='GROUP # OF NEEDLES';
IF INCOME <=25 THEN INCGRP=1;
IF INCOME >25 AND INCOME <=50 THEN INCGRP=2;
IF INCOME >50 AND INCOME <=75 THEN INCGRP=3;
IF INCOME >75 AND INCOME <=100 THEN INCGRP=4;
IF INCOME >100 THEN INCGRP=5;
IF NDLCNT<=5 THEN NDLGRP=1;
IF NDLCNT>5 AND NDLCNT<=10 THEN NDLGRP=2;
IF NDLCNT>10 THEN NDLGRP=3;
```

```
DATALINES;
01 10 23
02 10 24
03 8 32
04 4 45
05 7 38
06 0 150
07 4 85
08 10 19
09 10 20
10 5 52
11 4 54
12 3 78
13 0 144
14 6 36
15 7 15
16 4 80
17 3 95
18 6 70
19 4 90
20 0 101
;
PROC SORT; BY INCGRP;
PROC FREQ; TABLES NDLCNT;
TITLE1 'FREQ DIST FOR # NEEDLES FOUND ACROSS ALL SITES';
RUN;
PROC SORT; BY INCGRP;
PROC FREQ; TABLES NDLGRP*INCGRP;
TITLE1 'FREQ DIST FOR GROUP NEEDLES BY INCOME GROUP';
RUN;
```

FREQUENCY DISTRIBUTION FOR THE NUMBER NEEDLES FOUND ACROSS ALL SITES

Frequency of Needle Count				
Needle Count	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	3	15.00	3	15.00
3	2	10.00	5	25.00
4	5	25.00	10	50.00
5	1	5.00	11	55.00
6	2	10.00	13	65.00
7	2	10.00	15	75.00
8	1	5.00	16	80.00
10	4	20.00	20	100.00

The SAS commands to sort the data and run the PROC FREQ using the * between variables helps to summarize the data into 2-way frequency distribution tables. In this way, we can see at a glance, a summary of the dataset.

In the sequence of SAS processing commands, we first sort the data using PROC SORT, followed by the SAS commands PROC FREQ with the keyword TABLES and then the two variables that we wish to include in the 2-way table – NDLGRP * INCGRP.

Code Snippet for SAS Code to Demonstrate PROC SORT code added to PROC FREQ

```
PROC SORT; BY INCGRP;
PROC FREQ; TABLES NDLGRP*INCGRP;
TITLE1 'FREQ DIST FOR GROUP NEEDLES BY INCOME GROUP';
RUN;
```

The result of this sequence of commands enables us to produce the 2-way SAS table of the groups of needles found arranged by income groups. The problem with this table is that the delivery of information is not optimized for the reader if the reader does not know what an income group of 5, or an NDLGRP of 3 refers.

Using the PROC FORMAT command enables us to explain the categories within each variable. The code to explain the levels of each category uses the following two-step approach.

1.) At the start of the program add the PROC FORMAT statement and the VALUE for each categorical variable.

In our example, we have two categorical variables: INCGRP and NDLGRP. The variable INCGRP has 5 levels, while the variable NDLGRP has three groups.

```
PROC FORMAT;

VALUE INC 1='LESS THAN $25K' 2='$25K TO $50K' 3='$50K TO $75K' 4='$75K TO $100K' 5='MORE THAN $100K';
  VALUE NDL 1 = '<=5 NDLS' 2= '6 TO 10 NDLS' 3= '>10 NDLS';
  DATA TAB4_3;
  INPUT SITE NDLCNT INCOME;
```

Later in the program, after we call a SAS procedure, like in this case we call PROC FREQ, we then call the FORMAT function and assign the predefined format to each variable used by the SAS procedure.

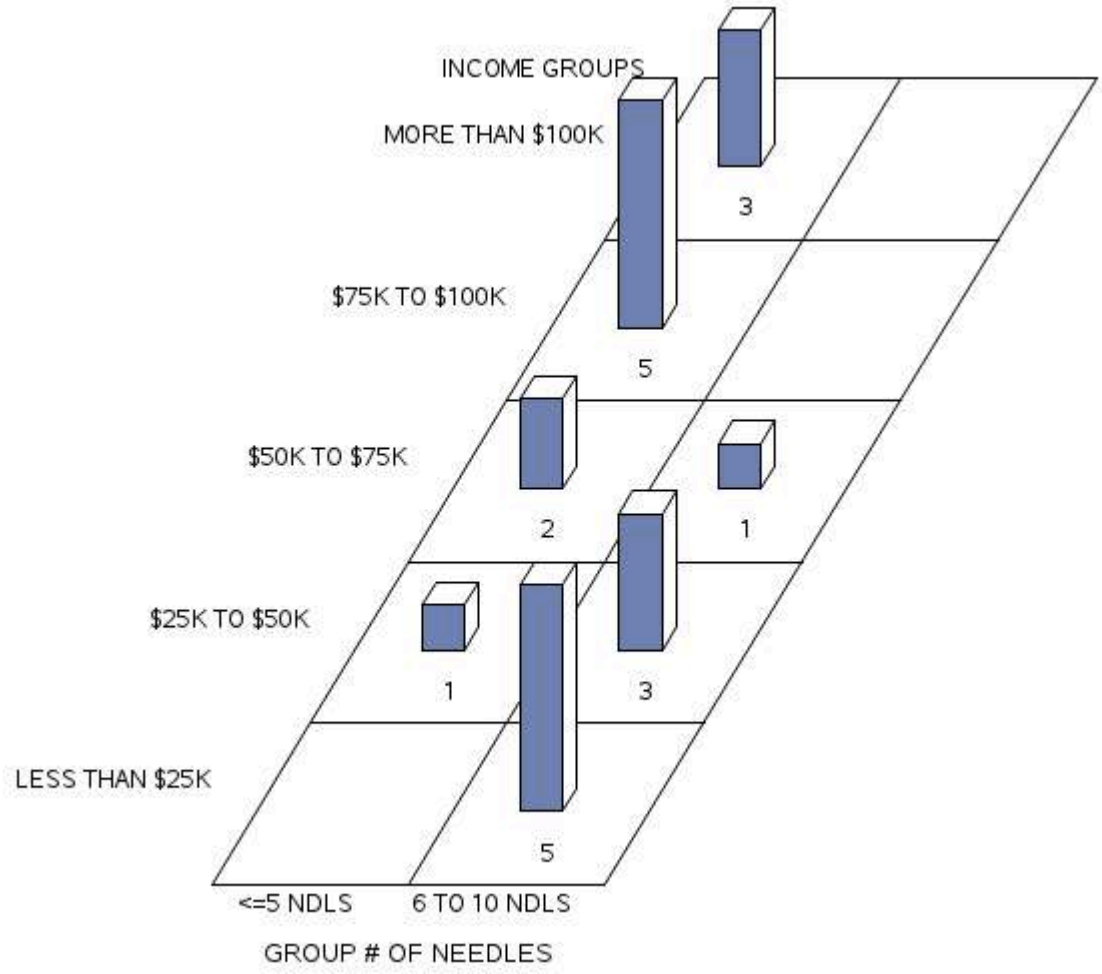
Notice we first call the variable – in this example the variable of interest is **NDLGRP** and this is followed by the PROC FORMAT VALUE name NDL. Notice also that when we include the VALUE name we follow it with a period(.). This command will place the full text for the variable category in the frequency distribution.

```
PROC SORT; BY INCGRP;
PROC FREQ; TABLES NDLGRP*INCGRP;
  FORMAT NDLGRP NDL. INCGRP INC. ;
TITLE1 'FREQ DIST FOR GROUP NEEDLES BY INCOME GROUP';
RUN;
```

The results of this analysis demonstrate that the highest number of needles found near the areas of safe injection sites tended to be higher among low-income neighborhoods than the number of needles found near the safe injection sites located in more affluent areas.

Figure 13.5 Features of Proc Freq: Adding Proc Format to the Frequency Procedure for a block chart

In the following output the SAS syntax is shown here.



At the top of the program add:

```
PROC FORMAT;
VALUE INC 1='LESS THAN $25K' 2='$25K TO $50K' 3='$50K TO $75K' 4='$75K TO $100K' 5='MORE THAN
$100K';
VALUE NDL 1 = '<=5 NDLS' 2= '6 TO 10 NDLS' 3= '>10 NDLS';
```

At the bottom of the program use

```
PROC GCHART; BLOCK NDLGRP/ GROUP = INCGRP DISCRETE;
FORMAT NDLGRP NDL. INCGRP INC. ;
TITLE1 'BLOCK CHART OF FREQ DIST FOR GROUP NEEDLES BY INCOME GROUP';
RUN;
```

13.2 Distribution for a categorical variable

As previously discussed, categorical variables involve grouping items, persons, or attributes, whereby the assignment of numbers to each group is arbitrary. For example, you might be interested in looking at the employment status of nursing home workers. The variable: *employment status* would be a categorical or grouping variable and might contain the following categories: *full-time*, *part-time*, *casual*, *and temporary*. You could assign any number you wish to represent the group label because the number is merely a label when applied to represent the category and doesn't hold any mathematical significance – the number simply enables you to group persons based on that variable (in this case, employment status).

It is important to remember that with categorical data our interest is not to compute measures of centrality or variance like means and standard deviations, and therefore we won't compare the distribution of items of persons to a normal distribution (i.e., the bell curve). Rather, the data that is held in the categories are counts and so our evaluation approach is to use statistical methods based on frequencies and ranks.

In the following steps, we calculate frequencies, relative frequencies, proportions, and percentages for categorical variables. Consider this simple data set.

Participant ID	Employment Status	Code
01	Casual	3
02	Full-time	1
03	Part-time	2
04	Casual	3
05	Casual	3
06	Full-time	1
07	Part-time	1
08	Part-time	2
09	Casual	3
10	Casual	3

We add up how many participants are in each *employment status* group and transfer the information to our chart:

Employment Status	Number of Participants	Frequency (f)	Relative Frequency (f/n)	Cumulative Percent
1	|||	3	3/10= 0.30	30%
2	||	2	2/10 = 0.20	50%
3	||||	5	5/10 = 0.50	100%

Better yet, here we will use SAS to produce a frequency distribution table.

SAS code to produce a frequency distribution for employment status

```
  PROC FORMAT;
VALUE EMP 1= 'FULL-TIME' 2 = 'PART-TIME' 3= 'CASUAL';
```

```
DATA EMPSTAT;
LABEL ID = 'PARTICIPANT ID' EMPSTAT = 'EMPLOYMENT STATUS';
INPUT ID 1-3 EMPSTAT 4;
DATALINES;
01 3
02 1
03 2
04 3
05 3
06 1
07 1
08 2
09 3
10 3
;
PROC FREQ; TABLES EMPSTAT;
FORMAT EMPSTAT EMP. ;
TITLE1 'FREQUENCY DISTRIBUTION OF EMPLOYMENT STATUS';
RUN;
```

This program produces the basic frequency distribution table for a set of categorical data and since we included the PROC FORMAT commands we can explain the data output clearly.

Features of Proc Freq: Distribution for a Categorical Variable

FREQUENCY DISTRIBUTION OF EMPLOYMENT STATUS

The FREQ Procedure for Employment Status

EMPSTAT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
FULL-TIME	3	30.00	3	30.00
PART-TIME	2	20.00	5	50.00
CASUAL	5	50.00	10	100.00

Distribution for a continuous variable

Now let's talk about analyzing data for continuous variables.

Suppose we recorded the heights (in inches) of 200 students. In this example, height is a *continuous variable* since the possible values include decimals (not just whole numbers), there are equal intervals between each line on a tape measure, and there is a meaningful 0.

While we can examine frequencies and create a histogram for a continuous variable, it is likely that we will have many different values in our dataset because each student will have a slightly different height. For example, Tom might be 61.5 inches tall while Cara is 61.6 inches tall. As a result, few students will record the exact same height. It may be, therefore, more meaningful to group these data and create categories. In other words, you can transform a continuous variable into a categorical variable simply by grouping the data with the IF-THEN **logic** statements.

In the following example, we will use the grouping approach so that we can create a more comprehensive frequency distribution.

Let's start with a dataset that includes two variables for our sample of 200 students (ID and HEIGHT). For each participant, we assign an ID and then record the height in inches for each of our participants. (note: despite that in Canada we use the metric scale for most of our measurements, we continue to refer to our heights in inches and feet – old habits die hard!)

Here we can use SAS to produce the frequency distribution table based on our grouping strategy for the data. We start by naming the working file and then include the appropriate SYNTAX to describe the variables and add the simple logic statements.

Raw Dataset 13.1 Two-hundred Height Measurements (inches)

001 58.5 002 58.8 003 60.1 004 61.3 005 61.75 006 61.96 007 62.32 008 62.67 009 62.65 010 62.66 011 63.12 012 63.23 013 63.56 014 64.20 015 64.25 016 64.41 017 64.51 018 64.58 019 64.67 020 64.68 021 64.70 022 64.74 023 64.78 024 64.80 025 64.85 026 64.92 027 64.95 028 65.00 029 65.41 030 65.45 031 65.52 032 65.55 033 65.59 034 65.63 035 65.68 036 65.70 037 65.72 038 65.78 039 65.90 040 66.00 041 66.22 042 66.24 043 66.46 044 66.51 045 66.53 046 66.76 047 66.82 048 66.91 049 66.91 050 67.04 051 67.05 052 67.06 053 67.03 054 67.13 055 67.25 056 67.25 057 67.30 058 67.31 059 67.42 060 67.45 061 67.53 062 67.65 063 67.60 064 67.62 065 67.65 066 67.67 067 67.69 068 67.70 069 67.71 070 67.72 071 67.72 072 67.72
073 67.75 074 67.76 075 67.77 076 67.78 077 67.80 078 67.82 079 67.91 080 67.94 081 67.95 082 67.97 083 67.99 084 68.01 085 68.03 086 68.05 087 68.07 088 68.10 089 68.12 090 68.15 091 68.17 092 68.20 093 68.23 094 68.31 095 68.32 096 68.38 097 68.65 098 68.75 099 68.87 100 69.00 101 69.37 102 69.50 103 69.56 104 69.60 105 69.70 106 69.75 107 69.78 108 69.80 109 69.83 110 69.87 111 69.90 112 69.94 113 70.00 114 70.05 115 70.09 116 70.10 117 70.14 118 70.15 119 70.16 120 70.18 121 70.23 122 70.27 123 70.30 124 70.49 125 70.51 126 70.65 127 70.72 128 70.77 129 70.80 130 70.82 131 70.85 132 70.90 133 70.95 134 70.97 135 71.00 136 71.05 137 71.10 138 71.15 139 71.20 140 71.23 141 71.25 142 71.31 143 71.35 144 71.38 145 71.40 146 71.44 147 71.48 148 71.50 149 71.53 150 71.56 151 71.59 152 71.63 153 71.67 154 71.70 155 71.75 156 71.80 157 71.81 158 71.83 159 71.87 160 71.90 161 72.00 162 72.07 163 72.09 164 72.10 165 72.13 166 72.20 167 72.30 168 72.23 169 72.34 170 72.45 171 72.50 172 72.57 173 72.65 174 72.69 175 73.26 176 73.28 177 73.30 178 73.37 179 73.40 180 73.45 181 73.48 182 73.53 183 73.65 184 73.75 185 73.79 186 73.83 187 74.00 188 74.25 189 74.35 190 74.53 191 74.67 192 74.78 193 74.95 194 76.25 195 76.34 196 76.45 197 78.59 198 79.25 199 79.40 200 79.47

Below is the SAS code required for the frequency analysis for the dataset above:

```
PROC FORMAT;
VALUE HT 1='LESS THAN 66.0′ 2='66.1 TO 68.0′
3='68.1 TO 70.0′ 4='70.1 TO 72.0′ 5='MORE THAN 72.0′;
DATA HEIGHTS;
LABEL ID = 'PARTICIPANT ID'
HEIGHT = 'PARTICIPANT HEIGHT' HTGRP='HEIGHT GROUP';
INPUT ID HEIGHT @@;
```

Notice some specific features of the INPUT statement above. Here we list two variables: ID and HEIGHT, followed by two @ symbols at the end of the list of variables. When two @ symbols are presented together SAS does not skip to a new line after reading the list of variables (in this case ID and HEIGHT), but rather reads across the page. This format enables us to read the data as a constant stream across the page for as many rows as is required to present the entire dataset. The computer reads the data in the order of the variables listed. That is, the computer reads through the dataset assigning the first value as the ID and the second value as the HEIGHT until all data are read.

Below is the paragraph of simple logic statements that follow the INPUT format statement. With these simple IF-THEN logic statements we organize the large unwieldy data set into six manageable groups.

```
   IF HEIGHT <=66.0 THEN HTGRP=1;
IF HEIGHT >66.0 AND HEIGHT <=68.0 THEN HTGRP=2;
IF HEIGHT >68.0 AND HEIGHT <=70.0 THEN HTGRP=3;
IF HEIGHT >70.0 AND HEIGHT <=72.0 THEN HTGRP=4;
IF HEIGHT >72.0  THEN HTGRP=5;
DATALINES;
001 58.5 002 58.8 003 60.1 004 61.3 005 61.75 006 61.96
. . .
   199 79.40 200 79.47
;
PROC FREQ; TABLES HEIGHT; RUN;
PROC FREQ; TABLES HTGRP;
FORMAT HTGRP HT. ; RUN;
```

As you see in the partial output presented in Figure 13.4 below, when we run the SAS command: PROC FREQ; TABLES HEIGHT; RUN; most of the values occur only once because height is a continuous variable which allows greater variation than categorical or count type variables. When reading this output, make sure that you screen for outliers by looking at the high and low values for the variable. SAS will also indicate the number of missing values which is also important when you are cleaning and screening your data.

Table 13.4 The output from PROC FREQ Applied to Continuous Data –Participant Height.

HEIGHT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
58.5	1	0.50	1	0.50
58.8	1	0.50	2	1.00
60.1	1	0.50	3	1.50
...				
79.25	1	0.50	198	99.00
79.4	1	0.50	199	99.50
79.47	1	0.50	200	100.00

13.3 Creating a Histogram in SAS

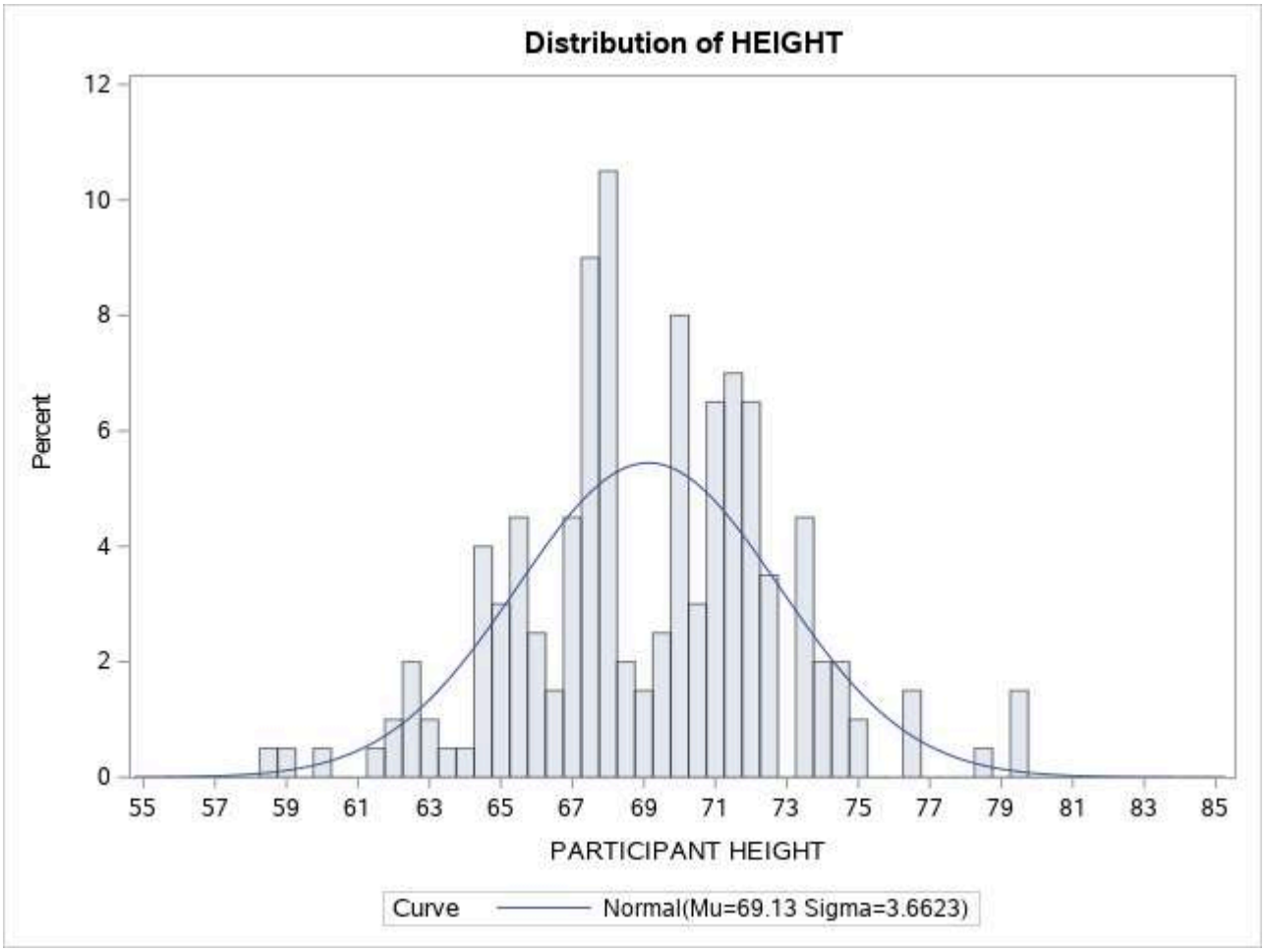
Producing graphs in SAS enables us to examine the distribution of the data visually rather than in a table. The SAS code shown here includes the option to produce a histogram. A histogram is more than a vertical bar chart. Histograms use rectangles to illustrate the frequency and interval, whereby the height of the rectangle is relative to the frequency (y axis) and the width of the rectangle is relative to the interval (x axis).

The SAS code used here provides the analysis for our sample of 200 measures of height within a cohort of children. In order to establish the appropriate number of intervals in our sample we calculate the range of our set of scores. The range refers to the spread of scores between the lowest estimate from our sample, and the highest estimate from our sample. We can estimate the range apriori by running the PROC UNIVARIATE command. When we include the command **MIDPOINTS=** we can customize the output. Here we include the command **HISTOGRAM /MIDPOINTS = 55 TO 85 BY 0.5,** to produce the expected RANGE of highest and lowest values and then plot the midpoints for all categories within the range.

```
PROC UNIVARIATE; VAR HEIGHT;
HISTOGRAM / MIDPOINTS=55 TO 85 BY 0.5 NORMAL;
RUN;
```

Notice in Figure 13.6 that the x-axis is a continuous variable. An overlay of the shape of the distribution is represented by the blue BELL-SHAPED normal curve.

Figure 13.6 Histogram for Heights of Students in Sample of 200 Participants

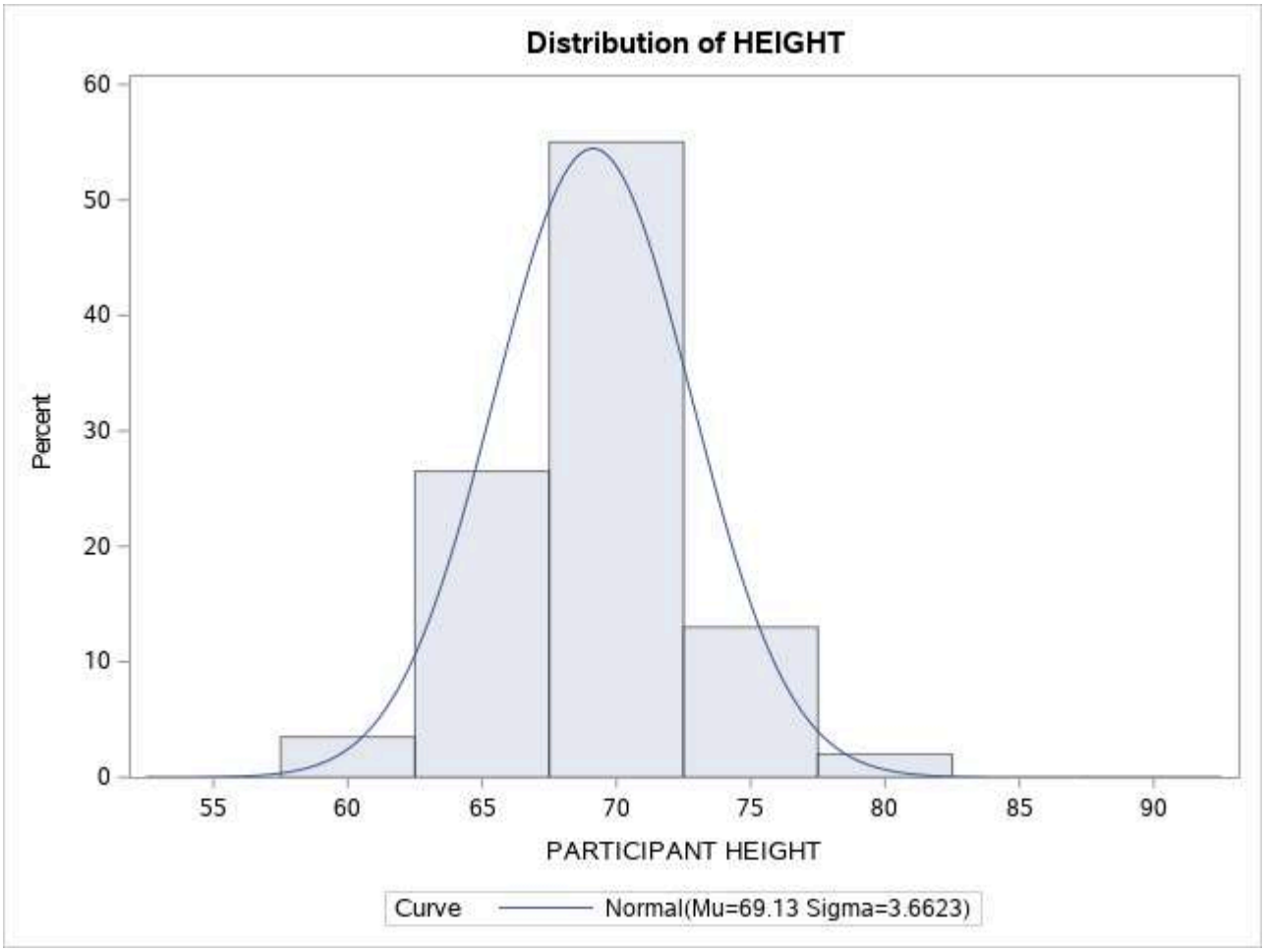


Use the following SAS Code to group the data into categories and add a representation of the shape of the distribution (CTEXT = BLUE) when plotting the histogram.

proc univariate; var HEIGHT; histogram HEIGHT /
normal midpoints = 55 60 65 70 75 80 85 90 CTEXT = BLUE;

Of you can use: **proc sgplot;**
histogram HEIGHT; density HEIGHT;

Figure 13.7 Histogram for Heights of Students N= 200 Grouped Data



Dividing a continuous variable into categories

In the example shown above, it is easy to see how helpful it is to arrange these continuous data into categories that represent a range of values rather than individual values on a continuum.

In our height example, we can optimize these data by creating height categories rather than exact heights because there is so much variance in exact heights. However, in this process of arbitrarily categorizing our response variable by grouping the data together we recognize that there will be a loss of information. When grouping data we are essentially saying that the responses are exactly the same even though differences are observed. For example, when you group a student who is 61.5 inches tall with a student who is 64 inches tall, the difference between the two individuals will be ignored within the category. There is absolutely nothing wrong with using categories that represent a range of continuous values but if you are planning to collect your own data, it is usually best to collect continuous data from the source and group the data later. Generally speaking, you can always convert data from continuous data to categories but without the original estimates, you cannot go the other way!

Once you deem it helpful to transform a continuous variable into categories, next you need to decide how to chop up your data. Ideally each you should have an equal range of values in each group. In our example, which you can see the would-be participants are quite tall, we decided to transform the continuous height data into categories that are each 2 inches wide. Group 1 includes students with a height of less than 66 inches, Group 2 starts at 66.1 and tops out at 68 inches, Group 3 starts at 68.1 and tops out at 70 inches … and so on.

Alternatively, we could have decided to group students in intervals of 5 inches or even 1 inch. As the researcher, you decide how to group the data based on what will make meaningful groupings. This might be based on past literature, clinical reference values, logic, or a combination of these factors. Once we create our categories we can sort students

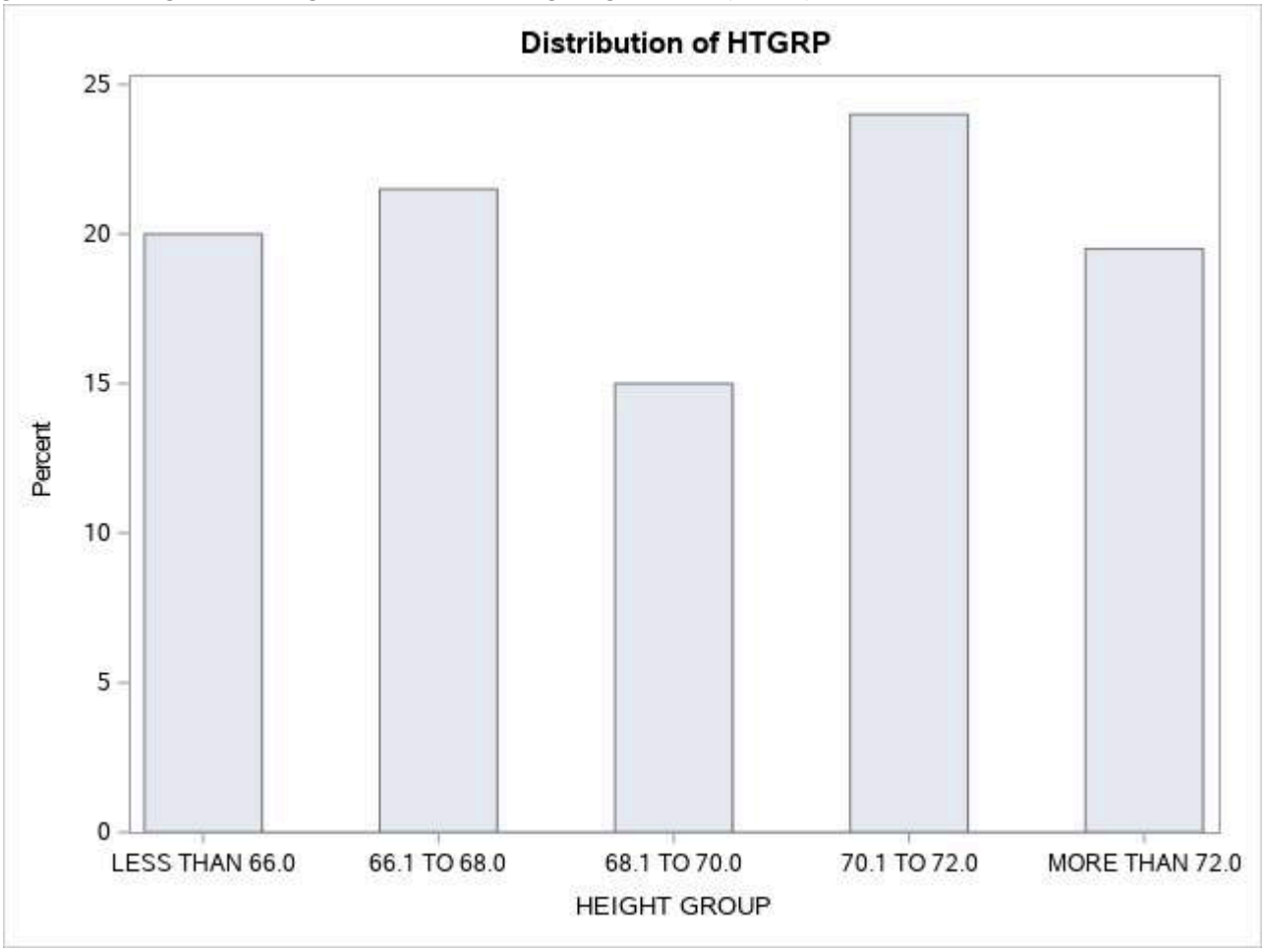
into each group based on their height. Then we can count how many students fall into each group and create a frequency distribution table (Table 13.5) and a histogram. Notice that the shape (distribution) of this histogram is the one we created using the continuous version of the height variable (Figure 13.6). This is because in the first histogram exact values are included and the data is divided into quintiles based on the mean. On the other hand, in the second histogram students are grouped together in 2-inch height intervals – depending on the cut-off values that we chose the distribution would be different.

Using the grouping routines with simple logic statements helps to simplify the organization of the data. The output for the grouped data is invoked with the SAS command: PROC FREQ; TABLES HTGRP; FORMAT HTGRP HT. ; RUN;

Table 13.5 Grouping PROC FREQ output for –Participant Height.

HTGRP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
LESS THAN 66.0	40	20.00	40	20.00
66.1 TO 68.0	43	21.50	83	41.50
68.1 TO 70.0	30	15.00	113	56.50
70.1 TO 72.0	48	24.00	161	80.50
MORE THAN 72.0	39	19.50	200	100.00

Figure 13.7 Histogram for Heights of Students using Height Group (N=200)



In addition to the histogram, SAS also includes a number of tables in the output with the PROC UNIVARIATE command. The data in these tables provide important information about our variable (height, in this example). As you can see in Table 13.8 the moments' table provides descriptive statistics about our variable including the mean, standard devia-

tion, and standard error, as well as the overall variance. Skewness and kurtosis are also provided which provide valuable information about how well the variable fits the normal distribution. Keep your critical thinking hat on when looking at these data because some of this information is not relevant for categorical variables. For example, you cannot have an average for a categorical variable and the normal distribution doesn't make sense.

Table 13.8 Descriptive Statistics for the Dataset of Heights

N	200	**Sum Weights**	200
Mean	69.1297	**Sum Observations**	13825.94
Std Deviation	3.66230697	**Variance**	13.4124924
Skewness	0.00184013	**Kurtosis**	0.45301492
Uncorrected SS	958452.17	**Corrected SS**	2669.08598
Coeff Variation	5.29773306	**Std Error Mean**	0.25896421

The next table presents the Basic Statistical Measures (Figure 13.9). In addition to the mean, standard deviation, and variance, this table also provides the median, mode, range, and interquartile range for the variable.

Table 13.9 Basic Statistical Measures Table

Location		**Variability**	
Mean	69.12970	**Std Deviation**	3.66231
Median	69.18500	**Variance**	13.41249
Mode	67.72000	**Range**	20.97000
		Interquartile Range	4.54000

Finally, Table 13.10 is the Extreme Observations Table which identifies the highest and lowest values of the variable. Here the data do not differ from what we would expect but when datasets contain outliers, this table is one way to identify the outlier data points easily. Notice that the table provides the case number along with the data point value, making it easy to revisit the original dataset and verify original values or consider adjustments to extreme values if needed.

Figure 13.10 Extreme Observations Table

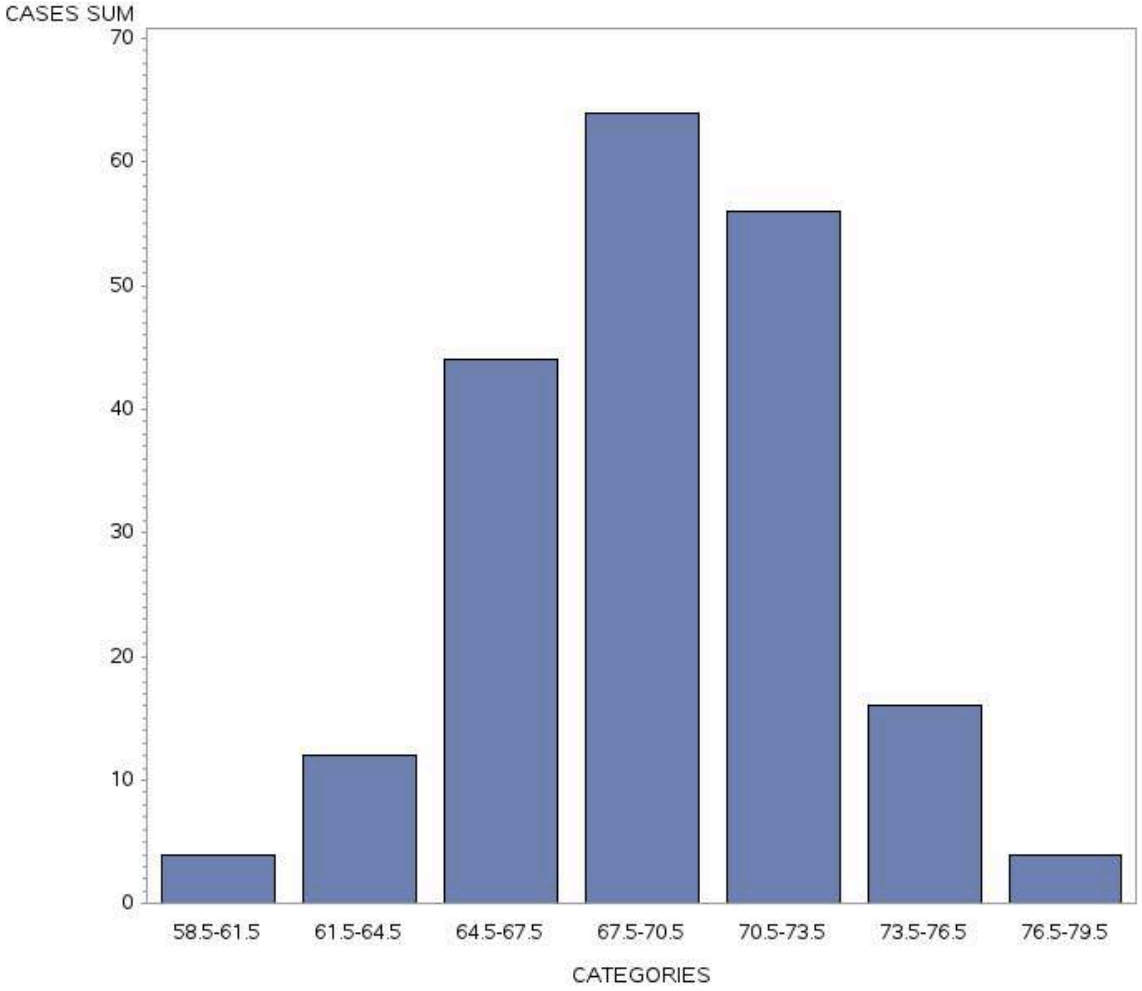
	Lowest		Highest
Value	Observation	Value	Observation
58.50	1	76.34	1
58.80	2	76.45	2
60.10	3	78.59	3
61.30	4	79.25	4
61.75	5	79.40	5

Below is a SAS program to produce a graphical presentation of the data while also creating an organized frequency distribution table.

```
OPTIONS PAGESIZE=55 LINESIZE=120 CENTER DATE;
DATA RPRT1;
INPUT DIVISION $ 1-12 CASES 14-16;
LABEL DIVISION='CATEGORIES';
DATALINES;
58.5-61.5   4
```

61.5-64.5 12
64.5-67.5 44
67.5-70.5 64
70.5-73.5 56
73.5-76.5 16
76.5-79.5 4
;
RUN;
PROC GCHART DATA=REPORTS.RPRT1; VBAR DIVISION/SUMVAR=CASES; RUN;

Figure 13.8. Vertical Bar Chart of height categories produced in SAS.



PROC FREQ DATA=REPORTS.RPRT1; WEIGHT CASES; TABLES DIVISION; RUN;

Table 13.11 Frequency table of height categories produced in SAS

division	Frequency	Percent	Cumulative Frequency	Cumulative Percent
58.5-61.5	4	2.00	4	2.00
61.5-64.5	12	6.00	16	8.00
64.5-67.5	44	22.00	60	30.00
67.5-70.5	64	32.00	124	62.00
70.5-73.5	56	28.00	180	90.00
73.5-76.5	16	8.00	196	98.00
76.5-79.5	4	2.00	200	100.00

13.4 Outliers

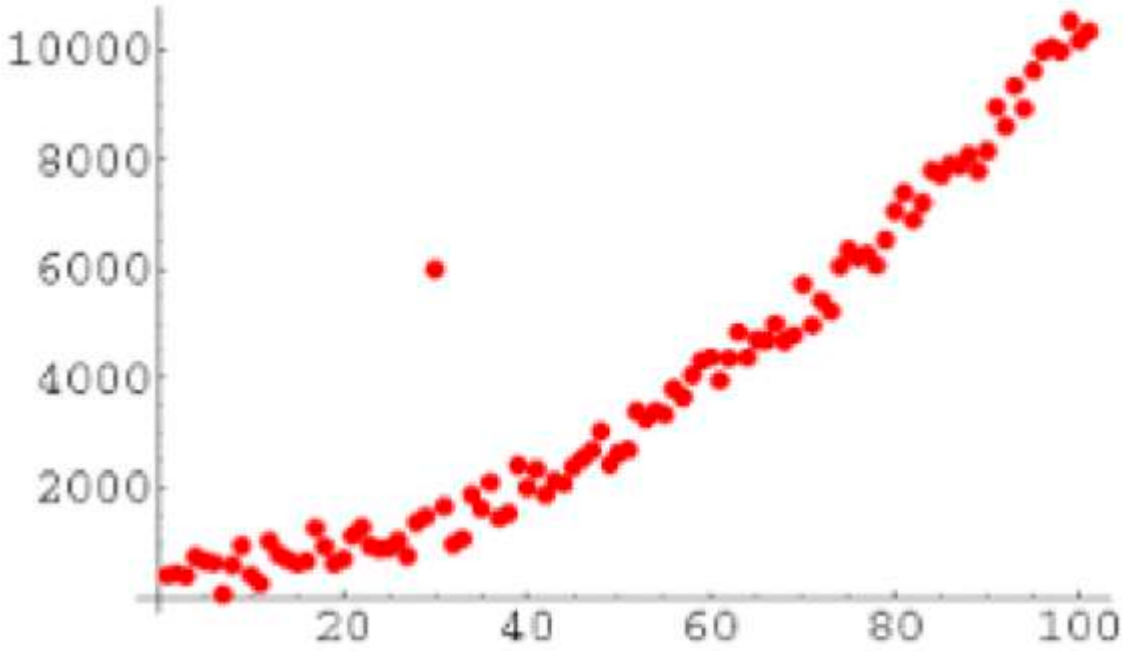
As depicted in Table 13.12. below, outliers can be defined as *Cases* with extreme values on one (univariate) or more variables (multivariate) (Tabachnick & Fidell, 2013). Outliers can be either error outliers (incorrect values) or "interesting" outliers (correct but unusual) (Orr, Sackett, & Dubois, 1991). Error outliers need to be checked against the original data for verification and then either corrected or removed from the data set. Interesting outliers are less easy to deal with (see Tabachnick and Fidell, 2013 for recommended strategies) but they are important to think about because they pull the mean towards them and have a stronger influence on the data than other values. The bottom line is that before you move on to further analysis or data transformation, it is *essential* to run a frequency analysis and screen your data for outliers.

Error outliers can be detected using PROC FREQ and checking for values that don't make sense. For example, if you had 976 as someone's age, that would be a red flag and you would investigate further.

Interesting outliers, while unusual, are still within the realm of possibility. They can be identified using the PROC GPLOT procedure outlined here This command produces a table with the five highest and five lowest values for a particular variable.

In the graph below, one person's data doesn't follow the same pattern as the rest of the sample. This is an example of an outlier.

Figure 13.12. Example of an Outlier within a Distribution


10000
8000
6000
4000
2000
20
40
60
80
100

14. Percentiles

What is a percentile?

The term “per cent” refers to “per 100”, and thus a percentile is a score representing a value relative to a base 100 scale.

The computation of percentiles is a useful way to evaluate scores within a frequency distribution, ie. the set of frequency scores.

The percentile provides a baseline at which a given proportion of scores will fall.

In other words, if we consider the 60th percentile, then we are suggesting that 60% of the scores in a distribution or set of scores will fall below that particular value.

Percentiles always refer to a specific position within a frequency distribution.

Formulas to compute percentiles for grouped data:

i) [latex]{k} = (\frac{frequency}{N} \times{100})[/latex]

ii) [latex]{\beta} = (\frac{\textit{Cumulative Frequency for all scores below the Category of Interest}}{N}) \times{100})[/latex]

iii) [latex]\textit{Percentile}={\beta} + (0.5 \times{k})[/latex]

The 0.5 is used to compute half of the number of scores within the category in which the number of interest resides.

Consider computing the percentile for the **score 71** in the frequency distribution shown in Table 14.1

Table 14.1 Frequency Distribution Output

Cell Boundaries	Freq (f)	[latex]{k} = (\frac{frequency}{N} \times{100})[/latex]	Cum. Freq.	β
58.5-61.5	4	4/200 * 100 = 0.02 * 100 = 2	4	4/200 * 100 = 2
61.5-64.5	12	12/200 * 100 = 0.06 * 100 = 6	16	16/200 * 100 = 8
64.5-67.5	44	44/200 * 100 = 0.22 * 100 = 22	60	60/200 * 100 = 30
67.5-70.5	64	64/200 * 100 = 0.32 * 100 = 32	124	124/200 * 100 = 62
70.5-73.5	56	56/200 * 100 = 0.28 * 100 = 28	180	180/200 * 100 = 90
73.5-76.5	16	16/200 * 100 = 0.08 * 100 = 8	196	196/200 * 100 = 98
76.5-79.5	4	4/200 * 100 = 0.02 * 100 = 2	200	200/200 * 100 = 100

The total sample of scores = 200. We are interested in the specific score with a value of 71. The score 71 resides within the category that has cell boundaries 70.5 to 73.5. This category has a corresponding frequency of 56, which indicates that there are 56 scores within the upper and lover boundaries of the category from 70.5 to 73.5. We can then enter 56 as the frequency value and 200 as the value of N in the following equation to determine the value of k in our series of percentile equations.

i) [latex]{k} = (\frac{frequency}{N} \times{100})[/latex]

[latex]{k} = (\frac{56}{200} \times{100}) = 28[/latex]

Here we see that in this scenario k= 28 where k represents the percent of scores in the category of interest. 56 of 200 scores represents 28% of all scores in our distribution.

Next we determine the value for [latex]{\beta}[/latex] based on the equation, [latex]{\beta} = (\frac{\textit{Cumulative frequency for all scores below the category of interest}}{N}) \times{100})[/latex]. The score for [latex]{\beta}[/latex] represents the cumulative proportion of scores in the data set up to the category in which our score of interest resides. In this example the Cumulative frequency for all scores below the category of interest refers to the cumulative frequency in the category that precedes the catergory in which our score (71) resides. Here the *Cumulative frequency for all scores below the category of Interest* is 124. Using the equation to compute [latex]{\beta}[/latex] shown here we see that the value is 62.

ii) [latex]{\beta} = (\frac{\textit{124}}{200}) \times{100}) = 62[/latex]

After we have determined k and [latex]{\beta}[/latex], we can then work through the steps in equation iii) to determine the percent of scores falling at or below our score of interest.

iii) [latex]\textit{Percentile}={62} + (0.5 \times{28})[/latex]

[latex]\textit{Percentile}={62} + (14)[/latex]

[latex]\textit{Percentile}=76^{th} \textit{percentile}[/latex]

The outcome indicates that 76 percent of the scores within this set (distribution) of scores fall below the score of 71.

Working through the computation of percentiles from a set of scores

Use *the table of frequency distributions for heights of Grade 5 elementary school children*, to compute the percentiles for the following values 123, 136, 138,149,152, indicate the values of k , and the percentile scores. Fill in the missing data in the following table to obtain a complete data set.

Table 14.2 Frequency Distribution For Heights Of Grade 5 Elementary School Children.

Category	Frequency	Cumulative Frequency
120-122	1	1
123-125	3	4
126-128	3	7
129-131	3	
132-134	1	11
135-137		13
138-140	1	14
141-143	2	
144-146	2	18
147-149	2	
150-152	3	
sum of freq=		

A SAS Application – The Scenario: ZIKA Virus at the Summer Olympics

In August 2016 Brazil hosted the Olympic Summer Games. However, several athletes decided to boycott the games because of the risk of exposure to the ZIKA virus. The ZIKA is a virus that can be transmitted through the bite of an infected Aedes mosquito. The ZIKA virus is extremely dangerous for young women as it can reside in the blood for up to 3 months and if the woman becomes pregnant, the virus can have negative consequences for the developing fetus. In particular, the ZIKA virus has been implicated in the development of microcephaly in newborn children.

In this example, we will use a series of random number generating commands to create a data set with four variables and 1000 cases. The variables are sex, sport and case and will use the following format: sex (1=m, 2=f), sport (1=golf, 2=equestrian, 3=swimming, 4=gymnastics, 5=track & field), case (1=yes, 2=no), and days which is a continuous variable representing the number of days since exposed to ZIKA virus-carrying mosquitoes.

A SAS Application – The Scenario: ZIKA Virus at the Summer Olympics

```
   PROC FORMAT;
VALUE SEXFMT 1 ='MALE'  2 ='FEMALE';
VALUE SPRTFMT 1 ='GOLF'  2 ='EQUESTRIAN'  3 ='SWIMMING'
4 ='GYMNASTICS'  5 ='TRACK & FIELD';
VALUE CASEFMT  1='PRESENT'  2='ABSENT';
   DATA SASRNG;
   /* Create 3 new variables labelled SCORE1 SCORE2 SCORE3 */
   ARRAY SCORES SCORE1-SCORE3;
   /* Set 1000 cases per variable */
   DO K=1 TO 1000;
   DAYS=RANUNI(13)*100;
   DAYS=ROUND(DAYS, 0.02);
   /* Loop through each variable to establish 1000 randomly generated scores */
   DO I=1 TO 3;
   SCORES(I)=RANUNI(I)*1000;
   SCORES(I)=ROUND(SCORES(I));
   SCORES(I)=1+(MOD(SCORES(I),105));
   /*  The variable sex will relate to score1, create a filter to establish the binary score for sex based on the ran-
domly generated output */
   IF SCORE1 > 55 THEN SEX = 2;
   IF SCORE1 >2 AND SCORE1<56 THEN SEX = 1;
   /* Sport Type   */
   IF SCORE2 >90 THEN SPORT = 5;
   IF SCORE2 >80 AND SCORE2<91 THEN SPORT = 4;
   IF SCORE2 >60 AND SCORE2<81 THEN SPORT = 3;
   IF SCORE2 >30 AND SCORE2<61 THEN SPORT = 2;
   IF SCORE2 >5 AND SCORE2<31 THEN SPORT=1;
   /* Case */
   IF SCORE3 > 48 THEN CASE = 1;ELSE CASE = 2;
```

```
END;
OUTPUT;
END; RUN;
PROC SORT DATA =SASRNG; BY SEX;
PROC FREQ; TABLES SEX SPORT CASE SEX*CASE;
FORMAT SEX SEXFMT. SPORT SPRTFMT. CASE CASEFMT. ;
PROC FREQ; TABLES SPORT*CASE;BY SEX;
FORMAT SEX SEXFMT. SPORT SPRTFMT. CASE CASEFMT. ;
PROC UNIVARIATE; VAR DAYS;
OUTPUT OUT=PCTLS PCTLPTS  = 30 60
PCTLPRE  = DAYS_
PCTLNAME = PCT30 PCT60;
PROC PRINT DATA= PCTLS;
RUN;
```

In SAS we can compute the specific percentiles using the PROC UNIVARIATE; feature on the continuous variable. The command PROC UNIVARIATE; VAR days; produces the following output table to produce a chart of percentiles for the variable: DAYS.

Table 14.3 Frequency Distribution Output Showing Percentiles

Level	Quantile
100% Max	99.94
99%	98.66
95%	94.34
90%	89.61
75% Q3	73.13
50% Median	46.83
25% Q1	24.75
10%	10.23
5%	4.86
1%	1.27
0% Min	0.02

However, we can also compute specific percentile values for a continuous variable using the PCTLPTS=, PCTLPRE=, and PCTLNAME= options.

Together these three commands help us to identify and label specific percentiles within a data set. For example, to select a specific percentile, such as the 30th percentile we use PCTLPTS= 30. The command PCTLPRE= provides the specific prefix in the label for a percentile. For example, here we use the prefix days_ and then follow the command with the PCTLNAME= command to list the label of the percentile. For example, the sequence of commands: PCTLPTS=

30, fPCTLPRE= DAYS_, and the PCTLNAME= pct30, identifies and labels the 30th percentile within the data set. In the following code we compute the 30th and 60th percentiles for the continuous variable: DAYS, using SAS Commands to identify specific percentiles.

SAS CODE to produce specific percentiles

```
output out=Pctls pctlpts = 30 60
pctlpre = days_
pctlname = pct30 pct60;
```

OUTPUT from the code above:

Obs	days_pct30	days_pct60
1	28.64	57.08

The **PROC FREQ** procedure in SAS enables us to create descriptive tables for the frequency distribution of the categorical variables. For example, we can compute the number of females and males in our sample, as well as the number of individuals across each of the sports, and then we can actually create a number to represent the number of cases of ZIKA in our randomly generated data set of 1000 participants.

TABLE 14.5 ZIKA Random Number Generated data for SEX

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
male	533	53.30	533	53.30
female	467	46.70	1000	100.00

TABLE 14.6 ZIKA Random Number Generated data for Sports

sport	Frequency	Percent	Cumulative Frequency	Cumulative Percent
golf	266	26.60	266	26.60
equestrian	286	28.60	552	55.20
swimming	192	19.20	744	74.40
gymnastics	96	9.60	840	84.00
track & field	160	16.00	1000	100.00

TABLE 14.7 ZIKA Random Number Generated data for Disease Present/Absent

case	Frequency	Percent	Cumulative Frequency	Cumulative Percent
present	505	50.50	505	50.50
absent	495	49.50	1000	100.00

This procedure also enables us to create cross-tabular tables for comparisons of variables.

TABLE 14.8 ZIKA Random Number Generated Cross Tabulations

Table of Frequencies for case by sex			
SEX	CASES		
	Present	Absent	Total
Male	275	258	533
Female	230	237	467
COLUMN TOTALS	505	495	1000

As in most SAS procedures, by including the PROC SORT command, we can arrange the processing and subsequent output of the data to control for the categorical variable(s). In this example we computed the cross-tabulation of the frequency distribution for the variables SPORT and CASE, controlling for SEX, to separate the output for Males and Females.

The table format provides the following data within each cell: frequency, followed by cell percent, followed by row percent, followed by column percent as shown in this example for the sport: golf.

TABLE 14.9 ZIKA Random Number Generated Cross Tabulations

Table of Frequencies for case by sports			
SPORT	CASES		
	Present	Absent	Total
MALE GOLF	Cell Freq = 73 Cell Pct = 13.70 Row Pct = 53.28 Col Pct = 26.55	Cell Freq = 64 Cell Pct = 12.01 Row Pct = 46.72 Col Pct = 24.81	Row Total = 137 Row Pct = 25.70
FEMALE GOLF	Cell Freq = 56 Cell Pct = 11.99 Row Pct = 43.41 Col Pct = 24.35	Cell Freq = 73 Cell Pct = 15.63 Row Pct = 56.59 Col Pct = 30.80	Row Total =129 Row Pct = 27.62
COLUMN TOTALS	505	495	1000

15. Introducing the Goodness of Fit Chi-Square

So you are asking yourself, "goodness of fitting what to what?"

The chi-square (pronounced "kie" square) is an extremely useful, non-parametric statistical technique, that allows a researcher to compare responses from a sample to expected responses in a – hypothetical distribution of responses for a population. Hence the name goodness of fit test.

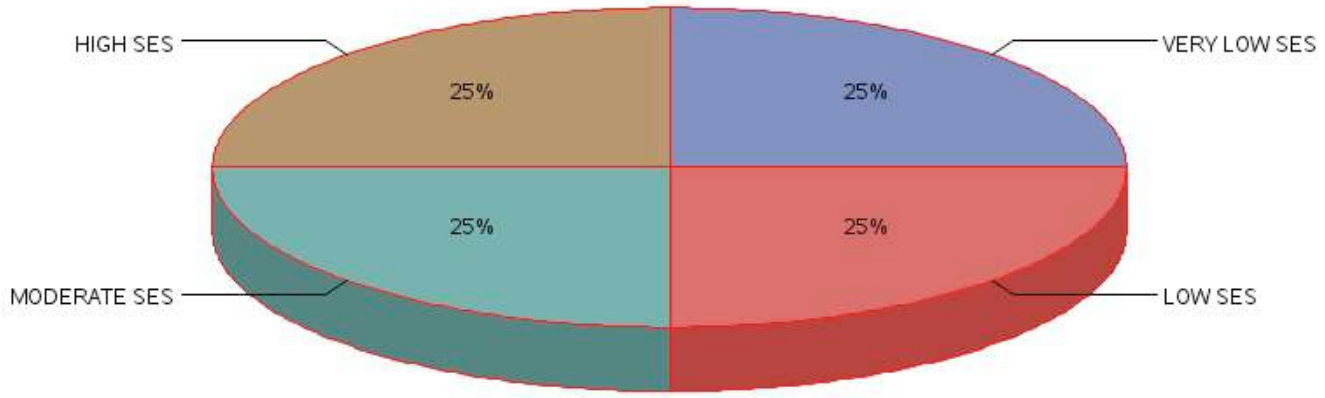
The chi-square goodness of fit test can be used to evaluate data at all variable levels, but because the currency of this test is count data, the goodness of fit test can be used to compute nominal and ordinal data.

The chi-square test evaluates data in the form of counts or frequencies, as in the number of responses within a given category, or the number of people who responded a given way to a specific question, or the number of cases across outcome categories.

The goodness of fit chi-square for one sample with four categories

In the following example, we consider the goodness of fit chi-square with four response categories. In this problem, we are studying a cohort of cancer patients to determine if cancer was more likely to be diagnosed in patients who are in a low-income category, based on socio-economic status (SES) quartiles. We begin by establishing that the expected distribution of cancer patients within the community is equally distributed across the four income categories so that in any community 25% of our population are in the highest SES category, 25% are in the moderate SES category, 25% are in the low SES income category, and 25% are in the very low SES category.

Proportional Distribution of Sample Across Socioeconomic Categories



However, in the observed data set for our sample of cancer patients, we recorded the following distribution of patients.

Highest SES 25%	Moderate SES 25%	Lower SES 25%	Very Low SES 25%
Data from the community sample of cancer patients collected over a 10 year period in a community with an average population of greater than 1 million households			
165 patients	283 patients	622 patients	980 patients

The null hypothesis for this study is stated in an unbiased way so that each SES quartile is expected to have an equal percentage of households with cancer patients. Therein, the term $f_{(k)}$ = refers to the frequency or number of patients within the quartile indicated by the subscript (k). Since we have four groups representing four quartiles then (k) ranges from 1 to 4.

H_0: $f_1 = f_2 = f_3 = f_4$

Since we have a total sample size of N = 2050, then each cell of the SES quartiles is expected to have a frequency (an expected number of patients) equal to 512.5 individuals.

The chi-square formula to test the null hypothesis is:

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

The equation measures how closely an observed set of responses (the"o" for "*observed*") matches an expected set of responses (the "e" for "*expected*").

So then how do we calculate the items that we use in the chi-square equation?

The observed frequencies are simply taken from the data recording sheet, but the expected frequencies are computed from the following formula:

$$\text{expected frequency} = \frac{\text{Total Frequency}}{\text{\# of response choices}}$$

Another way to view the computation of the expected frequencies is to consider the null hypothesis which stated that:

H_0: $f_1 = f_2 = f_3 = f_4$

and multiply the total frequency by the probability associated with each category, as in the following computations.

2050 x 0.25 = 512.5

The chi-square is then used to compute whether or not the observed distribution fits a hypothetical or expected distribution. This can be accomplished by setting up the following table below:

Response Category	Observed Frequency	Expected Frequency	$(Obs - Exp)^2 : Exp$
1: High SES	165	512.5	235.62
2: Moderate SES	283	512.5	102.77
3: Low SES	622	512.5	23.40
4: Very Low SES	980	512.5	425.45

$$\chi^2 = \sum \frac{(o-e)^2}{e} = \mathbf{788.24}$$

In this calculation for a one-sample scenario with 4 outcome categories, we see that the Here the chi-square statistic is: 788.24. So what does this mean?

To evaluate the meaning of the variable we calculated for the Chi-square we need to review the decision rule for the Chi-square statistic, and shown here.

Chi-Square decision rule (one-sample chi-square test):

The computed score is referred to as the chi-square observed. After computing the chi-square observed value, determine the chi-square critical score from a table of chi-square values. The chi-square critical score represents what we should expect to observe for a distribution with five responses. The critical value is determined by computing the degrees of freedom for our response set.

The computation of the degrees of freedom is:

degrees of freedom = k possible responses -1

degrees of freedom = 5-1

degrees of freedom = 4

and the chi-square critical value for degrees of freedom of 4 at p<0.05 = 9.49

If the chi-square observed value is GREATER THAN the chi-square critical value of 9.49, we must reject the null hypothesis and state that the distribution of responses across the four categories IS NOT EQUAL. A large chi-square value, that is a value that exceeds the chi-square critical value demonstrates that the outcome is less likely to occur by chance.

The chi-square statistic is computed as 788.24.

We, therefore, compare the chi-square observed value of 788.24 against a chi-square expected, based on the expected probability level and the degrees of freedom. In the k=4 chi-square, the degrees of freedom are: degrees of freedom = "k" possible responses -1, so that given k=4, then the degrees of freedom is 4-1 = 3 and at p<0.05 the chi-square critical value is 7.82. Therefore, since our chi-square observed value of 788.24 exceeds the chi-square critical (7.82) we reject the null hypothesis and state that the distribution of cancer patients is not equally distributed across the SES categories, and given the numbers we observed we can state that in this sample, the number of cancer patients in the very low SES group was significantly greater than the number of cancer patients in the high socio-economic category.

The following is the SAS code used to analyze the data in the scenario above.

```
  PROC FORMAT;
VALUE SLICE 1='HIGH SES' 2='MODERATE SES' 3='LOW SES' 4='VERY LOW SES';
DATA GFIT_1;
INPUT SESGRP N_PATNTS;
/* DEFINE THE AXIS CHARACTERISTICS */
AXIS1 LABEL=("SES CATEGORIES")
VALUE=(JUSTIFY=CENTER);
AXIS2 LABEL=(ANGLE=90 "ACTUAL NUMBER OF PATIENTS")
ORDER=(0 TO 1000 BY 100)
MINOR=(N=3);
AXIS3 LABEL=(ANGLE=90 "SES CATEGORIES");
AXIS4 LABEL=("ACTUAL NUMBER OF PATIENTS") ;
DATALINES;
1 165
2 283
3 622
4 980
```

```
;
/* HERE WE USE THE OPTION SUMVAR TO GRAPH THE SUM OF THE FREQ */
PROC FREQ ORDER=DATA; TABLES SESGRP/CHISQ CL CELLCHI2;
WEIGHT N_PATNTS;
FORMAT SESGRP SLICE. ;
TITLE 'FREQUENCY DISTRIBUTION FOR PROPORTION OF PATIENTS IN EACH SES GROUP';
TITLE2 'ONE SAMPLE GOODNESS OF FIT EXAMPLE FOR K=4';
RUN;
```

The output for Chi-square computation is shown here:

FREQUENCY DISTRIBUTION FOR PROPORTION OF PATIENTS IN EACH SES GROUP – ONE SAMPLE GOODNESS OF FIT EXAMPLE

The FREQUENCY procedure including the chi-square statistic to evaluate the null hypothesis H_0: $f_1 = f_2 = f_3 = f_4$.

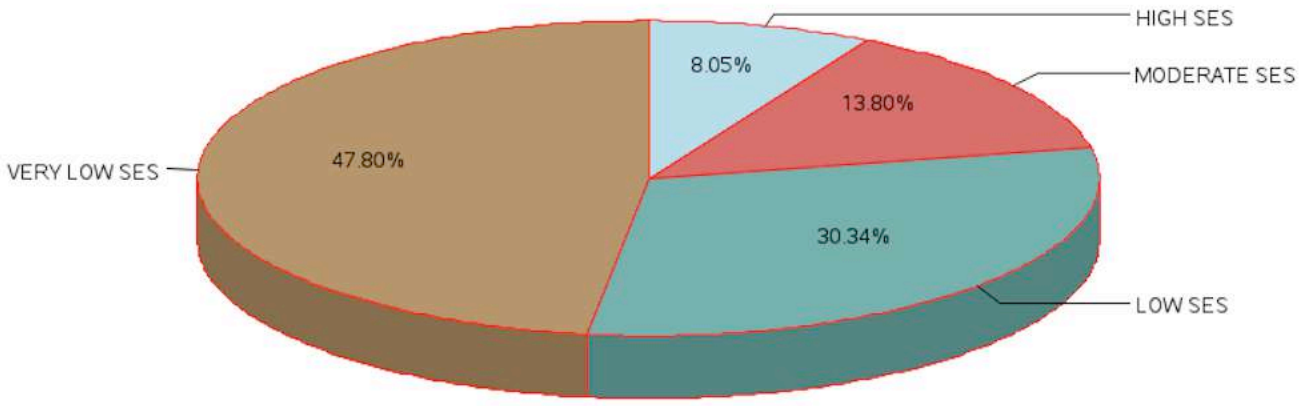
SES GRPS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
HIGH SES	165	8.05	165	8.05
MODERATE SES	283	13.80	448	21.85
LOW SES	622	30.34	1070	52.20
VERY LOW SES	980	47.80	2050	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	788.2400
DF	3
Pr > ChiSq	<.0001

The SAS code to produce the pie chart is as follows:

```
  PROC FORMAT;
VALUE SLICE 1='HIGH SES' 2='MODERATE SES' 3='LOW SES' 4='VERY LOW SES';
PROC GCHART DATA=GFIT_1;
PIE3D SESGRP/SUMVAR=N_PATNTS TYPE=SUM DISCRETE PERCENT=inside
COUTLINE=RED WOUTLINE=1 FILL=SOLID SLICE =ARROW CLOCKWISE
NOLEGEND NOHEADING VALUE=NONE;
FORMAT SESGRP SLICE. ;
TITLE1 'PIE CHART FOR PROPORTION OF PATIENTS IN EACH SES GROUP';
PATTERN1 COLOR = LIGHTBLUE;
RUN;
```

PIE CHART FOR PROPORTION OF PATIENTS IN EACH SES GROUP



Webulator Form 1:

The following is a Goodness of Fit Webulator for k= 4 responses In the table above we used the values for socioeconomic status:

Distribution of individuals across SES	
HIGH SES	165
MODERATE SES	283
LOW SES	622
VERY LOW SES	980

Enter these data into the webulator below for each of your four options and then click the button labelled ***compute expected frequencies***. This will produce the sum of the four values that you entered and compute the expected frequency for the values in the table.

An interactive or media element has been excluded from this version of the text. You can view it online here: https://pressbooks.library.upei.ca/montelpare/?p=375

The important value from this *Webulator* is the computed chi-square score. The computed score is referred to as the chi-square observed. After computing the chi-square observed value, determine the chi-square critical score from a table of chi-square values. The chi-square critical score represents what we should expect to observe for the distribution with "k" responses. The critical value is determined by computing the "degrees of freedom" for our response set.

The computation of the degrees of freedom is: degrees of freedom = "k" possible responses -1

degrees of freedom = 4-1 –> degrees of freedom = 3

and the "chi-square critical value" for degrees of freedom of "3" at $p<0.05$ = 7.815

If the "chi-square observed value " is **›** the "chi-square critical value of **7.815**", we must reject the null hypothesis and state that the distribution of responses across the response categories IS NOT EQUAL.

If you would you like to use the Webulators for your own applications, without this text visit: https://health.ahs.upei.ca/webulators/w_menu.php

This Webulator application to compute the one sample goodness of fit with k=4 is available at `https://health.ahs.upei.ca/webulators/4k_Gf.php`

16. Goodness of Fit Chi-Square for k=5

In the following example, we consider the goodness of fit chi-square test with **five** response categories.

The biweekly lottery – Lotto 649 provides players with an opportunity to win millions of dollars if they can select the set of six numbers that are randomly drawn from the set of numbers from 1 to 49. Since the lottery is purported to be random, the chance associated with a player's single ticket matching the six numbers drawn at random is based on the combinatorial formula for 49 choose 6 and has a probability of $1 : {}^{49}C_6$.

The probability associated with every single ticket is the same and is 1 in 13,983,816 possible combinations of 6 numbers. So then, what if we wanted to test the randomness of this lottery? In the following example, we will use the chi-square goodness of fit test to determine if each number is random with respect to selection, and that there is no apriori pattern of numbers from one range or another within the set of 49 occurring more frequently or with a systematic selection pattern.

To begin we need to organize the range of possible outcomes into manageable categories that can be processed with the chi-square goodness of fit test. Given that the range of all possible outcomes for the lotto is from 1 to 49, we can organize the potential sampling space (1 to 49) into 5 categories as follows 1-9, 10-19, 20-29, 30-39, 40-49.

Further, if we wanted to test randomness, then we would need to sample more than one single week of numbers, so considering that there are 104 draws per year we could use an entire year's worth of data to establish the frequency distribution of numbers drawn, and after organizing the outcomes into the 5 categories determine the chi-square goodness of fit, statistically.

Step 1: after establishing that there are 5 categories for the outcome frequency distribution we would expect that the distribution or organization of the responses should be equal across all of the possible responses categories as follows:

Data represent the actual numbers that are drawn in a single year for the lotto 649. That is, in any given year there are 104 draws, which is based on 2 draws per week for 52 weeks. Therefore, in the lotto 649 example we have 624 possible numbers drawn à (2 draws per week for 52 weeks = 6 numbers x 104). These data can then be organized into the following 5 categories to represent the set of all possible numbers drawn in the one year so that a frequency distribution chart of the responses might look like this:

1 – 9	124 numbers
10 – 19	125 numbers
20 – 29	125 numbers
30 – 39	125 numbers
40 – 49	125 numbers

Therefore, we can say from this chart that our responses to the research question should be evenly distributed across all of the possible responses.

Such a response pattern is consistent with our expected distribution. In other words, in an unbiased research study, we should expect that all possible responses are equally as likely to occur. We call this the unbiased null hypothesis, and state this in terms of frequencies of responses which are represented as f(k) = and is shown as follows:

$$H_0: f_1 = f_2 = f_3 = f_4 = f_5$$

Therefore, based on the null hypothesis, considering that each response category should have an equal number of responses, the formula to compute the expected responses might be as follows:

$$\text{expected frequency} = \frac{\text{Total Frequency}}{\text{\# of response choices}}$$

Now then let's consider the following example. We asked students to generate 52 weeks of biweekly draws of the lotto 649 and then to sort the data so that we could simulate a test of the outcome distribution to determine how random the simulated lottery is at selecting numbers. The response options for the data produced for my 104 draws are given in the following table.

1-9	146
10-19	155
20-29	282
30-39	12
40-49	29

From this data, it would appear that a large proportion of the scores were found in the 20-29 range and only a few scores were found in the 30-39 range 283 of 624 or 28.5%. However, the lowest proportions of the scores (choices drawn) 12 of 624 = 1.9%

The chi-square test is, therefore, a useful statistical test to determine if the overall distribution of the responses in the observed sample is similar to or matches the expected distribution of responses in the target population (the "target population" being defined as all scores drawn) in this one year of simulated data.. The equation below is the basic equation for the goodness of fit chi-square test.

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

The equation shown here measures how closely an observed set of responses (the "*o*" for "*observed*") matches an expected set of responses (the "*e*" for "*expected*").

So then how do we calculate the items that we use in the chi-square equation?

The observed frequencies are simply taken from the data recording sheet, but the expected frequencies are computed from the following formula:

$$\text{expected frequency} = \frac{\text{Total Frequency}}{\text{\# of response choices}}$$

Another way to view the computation of the expected frequencies is to consider the null hypothesis which stated that:

$$\mathbf{H_0: f_1 = f_2 = f_3 = f_4 = f_5}$$

and multiply the total frequency by the probability associated with each category, as in the following computations.

$$624 \times 0.20 = 124.8$$

The chi-square is then used to compute whether or not the observed distribution fits a hypothetical or expected dis-

tribution. This can be accomplished by applying the formula to each row of the response table. The computation of the first row is shown here:

$$\chi^2 = \frac{(obs-exp)^2}{exp} = \frac{(146-124.8)^2}{124.8} = \frac{(21.2)^2}{124.8} = \frac{449.44}{124.8} = 3.6$$

Response Category	Observed Frequency	Expected Frequency	(Obs – Exp)2 : Exp
1 – 9	146	124.8	3.60
10 – 19	155	124.8	7.31
20 – 29	282	124.8	198.01
30 – 39	12	124.8	101.95
40 – 49	29	124.8	73.54

In the calculation of the chi-square we see that in each row of the table, the observed score from the sample is subtracted from the expected score that represents the scores of the population. For example in ROW_1 of the table the observed score of 146 is subtracted from the expected score of 124.8. The difference of 21.2 is squared and the outcome is divided by 124.8, and the resulting value is 3.6. The calculation is repeated for each row of the table and the outcomes are added together to produce the chi-square value as shown below.

Response Category	Observed Frequency	Expected Frequency	(Obs – Exp)2 : Exp
$\chi^2 = \sum \frac{(o-e)^2}{e}$			3.60 + 7.31 + 198.01 + 101.95 + 73.54 **384.41**

Our next step is then to determine if the chi-square observed value is greater than the chi-square critical value, so that we can make a decision about the significance of the observed distribution.

Chi-Square decision rule for the one sample chi-square test.

The computed score is referred to as the chi-square observed. After computing the chi-square observed value, determine the chi-square critical score from a table of chi square values. The chi-square critical score represents what we should expect to observe for a distribution with five responses. The critical value is determined by computing the degrees of freedom for our response set.

The computation of the degrees of freedom is:

degrees of freedom = k possible responses -1

degrees of freedom = 5-1

degrees of freedom = 4

and the chi-square critical value for degrees of freedom of 4 at p<0.05 = 9.49

If the chi-square observed value is GREATER THAN the chi-square critical value of 9.49, we must reject the null hypothesis and state that the distribution of responses across the four categories IS NOT EQUAL. A large chi-square value, that is a value which exceeds the chi-square critical value demonstrates that the outcome is less likely to occur by chance.

Using the degrees for freedom for a one-sample chi-square, our degrees of freedom are:

degrees of freedom = "k" possible responses -1

degrees of freedom = 5-1

degrees of freedom = 4

and the " chi-square critical value" for degrees of freedom of "4" is 9.49

Therefore, because our chi-square observed value of 384.41 is › the chi-square critical value of 9.49, we must reject the null hypothesis and state that the distribution of responses across the four categories IS NOT EQUAL.

We can check our calculations with the following SAS Program. This program produces a frequency distribution with chi-square analysis to evaluate the null hypothesis (see above), as well as a pie chart to show the proportion of times a number from each category was drawn in the lotto.

ONE SAMPLE GOODNESS OF FIT CHI-SQUARE FOR K=5

```
   PROC FORMAT;
VALUE SLICE 1='#1 to #9' 2='#10 to #19' 3='#20 to #29'
4='#30 to #39' 5='#40 to #49';
DATA GFIT_2;
INPUT LOTTOGRP N_DRAWS;
/* DEFINE THE AXIS CHARACTERISTICS */
AXIS1 LABEL=("LOTTO CATEGORIES")
VALUE=(JUSTIFY=CENTER);
AXIS2 LABEL=(ANGLE=90 "N TIMES CATEGORY VALUE DRAWN")
ORDER=(0 TO 1000 BY 100)
MINOR=(N=3);
AXIS3 LABEL=(ANGLE=90 "LOTTO CATEGORIES");
   AXIS4 LABEL=("N TIMES CATEGORY VALUE DRAWN") ;
DATALINES;
1 146
2 155
3 282
4 12
5 29
;
/* HERE WE USE THE OPTION SUMVAR TO GRAPH THE SUM OF THE FREQ */
```

```
PROC FREQ ORDER=DATA; TABLES LOTTOGRP/CHISQ CL CELLCHI2;
WEIGHT N_DRAWS;
FORMAT LOTTOGRP SLICE. ;
TITLE 'FREQUENCY DISTRIBUTION FOR N TIMES CATEGORY VALUE WAS DRAWN';
TITLE2 'ONE SAMPLE GOODNESS OF FIT EXAMPLE K=5';
RUN;
PROC GCHART DATA=GFIT_1;
PIE3D LOTTOGRP/SUMVAR=N_DRAWS TYPE=SUM DISCRETE PERCENT=ARROW
COUTLINE=RED WOUTLINE=1 FILL=solid SLICE = arrow clockwise
noLEGEND noheading value=none;
FORMAT LOTTOGRP SLICE. ;
TITLE1 'PIE CHART FOR N TIMES CATEGORY VALUE WAS DRAWN';
PATTERN1 COLOR = LIGHTBLUE;
Run;
```

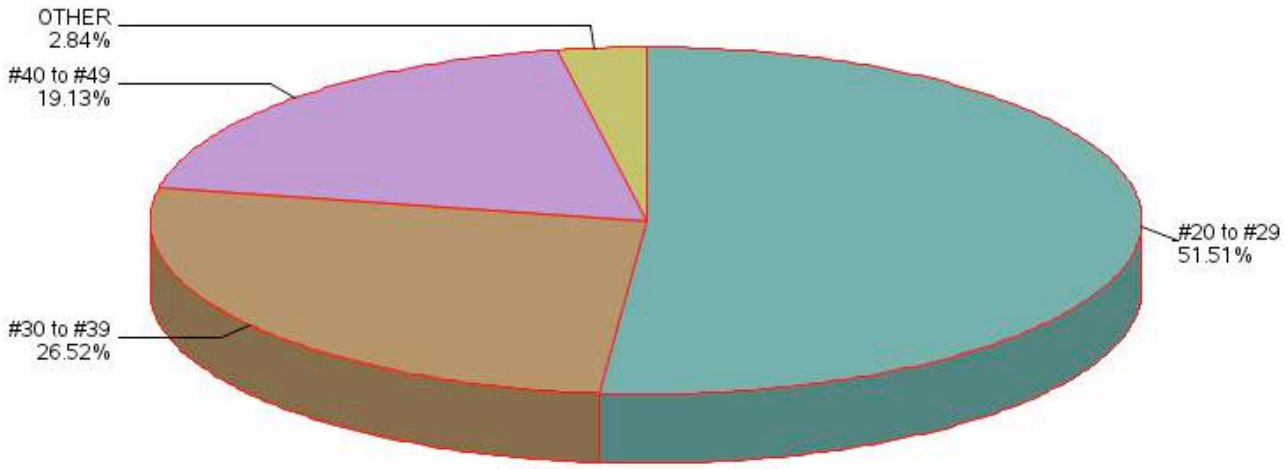
The output for the frequency distribution with corresponding chi-square is shown here:

The FREQ Procedure

LOTTOGRP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
#1 to #9	146	23.40	146	23.40
#10 to #19	155	24.84	301	48.24
#20 to #29	282	45.19	583	93.43
#30 to #39	12	1.92	595	95.35
#40 to #49	29	4.65	624	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	384.4135
DF	4
Pr > ChiSq	<.0001

The pie chart for the number of times a value was drawn within each category, expressed as a percent is shown here.



Webulator Form 1:

The following is a Goodness of Fit Webulator for k= 5 responses In the example above our raw data values for the cumulative times that a number was drawn from each category of the Lotto is shown here:

Distribution of Draws per Category	
#1 to #9	146
#10 to #19	155
#20 to #29	282
#30 to #39	12
#40 to #49	29

Enter these data into the webulator below for each of your category options and then click the button labeled ***CLICK ME***. This will produce the sum of the five values that you entered and compute the expected frequency for the values in the table.

An interactive or media element has been excluded from this version of the text. You can view it online here: https://pressbooks.library.upei.ca/montelpare/?p=382

The important value from this *Webulator* is the computed chi-square score. The computed score is referred to as the chi-square observed. After computing the chi-square observed value, determine the chi-square critical score from a table of chi-square values. The chi-square critical score represents what we should expect to observe for the distribution with "k" responses. The critical value is determined by computing the "degrees of freedom" for our response set.

The computation of the degrees of freedom is: degrees of freedom = “k” possible responses -1

degrees of freedom = 4-1 –> degrees of freedom = 3

and the “chi-square critical value” for degrees of freedom of “3” at p<0.05 = 7.815

If the “chi-square observed value ” is › the “chi-square critical value of **7.815**”, we must reject the null hypothesis and state that the distribution of responses across the response categories IS NOT EQUAL.

If you would you like to use the Webulators for your own applications, without this text visit: https://health.ahs.upei.ca/webulators/w_menu.php

This Webulator application to compute the one sample goodness of fit with k=5 is available at `https://health.ahs.upei.ca/webulators/goodfit2.php`

17. The Goodness of Fit Test for Two Groups

The Two-Sample Chi-Square Goodness of Fit Test

In this chapter, we will work through examples of the Goodness of Fit chi-square when we have two groups. Here we will use both SAS coding as well as the two sample webulator for a goodness of fit test. The two sample webulator enables us to compare the distribution of responses for one sample against the distribution of the responses for a second sample.

In the following example, we applied the goodness of fit test for a sample of individuals that were asked about their health status. The tool to collect the information was the RAND SF-36. In this example, we also added demographic information to represent sex, and although the response categories for SEX were (1=male, 2=female and 3=other) we processed the data as a binary outcome (males versus females). The data set was comprised of three variables which included id, sex and the individual's response to the five-item question: 1. In general, would you say your health is: i) Excellent, ii) Very good, iii) Good, iv) Fair, v) Poor.

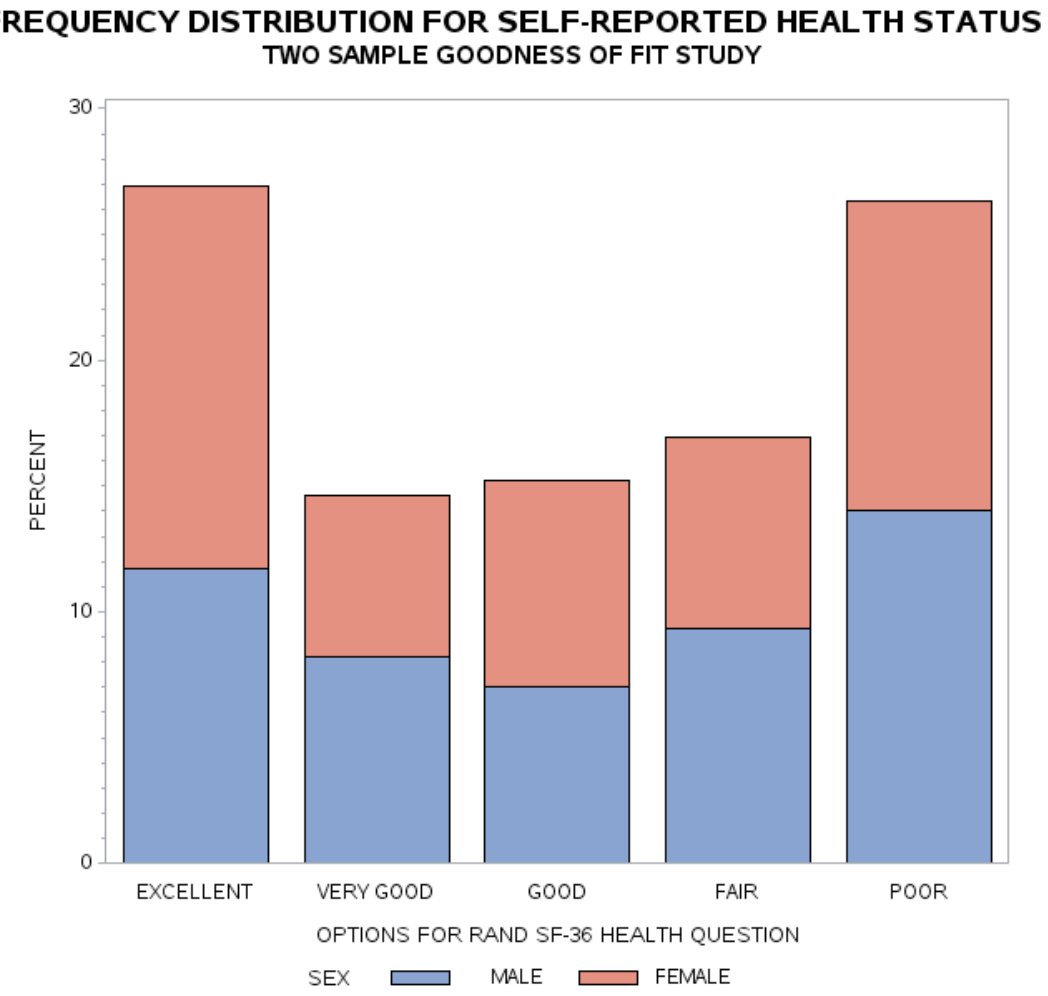
The relevant SAS code added to process the 2 group chi-square goodness of fit test

```
   PROC FORMAT;
VALUE HEAFMT 1 = EXCELLENT 2 = VERY GOOD 3 = GOOD 4 = FAIR 5 = POOR;
VALUE GENFMT 1=MALE 2=FEMALE 3=OTHER;DATA CHIGF2;
INPUT ID SEX HEALTH @@;LABEL HEALTH='OPTIONS FOR RAND SF-36 HEALTH QUESTION';
TITLE 'TWO GROUP GOODNESS OF FIT FOR HEALTH STATUS RAND SF-36 ';DATALINES;
001 1 1 002 1 2 003 1 3 004 1 4 005 1 5 006 2 1  007 2 2 008 2 3 009 2 4 010 2 5 011 3 1 012 1 2 013 3 3 014 1 4 015 1 5
016 2 1 017 2 2 018 2 3 019 2 4 020 2 5 021 1 1 022 1 2 023 1 3 024 1 4 025 1 5 026 2 1 027 2 2 028 2 3 029 2 4 030 2 5
041 1 1 042 1 1 043 1 1 044 1 1 045 3 1 046 2 1 047 2 1 048 2 1 049 2 1 050 3 1 031 1 5 032 1 5 033 1 5 034 3 5 035 1 5
036 2 5 037 3 5 038 3 5 039 2 5 040 2 5 101 1 1 102 1 2 103 1 3 104 1 4 105 1 5 106 2 1 107 2 2 108 2 3 109 2 4 060 3
5 061 3 1 062 1 2 063 1 3 064 1 4 065 1 5 066 2 1 067 3 2 068 2 3 609 2 4 700 3 5 081 1 1 082 3 2 083 1 3 084 1 4 085 1
5 086 3 1 087 2 2 088 2 3 089 2 4 090 2 5 081 1 1 082 1 1 083 1 1 084 1 1 085 1 1 086 2 1 087 2 1 088 2 1 089 2 1 080 2
1 051 3 5 052 3 5 053 1 5 054 1 5 055 1 5 056 2 5 057 3 5 058 3 5 059 2 5 100 2 5 160 2 5 161 3 1 162 1 2 613 1 3 641 1 4
651 1 5 166 2 1 167 3 2 168 2 3 169 2 4 170 2 5 181 1 1 182 1 2 183 3 3 184 1 4 185 3 5 186 2 1 187 2 2 188 2 3 189 3 4 190 2
5 181 3 1 182 1 1 288 3 1 289 2 1 280 2 1 251 1 4 252 1 3 253 1 4 254 3 3 255 1 4 256 2 3 257 2 4 258 2 4 259 2 3 100 2 2
160 2 5 161 1 1  162 1 2 613 1 3 641 1 4 651 1 5 166 2 1 167 2 2 988 2 1 389 2 3 380 2 1 351 3 5 352 1 5 353 1 5  354 1 5 355
1 5 356 2 5 357 3 5 358 2 5 359 2 5 100 2 5 160 2 5 161 1 1 162 3 2 613 1 3 641 1 4 651 1 5 166 2 1 167 2 2 560 2 5 561 1 1
562 1 2  563 1 3 564 1 4 565 1 5 566 2 1 567 2 2 568 2 3 569 2 4 570 2 5 581 1 1 582 1 2 583 1 3 584 3 4 585 1 5 586 2 1
587 2 2 588 2 3 589 2 4 590 3 5 581 1 1 582 1 2 583 1 2 584 1 2 585 1 2 586 3 2 587 2 1 588 2 1 589 2 3 580 3 3 551 1 3
552 1 5  553 1 4 554 1 5 555 1 4 556 3 5 557 2 4 558 2 4 559 3 3
;
/* PRODUCE A HISTOGRAM FOR THE ENTIRE SET OF DATA*/PROC SORT DATA=CHIGF2; BY SEX;
PROC SGPLOT; HISTOGRAM HEALTH;
FORMAT HEALTH HEAFMT. ;
RUN;/* CALCULATE CHI SQUARE GOODNESS OF FIT – MALES VS FEMALES */PROC FREQ;
TABLES HEALTH*SEX/CHISQ;
```

```
WHERE SEX<3; /* RESTRICT DATA TO A TWO GROUP COMPARISON */
FORMAT HEALTH HEAFMT. SEX SEXFMT. ;
TITLE 'FREQUENCY DISTRIBUTION FOR SELF-REPORTED HEALTH STATUS';
TITLE2 'TWO SAMPLE GOODNESS OF FIT STUDY';
RUN;/*CREATE A GRAPH USING COLORS */
/* Define the axis characteristics */
axis1 offset=(0,70) minor=none;
axis2 label=(angle=90);
pattern1 value=solid color=cx7c95ca;
pattern2 value=solid color=cxde7e6f;proc sort; by SEX;
proc gchart ;
vbar HEALTH / SUBGROUP=SEX TYPE=PERCENT
discrete raxis=axis2;
WHERE SEX<3; /* RESTRICT DATA TO A TWO GROUP COMPARISON */
FORMAT HEALTH HEAFMT. SEX GENFMT. ;
/* Define the title */
TITLE 'FREQUENCY DISTRIBUTION FOR SELF-REPORTED HEALTH STATUS';
TITLE2 'TWO SAMPLE GOODNESS OF FIT STUDY';
run;
proc sort; by SEX; RUN;/* ENDS SAS PROCESSING */
```

By separating the data by sex we can compare the distributions for males against the distributions for females.

Whereas the SGPLOT procedure produces a histogram for the entire set of data, notice the proc gchart procedure produces a vertical bar chart to compare the percent responses for males versus females. The data for the graphs are compared statistically using PROC FREQ with the Chi-square option; the results follow in the table below the graphs.



The statistical analysis that compares the distribution for the three groups of participants is shown in the following frequency distribution table.

Table 17.1 Frequency Distribution Table

Statistics for Table of HEALTH by SEX

Statistic	DF	Value	Prob
Chi-Square	4	1.8010	0.7723
Likelihood Ratio Chi-Square	4	1.8049	0.7716
Mantel-Haenszel Chi-Square	1	0.7697	0.3803
Phi Coefficient		0.1026	
Contingency Coefficient		0.1021	
Cramer's V		0.1026	

Sample Size = 171

These data suggest that there is no difference in the distributions for males versus females for the responses to the health status question (= 1.80 p=0.77). The chi-square output is highlighted in the summary table, above.

The SAS output produces a frequency distribution table that presents the data separately for males and females. There

is no data for subjects that declared other in this example because we restricted the SAS processing of the data with the command WHERE SEX<3;

Table 17.2 Frequency Distribution for Health by Sex

	Males	Females
EXCELLENT	20	26
VERY GOOD	14	11
GOOD	12	14
FAIR	16	13
POOR	24	21

These data can also be evaluated using the two-sample chi-square Webulator, for an ordinal scaled problem with 5 outcomes as shown below:

https://health.ahs.upei.ca/webulators/fiveby2.html

An interactive or media element has been excluded from this version of the text. You can view it online here: https://pressbooks.library.upei.ca/montelpare/?p=386

AN ANNOTATED EXAMPLE: Chi-Square Goodness of Fit Test For Two Samples

The following is an example of the two-group chi-square based on a study of the distribution of cell phone use by individuals relative to motor vehicle collisions.

In 2010, Issar, Kadakia, Tsahakis, Yoneda et al (2013), conducted a study to investigate the link between texting and motor vehicle collisions (MVC). Data were collected using a questionnaire sent to patients attending an orthopaedic trauma clinic. The responses were organized into two groups as follows: Group 1 included patients who were involved in a MVC and were driving the vehicle at the time of the collision, and Group 2 consisted of all other patients attending the orthopedic clinic between October 2010 to March 2011.

In Table 17.3 the frequency of general phone use by Group 1 and Group 2 is presented. Although both frequency data (counts) and percentages are reported, we can use a two-group chi-square goodness of fit analysis to evaluate the frequency data.

Table 17.3 General Phone Use Frequencies for MVC vs. non-MVC Phone use[1]

Phone use (hours/week)	Group 1: MVC	Group 2: Non-MVC
0 – 1	15 (26.3%)	32 (26.7%)
1 – 2	11 (19.3)	24 (20.0%)
2 – 3	10 (17.5%)	16 (13.3%)
3 – 4	6 (10.5%)	13 (10.8%)
>4	15 (26.3%)	35 (29.2%)

The data from Table 17.3 are used to determine if the two groups differ in their phone use, measured in hours per week.

In order to ensure that the research is not biased, the null hypothesis will be: "there is no association between MVC group and cell phone use in hours per week". Our first step in the evaluation process is to state the expected response pattern. The expected response pattern is consistent with our "expected distribution". In other words, in an unbiased research study, we should expect that all possible responses are equally as likely to occur within each of the samples. In the examples presented here, twenty percent of each group should answer each of the response options. We call this the unbiased null hypothesis and state it in terms of frequencies of responses. The null hypothesis for this set of examples is

H0: frequency response in $Group_{11...5}$ = frequency response in $Group_{21...5}$

The data responses for this example are presented in Table 17.4 below. The arrangement of these data forms a 2 x 5 contingency table and therefore is analyzed using the standard chi-square formula.

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

Table 17.4 Raw Data Used in the 2 x 5 Chi-Square Analysis

	MVC Group 1	Non-MVC Group 2
Option 1	15	32
Option 2	11	24
Option 3	10	16
Option 4	6	13
Option 5	15	35
Column Sums =	**57**	**10**

The chi-square test measures how closely the responses in two distributions match. That is, to what extent is the distribution for MVC Group 1 the same as the Non-MVC Group 2. Enter the frequency data for each option from the datasheet in Table 17.4 into the corresponding fields of the webulator below. Click through the frames to compute the 2 x 5 chi-square calculations.

An interactive or media element has been excluded from this version of the text. You can view it online here: https://pressbooks.library.upei.ca/montelpare/?p=386

Using the webulator for the 2 x 5 chi-sqaure we use a stepwise approach to compute the expected values for each cell using the formula (row sum * column sum) : grand total. These values are provided in the webulator and shown in the following table.

Table 17.5 Expected values for each cell based on the formula (row sum * column sum) : grand total.

Expected Cell Values	MVC Group 1	Non-MVC Group 2
Option 1	0.002	0.0006
Option 2	0.0065	0.003
Option 3	0.31	0.15
Option 4	0.002	0.001
Option 5	0.075	0.035

The chi-square score also referred to as the chi-square observed is produced in the final frame of the webulator. After computing the chi-square observed value, we next determine the chi-square critical score from a table of chi-square values. The chi-square critical score presented in these examples represents what we should expect to observe for two sample distributions each with five possible responses. The critical value is determined by computing the "degrees of freedom" for our response set. The computation of the degrees of freedom is:

degrees of freedom = (number of rows – 1) * (number of columns -1)

[latex]\therefore[/latex] degrees of freedom = (5-1) x (2-1)

degrees of freedom = (4) x (1)

degrees of freedom = 4

[latex]\therefore[/latex] the ***chi-square critical value*** for (d.f.=4) at $p<0.05 = 9.49$

Once we have calculated the chi-square observed and then determined the chi-square critical then we establish a decision about whether or not to accept or reject the null hypothesis for this comparison. Recall that our null hypothesis was initially set as: the distribution of responses for the MVC Group across the response options would be equal to the distribution of responses for the MVC Group across the response options. Therefore our decision to accept or reject the null hypothesis follows the decision rule: If the "chi-square observed value " is greater than **(›)** the "chi-square critical value of **9.49**", we would reject the null hypothesis and state that the two distributions ARE NOT EQUAL. However, if the "chi-square observed value " is less than **(‹)** the "chi-square critical value of **9.49**", we would ACCEPT the null hypothesis and state that the two distributions ARE EQUAL.

From our computations, we can see that the chi-square observed value is **0.59,** which is less than the chi-square critical value of **9.49** and therefore we accept the null hypothesis that the two distributions are equal. Restating this outcome with specific reference to texting and MVCs in the Issar study we conclude that the MVC group does not differ from the non-MVC group with respect to their phone hour use per week.

SAS Code used to verify the two group Chi-Square Goodness of Fit

In this example, we computed the differences in cell phone use by motor vehicle collisions. The following is the SAS code applied to the computations produced above.

The study intended to measure whether the group of individuals that were involved in motor vehicle collisions had the same profile of cell-phone use as the group that were not involved in motor vehicle collisions. The data set was comprised of three variables:

1) Phone Use: where 1 = '0 to 1 hrs/wk', 2 = '1 to 2 hrs/wk', 3 = '2 to 3 hrs/wk', 4 = '3 to 4 hrs/wk', 5 = '> 4 hrs/wk'; 2) Involvement in a motor vehicle collision: 1 = 'MVC', 2 = 'No-MVC'; and a third variable which was the number of events reported.

The relevant SAS code used to process this two group chi-square goodness of fit is shown here

```
   PROC FORMAT;
VALUE OPTFMT 1 = '0 TO 1 HRS/WK'
2 = '1 TO 2 HRS/WK'
3 = '2 TO 3 HRS/WK'
4 = '3 TO 4 HRS/WK'
5 = '> 4 HRS/WK';
   VALUE MVCFMT 1 = 'MVC'   2 = 'NO-MVC';
   DATA CHIMVC;
TITLE 'PHONE USE AND MOTOR VEHICLE COLLISIONS ';
INPUT PHONEUSE MVC NUM_RPRT @@;
   LABEL PHONEUSE = "HOURS OF PHONE USE PER WEEK";
LABEL MVC = "MOTOR VEHICLE COLLISION";
LABEL NUM_RPRT = "NUMBER OF EVENTS REPORTED";
   DATALINES;
1 1 15 1 2 32 2 1 11 2 2 24 3 1 10 3 2 16 4 1  6 4 2 13 5 1 15 5 2 35
;
   PROC SORT; BY  MVC;
/* Define the axis characteristics */
axis1 offset=(0,70) minor=none;
axis2 label=(angle=90);
pattern1 value=solid color=cx7c95ca;
pattern2 value=solid color=cxde7e6f;
PROC GCHART;
BLOCK PHONEUSE / SUBGROUP=MVC
discrete SUMVAR=NUM_RPRT
COUTLINE=RED WOUTLINE=1 raxis=axis2;
TITLE1 'BLOCK CHART OF MVC BY PHONE HOURS OF USE';
FORMAT PHONEUSE OPTFMT. MVC MVCFMT. ;
   PROC FREQ; TABLES PHONEUSE*MVC / CHISQ ;
WEIGHT NUM_RPRT;
FORMAT PHONEUSE OPTFMT. MVC MVCFMT. ;
TITLE 'COMPARISON OF MVCS BY WEEKLY CELL PHONE USE';
```

Figure 17.1 Block Chart of the Frequency Distribution for Number of MVCs

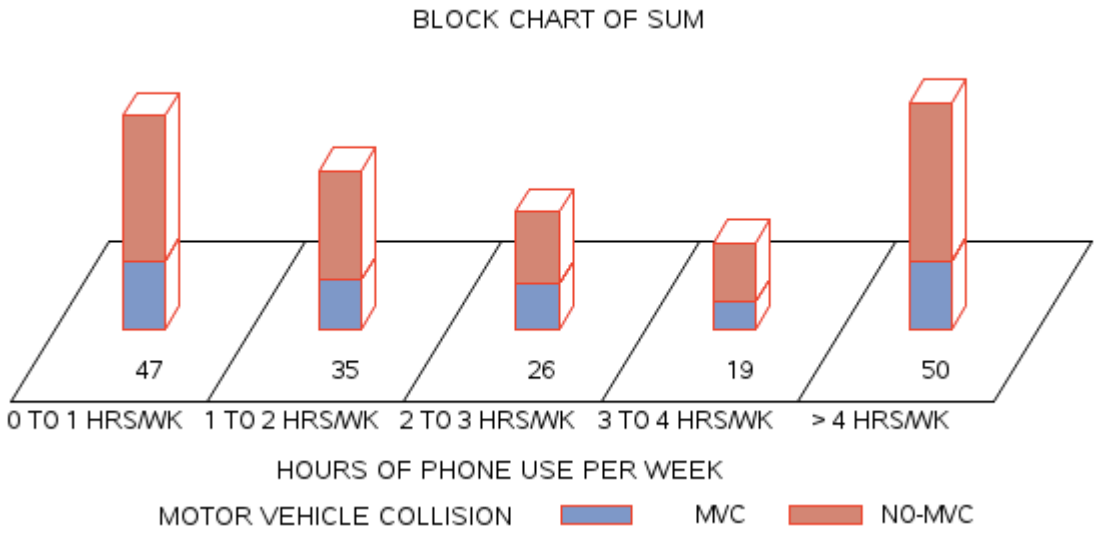


Table 17.6 Statistics for Table of Phone Use by Motor Vehicle Collision

Statistic	DF	Value	Prob
Chi-Square	4	0.5924	0.9639
Likelihood Ratio Chi-Square	4	0.5800	0.9653
Mantel-Haenszel Chi-Square	1	0.0327	0.8566
Phi Coefficient		0.0579	
Contingency Coefficient		0.0578	
Cramer's V		0.0579	

[1] From Issar, Kadakia, Tsahakis, Yoneda et al (2013): The link between texting and motor vehicle collision frequency in the orthopaedic trauma population. J Inj Violence Res. 2013 Jun; 5(2): 95-100.

18. Multi-way Contingency Table Chi-Square Analysis

Application of the Goodness of Fit Chi-square analysis to multi-way tables (3×3 and beyond)

Another form of the chi-square goodness of fit test is shown in the analysis of multi-way contingency tables. In the following example we show the use of a 3 x 3 contingency table to evaluate the association between visits to the emergency room in a cohort of COPD patients and the use of an online wellness program designed to provide customized programming for COPD patients.

In the following study a group of COPD patients were taught how to use an online program designed to provide up to date information about nutrition, exercise, stress and medications that could prevent the exacerbation of a dyspnea[1] response by the patient. The data were presented in several formats and included both direct and indirect communications between healthcare providers and the patients. The researchers organized the following contingency table to test the association between use of the online tools and visits to the emergency department in an 18-month period.

Table 18.1 Raw Data used to Evaluate the Association Between the Use of Online Tools and Visits to the Emergency Department

N= 375	**0 Visits to the emergency department**	**1-3 Visits to the emergency department**	**> 3 Visits to the emergency department**
Infrequent use of the online tools: less than once per week	12	55	100
Occasional use of the online tools: 1-3 times per week	21	37	19
Frequent use of the online tools: 4 or more uses per week	105	11	15
Column Totals	**138**	**103**	**134**

We can use the webulator presented below to compute the chi-square statistic for the multi-way (3 x 3) contingency table . Note that the equation for the 3 x 3 contingency table is the same as all chi-square tables.

[latex]{\chi}^2 = \sum\frac{(obs - exp)^2}{expected}[/latex]

In the data processing panels shown here the row and column sums (Panel 1) are used to compute the expected frequencies for each cell (Panel 2). The third panel provides the actual chi-square test. The sum of the variance computations is the chi-square statistic.

An interactive or media element has been excluded from this version of the text. You can view it online here:
https://pressbooks.library.upei.ca/montelpare/?p=390

The computed score is referred to as the chi-square observed. After computing the chi-square for the observed scores we next determine the chi-square critical score which represents the chi-square for the expected population. The chi-square critical score for a three by three frequency table is determined by computing the "degrees of freedom" for our response set.

The computation of the degrees of freedom is as follows:

degrees of freedom = (number of rows – 1) x (number of columns -1)

degrees of freedom = (3-1) x (3-1)

degrees of freedom = (2) x (2)

degrees of freedom = 4

and the "chi-square critical value" for degrees of freedom of "4" at $p<0.05 = 9.49$

Our null hypothesis in this scenario is that there is no association between the row and column variables.

If the "chi-square observed value" is **›** the "chi-square critical value of **9.49**" then we would reject the null hypothesis and state that there is an association between the row and column variables. However, if the "chi-square observed value " is **‹** the "chi-square critical value of **9.49**", we would ACCEPT the null hypothesis and state that the distributions ARE EQUAL.

The results of our analysis show that there is a relationship between the use of online tools and visits to the emergency room. That is, individuals that had a lower frequency of use of online tools were more likely to visit the emergency room than individuals that were considered frequent users of the online tools.

SAS Code used to demonstrate the computation of the 3 x 3 Chi-Square Goodness of Fit

In the example above we computed the differences in visits to the hospital by individuals that used (or chose not to use) online wellness resources. The following is the SAS code applied to the computations above. The study intended to compare the three distributions of hospital visits among online health resource users (or non-users).

The data set was comprised of three variables: Frequency of online health resource use: where 1 = 'infrequent', 2 = 'occasional', 3 = 'frequent';

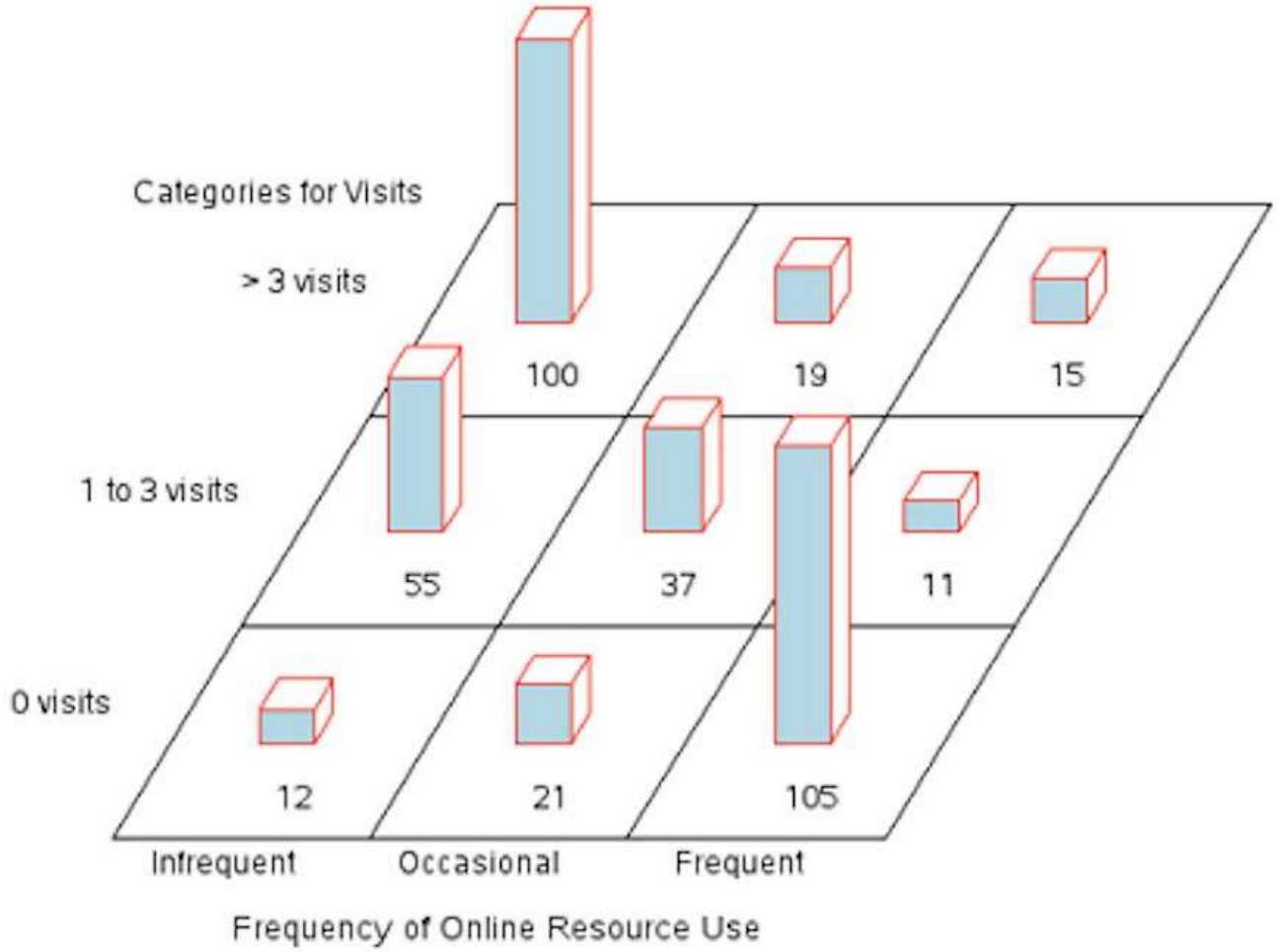
The category of the number of visits to the hospital: 1 = '0 visits', 2 = '1 to 3 visits'; and a third variable which was the number of cases reported to visit. The relevant SAS code used to process this two-group chi-square goodness of fit is shown here:

Two-Group Chi-Square Goodness Of Fit For A 3 X 3 Matrix

```
PROC FORMAT;
```

```
VALUE USEFMT 1 = 'INFREQUENT' 2 = 'OCCASIONAL'  3 = 'FREQUENT';
VALUE VISITFMT 1 = '0 VISITS' 2 = '1 TO 3 VISITS'. 3 = '> 3 VISITS';
DATA CHIVISIT;
TITLE 'ON LINE WELLNESS TOOLS REDUCE HOSPITAL VISITS';
INPUT TOOLS VISITS NCASES @@;
LABEL NCASES = 'NUMBER OF HOSPITAL VISITS REPORTED'
VISITS = 'CATEGORIES FOR VISITS'
TOOLS = 'FREQUENCY OF ONLINE RESOURCE USE';
DATALINES;
1 1 12 1 2 55 1 3 100 2 1 21 2 2 37 2 3 19
3 1 105 3 2 11 3 3 15
;
PROC SORT DATA= CHIVISIT; BY  VISITS;
PROC GCHART;
BLOCK TOOLS /SUMVAR=NCASES GROUP=VISITS  NOHEADER DISCRETE COUTLINE=RED WOUTLINE=1 ;
FORMAT TOOLS USEFMT. VISITS VISITFMT. ;
TITLE1 'HOSPITAL VISITS BY USE OF ONLINE HEALTH RESOURCES';
PATTERN1 COLOR = LIGHTBLUE;
PROC FREQ;
TABLES TOOLS*VISITS / CHISQ ; WEIGHT NCASES;
FORMAT TOOLS USEFMT. VISITS VISITFMT. ;
TITLE 'NUMBER OF HOSPITAL VISITS REPORTED';
TITLE2 'TWO SAMPLE GOODNESS OF FIT STUDY';
```

The SAS code above produced the following block chart of the distribution of the visits to the hospital related to the use of online resources.



Graph 18.1 Distribution of visits to the hospital related to the use of online resources

Below is the tabular output for the PROC FREQ procedure to produce the frequency distribution of the visits to the hospital by the use of online resources. The data represent a two-sample goodness of fit study design.

Frequency Percent Row Pct Col Pct

Table of tools by visits	
Online Resource Use	0 visits
Infrequent	
Occasional	
Frequent	
Total	

The following is a summary table generated by the PROC FREQ procedure. Here we can review the chi-square statistic and its corresponding p-value, and compare the value produced by SAS ([latex]{/chi^2}= 191.15 p<0.001[/latex] to that

which we produced above with our *Webulator* (also ([latex]{/chi^2}= 191.15 p<0.001[/latex]). Note, the sample size is provided at the end of the SAS output: **Sample Size = 375.**

Statistic	DF	Value	Prob
Chi-Square	4	191.1463	<.0001
Likelihood Ratio Chi-Square	4	202.0115	<.0001
Mantel-Haenszel Chi-Square	1	148.5705	<.0001
Phi Coefficient		0.7139	
Contingency Coefficient		0.5811	
Cramer's V		0.5048	

---

[1] Dyspnea is a sensation, referring to the sensation of shortness of breath or the feeling of having difficulty breathing.

19. All That From the 2 x 2 Table

Learner Outcomes:

After reading this chapter you should be able to:

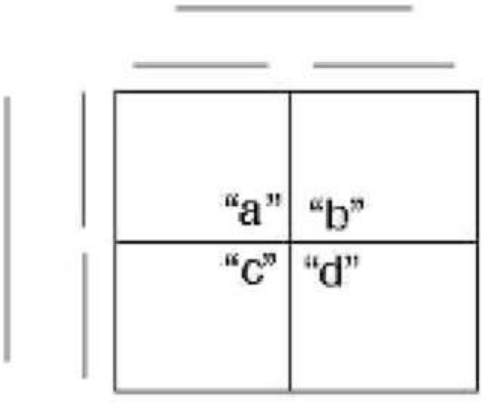
- Describe and compute the chi-square statistical procedure for a 2 x 2 research design to test for difference in two variables measured at the nominal level
- The chi-square calculation can be computed using the webulator embedded in this chapter or by submitting the SAS code for a 2×2 chi-square using the PROC FREQ commands shown in this chapter.

Part 1: Introduction to the 2 x 2 Chi-Square test

The 2 x 2 chi-square test is the most basic of the chi-square contingency tables and is often referred to as a fourfold table. The table can be used to test the association or differences between two variables when the data are presented as frequencies or counts. Frequency data or counts are often referred to as categorical data when analyzed with chi-square analyses because they represent the number of items (cases, individuals) within a designated category.

The total sample of individuals is distributed across the four outcome categories according to the response or outcome (column titles) and the exposure or stimulus group (row titles) to which the counts (participants) belong. Notice that each cell includes a letter. The letters a through d represent the commonly used labels for each of the outcome boxes in the two by two design.

Figure 19.1 Design of the 2 x 2 table



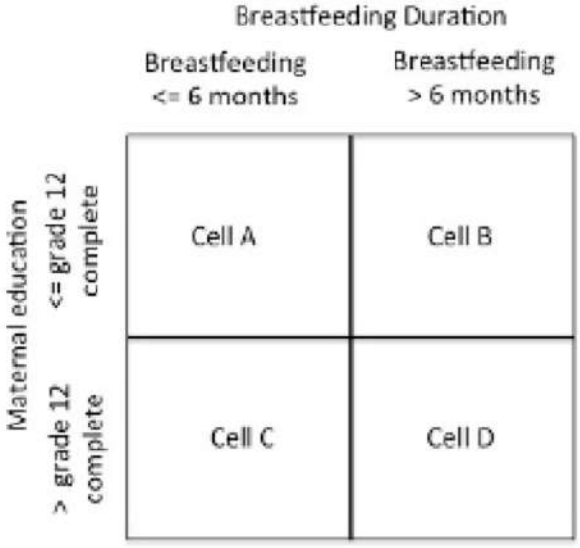
In an unbiased research study, we should expect that all possible responses are equally as likely to occur. We call this the unbiased null hypothesis and state this in terms of frequencies of responses. We can use these letters within each box to state the null hypothesis as follows:

[latex]H_{0}: f_{a} = f_{b} = f_{c} = f_{d}[/latex]

An annotated application of the 2 x 2 Chi-square to test for association between two variables

Consider a chi-square design to evaluate the relationship between maternal education and breastfeeding duration. To begin, we label the row and column headings within a fourfold table. In this example, we will use the row headings to represent the categories of maternal education and the column headings will be used to organize the breastfeeding duration as shown below in Figure 19.2.

Figure 19.2 Organization of data for 2 x 2 Chi-Square Test



The data in Cell A represent the observation of the number of mothers who have an education level equal to or less than grade 12 and breastfed their most recent child less than 6 months. In Cell B the data represent the observation of the number of mothers who have an education level equal to or less than grade 12 and breastfed their most recent child for more than 6 months. In Cell C the data represent the observation of the number of mothers who have an education level higher than grade 12 and breastfed their most recent child less than 6 months, and in Cell D the data represent the observation of the number of mothers who have an education level higher than grade 12 and breastfed their most recent child more than 6 months.

The null hypothesis for the chi-square test of association assumes that there is no association between the variables used in the two-way table. In the example presented here, this translates to stating that there is no association between the mothers' level of education and breastfeeding duration. In other words, breastfeeding duration does not vary in relation to the level of maternal education.

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

The following is an example of the application of the chi-square test of the association between breastfeeding duration and level of maternal education.

The data for 125 new mothers surveyed for breastfeeding duration and their highest level of education is shown in the following abbreviated table. The actual total observed scores for each condition of breastfeeding duration and the highest level of maternal education are provided in the table below.

Table 19.1 *Abbreviated table showing breastfeeding duration and the highest level of mothers' education for a sample of 125 new mothers.*

Subject ID	Breastfeeding duration in Months	Highest level of Maternal Education
01	2 months	Grade 9
02	4 months	College complete
03	9 months	Masters Complete
…	…	…
123	18 months	Grade 11
124	3 months	Grade 12
125	< 1 month	Ph.D.

Table 19.2 The observed counts of breastfeeding duration and the highest level of mothers' education for a sample of 125 new mothers arranged in the fourfold table

	BF duration <= 6 months	BF duration > 6 months	Row Totals
Education <= grade 12	Observed = 43	Observed = 27	n1.=70
Education > grade 12	Observed = 21	Observed = 34	n2.=55
Column Totals	n.1=64	n.2=61	Grand Total N=125

Once we have organized the observed data according to the appropriate marginal conditions, we next compute the expected values for each cell independently in the fourfold table. To compute the expected frequency independently for each cell we use the following formula:

[latex]\textit{Expected Scores} = {{\Sigma(\textit{Row frequency}}) \times {\Sigma(\textit{Column frequency}})\over \textit{Grand Total (N)}}[/latex]

Notice this formula computes the expected cell frequencies by first calculating the cross-product of the row sum multiplied by the column sum and dividing by the total sample size. The expected values in each cell are computed in Table 19.3 below.

Table 19.3 Computations of Expected Values in the 2 x 2 table

	BF duration	BF duration	Row Totals
	<= 6 months	> 6 months	
Education	Obs = 43	Obs = 27	
<= grade 12	Exp = (70 x 64)/125 Exp = 35.84	Exp = (70 x 61)/125 Exp = 34.16	n1.=70
Education	Obs = 21	Obs = 34	
> grade 12	Exp = (55 x 64)/125 Exp = 28.16	Exp = (55 x 61)/125 Exp = 26.84	n2.=55
Column Totals	n.1=64	n.2=61	Grand Total N observed =125

After calculating the expected cell values, we next compute the elements of the chi-square test for each cell of the 2 x 2 table, as follows:

STEP 1: Compute the difference between the observed and expected cell differences, then square this value and divide by the expected value within the cell.

a) in Cell A we compute: $(43\text{-}35.84)^2 : 35.84 = 1.43$

b) in Cell B we compute: $(27\text{-}34.16)^2 : 34.16 = 1.5$

c) in Cell C we compute: $(21\text{-}28.16)^2 : 28.16 = 1.82$

d) in Cell D we compute: $(34\text{-}26.84)^2: 26.84 = 1.91$

STEP 2: Sum the values computed in **STEP 1** for cell A, B, C, D, above. This is the chi-square statistic. [latex]{\chi^2} ={{\Sigma(\textit{observed - expected})}^2}\over \textit{expected}}[/latex]

[latex]{\chi^2} =[/latex] = (1.43 + 1.5 + 1.82 + 1.91)

[latex]{\chi^2} =[/latex] = 6.66

STEP 3: Compare the chi-square statistic observed for this sample of data ([latex]{\chi^2}_\textit{observed} =[/latex] = 6.66) against the chi-square critical value ([latex]{\chi^2}_\textit{critical} =[/latex] = 3.84) of χ2critical =3.84. The critical value represents the chi-square statistic expected for a 2 x 2 table at a probability level of $p<0.05$.

STEP 4: Decision Rule: If the [latex]{\chi^2}_\textit{observed} [/latex] > [latex]{\chi^2}_\textit{critical}[/latex] then we reject the null hypothesis.

In our example, we observed that the chi-square was 6.66 which is greater than 3.84 and therefore we reject the null hypothesis that

[latex]H_{0}: f_{(j)} \textit{observed} = f_{(j)}\textit{expected}[/latex] and suggest that there is a relationship between breastfeeding duration and mother's level of education.

Alternatively we could compute the chi-square statistic for the fourfold table using the following formula:

[latex]{\chi^2}_\textit{observed} ={ (ad-bc)^2 \times (a + b + c + d)\over{ (a+b)(c+d)(b+d)(a+c) }}[/latex]

In which case our estimate would be:

[latex]{\chi^2}_\textit{observed} ={ (43 \times 34 - 27 \times 21)^2 \times (43 + 27 + 21 +34)\over{ (43+27)(21+34)(27+34)(43+21) }}[/latex]

[latex]{\chi^2}_\textit{observed} ={ (1462 - 567)^2 \times (125)\over{ (70)(55)(61)(64) }}[/latex]

[latex]{\chi^2}_\textit{observed} ={ (100128125)\over{ (15030400) }}[/latex]

[latex]{\chi^2}_\textit{observed} = 6.66[/latex]

In the chi-square test statistic shown here, we were interested in measuring the association between breastfeeding duration and mother's level of education completed. This is not a causal model but a measure of association that lets us evaluate the relationship between two independent measures. We began with the null hypothesis that there was no association between the two variables, but after testing the association with the chi-square test, our conclusion is that there appears to be a relationship between maternal education and breastfeeding duration.

2 x 2 CHI SQUARE WEBULATOR

Using the Chi-square Webulator[1], you can enter the data from the table above into the webulator below to compute the Chi-square observed score.

1. <script src="https://pressbooks.library.upei.ca/montelpare/wp-content/plugins/h5p/h5p-php-library/js/h5p-resizer.js" charset="UTF-8"></script>

Enter the scores for cell “a” = 43, cell “b” = 27, cell “c” = 21, cell “d” = 34, and then click the buttons in the webulator to work through the calculations of the chi-square.

An interactive or media element has been excluded from this version of the text. You can view it online here: https://pressbooks.library.upei.ca/montelpare/?p=49

We can use SAS to confirm the calculations from the Webulator above.

SAS Program for the 2 x 2 chi-square:

```
  DATA CHISQR01;
TITLE '2 X 2 TABLE';
INPUT ROW COL OUTCOME;
/* NOTE DATA ARE ENTERED USING ROW-COLUMN ARRANGEMENT */
DATALINES;
1 1 43
1 2 27
2 1 21
2 2 34
;
PROC FREQ  ORDER=DATA; WEIGHT OUTCOME;
TABLES ROW*COL/ CHISQ EXPECTED;
RUN;
```

In the program above we use PROC FREQ with the TABLES keyword as the basic procedural statement. The TABLES statement requires the names of the row and column variables which we called ROW and COL in our example (not very imaginative!). Including the options, CHISQ and EXPECTED statements enable us to produce the statistical output for the chi-square as shown below. The other two important statements here are ORDER+DATA which maintains the position of the order in which the data were presented, respecting the row X column organization; and the keyword statement WEIGHT OUTCOME; which acknowledges that the data are not single scores but represent the sum of counts for the respective table cell. In the present example cell A: 43, Cell B: 27, Cell C: 21, Cell D: 34.

In the following output from the SAS program, both the chi-square value and the related p-value are the same as that which we calculated by hand – chi-square score = 6.66 with a p-value of <0.001. We can then make a decision to reject the null hypothesis that the count in Cell A = the count in Cell B = the count in Cell C = the count in Cell D.

Table 19.4 Results of the SAS PROC FREQ Procedure

	BF duration <= 6 months	BF duration > 6 months	Row Totals
Row 1	Frequency = 43 Expected Freq = 35.84 Percent = 34.40 Row Pct = 61.43 Column Pct = 67.19	Frequency = 27 Expected Freq = 34.16 Percent = 21.60 Row Pct = 38.57 Column Pct = 44.26	Row 1 Total = 70
Row 2	Frequency = 21 Expected Freq = 28.16 Percent = 16.80 Row Pct = 38.18 Column Pct = 32.81	Frequency = 34 Expected Freq = 26.84 Percent = 27.20 Row Pct = 61.82 Column Pct = 55.74	Row 2 Total = 55
Column Totals	Column Total = 64 Column Pct = 51.20	Column Total = 61 Column Pct = 48.80	Grand Total = 125

Table 19.5 Statistics for Table of ROW by COL

Statistic	DF	Value	Prob
Chi-Square	1	6.6617	0.0099
Likelihood Ratio Chi-Square	1	6.7197	0.0095
Continuity Adj. Chi-Square	1	5.7638	0.0164
Mantel-Haenszel Chi-Square	1	6.6084	0.0101
Phi Coefficient		0.2309	
Contingency Coefficient		0.2249	
Cramer's V		0.2309	

The Case for COVID-19 Testing

One important discussion in the midst of the COVID-19 Pandemic has been related to testing. The President of the United States followed the ill-logic that by doing more testing we will naturally see a rise in the number of reported cases-full stop! Meaning that if we stick our head in the sand – aka stop testing, then we won't see any cases. The number of problems associated with this kind of thinking is far too numerous to even begin debating the comment. How-

ever, the more important issue is, how do we use statistics to show the importance of the difference in the proportion of positive cases to all of those tested each week?

By using a 2 x 2 chi-square approach we can evaluate the significance of the changes in the proportionality between positive cases emerging from week to week. Consider the following scenario of data collection (testing and case identification) between two different weeks. Let's say that in week 1 there were 100 tests administered, and an incidence of 35 positive cases. Next, in week 2 there were 200 cases and an incidence of 110 positive cases.

Our first step is to arrange the data collection table as shown below with Outcome (Columns) by Weeks (Rows).

	Positive Case	Negative Case	
Week 1	Cell a: 35	Cell b: 65	100
Week 2	Cell c: 110	Cell d: 90	200
	145	155	300

We can use the 2 x 2 webulator to compute the chi-square statistic to determine if there is a significant difference in the proportion of cases from week 1 to week 2. Simply substitute the data from the table above into the webulator above.

Figure 19.3 First Form of the 2 x 2 Webulator

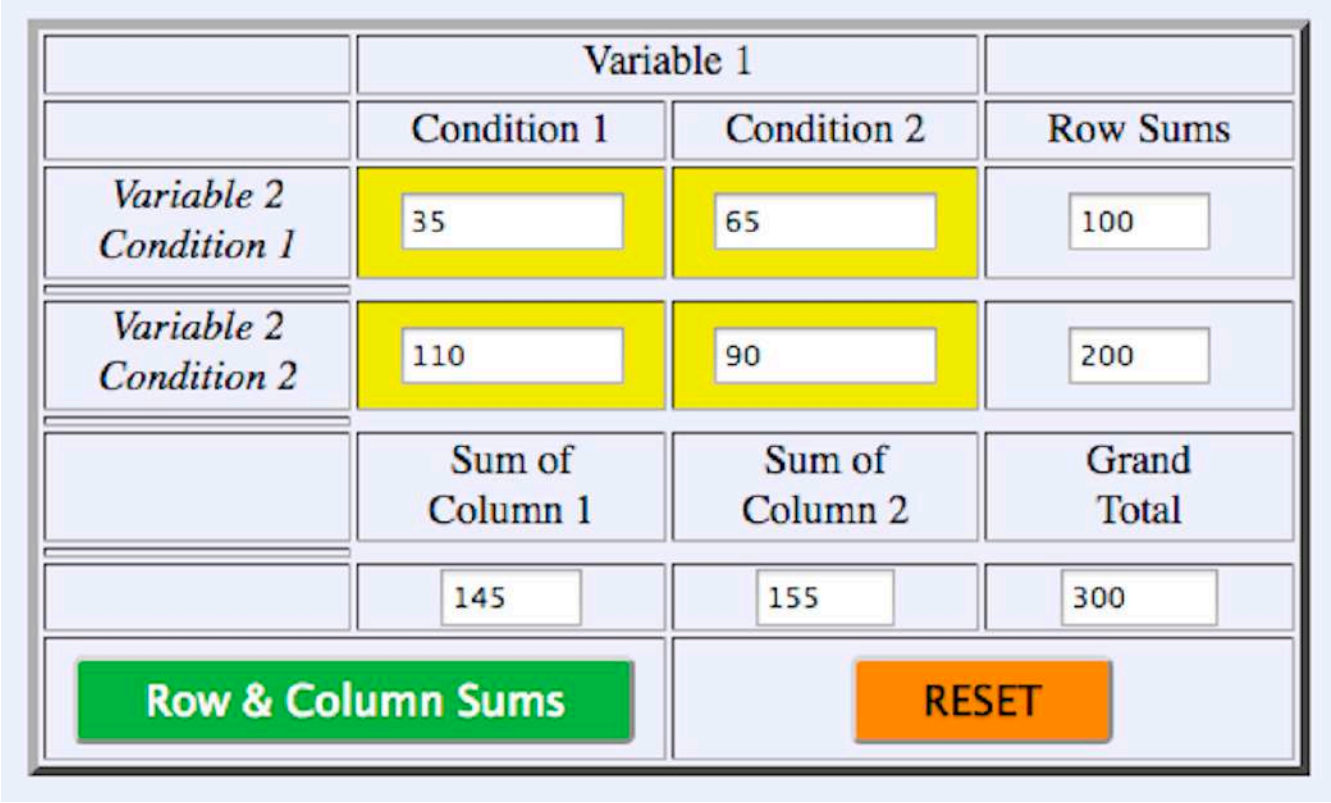


Figure 19.4 Second Form of the 2 x 2 Webulator

Chi Square Computations	Column 1	Column 2
$\frac{(observed- expected)^2}{expected}$	3.6781609	3.4408602
$\frac{(observed- expected)^2}{expected}$	1.8390804	1.7204301
Chi square formula	$\chi^2 = \sum \frac{(o-e)^2}{e}$	Chi-Square= 10.678531
	Reset	Compute chi square scores

As you can see the results of the chi-square statistic observed = 10.68 which is > chi-square critical of 3.84 means that there was a significant difference between the proportions. Next, we can compare these data using SAS as shown below.

Chi-square for COVID-19 cases versus tests

```
   DATA COVIDCHI;
TITLE 'COVID – 19 CHISQUARE 2 X 2 TABLE';
INPUT ROW COL OUTCOME;
/* NOTE DATA ARE ENTERED USING ROW-COLUMN ARRANGEMENT */
DATALINES;
1 1 35
1 2 65
2 1 110
2 2 90
;
PROC FREQ  ORDER=DATA; WEIGHT OUTCOME;
TABLES ROW*COL/ CHISQ EXPECTED;
RUN;
```

Table 19.6 Statistics for Table of ROW by COL

Statistic	DF	Value	Prob
Chi-Square	1	10.6785	0.0011
Likelihood Ratio Chi-Square	1	10.8101	0.0010
Continuity Adj. Chi-Square	1	9.8927	0.0017
Mantel-Haenszel Chi-Square	1	10.6429	0.0011

From these data, we observe that regardless of the number of tests administered, the estimate of interest is not merely the number of tests conducted but the proportionality of the number of positive tests from one week to the next.

20. Fisher's Exact and the Phi Coefficient

Part 2: Calculating Fisher's Exact Statistic

In the chi-square test statistic shown in the previous chapter, we were interested in measuring the association between breastfeeding duration and the mother's level of education that she had completed. This is not a **causal** model but **a measure of association** that lets us evaluate the relationship between two independent measures. We began with **the null hypothesis** that there was **no association** between the two variables, but after testing the association with the chi-square test and finding that the chi-square estimate that we calculated exceeded the chi-square estimate expected we rejected the null hypothesis and our conclusion was that there appears to be a relationship between the level of maternal education and breastfeeding duration.

Our decision to reject the null hypothesis was based on the chi-square estimate that we calculated is compared to a critical value associated with a 95% probability that our observed estimate was representative of that which we should find in a normal population. We could however be more precise than 95% and compute the exact probability of the chi-square statistic that we calculated by using the Fisher's Exact test.

The formula for the Fisher's Exact test is shown here as:

[latex]p = {(a+b)! (c+d)! (a+c)! (b+d)!\over (n! \times a! \times b! \times c! \times d!)}[/latex]

The term n! refers to "n"- factorial and it is computed by simply recursively calculating out the run of n x (n-1) x (n-2) x (n-3) … until (1). So that 6! is actually 6 x (6-1) x (6-2) x (6-3) x (6-4) x (6-5) or in simpler terms 6! is 6 x 5 x 4 x 3 x 2 x 1.

Applying the Fisher's exact test to our scenario we would compute the following exact probability for our chi-square statistic, where a=43; b=27; c=21; and d=34;

Note that you will have difficulty computing the exact probability with numbers as large as those represented by a, b, c, and d. Therefore, we can reduce this computation by using the following SAS code. This approach demonstrates the versatility of the SAS programming language to enable complex computations without requiring an apriori dataset.

SAS Code to compute Fisher's Exact Test from known values

```
   OPTIONS PAGESIZE=55 LINESIZE=120 CENTER DATE;
DATA FACT;
X1=FACT(70); X2=FACT(55); X3=FACT(64); X4=FACT(61);
Y1=FACT(125); Y2=FACT(43); Y3=FACT(27); Y4=FACT(21); Y5=FACT(34);
   /* REDUCE THESE FACTORIALS TO COMPUTE FISHER'S EXACT.
IMPROVE THE EFFICIENCY OF THE REDUCTION BY MATCHING LARGEST
NUMERATORS AND DENOMINATORS TO CANCEL NUMBERS WITHIN THE SEQUENCE */
   REDUCE1=(X1/Y1);  REDUCE2=(X2/Y2);  REDUCE3=(X3/Y5);  REDUCE4=(X4/Y3);
REDUCE5=(1/Y4);
   OUTCOME=ROUND (REDUCE1*REDUCE2*REDUCE3*REDUCE4*REDUCE5, 0.001);
   PROC PRINT DATA=FACT;
```

```
VAR X1 X2 X3 X4 Y1 Y2 Y3 Y4 Y5 REDUCE1 REDUCE2 REDUCE3 REDUCE4 REDUCE5 OUTCOME;
RUN;
```

The SAS code above produces the following results.

VARIABLE: X1	VARIABLE: X2	VARIABLE: X3	VARIABLE: X4
1.1979E100	1.2696E73	1.2689E89	5.0758E83

VARIABLE: Y1	VARIABLE: Y2	VARIABLE: Y3	VARIABLE: Y4	VARIABLE: Y5
1.8827E209	6.0415E52	1.0889E28	5.1091E19	2.9523E38

REDUCE_1	REDUCE_2	REDUCE_3	REDUCE_4	REDUCE_5	OUTCOME
6.3625E-110	2.1015E20	4.2979E50	4.6615E55	1.9573E-20	.005243165

The OUTCOME value shown in the table above is the EXACT P-value which when rounded is p = 0.005.

We could also compute the Fisher's Exact value by hand using the following formula with the following cell values: a=43, b= 27, c=21 and d=34.

[latex]p = {(a+b)! (c+d)! (a+c)! (b+d)!\over (n! \times a! \times b! \times c! \times d!)}[/latex]

[latex]p = {(43+27)! (21+34)! (43+21)! (27+34)!\over (125! \times 43! \times 27! \times 21! \times 34!)}[/latex]

[latex]p = {(70)! (55)! (64)! (61)!\over (125! \times 43! \times 27! \times 21! \times 34!)}[/latex]

[latex]p = {0.005243165}[/latex]

Part 3: Calculating Associations in 2 x 2 tables with the Phi Coefficient

In addition to computing the exact probability for statistical comparison, we can also determine the strength of the association between the two variables using a simple computation to produce the **phi-coefficient**. The phi-coefficient provides an estimate of association in a 2 x 2 table. If there is no association between the rows and columns then the outcome is 0. The maximum value of phi is 1, which indicates an extremely strong relationship. It is also common to observe that when there appears to be a very low probability associated with a chi-square outcome, the phi-coefficient may also appear to demonstrate a low estimate.

The formula for the phi coefficient is: [latex]{\phi^2} = {\chi^2 \over n} = {\phi ={\sqrt {\chi^2 \over n}}}[/latex]

In the chi-square example shown in the previous chapter and in the calculation of the Fisher's Exact Test shown above, the Phi Coefficient is reported in the output in that was generated by our SAS program.

Statistics for Table of ROW by COL from the original chi-square

Statistic	DF	Value	Prob
Chi-Square	1	6.6617	0.0099
Likelihood Ratio Chi-Square	1	6.7197	0.0095
Continuity Adj. Chi-Square	1	5.7638	0.0164
Mantel-Haenszel Chi-Square	1	6.6084	0.0101
Phi Coefficient		0.2309	
Contingency Coefficient		0.2249	
Cramer's V		0.2309	

The Phi Coefficient was reported as 0.2309 which is approximately the same as the value we can compute by hand from the formula shown here:

[latex]{\phi^2} = {\chi^2 \over n} = {\phi ={\sqrt {\chi^2 \over n}}} = {\phi ={\sqrt {6.66 \over 125}}} = 0.23[/latex]

The Phi Coefficient reported here demonstrates that while the chi-square result was significant, and thereby indicating a significant association, the actual measure of association is low at 0.23.

21. Estimating Relative Risk, the Odds Ratio, and Attributable Risk

Learner Outcomes:

After reading this chapter you should be able to:

- Assess risk for a 2 x 2 research design to test for difference in two variables measured at the nominal level
- Describe and compute the relative risk in a 2 x 2 design and specifically in a cohort study
- Describe and compute the odds ratio in a 2 x 2 design
- Describe and compute the measure of attributable risk

Assessing Risk

Computing associations with the 2 x 2 contingency table is just the beginning. In health research we are also interested in determining the extent to which an individual, having been exposed to a given circumstance or stimulus, will demonstrate an observable outcome or condition. As researchers we generate probabilistic estimates that we call risks to establish the likelihood of such stimulus-response relationships.

The term risk is classified into various estimators that help us to establish the chance of an event given a particular exposure – outcome scenario. In these next sections we will review the terms relative risk, the odds ratio, and the population attributable risk in relation to the specific webulators that assist in computing the relative estimates.

Relative Risk – Defining the term from the 2 x 2 table

A simple definition for the relative risk (RR) estimate is that it refers to the ratio of the risk of an outcome in individuals with a factor of interest to the risk of an outcome in individuals without the factor of interest. What this means is that the RR estimate is based on the relationship between two fractional estimates as shown in the following formula:

[latex]RR = {{ (a/(a+b)}\over{ (c/(c+d)}}[/latex] The relative risk formula presented here depicts the ratio of the outcomes observed among exposed individuals (a/(a+b)) to the outcomes observed among non-exposed individuals (c/(c+d)). Consider the following 2 x 2 contingency table as the starting point.

Relative risk computes the ratio of the data in ***cell a*** divided by the data in the row total of ***(cell a + cell b)* divided by** the data in ***cell c*** divided by the data in the row total of ***(cell c + cell d).*** The arrangement of the data in a, b, c, and d cells in relation to that, which is computed, is shown here.

Table 12.1 Arrangement of the data to compute Relative Risk

	Cases	Controls	
Exposed	CELL "a" + condition + exposed	CELL "b" – condition + exposed	Numerator (a/(a+b))
Not exposed	CELL "c" + condition – exposed	CELL "d" – condition – exposed	Denominator (c/(c+d))

Stated differently, the relationship between the cells in the 2 x 2 table can be explained as the ratio of the chance of an outcome among individuals who have a characteristic of interest or who have been exposed to a specific **risk** factor, to the chance of an outcome among individuals who lack the characteristic of interest or who have not been exposed to a specific **risk** factor. The relative risk estimate therefore suggests that:

The condition (or outcome) is **RR** times more likely to occur among those individuals that are exposed to the suspected risk factor (related to) THAN among those individuals with no exposure to the risk factor (unrelated to).

As a rule then, the larger the value of the relative risk, that is greater than 1, the stronger the association between the disease or disorder of interest and exposure to the risk factor.

Likewise, values of the relative risk estimate that are close to **1** indicate that the disease and exposure to the risk factor are unrelated (i.e., the risk of occurrence is the same for both exposed and non-exposed individuals).

Similarly values of RR less than **1** indicate a negative association between the risk factor and the disease. A relative risk estimate less than 1 is said to demonstrate a protective effect rather than a detrimental effect.

Application of The Relative Risk Estimate

In cohort studies the estimate of relative risk is used to show the ratio of the probability of those exposed versus the probability of those not exposed.

The formula for relative risk **(RR)** is given as the ratio of – the proportion of individuals within an exposed group showing a condition:

(cell a / (cell a + cell b))

versus the proportion of individuals within a non-exposed group showing a condition

(cell c / (cell c + cell d))

Consider The Following Example.

Researchers suggested that dental disease may be a risk factor for coronary heart disease. The suggestion in the.literature was that researchers observed the presence of a specific type of protein associated with dental disease (C-reactive protein), which may be a "cause" of myocardial infarction (i.e. heart attacks). You intend to test the relative risk of the presence of C-reactive protein on myocardial infarction by proposing the following **case-control** study.

Table for Relative Risk in a 2 x 2 case-control design

	+ve condition (CASES)	-ve condition (CONTROLS)	Row Incidence
Exposed	$a = 186$	$b = 93$	$a/(a + b)$ 186/(186+93) =0.67
Not Exposed	$c = 21$	$d = 41$	$c/(c + d)$ 21/(21+41) = 0.34
Row totals	**a + c = 207**	**b + d = 134**	**grand total= 341**

The data in this table are used to calculate the Relative Risk as:

[latex]RR = {{ (186/(186+93)}\over{ (21/(21+41))}[/latex]. [latex]RR = {{ (0.67)}\over{ (0.34)}}[/latex]. [latex]RR = {1.97}[/latex]

Estimating the Confidence Interval for Relative Risk in a 2 x 2 case-control design

Next we can estimate the 95% confidence estimate of this relative risk estimate (RR=1.97) using the following series of calculations.

1. Convert the relative risk to natural logarithm value ln(rr) = ln(1.97) = 0.68,
2. next we calculate the standard error of the ln(rr) estimate using:

[latex]{\pm z} \, {\sqrt{ { (b/a) \over (b + a)} + {(d/c) \over (d + c) } }}[/latex]

[latex]{\pm 1.96} \, {\sqrt{ { (93/186) \over (93 + 186)} + {(41/21) \over (41 + 21) } }}[/latex]

[latex]{\pm 1.96} \, {\sqrt{ { 0.002} + {0.03} }}[/latex]

Standard Error of ln(RR) = 0.35

This series of calculations produces the upper and lower limits of the 95% confidence interval for the natural logarithm of relative risk (ln(RR)).

Lower limt 95% CI ln(RR) = 0.68 − 0.35 = 0.33 and upper limt 95% CI ln(RR) = 0.68 + 0.35 = 1.03. By exponentiating the 95% confidence interval's lower and upper limits will return the estimated values to the original scale scores, as shown here. **exp(0.33)** lower limt 95% **CI (RR) = 1.37**; and **exp(1.03)** upper limt 95% **CI (RR) = 2.8**.

Considering our decision rule whereby the larger the value of the relative risk (greater than 1) then the stronger the association between the disease or disorder of interest and exposure to the risk factor. Given a relative risk of 1.97 with upper and lover 95% confidence limits of 1.37 and 2.8, the results of your study support that the exposure C-reactive protein increases the risk of myocardial infarction. In fact, you showed that comparing the two groups, individuals suffering an MI were twice as likely to have higher levels of C-reactive protein.

Using the webulator for relative risk, we can confirm these calculations by inserting the scores into the appropriate cells of the webulator and clicking on the button labeled "Compute".

https://health.ahs.upei.ca/webulators/rr_pb.html

An interactive or media element has been excluded from this version of the text. You can view it online here: https://pressbooks.library.upei.ca/montelpare/?p=47

We can also use the following SAS Code to evaluate the data for our C-reactive protein example.

```
   DATA RELRSK1;
TITLE 'SAS CALCULATION FOR RELATIVE RISK';
INPUT ROW COL OUTCOME @@;
DATALINES;
1 1 186 1 2 93 2 1 21 2 2 41
;
PROC SORT DATA=RROR; BY ROW COL;
PROC FREQ DATA=RROR ORDER=DATA;
TABLES ROW*COL/CHISQ RELRISK;
WEIGHT OUTCOME;
EXACT PCHI OR;
RUN;
```

The output generated by the SAS code is shown, below:

Table of ROW by COL

ROW	COL		
	1	**2**	**Total**
1	186 54.55 66.67 89.86	93 27.27 33.33 69.40	279 81.82
2	21 6.16 33.87 10.14	41 12.02 66.13 30.60	62 18.18
Total	207 60.70	134 39.30	341 100.00

Notable output from SAS:

Statistic	DF	Value	Prob
Chi-Square	1	22.8723	<.0001
Phi Coefficient		0.2590	

Relative Risk Estimates

Statistic	Value	95% Confidence Limits
Relative Risk	1.9683	1.3766 to 2.8143

Notice that the relative risk estimate and the upper and lower limits for the SAS program estimate of the 95% confidence interval of relative risk are similar to that which we calculated by hand and with the Webulator. Moreover, given that the estimate does not include a value of 1 then we can say that in comparisons between the two groups, individuals suffering an MI were nearly twice as likely to have higher levels of C-reactive protein.

Estimating the Odds Ratio

The odds ratio is another computation arising from the 2 x 2 table. Although earlier we described the odds ratio as part of the calculation for interpreting the case-control study, the odds ratio can also be used in cohort studies as well as in cross-sectional research designs, and as Bland and Altman (2002) describe are used in logistic regression analysis to evaluate the influence of measurable variables on binary relationships between variables.

Given a 2 x 2 design, as shown here, the data in ***Cell a*** refer to cases that demonstrate the condition of interest and were exposed to a suspected causal stimulus, while the data in ***Cell d*** refer to the cases that do not demonstrate the condition of interest and were likely not-exposed to the suspected causal stimulus.

Arrangement of the data to compute the Odds Ratio

	The outcome of interest Present (Cases)	**The outcome of interest Absent (Controls)**	
Suspected causal mechanism present (Exposed)	Cell "a" + case **+ exposed**	Cell "b" − case **+ exposed**	Numerator = (a * d)
Suspected causal mechanism absent (Not Exposed)	Cell "c" + case **− exposed**	Cell "d" − case **− exposed**	Denominator = (b * c)

The formula for the Odds Ratio is: OR = (a * d) : (b * c).

As stated previously, the odds ratio is computed here to compare the ratio of cases that were exposed versus not exposed (a/c) to the ratio of non-cases among exposed versus not exposed (the control group (b/d)). The odds ratio is then the ratio of the two ratios: [(a/c) : (b/d)] and can be computed by simple dividing the product of the main diagonal elements: (Cell a x Cell d) by the product of the off-diagonal elements (Cell b x Cell c).

[latex]{OR =} \, { (a \times d) \over (b \times c)} ={\textit{main diagonal elements}\over\textit{off diagonal elements} }={\textit{(+ cases, + exposed)} \times {(- cases, - exposed)}\over\textit{(- cases, + exposed)} \times {(+ cases, - exposed)}}[/latex]

So why do we call this an odds ratio?

To answer this question, let's begin by stating what the odds ratio is not. The odds ratio is not telling us about the relative risk. To compute relative risk we looked at the proportion of cases among exposed and compared that proportion (ratio) against the proportion of cases that were not exposed. The result of the relative risk gave us the fractional comparison of cases with exposure to cases without exposure. In this way, we can say that an individual exposed to a given stimulus is RR times more likely to be a case because of the exposure.

In the odds ratio we are able to describe the likelihood associated with the outcome. Let's put this conversation in the context of racing. To compute the odds of winning a race we need to compare the number of wins to the number of losses. If we ran five races and won three times then the odds would be calculated as: total race − number of wins = number of losses which in our example is: 5 -3 = 2. The odds are 3:2 or stated as 3 to 2.

Computing odds	The odds of winning to losing is 3:2
Number of races won	3
Number of races lost	2
Total number of races	5

Now let's return to our 2 x 2 table. The number in *cell a* (+cases, +exposed) are compared to the number in *cell c* (+cases,

-exposed). The value of the (a/c) fraction can be expressed as an odds in the form a:c. Likewise, number in *cell b* (-cases, +exposed) are compared to the number in *cell d* (-cases, -exposed). The value of the (b/d) fraction can be expressed as an odds in the form b:d. Therefore, because we are computing the ratio of two odds estimates, notably, (a:c : b:d) we call the estimate the ratio of the odds, or simply the odds ratio.

Estimating the Odds Ratio for a 2 x 2 table

Consider the following 2 x 2 table of the relationship between smoking status and lung cancer.

Arrangement of smoking status and lung cancer to compute the Odds Ratio

Smoking Status	+ve condition (CASES) lung cancer present	− ve condition (CONTROLS) no lung cancer present
Exposed (smoker)	***cell "a"*** = 23 + case, + exposed	***cell "b"*** = 8 − case, + exposed
Not Exposed (smoker)	***cell "c"*** = 11 + case, − exposed	***cell "d"*** = 25 − case, − exposed

In this example, an individual is a member of ***cell "a"*** if they are both a smoker and were observed to be positive for lung cancer. Similarly, an individual is a member of ***cell "d"*** if they are both a non-smoker and do not show any characteristics of lung cancer. Membership in each of these cells is intuitively expected given what we know from studies of smokers and the suspected cause of lung cancer. That is if you smoke you will develop lung cancer, if you don't smoke you won't develop lung cancer. Seems simple enough!

However, often when describing the relationship between smoking and lung cancer, someone will undoubtedly recall a story about their grandfather that smoked his entire life but never developed lung cancer. Grandpa would, therefore, be a member of ***cell "b"*** − classified as a smoker but did not develop lung cancer. Likewise, the grandfather's story is often countered by the story of a friend who never smoked a day in her life but died of lung cancer. This person would become a member of ***cell "c"*** − classified as a non-smoker but was observed to have developed lung cancer.

The odds ratio is computed with the following formula: OR = (***cell "a"*** × ***cell "d"***) : (***cell "c"*** × ***cell "b"***). Let's compute the ODDS RATIO by hand and then verify our computations with our webulator and with SAS.

OR = (cell A × cell D) : (cell C × cell B)

OR = (23 × 25) : (11 × 8)

OR = 575 : 88 = 6.5

An odds ratio estimate of 6.5 suggests that individuals are 6.5 times more likely to develop lung cancer if they are classified as smokers.

Decision-making with OR and 95% confidence intervals

The odds ratio enables the researcher to test the relationship between a suspected cause and a suspected outcome by considering whether to accept or reject the null hypothesis. In its simplest form, the evaluation of an odds ratio is that there is no relationship between the suspected risk factor and the outcome, and is given by an estimate of the odds ratio to equal 1 (H_0: OR=1). Some basic rules regarding the decisions about the magnitude of the odds ratio are given as follows:

1. The computed odds ratio indicates to the researcher the magnitude of the suspected risk factor on the outcome

condition. So that in our example we can say that a smoker is 6.5 times more likely to develop cancer if they smoke than they would by chance.
2. If the computed odds ratio is close to 1 then the researcher concedes that there is no relationship between the suspected cause and the outcome.
3. If the computed odds ratio is less than 1 then the researcher may suspect that the stimulus of interest is in fact demonstrating a protective effect on the sample observed.

In point 2 above, we suggested that if the odds ratio is close to 1 then the researcher concedes that there is no relationship, ***but how close is close to 1?***

By computing confidence intervals for the odds ratio, we can determine the upper and lower bounds of the odds ratio estimate that we computed for our sample. If the 95% confidence interval includes 1 then we would say that there is no relationship between the suspected risk factor and the outcome.

In order to compute the 95% confidence interval for the odds ratio we first convert the odds ratio to its equivalent as a natural logarithm [ln(OR)]; this is considered the point estimate for the computations of the CI. Then we compute the standard error of the natural logarithm ln(OR) by computing the square root of the sum of the inverse of each cell value as shown in the following formula (3) and then compute the 95% CI for each score. The specific calculations are shown below.

Ln(OR) = ln(6.53) = 1.88;

[latex]\textit{s.e. ln(OR) = } {\pm z} \, {\sqrt { (1/a + 1/b + 1/c + 1/d)}}[/latex]

[latex]\textit{s.e. ln(OR) = } {\pm 1.96} \, {\sqrt { (1/23 + 1/8 + 1/11 + 1/25)}}[/latex]

[latex]\textit{s.e. ln(OR) = }{\pm 1.96} \, {\sqrt{ 0.043+ 0.125 + 0.09 + 0.04}}[/latex]

[latex]\textit{s.e. ln(OR) = }{\pm 1.96} \, {\sqrt{ 0.299}} = (1.96\times 0.547) = 1.07[/latex]

[latex]\therefore[/latex] s.e. of ln(1.88) = 1.07 and the 95% C.I. for ln(OR) ± 1.96* SE ln(OR) = 1.88 +/- 1.07. Thus to compute the lower limit 95% CI we use 1.88-1.07 = 0.81 and to compute the upper limit 95% CI we use 1.88+1.07 = 2.95

Next, because the natural logarithm estimates are transformed from our original estimates we exponentiate the terms of the confidence interval estimates to return to the original scale of our data. Therefore: OR = exp(lnOR) [latex]\rightarrow[/latex] exp(1.88) so that the Odds Ratio is 6.53. Next we exponentiate the lower limit estimate: exp(LL95%) [latex]\rightarrow[/latex] exp(0.81) so that the lower limit of the 95%CI is 2.24; finally we exponentiate the upper limit estimate: exp(UL95%) [latex]\rightarrow[/latex] exp(2.95) so that the upper limit of the 95%CI is 19.10.

Elements Calculated for the 95% Confidence Interval of the Odds Ratio

Natural Log oddsRatio ln(6.53) = 1.88	Standard Error of lnOR = 1.07	95%CI lnOR lower limit = 1.88-1.07 = 0.81	95%CI lnOR upper limit = 1.88+1.07 = 2.95

Exponentiating the 95% Confidence Interval's Upper and Lower limits will return the estimated values to the original scale scores	exp(0.81) = 95%CI OR lower limit = 2.24	exp(2.95) = 95%CI OR upper limit = 19.10

Given that the reconstituted OR = 6.53 with a 95% confidence interval range of 2.24 to 19.10. **does not include 1** then we can say that there is a relationship between the suspected risk factor and the outcome. Further, because we used 95% as the measure for our confidence interval we can say that the relationship between the suspected risk factor and the outcome is significant at the $p<0.05$ level.

Using the webulator for the odds ratio, we can confirm these calculations by inserting the scores into the appropriate cells of the webulator and clicking on the button labeled "Compute".

https://health.ahs.upei.ca/webulators/or_pb.html

An interactive or media element has been excluded from this version of the text. You can view it online here: https://pressbooks.library.upei.ca/montelpare/?p=47

We can also use the following SAS Code to evaluate the data for our lung cancer and smoking example.

```
DATA ODDSRAT;
TITLE 'SAS CALCULATION TO ESTIMATE THE ODDS RATIO';
INPUT ROW COL OUTCOME @@;
DATALINES;
1 1 23 1 2 8 2 1 11 2 2 25
;
PROC SORT DATA=ODDSRAT; BY ROW COL;
PROC FREQ DATA=ODDSRAT ORDER=DATA;
TABLES ROW*COL/CHISQ RELRISK ODDSRATIO;
WEIGHT OUTCOME;
EXACT PCHI OR;
RUN;
```

The output generated by the SAS code is shown, below:

Table of ROW by COL

ROW	COL		
	1	**2**	**Total**
1	23 34.33 74.19 67.65	8 11.94 25.81 24.24	31 46.27
2	11 16.42 30.56 32.35	25 37.31 69.44 75.76	36 53.73
Total	34 50.75	33 49.25	67 100.00

Odds Ratio for Smoking and Lung Cancer

Statistic	**Value**	**95% Confidence Limits**
Odds Ratio	6.5341	2.24 [latex]\rightarrow[/latex] 19.1

Estimating Attributable Risk

According to Bruzzi, Green, Byar, Brinton and Schairer (1985)[1] a widely accepted definition of attributable risk is:

"*the fraction of total disease experience in the population that would not have occurred if the effect associated with the risk factor of interest were absent*".

Quite simply stated, the **attributable risk** is therefore the proportion of infirmity (disease, disorder, injury, outcome) within a cohort that can be attributed to **exposure to a suspected causal agent**.

The attributable risk is calculated as a fraction by subtracting the proportion of cases observed among the total group of non-exposed individuals from the proportion of cases observed among the group of exposed individuals.

A caveat of this estimate is that all other possible influences of cause are considered equal among the exposed and non-exposed groups so that the only difference between the two groups is the exposure.

Attributable risk can be computed from either prevalence or incidence data and is shown here using a 2 x 2 table.

Arrangement of the data to compute Attributable Risk

	+ve condition (CASES)	– ve condition (CONTROLS)	Row totals
Exposed	*a*	*b*	(*a*+*b*)
Not Exposed	*c*	*d*	(*c*+*d*)
Column totals	(*a*+*c*)	(*b*+*d*)	N= (*a*+*b*+*c*+*d*)

Formula for Attributable Risk

[latex]AR = P_{1} - P_{2} = {a \over(a+b)} - {c\over (c+d)}[/latex]

The formula for *Attributable Risk Fraction (exposed)*

$$AF_e = 1 - \frac{1}{RR} = 1 - \frac{1}{\left(\frac{\left(\frac{a}{a+b}\right)}{\left(\frac{c}{c+d}\right)}\right)}$$

The table above illustrates the elements required to calculate the ***Attributable Risk Fraction (exposed)***

The attributable risk fraction for exposure can be estimated from the formula shown above which includes the estimate for relative risk.

Estimating Attributable Risk for NAS Among Newborns

Consider a scenario in which 100 babies were born in the month of September, and in that cohort 11 babies were reported to show the signs Neonatal Abstinence Syndrome (NAS). As a researcher you suspect that the cause of NAS is related to the mother's use of drugs during her pregnancy. Without sub-classifying the data according to volume or type of drug used you created the following 2 x 2 table and sorted the outcomes.

Computing Attributable Risk for NAS In Newborns

	+ve condition (NAS CASES)	– ve condition (CONTROLS)	Row totals
Exposed – mother used a drug	10	20	30
Not Exposed – mother abstained from drug use	2	68	70
Column totals	12	88	N= 100

To calculate the attributable risk begin by calculating the Crude Risk Estimate for the Exposed Group: [latex]{P_{1}} = {\left(a\over{ (a+b)}\right)}= {\left(10\over{ (10+20)}\right)}= 0.33[/latex]

Next calculate the Crude Risk Estimate for the Reference Group (aka the non-exposed group): [latex]{P_{2}} = {\left(c\over{ (c+d)}\right)}= {\left(2\over{ (2+68)}\right)}= 0.029[/latex]

Attributable Risk is estimated from the proportions of outcomes based on the exposed versus non-exposed:*[latex]*AR = *P_{1} - P_{2} = {a \over(a+b)} - {c\over (c+d)}[/latex]*

*[latex]*AR = *P_{1} - P_{2} = {10 \over(10+20)} - {2\over (2+68)} = 0.33 - 0.0285 = 0.3045[/latex]*

Using the webulator shown here for attributable risk, we can confirm these calculations by inserting the scores into the appropriate cells of the webulator and clicking on the button labeled “Compute”.

https://health.ahs.upei.ca/webulators/ar_pb.html

An interactive or media element has been excluded from this version of the text. You can view it online here:
https://pressbooks.library.upei.ca/montelpare/?p=47

We can also use the following SAS Code to evaluate the data for our NAS example.

The SAS code statements shown below compute the Attributable Risk as a Crude Risk in both the exposed and non-exposed – AKA reference groups. The code presented here illustrates estimates related to the attributable risk estimates, as well as estimates of the attributable risk fraction and the population attributable risk.

```
   DATA ARP;
Title ‘Attributable Risk and Population Attibutable Risk’;
TITLE2 ‘Output includes Standardized Mortality Rate’;
input CELLA R1TOTAL CELLC R2TOTAL;
LABEL CELLA = ‘NAS POSITIVE’
R1TOTAL = ‘TOTAL EXPOSED’
CELLC = ‘NOT EXPOSED NAS POSITIVE’
R2TOTAL = ‘TOTAL NOT EXPOSED’;
DATALINES;
10 30 2 70
;
PROC STDRATE data=ARP
REFDATA =ARP
```

```
method=Indirect(AF)
STAT=RISK
Cl=normal
;
POPULATION EVENT=CELLA TOTAL = R1TOTAL;
REFERENCE EVENT = CELLC TOTAL = R2TOTAL ;
RUN ;
```

The SAS code above produces several measures associated with the estimates of attributable risk including the standardized mortality rate and the confidence intervals of the attributable risk.

Crude Risk Estimate from the SAS OUTPUT TABLE:

Standardized Risk			
RiskEstimate	Standard Error	95% Normal Confidence Limits	
0.3333	**0.0861**	**0.1646**	**0.5020**

Notice the values presented in the output table are similar to the values computed by hand and with the Attributable Risk Webulator, allowing for slight rounding differences.

SAS also reports the SMR in the table, and the calculation of the attributable risk fraction and the population attributable risk are based on the relative risk or risk ratio estimates as shown in the formula above. The method used to compute the confidence intervals for the attributable risk fraction and the population attributable risk in percentage terms are shown below.

SAS OUTPUT FOR ATTRIBUTABLE RISK

Observed Events	Number of Obs	Crude Risk	Reference Crude Risk	Expected Events	SMR*	Standardized Risk Estimate	Standard Error	95% Normal Confidence Limits
10	30	0.33	0.028	0.857	11.66	0.33	0.086	0.164 TO 0.502

This table presents the output from the SAS calculations of attributable risk. The format of the table presented here differs from that which is produced by SAS but the data are the same.

The 95% confidence interval in this instance provides the range in which we are 95% confident that the true population estimate for the attributable risk is captured within the estimated interval. To calculate the confidence interval associated with the estimated attributable risk we first identify the proportions of interest from our 2 x 2 table and then compute the standard error of the difference between the two estimates. In the example used here, our proportion estimates were: *[latex]P_{1} = {10 \over(30)} = 0.33[/latex] and [latex]P_{2} = {2 \over(70)} = 0.03[/latex] [latex]{\therefore} P_{1} -P_{2} = (0.33 - 0.03) = 0.30[/latex]*

Next we compute: [latex]q_{1} = (1 - P_{1})[/latex] = 1-0.33 = 0.67 and [latex]q_{2} = (1 - P_{2})[/latex] = 1-0.03 = 0.97

We can now estimate the Standard Error of the Attributable Risk using:

[latex]s.e._{p_{1} - p_{2}} = {\sqrt{{{ p_{1} \times q_{1}}\over{(a+b)} } + {{ p_{2} \times q_{2}}\over{(c+d)} }}}[/latex] = [latex]{\sqrt{{{ 0.33 \times 0.67}\over{(30)} } + {{ 0.03 \times 0.97}\over{(70)} }}}[/latex]

[latex]s.e._{p_{1} - p_{2}}[/latex]= [latex]{\sqrt{0.007 + 0.0004}}[/latex] = [latex]{\sqrt{0.0074}}[/latex] = 0.086

Notice that the standard error computed by hand supports the estimate provided by SAS and the webublator.

To compute the 95% confidence interval consistent with that which our SAS code provided for the AR=0.333, we use the standard error term to estimate the range of the 95% confidence interval for the attributable risk by multiplying 1.96 x 0.086 = 0.168 so that the estimated attributable risk can range from 0.333 ± 0.168 to be [latex]\rightarrow[/latex] lower limit = 0.164 and upper limit = 0.501.

Estimating Attributable Risk Fraction and the Population Attributable Risk

In addition, as noted above, we can also compute the Attributable Risk Fraction (ARF) for the exposed individuals by using the Relative Risk (RR), where RR is the ratio of the crude risks from the exposed to unexposed, as shown here:

Relative Risk [latex]= {\left(\textit{Crude Risk for Exposed}\over{ \textit{Crude Risk for Reference Group}}\right)}= {\left(P_{1}\over{ P_{2}}\right)}=[/latex][latex]{\left(a\over{a+b}\right)} \over{ {\left(c\over{c+d}\right)}}[/latex] [latex]= {\left(0.333\over{ 0.0285}\right)}=11.67[/latex]

Next we can compute the Attributable Risk Fraction (exposed) from the data in our 2 x 2 table using ; where a=10, b=20, c=2, d=68.

So that [latex]AF_{exposed} = 1 - {1 \over{\textit(RR)}}[/latex] = [latex]1 - {1 \over{11.67}}[/latex] = 0.9143 which we can convert to a percent value using [latex]\rightarrow[/latex] 0.9143 [latex]\times[/latex] 100.

Likewise the attributable risk can also be calculated in percentage terms using:

[latex]ARF = {P_{1}-P_{2} \over{P_{1}}} \times 100 = {0.33-0.0285 \over{0.33}} \times 100= {0.30 \over{0.33}} \times 100[/latex] = 91%

The estimated Attributable Risk Fraction (exposed) is useful because it can be interpreted relative to the effect of exposure. Here we can say that the estimate for the Attributable Risk Fraction (exposed) indicates that the risk of NAS among babies is approximately 91% higher when the mother uses a drug during pregnancy.

When we compute the Attributable Risk Fraction (exposed), a typical next step is to compute the population attributable risk (PAR). The population attributable risk is an estimate that extends the attributable risk fraction from the observed sample to the larger population.

Again, using the data from our 2 x 2 table the PAR can be calculated as follows:

[latex]PAR = p {\left((RR-1)\over{RR}\right)}; \textit{Where: } p= {\left(a\over{ (a+c)}\right)}=[/latex] [latex]{10\over{12}}= 0.833[/latex] and RR = 11.67 [latex]\therefore PAR = {10\over{12}} \times{11.67 - 1 \over{11.67}} = 0.833 \times{0.9143} =0.7642[/latex]. This value can also be expressed in percentage terms as approximately **76%.**

The SAS code statements used to compute the estimates for the ARF and PAR provide a summary table that includes the population estimates for each of the parameter values. It is important to notice, that although we can use simple algebra to compute the point estimates, the confidence intervals are not normally distributed and therefore require more advanced formulae that are available on the SAS Support website.

SAS Output for Attributable Fraction Estimates

Parameter	Estimate	95% Confidence Limits	
Attributable Risk	0.91429	0.82647	0.94309
Population Attributable Risk	0.76190	0.21732	0.92757

The important information to glean from the output table above is that because neither the estimate for the Attributable Risk Fraction nor the Population Attributable Risk includes 0, we can say that consistent with the test of the null hypothesis, these observed parameter estimates are significant at the p<0.05 level; and that taking drugs while pregnant **CAN** ***increase the risk*** of NAS in newborns.

[1] Bruzzi, P., S. B. Green, D. P. Byar (Biometry Branch, National Cancer Institute, NIH, Bethesda, MD 20205), L A. Brinton, and C. Schairer. Estimating the population attributable risk for multiple risk factors using case-control data. Am j Epidemiol 1985; 122:904-14.

PART IV

ANALYSIS OF NON-PARAMETRIC OUTCOMES

Often referred to as distribution-free statistics, non-parametric statistics are used when the data may not demonstrate the characteristics of normality (i.e. follow a normal distribution). Non-parametric statistics are used with nominal data, where the response set can be converted to counts of events, and the measurement scale is ignored.

Non-parametric statistics can be used when data are converted to ranks.

Non-parametric statistics are most useful when data are not normally distributed, or when sample sizes are so small that the representativeness of the sample to the population is questionable.

The standard scores (a.k.a. z scores)

In statistics, when we want to standardize scores within a distribution, we simply transform the scores using a common denominator to create a ratio level measurement. One of the simplest methods for standardizing scores is to produce an estimate referred to as the "z" score.

The z score is referred to as the standard normal value or standard normal deviate because it follows the standard normal distribution and represents the standardized estimate of difference of any score within a random variable from the mean of the random variable. The standard normal distribution is represented by the normal curve. The standard normal distribution has a mean = 0 and a standard deviation = 1.

Given that this exercise was to demonstrate to the research community how the set of sample scores are associated with the true set of population scores, then we need to find some way of relating the sample distribution to the population distribution (or how is the set of scores for the sample related to the set of scores for the population). One way to illustrate such a relationship is to standardize the scores for both the sample distribution and the population distribution. In statistics when we want to standardize an estimate we typically relate the estimate to a measurement standard called the normal curve.

The normal curve is a graphical representation of the standard normal distribution (ie. the frequency distribution graph of an expected distribution of scores within a "normal population"). By using the normal curve, researchers can describe how closely their sample distribution represents a population distribution.

Understanding the role of the normal curve is important to inferential statistics. The normal curve is a graphical presentation of the frequency distribution for a set of standardized (or adjusted) scores. For any set of z scores, a percentile estimate can be attributed to each z score. This has been shown several times and is commonly known as the Z table of estimates or the table for the normal curve. Conversely then for any percentile, we could determine a standardized estimate or a z score. That is, we could determine the z score for a percent of confidence such as the 95% confidence value.

The normal curve approximation

We call the computation of the z score, the normal curve approximation since we are trying to estimate where our events fall within the set of possible outcomes represented by the normal curve. The set of hypothetical expected outcomes for the normal curve is presented in the figure below. Notice the critical region in which we accept the null

hypothesis is within the boundaries (-1.96) to (+1.96). The region to accept the null hypothesis generally accounts for 95% of the expected outcomes, allowing 5% of the outcomes to be outside the region of acceptance (shown here as 2.5% in each of the tails).

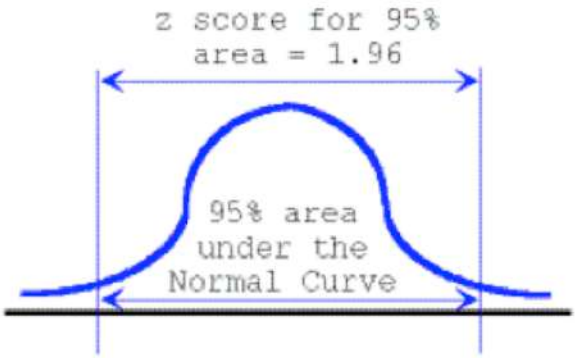


Illustration of the area under the normal curve

Decision rules for the normal curve approximation

If the calculated value for the z score (the Normal Curve Approximation) is within the boundaries of the critical values (-1.96) to (+1.96) {values greater than (-1.96) but less than (+1.96)} then we would say that the value falls within our region of accepting the null hypothesis and therefore state that the events were random. If the calculated value for the z score (the Normal Curve Approximation) is outside the boundaries of the critical values (-1.96) to (+1.96) {values less than (-1.96) but greater than (+1.96)} then we would say that the value falls outside our region of accepting the null hypothesis. Therefore, we must reject the null hypothesis and state that the events did not occur at random, rather, the events followed a distinct pattern.

The standardized scores (or **z scores**) are ratio scores based on the difference between any score within a set of scores and the measure of central tendency for that set of scores, divided by the standardized error attributed to that set of scores; as shown in the formula for z scores:

[latex]z = {{\left(x_{i}- \overline{x}\right) \over {s}}}[/latex]

22. Calculating Probabilities

Learner Outcomes

After reading this chapter you should be able to:

- apply probabilistic approaches to compute the likelihood of outcomes
- recognize and apply the binomial probability formula

1. Computing Bernoulli Trials

The rules of a Bernoulli trial are straight-forward. Given an independent process in which an outcome can be observed, the outcome can have only two possibilities and the chance or probability of the observed outcome is the same as the chance or probability of the non-observed outcome. Hence the fair toss of a fair coin is an excellent demonstration of a Bernoulli trial, because, as we observe in the tossing of a fair coin, there are only two possible outcomes: a head or a tail. Likewise, the probability of tossing a head is equal to the probability of tossing a tail, and this probability is equal to 0.5 or one-half. Further, if the coin is fair and the toss or flip is fair – without any external influence, then we can say that the process was independent.

When we are computing Bernoulli trials we often use the term *event* to refer to the process or test that we are conducting, and the outcome variable as the indicator variable. The outcome of an event in a Bernoulli trial is an element of the Bernoulli distribution, whereby the Bernoulli distribution is described as a discrete distribution with a possibility of one of two outcomes. The indicator variable sometimes referred to as a DUMMY variable or a BINARY variable, has two possible outcomes (success or failure). Further, when scoring the indicator variable we typically assign a value of 1 to the success and a value of 0 to the failure.

The notation used to represent the outcome of a Bernoulli trial is [latex]X_{i}[/latex], so that [latex]X_{1}[/latex] refers to a single Bernoulli trial and[latex]X_{n}[/latex] refers to ***n*** Bernoulli trials where ***n*** ranges from 1 to infinity. Further, the probability of success of an outcome in a Bernoulli trial is written as: ([latex]P(X_{i} =1)[/latex]) = p, while the probability of failure of an outcome in a Bernoulli trial is written as ([latex]P(X_{i} =0)[/latex]) = 1 – p.

We can also use p and q to represent the outcome of a Bernoulli trial, where p is representative of the probability of success and q is representative of the probability of failure. The probability of p is assigned in a fair and independent event as p = 0.5, and the probability of q is assigned as (1 – p) = (1 – 0.5) = 0.5.

In the following example, we can use SAS and a set of probability outcomes that range from 0 to 1 and are based on an interval of 0.025 to plot the variance of a Bernoulli trial. In this example, the outcome is based on the assumption that the mean [latex]X_{i}[/latex] = p and the variance of [latex]X_{i}[/latex] = p(1-p).

The data set for this example will be based on[latex]X_{1}[/latex] = p: 0.00, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25, 0.275, 0.3, 0.325, 0.35, 0.375, 0.4, 0.425, 0.45, 0.475, 0.5, 0.525, 0.55, 0.575, 0.6, 0.625, 0.65, 0.675, 0.7, 0.725, 0.75, 0.775, 0.8, 0.825, 0.85, 0.875, 0.9, 0.925, 0.95, 0.975, 1

These data are entered as follows:

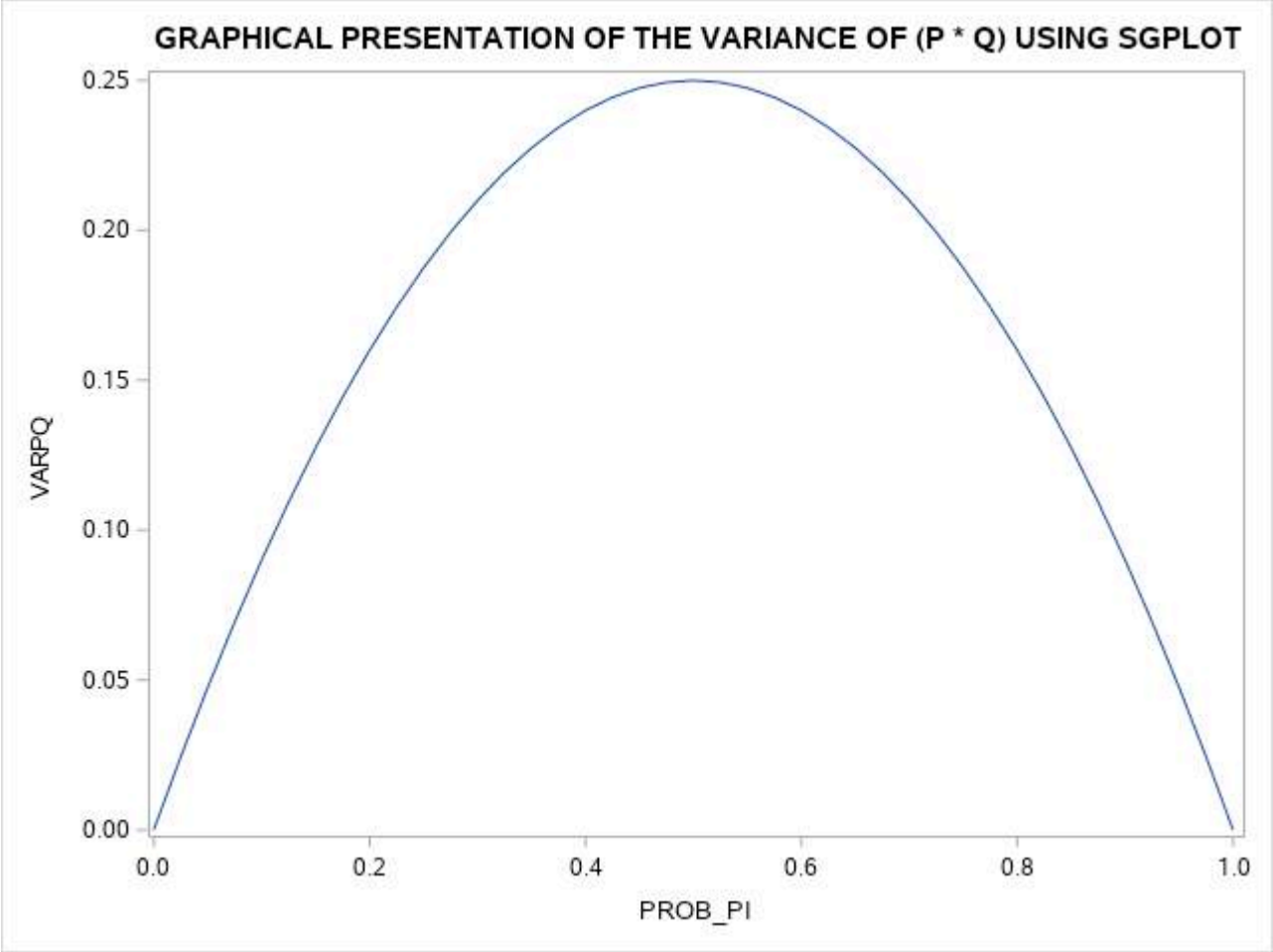
1 0.00
2 0.025
3 0.05
. .
. .
. .
39 0.95
40 0.975
41 1.00

The SAS code to produce the variance [latex]\color{dbeeff}\rightarrow {varX_{1}}[/latex]= p(1-p) based on these data is shown here.

SAS Program to compute variance of a Bernoulli Trial

```
   DATA BERNOULLI;
INPUT ID PROB_PI @@ ;
PROB_QI=(1-PROB_PI);
VARPQ=(PROB_PI * PROB_QI);
   DATALINES;
1 0.00 2 0.025 3 0.05 4 0.075 5 0.1 6 0.125 7 0.15
8 0.175 9 0.2 10 0.225 11 0.25 12 0.275 13 0.3
14 0.325 15 0.35 16 0.375 17 0.4 18 0.425 19 0.45
20 0.475 21 0.5 22 0.525 23 0.55 24 0.575 25 0.6
26 0.625 27 0.65 28 0.675 29 .7 30 0.725 31 0.75
32 0.775 33 0.8 34 0.825 35 0.85 36 0.875 37 0.9
38 0.925 39 0.95 40 0.975 41 1
;
PROC SGPLOT;
SERIES X=PROB_PI Y=VARPQ;
* XAXIS TYPE = DISCRETE;
TITLE1 "GRAPHICAL PRESENTATION OF THE VARIANCE OF (P * Q) USING SGPLOT ";
   RUN;
PROC PRINT;
VAR PROB_PI PROB_QI VARPQ;
TITLE1 'PRINT OF DATA FOR COMPLETE BERNOULLI TRIAL';
RUN;
```

The SAS statements: proc SGPLOT and PLOT varpq*prob_pi produced the following graph which shows the distribution of variance across all estimates of [latex]p_{success}[/latex] and [latex]q_{failures}[/latex] from 0.00 to 1.00.



With the PROC PRINT statement, we can produce a complete listing of the data set for probabilities of success (**[latex]p_{i}[/latex]**) and the probability of failures (**[latex]q_{i}[/latex]** along with the variance of the success and failures (variance of p*q). These results are shown in the 22.1 below.

Table 22.1 Discrete Probability Distribution of the Bernouli Trial for all possible outcomes for the data set [latex]X_{i}[/latex] where i = 0 to 41.

Obs(i)	[latex]p_{i}[/latex]	[latex]q_{i}[/latex]	variance of p*q
1	0.000	1.000	0.00000
2	0.025	0.975	0.02438
3	0.050	0.950	0.04750
4	0.075	0.925	0.06938
5	0.100	0.900	0.09000
6	0.125	0.875	0.10938
7	0.150	0.850	0.12750
8	0.175	0.825	0.14438
9	0.200	0.800	0.16000
10	0.225	0.775	0.17438
11	0.250	0.750	0.18750
12	0.275	0.725	0.19938
13	0.300	0.700	0.21000
14	0.325	0.675	0.21938
15	0.350	0.650	0.22750
16	0.375	0.625	0.23438
17	0.400	0.600	0.24000
18	0.425	0.575	0.24438
19	0.450	0.550	0.24750
20	0.475	0.525	0.24938
21	0.500	0.500	0.25000
22	0.525	0.475	0.24938
23	0.550	0.450	0.24750
24	0.575	0.425	0.24438
25	0.600	0.400	0.24000
26	0.625	0.375	0.23438
27	0.650	0.350	0.22750
28	0.675	0.325	0.21938
29	0.700	0.300	0.21000
30	0.725	0.275	0.19938
31	0.750	0.250	0.18750
32	0.775	0.225	0.17438
33	0.800	0.200	0.16000
34	0.825	0.175	0.14438
35	0.850	0.150	0.12750
36	0.875	0.125	0.10938
37	0.900	0.100	0.09000
38	0.925	0.075	0.06937
39	0.950	0.050	0.04750
40	0.975	0.025	0.02438

Obs(i)	[latex]p_{i}[/latex]	[latex]q_{i}[/latex]	variance of p*q
41	1.000	0.000	0.00000

The proc freq procedure produced a complete frequency distribution independently for each of the variables: prob_pi , prob_qi , and varpq.

The output shown below is identical for the frequency distributions of the variables prob_pi and prob_qi. Therefore, only the data for prob_pi is shown here.

prob_pi	Freq	PCT	Cumulative Frequency	Cumulative Percent
0	1	2.44	1	2.44
0.025	1	2.44	2	4.88
0.05	1	2.44	3	7.32
0.075	1	2.44	4	9.76
0.1	1	2.44	5	12.20
0.125	1	2.44	6	14.63
0.15	1	2.44	7	17.07
0.175	1	2.44	8	19.51
0.2	1	2.44	9	21.95
0.225	1	2.44	10	24.39
0.25	1	2.44	11	26.83
0.275	1	2.44	12	29.27
0.3	1	2.44	13	31.71
0.325	1	2.44	14	34.15
0.35	1	2.44	15	36.59
0.375	1	2.44	16	39.02
0.4	1	2.44	17	41.46
0.425	1	2.44	18	43.90
0.45	1	2.44	19	46.34
0.475	1	2.44	20	48.78
0.5	1	2.44	21	51.22
0.525	1	2.44	22	53.66
0.55	1	2.44	23	56.10
0.575	1	2.44	24	58.54
0.6	1	2.44	25	60.98
0.625	1	2.44	26	63.41
0.65	1	2.44	27	65.85
0.675	1	2.44	28	68.29
0.7	1	2.44	29	70.73
0.725	1	2.44	30	73.17
0.75	1	2.44	31	75.61
0.775	1	2.44	32	78.05
0.8	1	2.44	33	80.49
0.825	1	2.44	34	82.93
0.85	1	2.44	35	85.37
0.875	1	2.44	36	87.80
0.9	1	2.44	37	90.24
0.925	1	2.44	38	92.68
0.95	1	2.44	39	95.12
0.975	1	2.44	40	97.56

prob_pi	Freq	PCT	Cumulative Frequency	Cumulative Percent
1	1	2.44	41	100.00

However, the frequency distribution for var(PQ) is unique and is shown here.

Var(p*q)	Frequency	PCT	Cumulative Frequency	Cumulative Percent
0	2	4.88	2	4.88
0.024	2	4.88	4	9.76
0.048	2	4.88	6	14.63
0.069	2	4.88	8	19.51
0.09	2	4.88	10	24.39
0.109	2	4.88	12	29.27
0.128	2	4.88	14	34.15
0.144	2	4.88	16	39.02
0.16	2	4.88	18	43.90
0.174	2	4.88	20	48.78
0.188	2	4.88	22	53.66
0.199	2	4.88	24	58.54
0.21	2	4.88	26	63.41
0.219	2	4.88	28	68.29
0.228	2	4.88	30	73.17
0.375	2	4.88	32	78.05
0.24	2	4.88	34	82.93
0.244	2	4.88	36	87.80
0.248	2	4.88	38	92.68
0.249	2	4.88	40	97.56
0.25	1	2.44	41	100.00

2. The Coin Toss That Might Mean Something

The American Football league's national championship: the Super Bowl begins with a coin toss. At the start of the game, the captain's of each team meet in the centre of the field to toss a coin to determine which one of the teams will start the game as the kicking team and which team will start the game as the receiving team. Since the outcome of either kicking the ball to the opposing team to start the game or receiving the ball from the opposing team to start the game may have consequences on the final score, there is an attempt to make this decision an unbiased and fair process. The National Football League has chosen to render this decision to a Bernoulli trial.

Considering that a fair toss of a fair coin has a 50% chance of turning up heads and a 50% chance of turning up tails then the use of a coin toss to determine outcomes is a good approach.

The Binomial Formula establishes the probability using the following formula:

[latex]P_{x} = \frac{n!}{x!(n-x)!} \times p^{x} q^{n-x}[/latex]

The elements of this probability prediction formula are explained as follows:

[latex]P_{x} =[/latex] the probability of exactly *x* events of a given outcome appearing in ***n* trials.**

p = the probability of an event on any given trial (if we are flipping a coin then this value is ½ with a fair coin).

q = the probability of an event on any given trial *q* = 1- *p* (usually this value is ½ if we were flipping a coin).
n = the number of events.
x = the number of a given outcome (e.g. heads) being evaluated.

Consider an example for the Probabilities associated with tossing a fair coin

The coin tossing exercise is a useful way of demonstrating the probability of an outcome within a given set of trials when the expected chance of an outcome is fixed (known) or expected. For example, if we have a "*fair*" coin then the expected probability or chance of tossing a given outcome (i.e. heads) is 0.5 or ½. Therefore, given ten tosses of the fair coin we could predict the number of times we should expect to see the outcome as heads or tails.

In other words, to compute the proportion of outcomes observed we can predict the chance that an outcome or event will occur.

In the following example, we can determine the probability associated with flipping a "head" four times in ten tosses of a fair coin. That is, if we flip a fair coin ten times then we could predict the number of times we should expect to see "heads" appear in four of the ten flips.

The formula used to resolve this question is the binomial and is worked through as follows. Let x=4 (the number of heads), n=10 (the number of throws), and P=probability of 4 heads in 10 throws, where *p* is the starting probability and *q* is 1 – *p*. We begin with the binomial formula:

[latex]P_{x} = \frac{n!}{x!(n-x)!} \times p^{x} q^{n-x}[/latex]

Step 1: [latex]P_{4} = \frac{10!}{4!(10-4)!} \times (0.5)^{4} (0.5)^{10-4}[/latex]

Step 2: [latex]P_{4} =[/latex][latex]({10*9*8*7*6*5*4*3*2*1})\over({4*3*2*1}) \times ({6*5*4*3*2*1})[/latex][latex]\times (\frac{1}{2})^{4+6}[/latex]

Step 3: [latex]P_{4} =[/latex][latex]({10*9*8*7} \times{\enclose{horizontalstrike}{6*5*4*3*2*1}}) \over({4*3*2*1}) \times {\enclose{horizontalstrike}{6*5*4*3*2*1}})[/latex][latex]\times (\frac{1}{2})^{10}[/latex]

Step 4: [latex]P_{4} ={5040 \over{24} } \times {(1)^{10} \over (2)^{10} }[/latex]

Step 5: [latex]P_{4} ={210\times{1\over{1024} }}[/latex]

Step 6: [latex]P_{4} ={210\over{1024} }[/latex] = **0.206**

The calculation table above shows us that in ten tosses of a fair coin there is roughly a 20 percent chance of tossing 4 heads. Further, we can use the binomial formula to compute all possible outcomes for a given series of events when we establish the beforehand (a priori) probability of an outcome in a defined set.

For example, let's use the binomial to compute all possible outcomes for ten tosses of a fair coin. That is, how many times in 10 tosses would 0 heads appear? Likewise, how many times in 10 tosses would 1 through to 10 heads appear?

After working through each application of the binomial equation we could create a table of all possible events in the outcome space. This table is referred to as the Probability Density Chart, and is shown below.

The Probability Density Chart for the outcome space when determining the likelihood of tossing a head in 10 tosses of a fair coin

(x = number of a given outcome; p = ½ and q = 1 − p= ½)

x	The probability expressed as a ratio	The probability expressed as a decimal
0	1:1024	0.0009765
1	10:1024	0.0097656
2	45:1024	0.0439453
3	120:1024	0.1171875
4	210:1024	0.2050781
5	252:1024	0.2460937
6	210:1024	0.2050781
7	120:1024	0.1171875
8	45:1024	0.0439453
9	10:1024	0.0097656
10	1:1024	0.0009765
SUM	1024:1024	1.00

3. Patients as Coins – An Application of the Coin Toss

We can use the example of tossing a fair coin as a proxy for estimating the likelihood of identifying individuals to develop health conditions.

For example, consider the hypothetical situation where it is suspected that families within certain rural environments may be exposed to carcinogenic compounds in their drinking water as a result of run-off from farm fields into their wells. Let's start with the following scenario in which you are asked, "What is the likelihood of observing 15 blood screens that test positive for a given carcinogenic substance in blood samples drawn from 25 mothers attending a prenatal health program?"

To compute the likelihood of observing 15 positive blood screens for the 25 mothers sampled, we decide to use the binomial formula with the following elements. Since we observed 15 positive cases then the term [latex]P_{x}[/latex] refers to the probability of observing this outcome in the 25 mothers, where **x=15** and **n=25**.

[latex]P_{x} = \frac{n!}{x!(n-x)!} \times p^{x} q^{n-x}[/latex]

Step 1: [latex]P_{15} = \frac{25!}{15!(25-15)!} \times {\left({1\over{2} }\right)^{15}}{\left({1\over{2} }\right)^{25-15}}[/latex]

Step 2: [latex]P_{15} = \frac{25!}{15!\times(10!)} \times {\left({1\over{2} }\right)^{15+10}}[/latex]

While this formula looks neat in the arrangement of terms it can become quite unwieldy quickly because we are multiplying and dividing such large numbers. Note the term (n!) is 25! which indicates that we use a series of multiplication

steps that are (n * (n-1)) repeatedly until we converge to (2 * (2-1)). Given that we include factorials in the numerator and the denominator our challenge is to organize all of the operations while respecting the BEDMAS principle and arriving at the appropriate solution to the formula. While we can do this with a handheld calculator it is so much easier to simply write a program to analyze this scenario using the following SAS code:

Use SAS to do the work in our computations of probability to identify individuals to develop health conditions.

```
   DATA BERN2;
/* create the variable to represent the numerator (n!) */
NUM1 = FACT(25);/* create the variable to represent the denominator (x!(n-x)!) */
DEN1= (FACT(15)*FACT(10));/* notice in the statements above, the function to produce a factorial of a number
is FACT(#), as in FACT(25) will produce the factorial of the number 25. *//* create the variable to represent the
first fraction  */
FRACTION1= NUM1 / DEN1;/* create variable to represent the combined probability estimates, then include
PUT statement to use the outcome in the subsequent calculations */
PQ1= (0.5)**25; PUT PQ1;/* variable to represent the expected outcome */
ANSWER1 = (FRACTION1 * PQ1);
RUN;/* print the important variables */
PROC PRINT; VAR  NUM1 DEN1 FRACTION1 PQ1 ANSWER1;
RUN;
```

The SAS code above produced the following table of results.

1	1.5511_E25	4.7453_E18	3268760	2.9802_E-8	0.097417

Let's walk through this SAS Output to explain each of the parts of the exercise in calculating the probability of identifying 15 Cases from a sample of 25 women visiting the health clinic. The following elements of the Bernoulli equation (binomial equation) were computed with the SAS program above.

(i) The Numerator term is:**[latex]\color{#330099}(n!)[/latex]**

1.5511(E25) [latex]\rightarrow[/latex] reminds us to add 21 trailing zeros and move the decimal place to the right by 25 spaces, since E25 refers to 1.5511 times ten to the twenty-fifth power

(ii) The Denominator term is:**[latex]\color{#330099}(x!)\times(n-x)![/latex]**

4.7453(E18) [latex]\rightarrow[/latex]add 13 leading zeros since E18 refers to 4.7453 times ten to the eighteenth power

(iii) The fraction of **[latex]\color{#330099}\frac{(n!)}{(x!)\times(n-x)! }[/latex] is:**

[latex]{\textit{1.5511(E25)}\over{\textit{4.7453(E18)}}} = 3,268,760[/latex]

(iv) The unbiased expected probability terms [latex]p^x q^{n-x}[/latex] are:

[latex]{\left({1\over{2} }\right)^{15+10}}[/latex]
2.98 E-8 which represents 0.0000000298 because E with a – sign indicates the number by which we move the decimal to the left of the whole number and add leading zeros (0)

(v) The resulting probability expressed in terms of a percentage:

[latex]P_{15}\rvert (N=25)[/latex]
0.097 can be expressed as 10%

In example 1, we found that there was a 10% chance of observing 15 positive screens for the suspected carcinogen in the sample of 25 women attending the prenatal class.

What is the likelihood of identifying 5 positive cases for the suspected carcinogen in a sample of 50 women selected at random from the corresponding urban environment?

The SAS code to resolve this question is:

```
DATA BERN3;
EX2A= FACT(50); EX2B= ((FACT(5)*FACT(45)));
EX2C= EX2A/EX2B; EX2D = (0.5)**50; EX2E = EX2C * EX2D;
RUN;
PROC PRINT; VAR EX2A EX2B EX2C EX2D EX2E ;
RUN;
```

The SAS code produced the following output:

Numerator [latex](50!)= 50 \times 49 \times 48 \times ... \times 2 \times 1[/latex]	Denominator [latex](5!) \times (45!)[/latex]	[latex]{Numerator}\over{Denominator}[/latex]	[latex]{\left({1\over{2} }\right)^{5+45}}[/latex]	[latex]P_{5}\rvert (N=50)[/latex]
3.0414E64	1.4355E58	2118760	8.8818E-16	1.8818E-9

In example 2, we suggest that the likelihood of observing 5 cases in 50 patients was extremely unlikely and is less than 1% as shown by the answer à 1.88 E-9 which translates to a probability of 0.00000000188 given a sample size of 50 women, and can be written as: [latex]P_{5}\rvert (N=50) = 0.00000000188[/latex]

Computing the probabilities of tossing a single die

Considering a single die – what is the probability (or chance) of rolling a given number? For example, for a single die, estimate the probability of rolling a number less than "5".

Step 1: determine the set of all possible outcomes.

1 roll of a single die = {1, 2, 3, 4, 5, 6} = 6

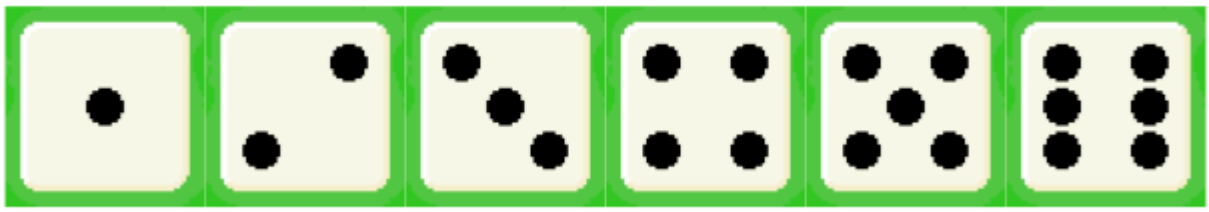


Figure 22.2 Image of all possible outcomes for a single die

On the roll of the dice . . . A single die has six sides, each side with a different number from 1 to 6.

Therefore, the set of all possible outcomes is:

– 1 die = {1, 2, 3, 4, 5, 6}

– the probability of rolling any "given number" is 1/6 or p(roll) = 0.17.

*Therefore, with a **single die**, estimate the probability of rolling a number less than "5".*

Step 1: determine the set of all possible outcomes.

1 roll of a single die = {1, 2, 3, 4, 5, 6} = 6

Step 2: determine the set of favourable outcomes.

Numbers less than 5 = {1, 2, 3, 4} = 4

Step 3: divide the number of favourable or anticipated outcomes by the number of possible outcomes to estimate the probability. Therefore, there is a 67% chance of rolling a number less than 5 as shown here:

Probability = 4/6 = 2/3 = 0.6666 = 67%

***HOWEVER,** what if we were asked to consider rolling a number less than 5, in four of ten tosses of a single die? To answer this question we would apply the binomial formula, using the following apriori estimates: n=10, x=4, p=0.67, q=0.33.*

[latex]P_{x} = \frac{n!}{x!(n-x)!} \times p^{x} q^{n-x}[/latex]

[latex]P_{4} = \frac{10!}{4!(10-4)!} \times {\left({0.67}\right)^{4}}{\left({0.33}\right)^{10-4}}[/latex]

[latex]P_{4} = {210} \times {0.000259}[/latex]

[latex]P_{4} = {0.05465}[/latex] = roughly 5%

---

4. Computing Probabilities Associated With Lottery Number Selection

So what is the probability of winning from the purchase of a single lottery ticket?

The chance of any single combination of six numbers from 1 to 49 is extremely low [latex]{1 \over {49 \choose{6}}}[/latex] which is read as 1 ticket divided by the binomial coefficient of (n choose k) or (49 choose 6) and our likelihood of winning the lottery is 1 chance in 13,983,816 combinations.

Let's say you wanted to buy a lottery ticket on the *lotto* **649**. You pay one dollar and pick 6 numbers from 49 on a specific computer scan sheet. Your first expectation after (or maybe prior to) purchasing the lottery ticket is that every

number on the lottery card between 1 and 49 has an equally likely chance of being selected. Therefore, if every number on the card has an equally likely chance of being selected, then every combination of 6 numbers that can be made from the 49 numbers on the lottery card, has an equally likely chance of being selected. This is an expectation that the selection of the numbers from the lottery card is truly random.

How many combinations of six numbers are we really talking about?

To compute the number of possible combinations of 6 numbers from the 49 numbers, we need to use the following combinatorial (or factorial) formula. We have 49 numbers choose 6. The number 49 represents the population from which the sample "6" was chosen. We write the formula for determining the combinations using the following combinatorial equation or the binomial coefficient:

[latex]{N \choose{n}} = {49 \choose{6}}[/latex]

or we may wish to write the formula using a factorial format as:

[latex]{N! \over {n!{(N - n)}!}}[/latex] = [latex]{49! \over {6!{(49 - 6)}!}}[/latex]

Therefore the number of all possible combinations of 6 numbers from a set of 49 consecutive numbers is:

[latex]{(49 \times 48 \times ... 2 \times 1) \over{(6 \times 5 \times ... 1) \times (49 \times 48 \times ... 2 \times 1) }}[/latex][latex]= {(10,068,347,520) \over{720 }} = 13,983,816[/latex]

Yet you won't be happy unless all of your numbers were chosen, but REALLY what is the chance that all six of your numbers will be selected by the lottery machine. Well since you only bought one ticket, then your chance of winning the lottery is 1 in 13,983,816 chances, or [latex]1 \over{49 \choose{6}}[/latex] [latex]\rightarrow { 1 \over{13,983,816} }[/latex] where the value 0.000000071 represents the probability associated with your set of scores.

Given this large set of possible outcomes, how might we evaluate the data that are generated from one year of twice-weekly draws for any patterns that seem to be emerging?

One of the simplest ways to present these data is to combine all of the numbers and present the outcome data in a chart of the **frequency of outcomes**. This organizational strategy would show that 6 unique numbers are drawn from the set of possible numbers ranging from 1 to 49, each week for 104 picks (52 weeks with draws held twice weekly). This approach considers that we are using **sampling without replacement,** which means that once a number has been selected from the set of 49 possible outcomes each week, that number cannot be selected again in that week. As shown below, the set of outcomes can be organized by the order of choices per week. That is, for any given lottery we can chart the first number drawn, the second number drawn, the third number drawn, the fourth number drawn, the fifth number drawn, or the sixth number drawn, each week.

draw #	1st pick	2nd pick	3rd pick	4th pick	5th pick	6th pick
1	**13**	**21**	**7**	**32**	**47**	**11**
2	**5**	**34**	**28**	**2**	**14**	**44**
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
103	**33**	**16**	**21**	**48**	**15**	**1**
104	**18**	**49**	**28**	**3**	**26**	**37**

The set of outcomes will then generate a table with 104 rows representing the six numbers drawn each week. However,

this table is far too cumbersome and will not help us to make sense of the choices. Using SAS and the PROC FREQ command we can generate a set of six unique outcomes for 104 draws to replicate the twice-weekly draws of the lottery in a given year (52 weeks x 2 draws per week).

Copy the following program to your SAS space and run the program to see which lucky lottery numbers you can produce from [latex]{49 \choose{6}}[/latex]. Using if-then logic statements will enable you to group the data for each ball drawn each week and thereby provide simple categories to graph the outcomes.

SAS PROGRAM TO GENERATE 104 LOTTERY PICKS from 49 choose 6 combinations

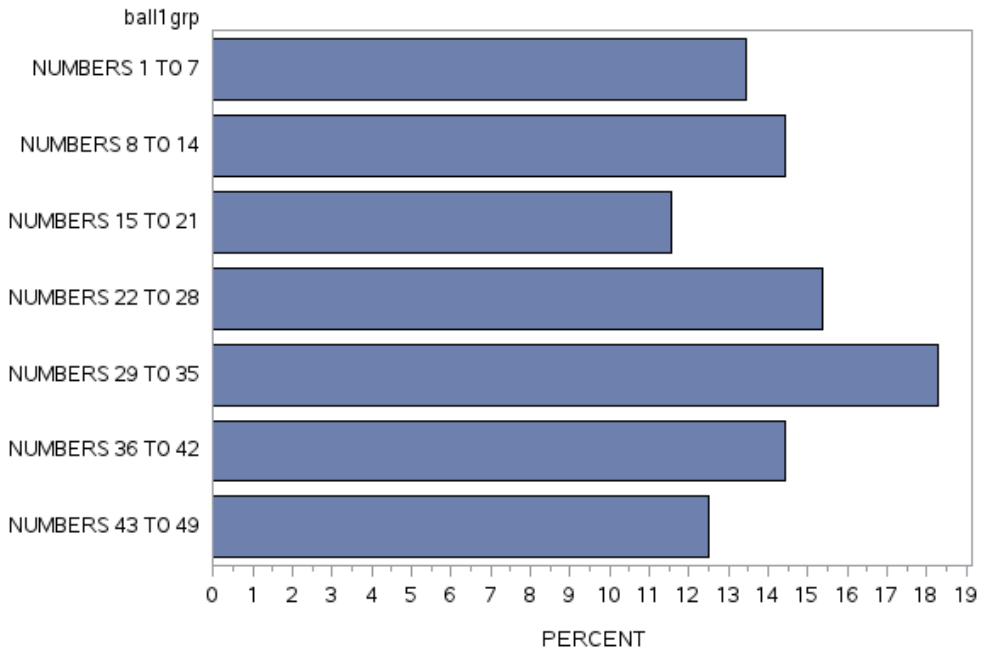
```
options pagesize=60 linesize=80 center date;
PROC FORMAT;
VALUE GRPFMT 1 = 'NUMBERS 1 TO 7'
             2 = 'NUMBERS 8 TO 14'
             3 = 'NUMBERS 15 TO 21'
             4 = 'NUMBERS 22 TO 28'
5 = 'NUMBERS 29 TO 35'
6 = 'NUMBERS 36 TO 42'
7 = 'NUMBERS 43 TO 49';
data sasrng1;
call streaminit(13);
/* this is the seed for the RNG */
array balls ball1-ball6;
   do k=1 to 104;
        do i=1 to 6;
           balls(i) = RAND("normal")*1000000000000;
           balls(i)=ROUND(balls(i));
           balls(i)=1+(mod(balls(i),49));
           balls(i) = ABS(balls(i));
 if ball1 = 0 then ball1 = 1;
if ball1 >0 and ball1<8 then ball1grp=1;
if ball1 >7 and ball1<15 then ball1grp=2;
if ball1 >14 and ball1<22 then ball1grp=3;
if ball1 >21 and ball1<29 then ball1grp=4;
if ball1 >28 and ball1<36 then ball1grp=5;
if ball1 >35 and ball1<43 then ball1grp=6;
if ball1 >42 and ball1<50 then ball1grp=7;
end;
call streaminit(999);
do until (ball2 ne ball1);
      ball2 = RAND("normal")*1000000000000;
      ball2 = ROUND(ball2);
      ball2 = 1+(mod(ball2,49));
      ball2 = ABS(ball2);
 if ball2 = 0 then ball2 = 1;
```

```
if ball2 >0 and ball2<8 then ball2grp=1;
if ball2 >7 and ball2<15 then ball2grp=2;
if ball2 >14 and ball2<22 then ball2grp=3;
if ball2 >21 and ball2<29 then ball2grp=4;
if ball2 >28 and ball2<36 then ball2grp=5;
if ball2 >35 and ball2<43 then ball2grp=6;
if ball2 >42 and ball2<50 then ball2grp=7;
end;
call streaminit(28);
do until (ball3 ne ball2 and ball3 ne ball1);
      ball3 = RAND("normal")*1000000000000;
      ball3 = ROUND(ball3);
      ball3 = 1+(mod(ball3,49));
      ball3 = ABS(ball3);
 if ball3 = 0 then ball3 = 1;
if ball3 >0 and ball3<8 then ball3grp=1;
if ball3 >7 and ball3<15 then ball3grp=2;
if ball3 >14 and ball3<22 then ball3grp=3;
if ball3 >21 and ball3<29 then ball3grp=4;
if ball3 >28 and ball3<36 then ball3grp=5;
if ball3 >35 and ball3<43 then ball3grp=6;
if ball3 >42 and ball3<50 then ball3grp=7;
end;
call streaminit(218);
do until (ball4 ne ball3 and ball4 ne ball2 and ball4 ne ball1);
      ball4 = RAND("normal")*1000000000000;
      ball4 = ROUND(ball4);
      ball4 = 1+(mod(ball4,49));
      ball4 = ABS(ball4);
 if ball4 = 0 then ball4 = 1;
if ball4 >0 and ball4<8 then ball4grp=1;
if ball4 >7 and ball4<15 then ball4grp=2;
if ball4 >14 and ball4<22 then ball4grp=3;
if ball4 >21 and ball4<29 then ball4grp=4;
if ball4 >28 and ball4<36 then ball4grp=5;
if ball4 >35 and ball4<43 then ball4grp=6;
if ball4 >42 and ball4<50 then ball4grp=7;
end; call streaminit(28);
do until (ball5 ne ball4 and ball5 ne ball3 and ball5 ne ball2 and ball5 ne ball1);
      ball5 = RAND("normal")*1000000000000;
      ball5 = ROUND(ball5);
      ball5 = 1+(mod(ball5,49));
      ball5 = ABS(ball5);
   if ball5 = 0 then ball5 = 1;
if ball5 >0 and ball5<8 then ball5grp=1;
if ball5 >7 and ball5<15 then ball5grp=2;
```

```
if ball5 >14 and ball5<22 then ball5grp=3;
if ball5 >21 and ball5<29 then ball5grp=4;
if ball5 >28 and ball5<36 then ball5grp=5;
if ball5 >35 and ball5<43 then ball5grp=6;
if ball5 >42 and ball5<50 then ball5grp=7;
end; call streaminit(68);
do until (ball6 ne ball5 and ball6 ne ball4 and ball6 ne ball3 and ball6 ne ball2 and ball6 ne ball1);
      ball6 = RAND("normal")*1000000000000;
      ball6 = ROUND(ball6);
      ball6 = 1+(mod(ball6,49));
      ball6 = ABS(ball6);
   if ball6 = 0 then ball6 = 1;
if ball6 >0 and ball6<8 then ball6grp=1;
if ball6 >7 and ball6<15 then ball6grp=2;
if ball6 >14 and ball6<22 then ball6grp=3;
if ball6 >21 and ball6<29 then ball6grp=4;
if ball6 >28 and ball6<36 then ball6grp=5;
if ball6 >35 and ball6<43 then ball6grp=6;
if ball6 >42 and ball6<50 then ball6grp=7;
end;output; end;
run;
proc freq; tables ball1grp ball2grp
ball3grp ball4grp ball5grp ball6grp;
FORMAT ball1grp – ball6grp GRPFMT. ;run;
/* CALCULATE CHI SQUARE GOODNESS OF FIT
PROC FREQ;
TABLES ball1grp/CHISQ;
FORMAT ball1grp GRPFMT. ;
TITLE 'CALCULATING THE GOODNESS OF FIT FOR ball1grp';
RUN; */
/* Define the axis characteristics */
  axis1 offset=(0,50) minor=none;
  pattern1 value=solid color=cx7c95ca;
proc sort; by ball1;
proc gchart ;
    Hbar ball1grp /  TYPE=PERCENT
    discrete ;
    FORMAT ball1grp  GRPFMT. ;run;
/* Define the title */
TITLE 'FREQUENCY DISTRIBUTION FOR OUTCOME GROUPS FOR BALL1';
  run;
```

A sample of the output from this procedure is shown below:

The HORIZONTAL BARCHART WITH FREQ TABLE



	FREQ.	CUM. FREQ.	PCT.	CUM. PCT.
NUMBERS 1 TO 7	14	14	13.46	13.46
NUMBERS 8 TO 14	15	29	14.42	27.88
NUMBERS 15 TO 21	12	41	11.54	39.42
NUMBERS 22 TO 28	16	57	15.38	54.81
NUMBERS 29 TO 35	19	76	18.27	73.08
NUMBERS 36 TO 42	15	91	14.42	87.50
NUMBERS 43 TO 49	13	104	12.50	100.00

23. Computing the Sign Test

A good place to begin the application of non-parametric statistical applications is with the sign test. Just as the name implies, the sign test is a non-parametric statistical procedure that evaluates the number of differences between paired comparisons using + and – signs to represent the direction of the differences between the pairs.

We can think of the sign test as a statistical method that compares pairs of outcomes and uses the (+) sign to describe the agreement and the (-) sign for the disagreement.

Consider that we wish to compare the number of symptoms reported by concussed patients in two different groups. The first group will be comprised of individuals who are asked to sit quietly in their room for seven days following their concussion injury. The second group will be comprised of individuals that are asked to participate in seven days of controlled and monitored exercise beginning 24 hours after the individual is asymptomatic.

Given this scenario, the sign test can be used to test the null hypothesis in the two-group comparison, where the null hypothesis is written as:

The chance for the number of symptoms in group 1 (the sedentary group) to be greater than the number of symptoms in group 2 (the exercise group) is equal to the chance for the number of symptoms in group 1 (the sedentary group) to be less than the number of symptoms in group 2 (the exercise group).

The numeric value associated with the aforementioned chance is equal to 0.5 or one half and is written as:

The **probability** of (the number of symptoms in group 1) **being greater than** the **probability** of (the number of symptoms in group 2) **is equal to** the **probability** of (the number of symptoms in group 1) **being less than** the **probability** of (the number of symptoms in group 2) **is equal to one half**

$$p(scoregroup1 > scoregroup2) = p(scoregroup1 < scoregroup2) = ½$$

We can simplify the computations of the sign test by evaluating the outcomes using the binomial equation. The binomial equation is a formula that computes exact probabilities for an outcome with two possible options.

In this first example, when comparing symptoms between a patient in group 1 and a patient in group 2, the two possible outcomes are either to have more symptoms or less symptoms.

In the **SIGN TEST** the outcome options are represented with a (+) or a (-) sign, and as such, the binomial equation can be used to estimate exact probabilities for any pairwise comparison.

$$P_x = \frac{n!}{x!\,(n-x)!}\, p^x\ q^{n-x}$$

NOTE: The following is a quick overview of the elements of the binomial equation used to establish probability.

P_x refers to the probability of exactly x events appearing in n trials. For example – the probability of turning up three heads in 5 tosses of a fair coin.

p^x refers to the expected probability of the event associated with the x term on any given trial (if we are flipping a coin then this value is ½ with a fair coin).

q^{n-x} refers to the probability of an event on any given trial $q = 1- p$ (usually this value is ½ if we were flipping a coin).

n refers to the number of events.

x refers to the number of a given outcome being evaluated (e.g. how many heads were you expecting in the total number of tosses).

In the application of the sign test, the binomial equation can be used to determine if the number of (+) signs occurs more or less often than the number of (-) signs.

Application 1:

Consider that you are responsible to test the efficacy of exercise treatment for concussion recovery at your rehabilitation clinic. You decide for the next 20 patients that arrive following a concussion, alternate assignment of the patient to either the sedentary group or the activity group. After 7 days of treatment within their respective groups, you survey the patient to determine if the number of symptoms reported by the group assigned to sedentary therapy is higher (or lower) than the number of symptoms reported by the group assigned to the physical activity intervention.

The data are presented in the table below, and the sign test is used to determine if the p(symptom reports in group 1 > symptom reports in group 2) = p(symptom reports in group 1 < symptom reports in group 2) = ½.

In the table, the number of symptoms reported by each pair of members from either the sedentary group or the activity group are compared, and the difference is recorded as being either greater than or less than.

Table of Matched group data example of symptom reporting in exercising versus sedentary patients recovering from concussion.

Number of Symptoms (Exercise Group)	Number of Symptoms (Sedentary Group)	Comparing Exercise to Sedentary patients	Sign (+) of difference between symptom reports
16	19	<	+
14	18	<	+
18	21	<	+
10	09	>	–
14	21	<	+
16	20	<	+
10	08	>	–
12	18	<	+
08	18	<	+
10	20	<	+
16	19	<	+
14	18	<	+
18	21	<	+
10	19	<	+
14	21	<	+
16	20	<	+
10	08	>	–
12	18	<	+
08	18	<	+
10	20	<	+

Note: The sign is recorded as (+) when the sedentary group member has more symptoms reported than the matched exercise group member.

These data can be compared using the binomial formula, shown here for a sample of n=20 pairs of patients. In this equation x refers to the number of patients from the sedentary group reporting more symptoms than the patients in the exercise group (x = 17), and the expected probability, based on the null hypothesis where we expected that the number of pairs of patients reporting more or fewer symptoms would be equal is p=0.5.

In the sign test, the test statistic is the number of (+) signs.

$$P_x = \frac{n!}{x!\,(n-x)!}\, p^x\ q^{n-x}$$

$$P_x = \frac{20!}{17!\,(20-17)!}\, 0.5^{17}\ 0.5^3$$

$$P_x = \frac{20 \times 19 \times 18 \times 17 \times \ldots \times 4 \times 3 \times 2 \times 1}{17 \times 16 \times 15 \times \ldots \times 4 \times 3 \times 2 \times 1(3 \times 2 \times 1)}\ 0.000008 \times 0.125$$

$$P_x = \frac{6840}{6} \times 0.000001$$

$$P_x = 0.001$$

By convention, we typically accept the null hypothesis when the resulting P_x is greater than 0.05. In this example where we compared the number of patients from the sedentary group reporting more symptoms than the patients in the exercise group, the probability is 0.001 suggesting that the null hypothesis is false and therefore can be rejected.

In SAS, we can evaluate the null hypothesis for this example using the following program. The problem with the application of the PROC UNIVARIATE procedure here is that the sample size is too small to meet the assumptions of normality and therefore computing the significance of the difference requires that we consider an alternative to the parametric tests. Hence the comparison is an excellent opportunity to consider a non-parametric test like the SIGN-TEST.

SAS program to compute the SIGN TEST using PROC UNIVARIATE

```
   DATA SIGN;
INPUT @1 PAIRNUM EXGRP SEDGRP SIGNPOS;
PAIRDIFF=(SEDGRP-EXGRP);
LABEL EXGRP = "NUMBER OF SYMPTOMS REPORTED BY EXERCISE PATIENT"
SEDGRP = "NUMBER OF SYMPTOMS REPORTED BY SEDENTARY PATIENT"
PAIRDIFF = "DIFFERENCE BETWEEN"
SIGNPOS = "SIGN IS POSITIVE";
DATALINES;
01 16 19 1
02 14 18 1
03 18 21 1
04 10 09 0
05 14 21 1
06 16 20 1
07 10 08 0
08 12 18 1
```

```
09 08 18 1
10 10 20 1
11 16 19 1
12 14 18 1
13 18 21 1
14 10 19 1
15 14 21 1
16 16 20 1
17 10 08 0
18 12 18 1
19 08 18 1
20 10 20 1
;
PROC UNIVARIATE FREQ; VAR PAIRDIFF;
TITLE “COMPUTING THE SIGN TEST FOR PAIRDIFF”;
RUN;
```

The output table is shown below. The value produced for the sign test shown here is an estimate for a two-tailed test and therefore, the p-value should be divided by 2. Notice that the SAS table produced a p-value of **0.0026** for the two-tailed test which we convert to p=0.001 for a one-tailed test.

Tests for Location: Mu0=0

Test	Statistic		p Value	
Student’s t	t	5.710367	Pr > |t|	<.0001
Sign	**M**	**7**	**Pr >= |M|**	**0.0026**
Signed Rank	S	99	Pr >= |S|	<.0001

The information that is relevant from the SAS output produced above is the SIGN Test result Pr >=|M| (0.0026/2)=0.001. This result indicates that there is a significant difference in the responses of the two groups as determined by the SIGN test.

Application 2:

Consider that you have two samples of 10 individuals in each sample. You identify a group of 10 smokers and a matched group of 10 non-smokers – and you are interested in determining if the outcome on the dependent measure (resting heart rate) is the same in both groups. You begin by matching the participants in your groups on the variables: age, sex, socio-economic status, education, and smoking status where you separate the groups according to: never smoked versus smoke at least 3 cigarettes per day. Then you measure the resting heart rate for individuals after asking them to sit quietly for 5 minutes (without smoking!!).

Table of matched group data example of smokers versus non-smokers with the outcome heart rate (bpm)

Non-smoker heart rate (bpm)	Smoker heart rate (bpm)	Compare Non-Smoker to Smoker	Sign (+) when NShr < Shr
76	**99**	**<**	**+**
54	**78**	**<**	**+**
68	**68**	**tied**	**tied**
60	**54**	**>**	**–**
54	**82**	**<**	**+**
86	**92**	**<**	**+**
90	**80**	**>**	**–**
62	**78**	**<**	**+**
58	**88**	**<**	**+**
60	**90**	**<**	**+**

Notice in the table above there are 7 occurrences where the heart rate measures for non-smokers showed fewer beats per minute than the smokers. Notice also that there was one occurrence in which the heart rate for the smoker was the same as the heart rate for the non-smoker. In the SIGN test, the number of **ties is not included** in the binomial computation.

Applying the binomial formula for computations in the Sign-Test

In our sample data to compare heart rates in smokers versus non-smokers for n=10, x=7, p=0.5 and q=0.5 we begin by

$$P_x = \frac{n!}{x!\,(n-x)!}\,p^x\;q^{n-x}$$

applying the binomial formula.

To test the difference in heart rates we counted the number of times that non-smokers' heart rates were lower than smokers' heart rates. Our count was 7 out of 10, and the probability of observing 7 out 10 in any random selection of smokers and non-smokers, matched on age, sex, socio-economic status, education, and smoking group where the groups were separated according to: **never smoked** versus **smoke at least 3 cigarettes per day** was found to be **Px = 0.12.** The arithmetic is shown below:

$$P_x = \frac{10!}{7!\,(10-7)!}\,0.5^7\;0.5^{10-7}$$

$$P_x = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1\,(3 \times 2 \times 1)}\,0.5^7\;0.5^{10-7}$$

$$P_x = \frac{10 \times 9 \times 8 \times \cancel{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}}{\cancel{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}(3 \times 2 \times 1)}\,0.008 \times 0.125$$

$$P_x = \frac{720}{6}\,0.001$$

$$P_x = 120 \times 0.001 \rightarrow P_x = 0.12$$

Recall that when evaluating the null hypothesis: H0: Group1 = Group2 we can use the standard normal distribution. We generally accept the null hypothesis when the probability associated with the difference between our estimates is less than 0.05.

Therefore, we conclude from these data that the two groups are equal for their resting heart rates.

However if the number of times we observed a non-smoker's heart rate to be less than a smoker's resting heart rate was **8 times** (or 8 out of the 10 participants) then the value of ***Px*** would have been ***Px* =0.043** and by convention we would have said that there was a difference in resting heart rates in this sample of smokers and non-smokers.

24. Computing the Wilcoxon-Mann-Whitney U Test

The Wilcoxon Mann-Whitney is a 2 group non-parametric comparison test equivalent to the Parametric t-test that can be used to test treatment effects when data are not normally distributed.

- The Mann-Whitney U test, which may also be referred to as the Wilcoxon-Mann-Whitney test, or the Wilcoxon Rank-Sum test, evaluates the ranks of the combined scores from two independent groups.
- The Wilcoxon rank-sum test statistic (referred to as **Ws** if using the name Wilcoxon rank-sum) is based on using the sum of the ranks for observations drawn from one of the groups within the sample of data.

Generally, the groups being studied are designated as GROUP 1 = treatment group and GROUP 2 = control group. This statistic – regardless of whether you refer to it as the Mann-Whitney test or the Wilcoxon rank-sum test, is considered to be among the more powerful of the non-parametric statistical procedures; and when using large samples, the computational result of this test is generally the same as the parametric t-test for two independent groups.

When evaluating the outcome of this statistic, we can test the Wilcoxon rank-sum test statistic against a critical value from a table of standard values, or we can compute a z-score for the comparison of ranks.

In the Mann-Whitney U– Wilcoxon rank-sum test we compute a "z score" (and the corresponding probability of the "z score") for the sum of the ranks within either the treatment or the control group. The "U" value in this z formula is the sum of the ranks of the "group of interest" – typically the "treatment group".

Essential Formulae

Below is the formula to compute z score for the Wilcoxon-Mann-Whitney test:

$$z = (U_1 + 0.5) - (\frac{U_1 + U_2}{2}) / \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

The formula to compute the probability of arriving at the z that you computed under the standard normal distribution (SND) is shown in this next formula. We use this probability value to evaluate the outcome of the Mann-Whitney test. In this formula, replace the x term with z from the formula above. The value for π is 3.14, the value for σ^2 is 1, the value for μ is 0.

$$prob\ under\ the\ SND = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

A working example:

You conducted a study to determine if a new treatment procedure was better than the standard method. 30 participants were recruited from a population of students and randomly allocated to either the new treatment procedure (T) or the standard method (S) so that the initial distribution was set at (n_1 = 15 and n_2 = 15).

After applying the treatments to each group respectively the students were ranked on a specific measure that demonstrates the influence of the two treatment methods. The ranks for each response score are given in the following table while maintaining the student's group membership.

NT= New treatment; ST = Standard Treatment

Row 1: Participant ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Row 2: Dependent Variable Scores	8	12	13	15	19	21	22	28	31	36	37	39	40	41	43
Row 3: Group codes	NT	NT	NT	ST	ST	NT	NT	ST	ST	ST	NT	NT	NT	NT	NT
Row 4: Rank of score	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Row 1: Participant ID	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Row 2: Dependent Variable Scores	48	52	53	55	59	61	62	68	71	76	77	79	80	81	83
Row 3: Group codes	NT	NT	NT	ST	ST	ST	ST	ST	ST	ST	ST	ST	ST	NT	NT
Row 4: Rank of score	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

Values in Row 2 in the table above represent the score on the dependent variable measuring the response to the two treatment types.

Values in Row 3 in the table above represent the codes for the group membership, where T=new treatment method group and S=standard method group.

Values in Row 3 in the table above represent the Rn= rank of participants within the total data set.

Ranks begin from the lowest score to the highest score.

The sum of the ranks (U_1) in the NEW Treatment group are: (1+2+3+6+7+11+12+13+14+15+16+17+18+29+30) = 194

The sum of the ranks (U_2) in the STANDARD Treatment group are: (4+5+8+9+10+19+20+21+22+23+24+25+26+27+28) = 271

$$z = (U_1) - (\frac{U_1 + U_2}{2}) / \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

$$z = (194.00 + 0.5) - (\frac{194 + 271}{2}) / \sqrt{\frac{15 \times 15(15 + 15 + 1)}{12}}$$

$$z = \frac{(194.5) - (232.5)}{\sqrt{581.25}}$$

$$\mathbf{z = \frac{-38}{24.11}} \qquad \mathbf{z = -1.576}$$

What about ties?

In the case of a tie, we simply organize all of the data as in the table above, and then we assign each observation in a tie its average rank. So if we had two scores 12 and 12 and they had a rank of 3 and 4 then we would simply give the first value of 12 a rank of 3.5 and the second value of 12 a rank of 3.5.

Verifying the Computations with SAS

```
DATA MWW;
INPUT ID GROUP SCORE @@;
CARDS;
01 1 8 02 1 12 03 1 13 04 2 15 05 2 19 06 1 21 07 1 22 08 2 28 09 2 31
10 2 36 11 1 37 12 1 39 13 1 40 14 1 41 15 1 43 16 1 48 17 1 52 18 1 53
19 2 55 20 2 59 21 2 61 22 2 62 23 2 68 24 2 71 25 2 76 26 2 77 27 2 79
28 2 80 29 1 81 30 1 83
;
PROC PRINT; VAR ID GROUP SCORE;
PROC NPAR1WAY DATA=MWW WILCOXON;
CLASS GROUP; VAR SCORE; EXACT;
RUN;
```

The NPAR1WAY PROCEDURE OUTPUT

Wilcoxon Scores (Rank Sums) for Variable score: Classified by Variable group

GROUP	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
New Treatment	15	194.0	232.50	24.109127	12.933333
Standard Treatment	15	271.0	232.50	24.109127	18.066667

Wilcoxon Two-Sample Test: Z includes a continuity correction of 0.5 –> Statistic (S) = 194.00

Normal Approximation: Z = -1.5762; One-Sided Pr < Z = 0.0575; Two-Sided Pr > |Z| = 0.1150. EXACT TEST: One-Sided Pr <= S = 0.0580; Two-Sided Pr >= |S – Mean| = 0.1160.

Your Turn

Compute the Sign Test and the Mann-Whitney Test

You are interested in the effects of daily exercise on fitness levels.

You create an experiment in which individuals are allocated to either a twelve-week exercise program or a sedentary control group. You were successful in recruiting 30 subjects, and you matched these individuals on gender and exercise profiles to balance the groups that will either participate or remain sedentary.

You arrange the participants into two groups (n1=15, and n2=15). Group 1 receives a 12-week regimen of noon-hour exercises while Group 2 is considered the control group and does not receive any exercise programming or any related information. Both groups maintain very nearly similar profiles for sleep and diet. The probability of being selected as an exerciser is 50% or "p= ½", and conversely the probability of being selected to the control group is 50% or "p= ½"

The dependent variable for the experiment is the measure of the individual's predicted VO2 max as determined by a sub-maximal walking test. Given this scenario and research design, you decide to use the "SIGN TEST" as the statistical procedure to determine if the exercise regimen caused significant changes in the fitness levels of the participants compared to the control group members. The data for this experiment are presented in DATA SET #1 below. COMPLETE THE TABLE, and compute the significance of the sign test.

Data Set #1 – Computing the significance of the Z statistic in the "Sign test"

VO2 Grp1 (ml/ kg·min^{-1})	VO2 Grp2 (ml/ kg·min^{-1})	Comparison of VO2 max test scores between the two groups	Sign of difference
43	40	Subject1Group1 > $Subject_1Group_2$	+
48	42	Subject2Group1 $Subject_2Group_2$	
39.4	43.8	Subject3Group1 < $Subject_3Group_2$	–
32.7	31.9	Subject4Group1 > $Subject_4Group_2$	
36.9	48.4	Subject5Group1 $Subject_5Group_2$	
50.2	41.4	Subject6Group1 $Subject_6Group_2$	+
39.9	31.9	Subject7Group1 $Subject_7Group_2$	
45.3	33.2	Subject8Group1 $Subject_8Group_2$	
40.8	39.6	Subject9Group1 $Subject_9Group_2$	
39.8	41.2	Subject10Group1 $Subject_{10}Group_2$	
45.4	43.5	Subject11Group1 > $Subject_{11}Group_2$	
57.3	52.5	Subject12Group1 $Subject_{12}Group_2$	+
58.7	40.6	Subject13Group1 $Subject_{13}Group_2$	
35.4	39.5	Subject14Group1 $Subject_{14}Group_2$	
58.4	49.5	Subject15Group1 > $Subject_{15}Group_2$	+

Null hypothesis	*Number of (+) or (-) SIGNS*	*Z sign test*	*The decision concerning the null hypothesis*

Recall that in our study we had N = 30 where we created two matched groups of 15 subjects per group. Treatment group is GROUP 1 and Control group is GROUP 2.

As a follow-up to the study, you decided to compare average resting heart rate responses for the group of individuals who participated in the "lunch-hour exercise group", against the sedentary control group, over the twelve-week timeline.

The data for this study are presented in Data Set #2 below. Given the arrangement of data, recall that you are attempting to measure if the heart rates for the "lunch-hour exercise group" are generally higher or lower than the heart rates for the sedentary control group. Since you expect that twelve weeks of exercise at lunch hour should have a positive effect on the cardiovascular system, you also expect that the resting heart rates for the "lunch-hour exercise group" would be generally lower than the heart rates for the sedentary control group.

Use the Mann-Whitney test to compute a "z score" for the sum of the ranks within either the treatment or the control group. Include the null hypothesis and your decision about the null hypothesis based on the computations.

Data Set #2 – Computing the Mann Whitney Statistic

$n_{1\ (the\ exercise\ group)}$ = 11 (use **EG** as the exercising group code).

EG = 89, 95, 103, 105, 109, 113, 114, 115, 117, 123, 128

$n_{2\ (the\ sedentary\ control\ group)}$ = 9 (use **CG** as the control group code)

CG = 100, 101, 107, 119, 126, 134, 135, 136, 139

ROW 1: Scores arranged from lowest to highest

ROW 2: Group membership for scores (E= experimental, C= control)

ROW 3: Rank position for scores (beginning lowest score to highest)

ROW 1	89			103		113		126		139
ROW 2	E	E	C	E		E		C	C	C
ROW 3	1	2	4	5		9		15	19	20

Null hypothesis	U1	Sum of ranks in the control group	Z Mann-Whitney	The decision concerning the null hypothesis

25. Computing the Z Statistic for the One Sample Runs Test

Often in quantitative methods, we expect that any score we observe occurs at random and is not a result of selection bias. This expectation is of particular importance when we are dealing with strings of binary events, such as viewing the change in a particular measure over time or counting the sequence of similar outcomes without a break.

Wald and Wolfowitz referred to such strings or sequences of similar events as ***runs***. A run is defined as a sequence of similar data values. A run of an event occurs when a particular outcome of interest is observed within a sampling space. A run can have a sequence of 1 or a run can have a sequence of > 1.

For example, consider the toss of a fair coin. If we toss the coin 20 times we could expect to observe the following extreme outcomes:

1. H,T,H,T,H,T,H,T,H,T,H,T,H,T,H,T,H,T,H,T
2. H,H,H,H,H,H,H,H,H,H,T,T,T,T,T,T,T,T,T,T

In the first sample space above (1), the outcome was a complete interspersing of each toss of H followed by a T (or T followed by H). In the second sample space above (2) we observe the complete clumping of ten heads followed by ten tails, both of these events can be considered random, but they represent the extremes of what we might observe.

Consider that the purpose of the runs test is to determine, within a string of events, the randomness of fluctuations. That is, do the observed fluctuations (if any) occur at random or do the fluctuations of observations exhibit some form of clumping together? Does the sequence observed within a sampling space represent a pattern over a given sampling space or period of time.

The formula for the runs z-statistic is shown here:

$$z = \frac{r - \left(\frac{2n_1n_2}{n_1 + n_2} + 1\right)}{\sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1 + n_2 - 1)}}}$$

There are three parts to the z formula for the One Sample Runs Test.

1) The first part is to count the number of runs of a given type of events. For example, in the coin toss example, there were 20 tosses of a fair coin, which resulted in **10 runs** as shown here:

H, T, T, T, H, T, H, H, T, H, H, H, T, T, H, H, H, H, T, T

2) The second part is to compute the mean number of expected runs using the formula:

$$\mu_r = \frac{2n_1n_2}{n_1 + n_2} + 1$$

In this scenario, there were 21 reported outcomes whereby we consider the number of heads were counted as n_1 and the number of tails were counted as n_2, so that n_1=11 and n_2=9.

$$\mu_r = \frac{2(11 \times 9)}{11 + 9} + 1 = \frac{198}{20} + 1 = 10.9$$

3) The third part of the calculation is to compute the standard deviation of the estimate of runs using the formula:

$$\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1 + n_2 - 1)}}$$

$$\sigma_r = \sqrt{\frac{2(11 \times 9)(2(11 \times 9) - 11 - 9)}{(11 + 9)^2\,(11 + 9 - 1)}}$$

$$\sigma_r = \sqrt{\frac{198 \times 178}{400\,(19)}}$$

$$\sigma_r = \sqrt{\frac{35244}{7600}} = \sqrt{4.64} = 2.15$$

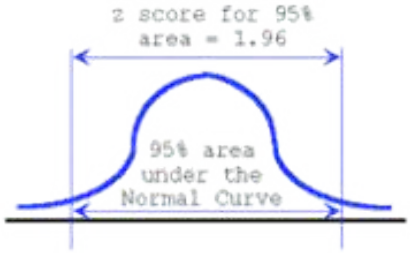
$$\mathbb{Z}_r = \frac{N_{runs} - \mu_r}{\sigma_r}$$

$$\mathbb{Z}_r = \frac{10 - 10.9}{2.15} = 0.42$$

4) the calculation of z for the runs test is then simplified to:

The evaluation of runs of events is a z test, which means that the evaluation of the null hypothesis associated with this test is based on a normal (z) distribution.

The value of z = 0.42 is within the region of acceptance of the null hypothesis, as shown with this graph. The null hypothesis: is accepted if the z observed > -1.96 and <1.96.



Therefore, we accept the null hypothesis that there is no pattern or sequence to **THIS** toss of a fair coin.

Your Turn: Compute the One Sample Runs Test

Consider the runs of increases and decreases in the daily weather pattern for one month in the seaside Village of Cavendish, Prince Edward Island. A run is defined as a sequence of similar data values. The sequence can be a single entry, or a string of entries occupying the entire set of observations. Since you are a golfer, who likes to play when the weather is hot, you hope that there is only one run and that it is a positive increase to warmer weather each day.

Use the approach explained for the "one sample runs test" to compute the significance of the runs of temperatures in

the following example. In your response state the null hypothesis for this question, and in addition to the results of your computations, include a statement about your decision of whether to accept or reject the null hypothesis.

Data Set #3 One sample runs test data for the month of July

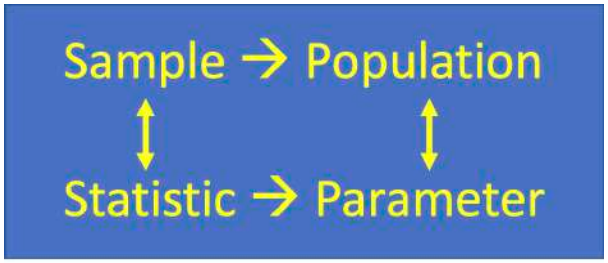
Date & Temperature	Change	Date & Temperature	Change
June 30th 20° C	·	July 16th 20° C	–
July 1st 21° C	+	July 17th 21° C	+
July 2nd 22° C	+	July 18th 22° C	+
July 3rd 22.5° C	+	July 19th 23° C	+
July 4th 23° C	+	July 20th 25° C	+
July 5th 24° C	+	July 21st 23° C	–
July 6th 25° C	+	July 22nd 23.5° C	+
July 7th 26° C	+	July 23rd 22° C	–
July 8th 24° C	–	July 24th 21° C	–
July 9th 21° C	–	July 25th 20° C	–
July 10th 19° C	–	July 26th 24° C	+
July 11th 18° C	–	July 27th 25° C	+
July 12th 21° C	+	July 28th 26° C	+
July 13th 22° C	+	July 29th 27° C	+
July 14th 22.5° C	+	July 30th 25° C	–
July 15th 21° C	–	July 31st 24° C	–

Null hypothesis	Average run mr	Standard deviation (sr)	Z runs test	Decision concerning the null hypothesis

PART V

PARAMETRIC STATISTICS

Parameters, statistics, populations, and samples



We read the illustration above as: a sample is to a population as a statistic is to a parameter.

Here we demonstrate the transitive relationship between samples and populations using parameters and statistics. That is, when we are working within a parametric statistical paradigm we apply inference to establish the relationship between that which we calculate for a sample and that which we intend to represent for a population.

We never know a parameter, because we never measure the population, but we estimate the parameter with the statistics that we compute. Further, we can never measure the entire population rather, we collect a sample, which we deem to be representative of the population and then calculate statistics that represent the parameter of interest within the population.

For example, let's say we wish to calculate the average age of all residents in Long Term Care Homes in the Province of Prince Edward Island. In such an example, the group of **all residents** will be considered the population. However, knowing that we don't have the capability (time or permission) to visit every resident in the Long Term Care homes in PEI, instead, we randomly choose a few residences and arrange to visit these specific locations so that we can ask the residents their age.

Moving along in this quest we realize that not every resident is interested in our exercise and so has no intention to tell us their age. At the end of the day, the ages that we actually recorded represent our sample of data, and we assume that, given our best attempts to record the ages of all residents in the Long Term Care Homes, our final data set is a representative sample of the population of all residents in Long Term Care Homes in the Province of Prince Edward Island.

Therefore, the average age that we estimate for this sample is referred to as a statistic and is presented with the symbol [latex]\bar{x}[/latex]. Further, given that we are working within a parametric statistical paradigm we infer that the calculated statistic is a representative estimate of the parameter for the average age of the population which is represented by the symbol **[latex]{\mu}[/latex].**

In the following sections, we will explore parametric applications in the context of applied statistics in healthcare.

26. Measures of Central Tendency

PART 1: Measures of Central Tendency

The most common measure of central tendency is the **mean** or **average** score. The mean is a calculated score that is intended to represent all of the scores in the distribution (set of scores).

The formula for the mean of a sample is shown here:

[latex]{\overline{x}} = \Sigma{(x_i)\over{n}}[/latex]

Where:

- [latex]{\overline{x}}[/latex] refers to the sample mean
- [latex]\Sigma{(x_i)} efers to the sum of all the scores
- i refers to the "ith" case within the distribution
- n refers to all of the cases within the distribution.

To calculate the mean for a continuous variable, add up all of the values and divide the sum of values by the number of values. Below is a set of blood glucose measures for 5 patients. These data are represented in millimoles per litre (mmol/L). P_n represents the nominal value label for each patient, so that P1 is patient 1. P_1 4.2 mmol/L, P_2 5.6 mmol/L, P_3 7.9 mmol/L, P_4 10.2 mmol/L, P_5 7.5 mmol/L, Follow these steps to calculate the mean:

- First add the values together: 4.2 + 5.6 + 7.9 + 10.2 + 7.5 = 35.4.
- Next, divide by the number of values (to produce the average): 35.4/5 = 7.08 mmol/L

We can also use SAS to compute the mean for a set of scores. Two specific SAS programs that process measures of central tendency are PROC MEANS, and PROC UNIVARIATE. Each of these programs was designed to produce descriptive statistics for a sample of scores. Below are the SAS commands to compute the mean for a set of 10 resting heart rate scores. In this first program we used the SAS procedural command PROC MEANS to compute three basic estimates: the mean, the standard deviation and the minimum/maximum scores for the sample dataset of 10 numbers.

SAS PROC MEANS to Produce Descriptive Statistics for a Sample of 10 Numbers

```
DATA MN_HR; INPUT ID SCORE @@; DATALINES; 01 48 02 54 03 66 04 72 05 56 06 68 07 48 08 67 09 55 10 84
; PROC MEANS DATA=MN_HR; VAR SCORE; RUN;
```

Notice in the code written above, the semi-colon (;) is placed on a separate line below the set of scores. While PROC MEANS, in its simplest form (without options) provides three basic estimates that describe estimates within a distribution, the SAS procedural command PROC UNIVARIATE not only computes the mean but also creates the Basic Statistical Measures Table which provides an entire summary of descriptive statistics. The output generated by the SAS program above – using the PROC MEANS statement without options – produced a table of summary estimates that included the mean and standard deviation as well as the minimum and maximum values for the dataset. **SAS Output from the MEANS Procedure: Variable of interest was Heart Rate**

N	Mean	Std Dev	Minimum	Maximum
10	61.80	11.56	48.00	84.00

When we call the PROC UNIVARIATE procedure of SAS, the output is a more complete table of summaries that include estimates of centrality but also the moments, measures of variance, and the tests of the location of the mean, as shown below.

SAS PROC UNIVARIATE to Produce Descriptive Statistics for a Sample of 10 Numbers

```
PROC UNIVARIATE DATA=MN_HR; VAR SCORE; RUN;
```

The UNIVARIATE Procedure -- Variable: SCORE

MOMENTS			
N	10	**Sum Weights**	10
Mean	61.8	**Sum Observations**	618
Std Deviation	11.5547008	**Variance**	133.511111
Skewness	0.55954538	**Kurtosis**	-0.2284272
Uncorrected SS	39394	**Corrected SS**	1201.6
Coeff Variation	18.6969269	**Std Error Mean**	3.65391723

Tests for Location: Mu0=0				
Test	**STATISTIC**	**ESTIMATE**		**p Value**
Student's t	**t**	16.91336	**Pr > |t|**	.0001
Sign	**M**	5	**Pr >= |M|**	0.0020
Signed Rank	**S**	27.5	**Pr >= |S|**	0.0020

Comparing the Mean for a Sample to the Expected Mean for a Population

In the output from the PROC UNIVARIATE procedure, SAS includes a table in which the mean for the variable: SCORE is compared to the mean for the Standard Normal Distribution (SND). The SND represents the hypothetical population mean and has a value of 0 with a standard deviation of 1. In the SAS table shown above, entitled **Tests for Location: Mu0=0** the comparison of the sample mean ([latex]{\overline{x}}[/latex]) to the population ([latex]{\mu}[/latex]) is evaluated with the Student's t-Test.

The results presented in the table above show that the Student's t-Statistic value is 16.91 and the probability associated with this estimate is <0.001. Together these values indicate that the observed sample mean is significantly different than the hypothesized expected mean for the population (set at Mu_0=0) from which the sample was drawn.

However, what if we wanted to establish a suggested value for the population mean that is not 0, but that is based on value reported in the literature? In this case, we could assign a suggested value to the population mean and then

compare the observed mean for the sample to the expected value for a population. In the following code, we test this notion.

Assign a suggested value to the population mean

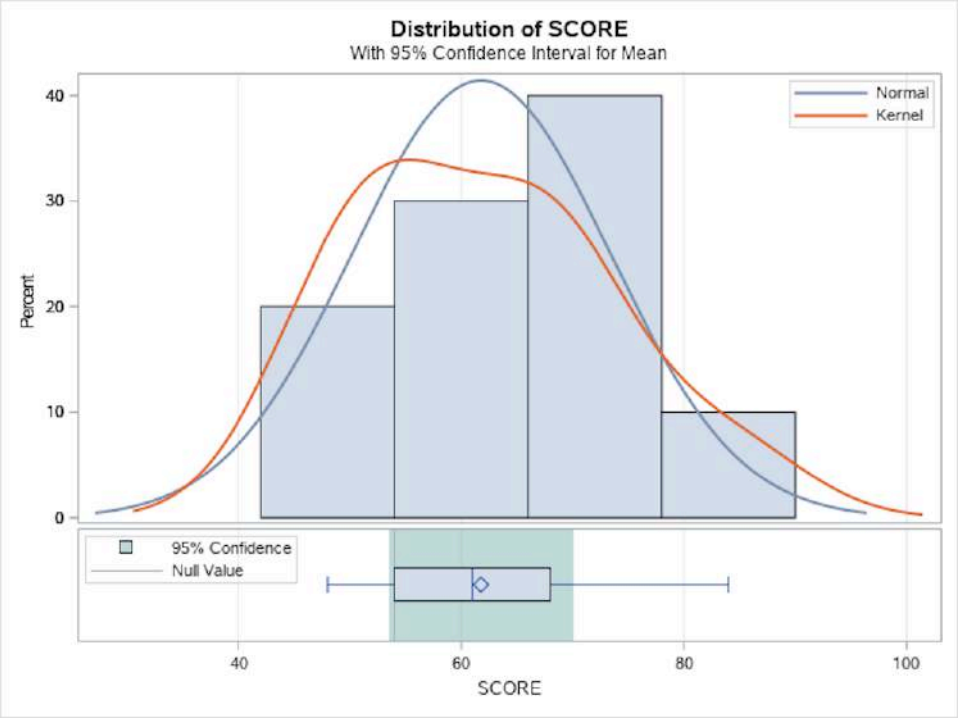
```
   PROC TTEST H0=54
PLOTS(SHOWH0)
ALPHA=0.05;
VAR SCORE;
RUN;
```

The SAS output is given below. The results indicate that the average score for the sample ([latex]{\overline{x}}[/latex] = 61.80) is not significantly different at the probability level of p < 0.05 than the expected score of ([latex]{\mu}[/latex] =54). Notice, in addition to the table of output SAS also includes a graph illustrating the shape of the distribution and the comparison of the sample estimate to the expected population estimate of centrality.

The t-test Procedure		
DF	**t Value**	**Pr > |t|**
9	2.13	0.0615

Parameter estimates		
Mean	**95% CL Mean**	
61.8000	Lower limit: 53.5343	Upper Limit: 70.0657

Considering that the confidence interval shown here includes the mean for the sample (61.8) and the mean for the population which we set apriori as 54, no significant difference is observed, between that which is expected and that which was observed. This estimate is illustrated in the following graph.



Calculate the Mean for A Frequency Distribution

In the following example, we compute the mean for frequency distribution. The formula to compute the mean of a frequency distribution is shown here as:

[latex]{\overline{x}} = {\Sigma{fx_i}\over{n}}[/latex]

Where:

- f refers to the frequency in each interval
- xi refers to the mid-point of the interval
- i refers to the “ith” case within the distribution
- n refers to all of the cases within the distribution.

Below is the frequency distribution table for the heights of 200 individuals. The data represent heights recorded in centimetres and organized into seven categories. The SAS code to compute the mean for this set of data is shown below the table. Notice that the table is reduced to a simple composition of two variables which includes the mid-point of the category represented by the variable: GRPMDPT, and the number of individuals, whose height scores fall within the specific category, represented by the variable: COUNTS.

Column 1 cell boundaries	Column 2 frequency (f)	Column 3 cell mid-point	Column 4 (f) x cell midpoint	Column 5 (col 4 : n)
158.5 – 161.5	4	160	4 x 160 = 640	640/200 = 3.2
161.5 – 164.5	12	163	12 x 163 = 1956	1956/200 = 9.78
164.5 – 167.5	44	166	44 x 166 = 7304	7304/200 = 36.52
167.5 – 170.5	64	169	64 x 169 = 10816	10816/200 = 54.08
170.5 – 173.5	56	172	56 x 172 = 9632	9632/200 = 48.16
173.5 – 176.5	16	175	16 x 175 = 2800	2800/200 = 14.00
176.5 – 179.5	4	178	4 x 178 = 712	712/200 = 3.56
	[latex]{\overline{x}} = {\Sigma{fx_i}\over{n}}[/latex]	[latex]{\overline{x}} = {33860\ over 200}[/latex]	= 169.3	The [latex]{\overline{x}}[/latex] is the sum of column 5

The SAS code to compute the mean for data in the table above

```
   DATA FREQMN;
INPUT GRPMDPT COUNTS @@;
CRSPRDCT= GRPMDPT*COUNTS;
/* COMPUTE RATIO FOR THE CROSS PRODUCT USING GROUP MIDPOINT X CELL FREQUENCY */
XP_RATIO=CRSPRDCT/200;
LABEL GRPMDPT = ‘GROUP MIDPOINT’
COUNTS = ‘NUMBER OF CASES PER CELL’
CRSPRDCT = ‘CROSS PRODUCT PER CELL’
XP_RATIO = 'CROSS PRODUCT RATIO';
DATALINES;
160 4 163 12 166 44 169 64 172 56 175 16 178 4
;
PROC PRINT;
VAR GRPMDPT COUNTS CRSPRDCT XP_RATIO;
SUM CRSPRDCT XP_RATIO;
FOOTNOTE1 "* THE MEAN IS PRODUCED AS THE SUM OF THE VARIABLE XP_RATIO";
FOOTNOTE2 "** THE MEAN CAN ALSO BE CALCULATED FROM THE SUM OF THE VARIABLE CRSPRDCT :
200";
RUN;
```

The output generated by the SAS program above is the table of raw data presented in column form and includes the sums of the columns used to compute the mean for the frequency distribution.

Obs	grpmdpt	counts	crsprdct	cp_ratio
1	160	4	640	3.20
2	163	12	1956	9.78
3	166	44	7304	36.52
4	169	64	10816	54.08
5	172	56	9632	48.16
6	175	16	2800	14.00
7	178	4	712	3.56
			33860	**169.30**

* The mean is produced as the sum of the variable **XP_RATIO**

** The mean can also be calculated from the sum of the variable ***crsprdct*** : 200

The Weighted Mean Score

In some situations, we may wish to combine means from several samples. Under such circumstances, we need to consider the sample size (or weight) of the distribution from which the means were drawn. By adjusting each independent sample mean by the number of subjects in the respective sample from which the means were drawn, we are able to provide different relative contributions of each mean to the total mean of all samples combined. The formula for a weighted mean from two samples is shown here. The formula for the mean of a sample is shown here:

[latex]{\overline{x}}={n_i\times{\overline{x_1}}+n_2{\overline{x_2}}\over{n_1 + n_2}}[/latex]

The Median Score

The median score is also a measure of central tendency, and it is defined as the middle score in a set of ordered scores. In the example below, we begin with a set of scores (an array), we next sort the scores from lowest to highest. Then we identify the number that is in the middle of the ordered set of scores where half the numbers are above the identified middle score, and half the numbers are below the identified middle score.

Example: Median

The *median* is the middle score. Considering the heart rate values again, we put these readings in order of magnitude and then identify which value is in the middle:

- 57
- 59
- 59
- 75
- 78
- 78
- 85
- 88

- 88
- 88

In this case, we have an even number of values (n = 10) so we can calculate the average of the two values in the middle. It just so happens that they are the same value in this example (78) so the median is 78.

- initial array of scores: {12, 72, 56, 34, 35, 13, 36, 16, 67}
- sorted array of scores: {12, 13, 16, 34, 35, 36, 56, 67, 72}
- sorted array of scores: {12, 13, 16, 34, 35, 36, 56, 67, 72}

Notice in the example above, regardless of the actual scores, the middle score in the ordered set of scores is the median, which in this set is 35.

When we have an even number of scores in our array there is a special caveat to identifying the median score in the distribution (set of scores). When we have two scores selected as the identified middle score we simply compute the average between the two identified middle scores and use that number as the median score. That is, we add the two middle scores together and divide by 2.

- initial array of scores: {22, 32, 86, 44, 25, 13, 16, 18, 47, 11}
- sorted array of scores: {11, 13, 16, 18, 22, 25, 32, 44, 47, 86}
- computed median for the array: {11, 13, 16, 18, 22, 23.5, 25, 32, 44, 47, 86}

The Mode Score

The mode score is the third measure of central tendency, and it is defined as the most frequently occurring score in a set of scores. In the example below, we simply count the number of scores that are the same within a set of scores, within an array or within a distribution.

Below are 10 resting heart rate values:

78, 88, 57, 59, 75, 85, 88, 78, 59, 88

The mode is 88 because it appears most often.

In the following example of 16 scores, the number 2 occurs 3 times, but the number 27 occurs 4 times therefore we would identify 27 as the mode score.

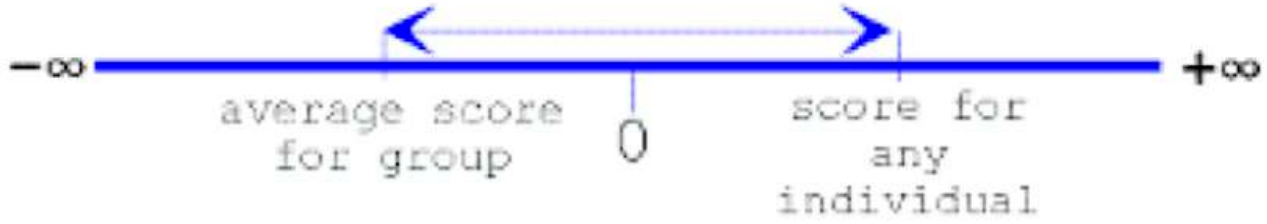
2, 2, 2, 5, 6, 14, 15, 23, 26, **27, 27, 27, 27,** 28, 37, 41

---

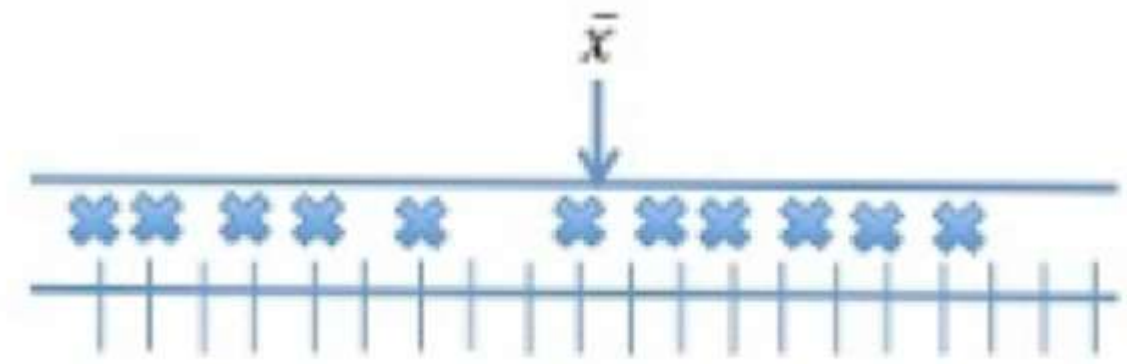
27. Measures of Variance

PART 2: Measures of Variance

In quantitative methods, we are often interested in the spatial relationship between responses within a set of numbers. We can describe this relationship as the dispersion of scores around the mean, or state that the dispersion of scores is represented by the variability or variance of the scores around the mean. For example, consider a straight line with boundaries at negative infinity and positive infinity. Notice that the straight line is continuous between the two boundary points, and within this space, we can measure the distance between an individual's score and the estimated average score of a group of scores.

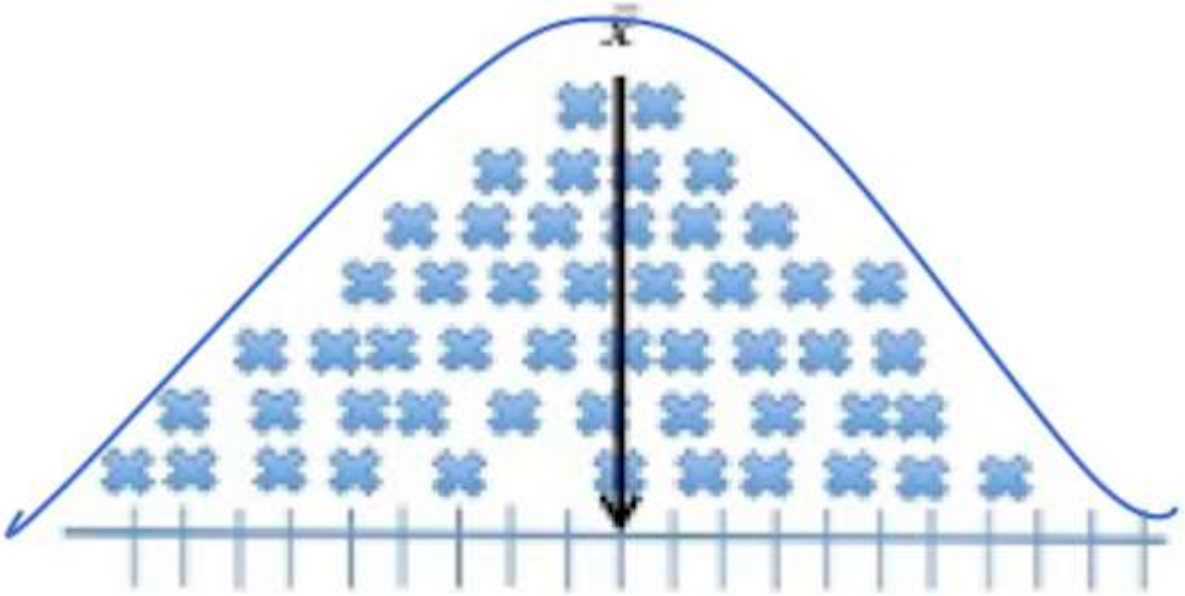


The differences between each observed score and the mean score within a set of scores are referred to as the variability of scores. In the following three images we can see the relationship between scores within a distribution and the mean of the distribution. In the first image, shown below, the scores – represented by X – are plotted on a scale from lowest to highest, and the algebraic midpoint is shown as the mean ([latex]\overline{x}[/latex]).

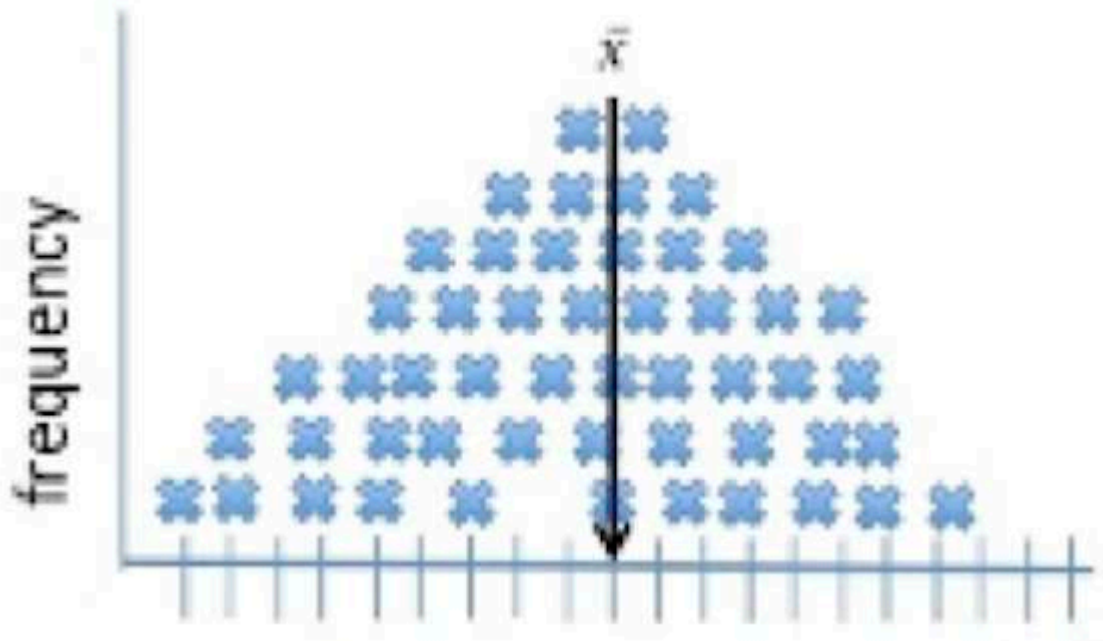


In the second image, we see the mean for the distribution is included in the image as a reference point to show the algebraic center of the distribution. In this second figure, we see the accumulation of all scores in the set of data. Notice how they create a shape for the distribution.

If we drew a smooth line over the top of this pile of scores it may indicate a bell shape.



In the third image, we see the scale of the frequency of the scores within the distribution is added to the left side of the image. The Y-axis is labeled frequency and can provide a count of scores at each of the values on the x-axis. The frequency is used to compute the proportion of scores within the entire set of scores.



Since we have a set of data, we can compute both the measure of centrality (the mean) as well as the average distance between the individual scores and the estimated mean. The average distance of scores from the mean is referred to as the variance.

Calculating Variance

Although we can conceptualize variance as an average score, this average score is not simply calculated by summing the difference scores and dividing by the number of scores. Rather, the variance score can be calculated for a "population" and for a "sample", where the terms in the denominators of the two calculations differ. The formulas to compute variance, first for a population and then for a sample are shown in the following two equations, below. Notice that the variance for the population is the average squared difference between the scores and the mean [latex]\mu[/latex]. However, in the calculation of variance for the sample, we subtract 1 in the denominator to enable us to produce an estimate which will be large enough to capture the true population estimate.

i) Population Variance: [latex]\sigma^2 = {{\sum(x_{i}- \mu)^2 \over {N}}}[/latex]

ii) Sample Variance: [latex]s^2 = {{\sum(x_{i}- \overline{x})^2 \over {n - 1}}}[/latex]

Standard Deviation

The term deviation refers to differences. When we consider deviation in a sample or a set of scores we are really interested in the dispersion of the scores around the mean. The term standard deviation refers to the "standardized estimate of variance". The standard deviation presents the variance in the original units of measurement.

The standard deviation is derived from the estimate of variance and is computed by calculating the square root of the variance as shown in equation iii) – standard deviation for a population and equation iv) – standard deviation for a sample, as shown below.

iii) Population Standard Deviation: [latex]\sigma^2 = {\sqrt{{\sum(x_{i}- \mu)^2 \over {N}}}}[/latex]

iv) Sample Standard Deviation: [latex]s^2 = {\sqrt{{\sum(x_{i}- \overline{x})^2 \over {n - 1}}}}[/latex]

In quantitative methods, we often consider the spreads of distributions and the shape of distributions. Estimates of difference between an individual's score and the measure of central tendency, the shape of the distribution, and the size of the distribution, are all elements of VARIANCE.

The Coefficient of Variation

We use the coefficient of variation to compare the dispersion of scores around the mean. The coefficient of variation is computed by dividing the standard deviation by the mean and multiplying by 100, as shown in equation v).

v) Coefficient of Variation: [latex]cv = {s \over {\overline{x}}} \times 100[/latex]

The coefficient of a variation is a useful measure to show the dispersion of a sample of scores around the mean. When used as a single measure it may not be as effective as when it is used as a comparative measure between two samples. The coefficient of variation is an effective measure to compare data from different samples. In the following example, we demonstrate the usefulness of the coefficient of variation by comparing the spread of scores around the mean in two samples measuring systolic blood pressure at rest. Here we recorded resting systolic blood pressure measures for a sample of males and a sample of females in a graduate course in biostatistics. The SAS program to compute the coefficient of variation for the comparison is shown in the example below.

SAS program to compute the coefficient of variation

```
   OPTIONS PAGESIZE=65 LINESIZE=80;
DATA SPREAD;
INPUT ID GRP SCORE @@;
DATALINES;
01 1 121 02 1 134 03 1 133
04 1 128 05 1 125
06 1 127 07 1 124 08 1 123
09 1 126 10 1 128
11 2 134 12 2 156 13 2 129
14 2 141 15 2 142 16 2 139 17 2 138
18 2 145 19 2 133 20 2 145
;
PROC SORT DATA=SPREAD; BY GRP;
PROC MEANS N MEAN STDDEV MEDIAN MODE CV;
VAR SCORE; BY GRP;
RUN;
```

Output from the SAS MEANS Procedure

Analysis Variable: SCORE for GRP=1					
N	**Mean**	**Std Dev**	**Median**	**Mode**	**Coeff of Variation**
10	126.90	4.12	126.50	128.00	3.25

Analysis Variable: SCORE for GRP=2					
N	**Mean**	**Std Dev**	**Median**	**Mode**	**Coeff of Variation**
10	140.20	7.61	140.00	145.00	5.43

Notice in the comparison of the scores in the two groups above, the mean for group 1 was 126.9 with a standard deviation of 4.12. In group 2, the mean score was 140.2 with a standard deviation of 7.61. These scores indicate that group 1 had less dispersion of scores around the mean (C.V. = 3.25), while group 2 had a much larger spread of scores around the mean (C.V. = 5.43). We can interpret these data to say that the data in group 2 was less homogenous than the data for group 1.

PART 3: Shapes of Distributions

Skewness

Skewness is an estimate of variability and as such, skewness is considered the "third moment" of scores within a distribution. This is easily recognized in the formula for skewness shown below.

Skewness: [latex]skew = {{\sum(x_{i}- \overline{x})^3 \over {N}}}[/latex]

The estimate of skewness is the ratio of the sum of the cubed differences between the observed scores and the mean

to the number of scores in the sample. Unlike variance, which squared the difference scores, skewness computes the cube of the difference scores. That is, skewness is computed by:

- subtracting each observation from the average score,
- raising the difference to the exponent "3",
- summing the cubes and then
- dividing by the number of observations (or the number of observations minus one).

A negative skewness score indicates a negatively skewed distribution, while a positive skewness score indicates a positively skewed distribution. Examples of skewed distributions are shown below.

Illustration of negative skewness shown here – notice the tailing off of the distribution toward negative infinity.

Illustration of a normal distribution (no skewness) – notice the symmetry of the distribution, there is no tailing-off toward either negative or positive infinity.

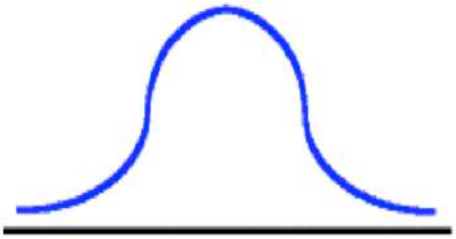
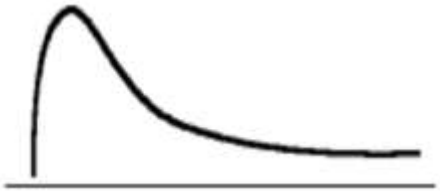


Illustration of positive skewness – notice the tailing off of the distribution toward positive infinity



In most applications, the raw skewness is not used. Rather the reader is directed to the "standardized skewness" as shown in the formula presented in the equation, below.

Standardized Skewness: [latex]skew = {{\sum(x_{i}- \overline{x})^3 \over {N}}} \times({ 1\over {s^2 *s}})[/latex]

Kurtosis

Kurtosis is also an estimate of variability. Kurtosis is considered the "fourth moment" of scores within a distribution. Again, this is easily recognized by the formula for kurtosis shown in the equation below. Notice that similar to estimates of variance and skewness, the estimate of kurtosis is computed as a ratio of the difference between the observed scores and the mean scores raised to an exponent – in the case of kurtosis the exponent is 4.

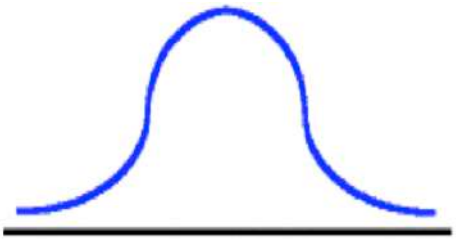
[latex]kurtosis = {{\sum(x_{i}- \overline{x})^4 \over {N}}}[/latex]

However, unlike variance, in which the difference scores were squared, and skewness, in which the difference scores were cubed, kurtosis computes the difference scores to the 4th power. That is, kurtosis is computed by:

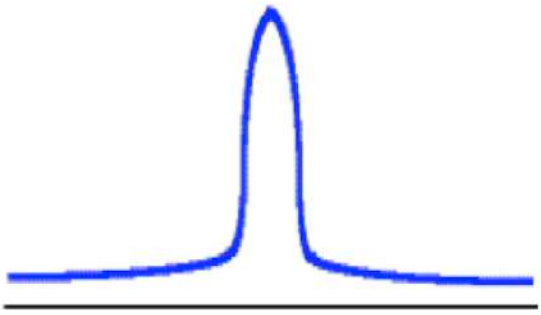
- subtracting each observation from the average score,
- raising the difference to the exponent "4",
- summing the products and then
- dividing by the number of observations (or the number of observations minus one).

Examples of kurtosis are shown here:

The figure above is an illustration of a **platykurtic distribution**. Notice the flatness of the distribution (remember the term platy refers to flat). In order to achieve a platykurtic distribution, all scores within the distribution are unique. Below is an illustration of the **mesokurtic distribution**.



Notice in the **mesokurtic distribution** the shape is that of a normal distribution – neither flat nor demonstrating any peakedness in the distribution. Below is an image illustrating the **leptokurtic distribution**.



Notice the peak-ness of the distribution (remember the term lepto refers to leaping). In order to achieve a leptokurtic distribution, all scores within the distribution are located close to the mean. There is very little deviation of scores from the measure of the central tendency within the distribution.

In most applications, the raw kurtosis is not used. Rather the reader is directed to the "standardized kurtosis" as shown in the equation below.

Standardized Kurtosis =[latex]{{\sum(x_{i}- \overline{x})^4 \over {N}}} \times({ 1\over {s^2 * s^2}})[/latex]

When computing the standardized skewness and standardized kurtosis a check of the distribution characteristics are as follows:

1. When a negative skewness value is observed, the distribution is negatively skewed. (tailing to the negative – the left)
2. When a positive skewness value is observed, the distribution is positively skewed. (tailing to the positive – the right)
3. When a skewness value is zero, the distribution has no skewness.
4. When a kurtosis value is less than three, the distribution is platykurtic.

5. When a kurtosis value is greater than three, the distribution is leptokurtic.
6. When a kurtosis value is equal to three, the distribution is mesokurtic.

28. Estimating Confidence Intervals for a Sample Mean

Putting it all together

In research, we often collect information from a collection of individuals that are drawn from a larger group. Since the procedures for information collection have real costs, researchers are forced to make the assumption that the individuals selected for study are a true representation of all of the individuals within the larger group. In quantitative analyses the larger group is the population, (represented by the letter 'N'), and the selected subgroup is referred to as the sample (represented by the letter 'n'). Given the assumption that the sample represents the population from which it was drawn, then it is also assumed that any computations, estimates, or inferences based on the measures from the sample, must also represent the population from which the sample was selected.

As such, the average score computed for the sample is assumed to represent the average score for the population. Likewise, the variability of scores within the sample (the subgroup) is expected to represent the variability of the scores within the population (the larger group). Similarly, the standardized estimate of the differences computed for the sample should represent the standardized estimate of differences computed for the population.

The confidence interval is based on the following relationship between the sample means and the true population mean or [μ]:

the lower limit of the sample mean < μ < upper limit of the sample mean

This sentence is read as: The lower limit of the sample mean is less than the true population estimate which is less than the upper limit of the sample mean. Therefore: the population mean = sample mean ± sampling error : μ = [latex]\overline{x}[/latex] ± (SE),

where:

- μ refers to the measure of central tendency for the population
- is the sample mean and refers to the measure of centrality in a sample
- Standard Error (SE) – error due to randomness

There are two basic assumptions in this approach:

First, we assume that the sample mean is only our best estimate of the true population mean.

Second, we assume that the chance associated with the sample mean's ability to represent the true population mean is dependent upon the ability of the sample scores to represent the population scores.

So that by adding or subtracting the sampling error to or from the sample mean we will be able to identify the range within which the true population estimate falls.

Standard Error

The term sampling error, which is also called the standard error of the mean, is a measure of the extent to which the sample means can be expected to vary due to chance. In other words, the standard error of the mean provides an estimation of the variance (or error) of the mean in the sample and can be attributed to the sampling characteristics associated with the sample.

The Confidence Interval (95%)

Confidence intervals help the researcher determine the accuracy that a sample estimate represents a true population parameter. In most studies, the researcher has an implicit expectation that the sample is representative of the larger group (i.e. the population). Therefore, if we assume that our sample represents a population, then we must also assume that any computations, estimates, or inferences based on the numbers from the sample, must also represent the population from which the sample was selected. As such, the average score computed for the sample is assumed to represent the average score for the population; similarly, the variability of scores within the sample (the subgroup) should represent the variability of the scores within the population (the larger group).

Given that the sample is an accurate representation of the population, standardized estimates of the differences computed for the sample should represent the standardized estimates of differences computed for the population. One can also expect that the measure of central tendency for the population (μ) can only be estimated by the measure of central tendency for the sample, whereby = μ and therefore it is accepted that there will always be some amount of error due to known and unknown factors. While the sample mean, variance and standard deviation each represent estimates of the true population values, the value that represents the accuracy of our projected estimates are expressed as a measure of confidence. We can, therefore, assume that the confidence interval is an accurate representation of the actual space within which we could expect to find the true population measures. Such expectations are based on the following principles: we assume that the sample mean is only our best estimate of the true population mean. We assume that the chance associated with the sample mean's ability to represent the true population mean is dependent upon the ability of the sample scores to represent the population scores. By adding or subtracting the sampling error to or from the sample mean we will be able to identify the range within which the true population estimate falls.

The term sampling error refers to the errors that occur in the process of data collection. Sampling error is expected and should thus be accounted for in the computation of the estimates that represent the data. Researchers state that the estimates (measures of central tendency, frequencies, or ratio estimates) produced from a selected sample are expected to represent the true population estimate within a specific range. For example, the researcher states that: They are 95% confident that the sample mean represents the true population mean within 10% error.

Typically, researchers indicate that they would like to be at least 95% confident that the sample mean is an estimate of the population mean. Therefore, the researcher is suggesting that 19 out of 20 times the sample mean ± sampling error will include [μ]. The confidence interval is based on the following relationship between the sample mean and the true population mean or [μ]: This sentence is read as: The lower limit of the sample mean is less than the true population estimate which is less than the upper limit of the sample mean.

The standard error of the mean, is computed by the following formula shown here: [latex]s.e. = {s \over{\sqrt{n}}}[/latex]

Example: Compute the confidence interval at 95% for a given sample mean where the mean = 58 ± standard deviation (s) = 13 and n=25 participants. To compute the standard error (se) we use: [latex]s.e. = {s \over{\sqrt{n}}}[/latex] = [latex]s.e. = {13 \over{\sqrt{25}}}[/latex] = 2.6

The upper and lower limit of the 95% confidence interval for the mean = 58 is computed with the basic formula: [latex]CI_{95} = { \overline{x} \pm }(1.96 \times{se})[/latex]

[latex]58 \pm [1.96 \times{2.6}] {\rightarrow}\textit{95% confidence interval is 58} \pm 5.1[/latex]. Which means that there is a 95% probability or chance that the range 52.9 and 63.1 will capture the true population mean μ.

THE BASIC PREMISE OF ESTIMATION AND CONFIDENCE INTERVALS

μ = sample mean ± (1.96 x standard error of the mean)

where:

- μ refers to the measure of central tendency for the population
- sample mean refers to the measure of central tendency for the sample
- standard error of the mean also known as sampling error is the estimate by which the mean can vary

In computing confidence intervals, we determine the estimate of the error of the sample selected or the sampling error. This error is also called the standard error of the mean and is a measure of "the extent to which the sample means can be expected to vary due to chance". In other words, the standard error of the mean is "an estimate of the error associated with the observed mean in this specific sample" and is due to the sampling characteristics associated with this sample.

29. Applying the Student's t-test for Single and Paired Samples

Learning Objectives

After reading this chapter you should be able to:

- Calculate differences when using the student's t-test and when using a matched pairwise comparison t-test
- Apply and test the null hypothesis for t-test calculations
- Apply the decision rule to evaluate a null hypothesis
- Write a SAS program to evaluate the student's t-test and a matched pairwise comparison t-test

Introduction to the Tests of Significance for Mean Differences

In this next section, we will test the significance of differences between the mean for a set of data and the mean for a comparison set of data. In the first instance, we will compare the mean score from a single set of data to the expected mean for a population. Next, we will compare the mean score for a change from one day to another; and then, in the second section, we will compare the means from two different groups.

An important consideration in calculating any statistic is that the mathematics of the computation doesn't consider the subject of the scores. That is, the data can be extracted from any source and the mathematics will be the same. For example, consider the following two variables: annual income, and waiting time in the emergency department. The mathematics to calculate the means for each set of data are the same, despite that the two variables hold different information and different meaning.

In a situation, where you have data for a single group (as in a one-sample design), you can also compute a difference score between the mean for the set of observations and then compare your single group mean score against the population score. The population score is referred to as the parameter estimate since we never really know the true population score.

Computing the Student's single sample t-test:

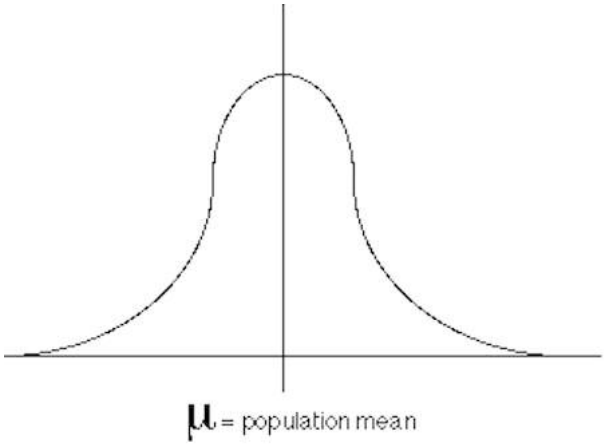
Let's consider first the Student's single-sample t-test. This means that we are going to compare the mean from a sample of data against the mean for the population (which we generally accept to be 0). The statistical test that we will use in this computation is the Student's t-test.

The purpose of the Student's t-test is to evaluate the null hypothesis of the sample mean against the population mean.

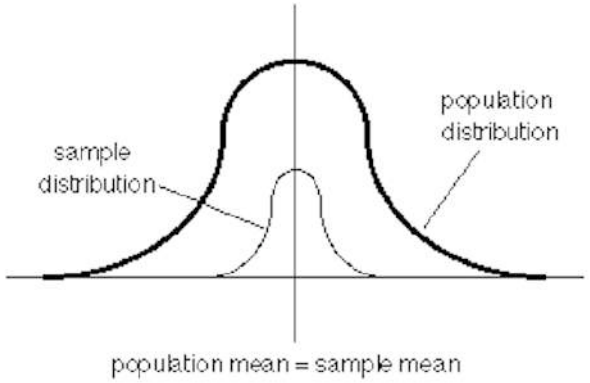
[latex]H_{0}: \overline{x} = \mu[/latex]

- where: [latex]H_0[/latex] refers to the null hypothesis
- [latex]\overline{x}[/latex] refers to the mean of the sample or the distribution
- [latex]\mu[/latex] refers to the mean of the population

To evaluate the significance of the difference between **a sample mean** (the mean for the set of observed scores) and **an expected mean** (the mean for the population), use the Student's t-test. That is to ask, "Is the mean for a sample of scores the same as the mean expected to represent a population?"



Above is an illustration of the population distribution. As shown in the illustration below, we draw (select, observe, record) the sample distribution from the population distribution.



Under the null hypothesis, we test if the mean of the sample is truly representative of the mean for a population. This test is based on the null hypothesis question expressed here: $H_0: \bar{x} = \mu$

Considering an approach in which we use randomized representative sampling to collect our data, we should expect that the sample mean should be the same as the population mean. If this outcome were observed, then we would accept the null hypothesis.

To evaluate the null hypothesis, which is to say, to determine if, in fact, the null hypothesis is true: that the sample represents the population; we compare the sample mean to the population mean, whereby our sample mean is the mean for the set of observations and the population mean is the mean for the set of expected scores.

Therefore, to evaluate the difference between a sample mean and an expected mean we use the Student's t-test, as shown in the following formula.

[latex]\textit{Student's t-test} = {\overline{x}_{observed}-\overline{x}_{expected}\over{\frac{S_{observed}}{\sqrt{n_{observed}}}}}[/latex]

SAMPLE CALCULATION 29.1

In this first example, we will use the student's t-test to evaluate the mean score for a continuous random variable for a selection of individuals drawn from a population. In Table 29.1, data for the number of hours in a given month that individuals waited to see their healthcare provider are presented. You are asked to compute the average wait time for this sample and determine if the average for the sample is significantly different than that which is expected in the population.

Table 29.1 Monthly wait time in hours for a random sample of individuals

Patient ID	Hours Waited
01	**12**
02	13.5
03	8.5
04	4.5
05	10.5
06	15.5
07	12.5
08	9.5
09	11
10	7.5

Here we use the Student's t-test to evaluate the null hypothesis that the observed mean is equal to the expected mean in this set of scores. In the application of the t-test, we assume that the mean score for the sample is representative of the mean score for population, and likewise, that the variance for this sample of scores is an estimate of the variance for the population. We also assume, unless we have some other information, that these data were drawn from a normal distribution.

In this example, we are assessing the null hypothesis: H0: [latex]\overline{x} = \mu[/latex], also presented as H0: [latex]\overline{x} = 0[/latex]. In using the t-test we assume that the population mean [latex][\mu][/latex]has a value of 0 and a variance (standard deviation) of 1. When we are applying the t-test, we are comparing our observed sample mean against the expected population mean of 0 with a variance of 1. The SAS code below produces a test of the null hypothesis using the Student's t-test.

SAS Code to Compute the Student's t-Test for a Single Sample

```
   DATA STUDENT;
INPUT ID SCORE @@;
DATALINES;
001 12.0 002 13.5 003 08.5 004 04.5 005 10.5 006 15.5 007 12.5 008 09.5 009 11.0 010 07.5
;
PROC SORT DATA=STUDENT; BY ID;
TITLE "STUDENT'S t-TEST TO EVALUATE WAIT TIMES";
PROC UNIVARIATE; VAR SCORE;
PROC MEANS N MEAN MEDIAN T STDERR PRT;
VAR SCORE; RUN;
```

The SAS code above produces several important output tables to help us explore our dataset and to evaluate the null hypothesis: H0: [latex]\overline{x} = 0[/latex]. In the output shown below, we can review the descriptive statistics for the dependent variable which we labeled SCORE. In this output we see that the mean score for the sample was 10.5 with a standard deviation of ± 3.17.

Table 29.2 SAS Output for Dependent Variable: SCORE

N	10	**Sum Weights**	10
Mean	10.5	**Sum Observations**	105
Std Deviation	3.1710496	**Variance**	10.0555556
Skewness	-0.3854798	**Kurtosis**	0.23914105
Uncorrected SS	1193	**Corrected SS**	90.5
Coeff Variation	30.2004724	**Std Error Mean**	1.00277393

Basic Statistical Measures are given here as location – mean, median and mode, and measures of variability – standard deviation, variance, tang and interquartile range.

Table 29.3 Measures of Location and Variability

Mean	10.50000	**Std Deviation**	3.17105
Median	10.75000	**Variance**	10.05556
Mode	.	**Range**	11.00000
		Interquartile Range	4.00000

The SAS code also produced statistics that help us to determine a decision with regard to accepting or rejecting the null hypothesis. In Table 29.4 below, we see that the Student's t value is given as 10.47, and the corresponding p-value was Prt > |t| <0.0001. Because the probability value (Prt) associated with the Student's t score is <0.05 the outcome indicates that the mean value (10.5 ± 3.17) was significantly different than 0, and therefore we would reject the null hypothesis that H0: [latex]\overline{x} = 0[/latex].

Table 29.4 SAS Output for Tests of [latex]\overline{x} = 0[/latex] Tests for Location: Mu=0

Test	Statistic	p Value		
Student's t	**t**	10.47095	**Pr > |t|**	<.0001
Sign	**M**	5	**Pr >= |M|**	0.0020
Signed Rank	**S**	27.5	**Pr >= |S|**	0.0020

SAMPLE COMPUTATION 29.2

So, what if in this specific example we knew that the real population mean wasn't **0** but that the mean value was actually a specific number that we determined from the literature. Let's say that the expected mean should be a value of **10** for this sample. In the following example, we can redo the computations of the Student's t-test and set the population mean to the specific value of 10. ***In practice, the value against which the mean is compared should be based on theoretical considerations and/or previous research so that for many experiments the outcome range of the dependent variable can be known.***

To test the null hypothesis for the one sample scenario that the expected mean = 10, we compare the *t-score observed* against the *t-score critical*. The decision rule then becomes: if the *t-test score observed* is greater than the *t-test score critical* then we reject the null hypothesis and state that the sample mean is significantly different than the population mean.

Code to Compute Student's or Single Sample t-test Using Proc t-test and Setting μ =10

```
   OPTIONS PAGESIZE=65 LINESIZE=80;
DATA STUDENT;
INPUT ID SCORE @@;
DATALINES;
001 12.0 002 13.5 003 08.5 004 04.5 005 10.5 006 15.5 007 12.5 008 09.5 009 11.0 010 07.5
;
PROC SORT DATA=STUDENT; BY ID;
PROC TTEST H0=10;
VAR SCORE;
TITLE1 "t-TEST WHEN EXPECTED MEAN NOT 0";
TITLE2 "EXPECTED MEAN = 10";
RUN;
```

Table 29.5 "t-TEST WHEN EXPECTED MEAN NOT 0" – "EXPECTED MEAN = 10"

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	10.5	3.17	1.00	4.50	15.50

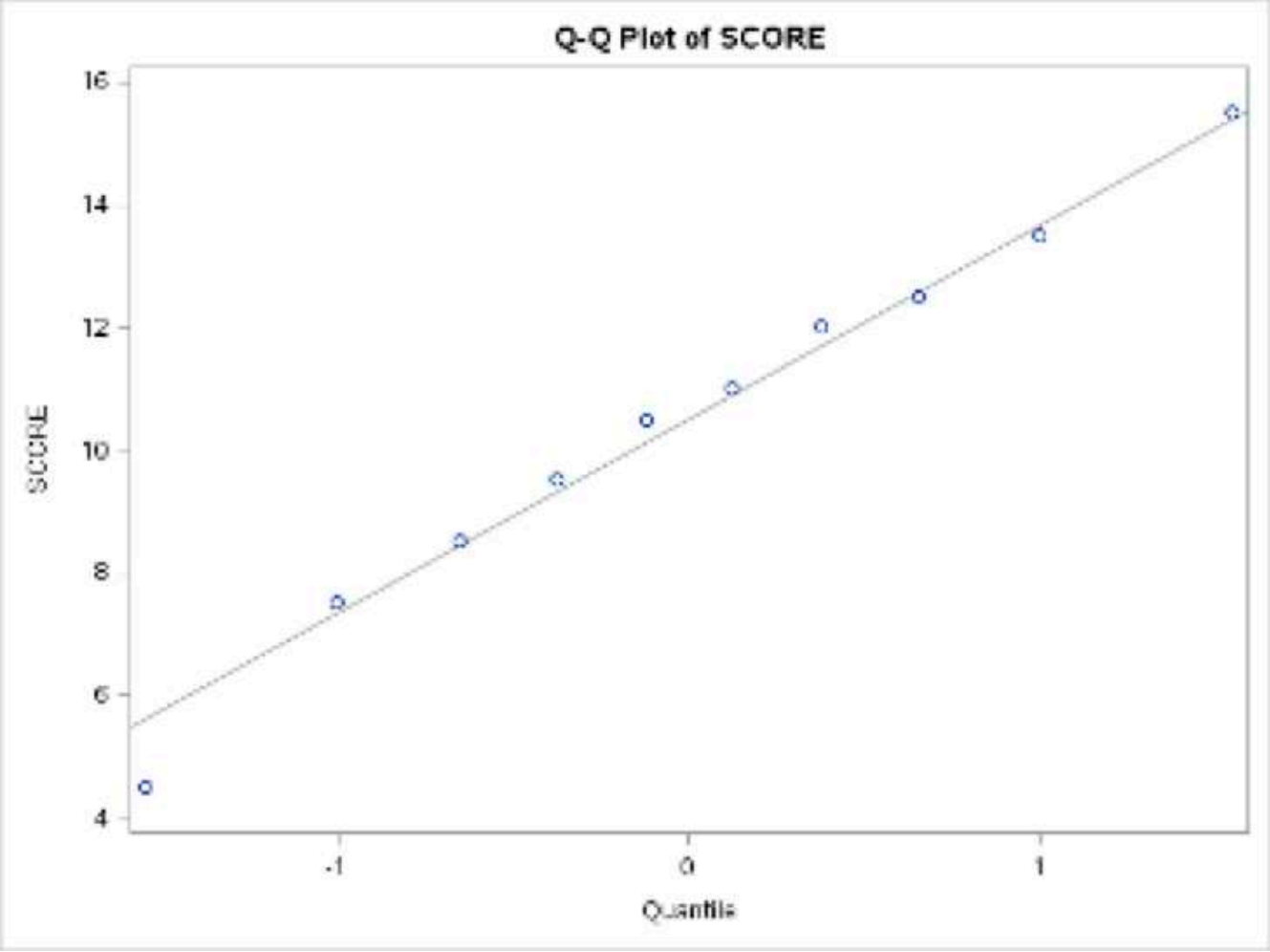
Mean	95% CL Mean
10.5±3.17	8.23[latex]\rightarrow[/latex]12.77

DF	t Value	Pr > |t|
9	0.50	0.6300

In the data presented above, the computed mean was 10.5 with a standard deviation of 3.17. This was compared against an expected mean of 10 and thus produced a t-test score, labeled here as t Value, equal to 0.50. The probability associated with this t-test score is given in the table as Pr >|t|, and is equal to 0.6300. Because the p-value is greater than 0.05 we accept the null hypothesis that the observed mean for the sample is not different than the apriori mean we had set at [latex]\overline{x} = 10[/latex].

This SAS output for the t-test procedure also produces a Q-Q plot. The Q-Q plot is a visual representation that the data used in this example were drawn from a normal population. If the Q-Q plot presents a relatively straight line then we can assume that the data were drawn from a population that was normally distributed. However, if the data are not representative of a somewhat straight line then we can suggest that the data were not drawn from a population that we would consider to be normally distributed. Notice in Figure 29.1, below, the Q-Q plot shows a straight line and thereby indicates that the data were drawn from a population that was normally distributed.

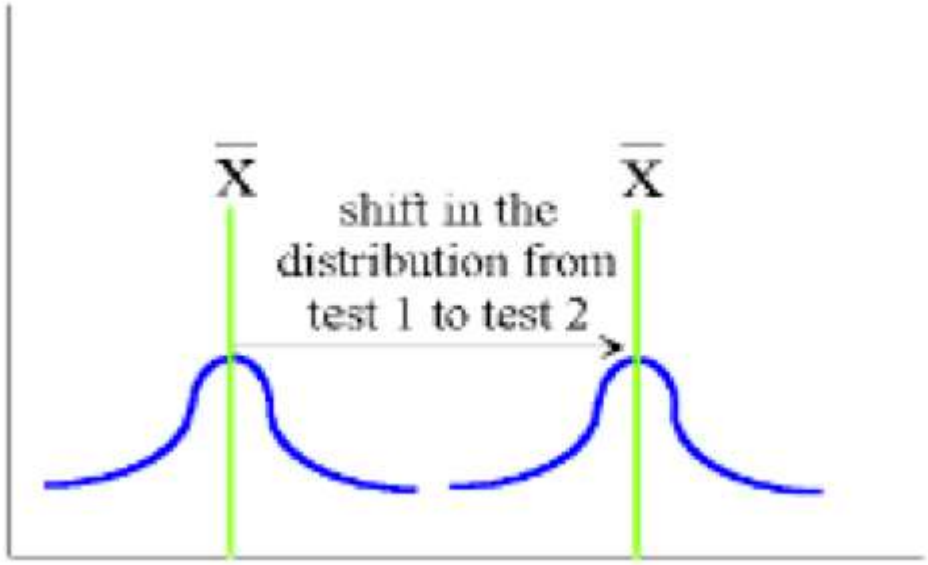
Figure 29.1 Q-Q plot of data to demonstrate normality in the sample distribution



The Pairwise t-test

To evaluate the significance of the difference in a pre to post design t-test, use the pairwise t-test formula.

Figure 29.2 Evaluating the shift from pretest to posttest



The pre to post (pairwise) t-test measures the amount of shift in the data over the time interval. The formula to compute the pairwise t-test uses the average difference in the measure of interest, from the pre-test score to the post-test score, and then divided by the standard error of the average difference. The standard error of the average difference is computed by dividing the standard deviation of the average difference by the square root of the number of cases in the pairwise comparison.

Figure 29.3 Formula to evaluate the Pre to Post-test design

$$t_{pairwise} = \frac{\overline{x}_{\Delta} - 0}{s.e.} \qquad s.e. = \frac{\sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}}{\sqrt{n}}$$

Computing a Pairwise Comparison T-Test to Evaluate Change

We can use the formula for the pairwise t-test shown above, to compute the significance of the change in the mean score from the pre-test data to the post-test data for the following two sets of 10 numbers.

Pre test data

Participant Code	001	002	003	004	005	006	007	008	009	010
Score	34	54	60	68	53	70	81	87	65	55

Post test data

Participant Code	001	002	003	004	005	006	007	008	009	010
Score	44	75	72	98	73	80	91	99	69	76

Analyzing these data as a paired t-test.

In this analysis, we will treat the difference scores produced by subtracting the pre-test scores from the post-test scores. In order to do this, we will need to create a new variable. We can create a variable DIFFSCR by subtracting the post-test score from the pre-test scores. As a rule, it is always best to keep our variable names simple and to restrict the label to 8 characters or less. The SAS code to set up this analysis is shown below. However, it is important to note that the degrees of freedom for the paired t-test uses a single sample mean and therefore is df = n – 1, and in this case, df = 10 – 1 = 9 so that the t critical value for $\alpha = 0.05$ is 2.262 for a two-tailed test and 1.833 for a one-tailed test.

SAS Code to Compute pairwise difference with a t-test

```
   OPTIONS PAGESIZE=65 LINESIZE=80;
DATA PAIRWISE;
INPUT @1 ID PRETEST POSTTEST;
DIFFSCR=(POSTTEST – PRETEST);
DATALINES;
001 34 44
002 54 75
003 60 72
004 68 98
005 53 73
006 70 80
007 81 91
008 87 99
009 65 69
010 55 76
;
```

In this analysis, we will evaluate the set of difference scores as a single group (as in a one-sample design) just as we did for the student's t-test. Under this design, we can use the ***single-sample t-test*** *or* ***Student's t-test*** procedure to compute the significance of the difference scores. Therefore, under this design, we are testing the null hypothesis that the difference between the pre and post scores = 0 (i.e. no difference).

Step 1 in computing the difference scores between the results in the pre test and the results in the post test was to create the variable: DIFFSCR. In this example uses POSTTEST – PRETEST to show a positive change (an increase in score values). The production of the difference score was accomplished with the SAS. command: DIFFSCR=(POSTTEST – PRETEST);

Using the dependent variable DIFFSCR, and the following SAS Code we produced three different SAS outputs: i) **PROC UNIVARIATE**, *ii*) **PROC MEANS**, iii) **PROC TTEST.**

```
   PROC SORT DATA=PAIRWISE; BY ID;
PROC UNIVARIATE;  VAR DIFFSCR;
PROC MEANS N MEAN MEDIAN T STDERR PRT;  VAR DIFFSCR;
PROC TTEST; VAR DIFFSCR;
RUN;
```

The results for each of these procedures are shown below the descriptive statistics produced by the Proc Univariate procedure.

N	10	**Sum Weights**	10
Mean	15	**Sum Observations**	150
Std Deviation	7.7172246	**Variance**	59.5555556
Skewness	0.65636272	**Kurtosis**	0.00011636

Evaluation of the Null Hypothesis based on output for the UNIVARIATE Procedure

Test	**Statistic**	**p Value**		
Student's t	**t**	6.146532	**Pr > |t|**	0.0002
Sign	**M**	5	**Pr >= |M|**	0.0020
Signed Rank	**S**	27.5	**Pr >= |S|**	0.0020

Evaluation of the Null Hypothesis based on output for the MEANS Procedure

Analysis Variable: DIFFSCR					
N	**Mean**	**Median**	**t Value**	**Std Error**	**Pr > |t|**
10	15.0000000	12.0000000	6.15	2.4404007	0.0002

Evaluation of the Null Hypothesis based on output for the t-Test Procedure

DF	**t Value**	**Pr > |t|**
9	6.15	0.0002

The important estimates from this output are the mean and variances, which are listed within each of the three SAS procedures. However, each procedure presents the outcome values in different ways. In the PROC UNIVARIATE procedure, all moments of variance are given – standard deviation, variance, skewness, kurtosis, and standard error.

However, in the PROC MEANS procedure only the mean, median, and standard error are reported when the coding above is used. Finally, the PROC TTEST procedure provides the mean score along with the standard deviation and the

computation of the confidence limits. In each basic SAS procedure used above, the Student's t-test computes the comparison between the sample (based on the difference scores) against the population.

Application of the General Decision Rule for the Evaluation of the t-Test

Student's t-test: H0: sample mean = population mean.

To test this null hypothesis, we compare the *t-score observed* against the *t-score critical.* The decision rule is as follows: *If the t-score observed is greater than the t-score critical then we reject the null hypothesis and state that the sample mean is significantly different than the population mean.*

Pairwise comparisons t-test: H0: mean difference for the sample = mean difference for the population

To test this null hypothesis, we compare the *t-score observed* against the *t-score critical.* In this case, the t-score is based on an evaluation of the average difference from pre-test to post-test scores against the expected average difference, in a population of scores, should be 0. The decision rule then becomes: *If the t-score observed is greater than the t-score critical then we reject the null hypothesis and state that the change from pre-test scores to post-test scores is significantly different than that which is expected.*

t-tests for two independent samples: H0: mean for group1 = mean for group2

Again, to test this null hypothesis we compare the *t-score observed* against the *t-score critical.* In this case the t-score is based on a comparison of the difference of the mean for ***group1*** against the mean for ***group2***. The decision rule is as follows: if the *t-score observed* is greater than the *t-score critical* then we reject the null hypothesis and state that the two means are significantly different from each other.

Computing a Pairwise Comparison t-Test to Evaluate Change in Resting Blood Pressure

The null hypothesis for the paired or pairwise t-test is given below:

$$H0: \mu1 = \mu2 \quad \text{or} \quad H0: \mu1-\mu2=0$$

In this experiment, a group of 10 individuals agreed to participate in a study of blood pressure changes following exposure to halogen lighting. Resting systolic blood pressure was recorded for each individual. The participants were then exposed to 20 minutes of halogen lighting by playing a video game in a room, which was lit only by halogen lamps. A post-exposure systolic blood pressure reading was recorded for each individual. The results are presented in the following data set.

ID	01	02	03	04	05	06	07	08	09	10
SEX	M	F	M	F	M	F	M	F	M	F
Pre exposure Systolic BP	120	132	120	110	115	128	120	112	110	100
Post exposure Systolic BP	140	156	145	130	117	148	137	119	127	135

In the following worksheet we compute the significance of the difference between the pre test blood pressures and the post-test blood pressures..

The null hypothesis is: Ho: μ1 = μ2 and the computation of the variance is shown with the raw data in the table below.

ID	01	02	03	04	05	06	07	08	09	10
SEX	M	F	M	F	M	F	M	F	M	F
Pre exposure Systolic BP	120	132	120	110	115	128	120	112	110	100
Post exposure Systolic BP	140	156	145	130	117	148	137	119	127	135

In the following worksheet, we compute the significance of the difference between the pre-test blood pressures and the post-test blood pressures.

The null hypothesis is: Ho: μ1 = μ2 and the computation of the variance is shown with the raw data in the table below.

ID	sex	pretest BP	post-test bp	pre-post difference [latex]\Delta(x_{i})[/latex]
01	m	120	140	20
02	f	132	156	24
03	m	120	145	25
04	f	110	130	20
05	m	115	117	2
06	m	128	148	20
07	f	120	137	17
08	m	112	119	7
09	f	110	127	17
10	m	100	135	35
				[latex]\overline{\Delta}=\sum{\Delta(x_{i})}[/latex]

[latex]\Delta(x_{i}) - \overline{\Delta}[/latex]	[latex]\left(\Delta(x_{i}) - {(\overline{\Delta})}\right)^2[/latex]	Squared difference scores
20-18.7=1.3	[latex](1.3)^2[/latex]	1.69
24-18.7=5.3	[latex](5.3)^2[/latex]	28.09
25-18.7=6.3	[latex](6.3)^2[/latex]	39.69
20-18.7=1.3	[latex](1.3)^2[/latex]	1.69
2-18.7=-16.7	[latex](-16.7)^2[/latex]	278.89
20-18.7=1.3	[latex](1.3)^2[/latex]	1.69
17-18.7=-1.7	[latex](-1.7)^2[/latex]	2.89
7-18.7=-11.7	[latex](-11.7)^2[/latex]	136.89
20-18.7=-1.7	[latex](-1.7)^2[/latex]	2.89
35-18.7=16.3	[latex](16.3)^2[/latex]	265.69
Sum of differences = [latex]\sum \left(\Delta(x_{i}) - \overline{\Delta}\right)[/latex] = 0	Sum of squared difference scores = [latex]\sum \left(\Delta(x_{i}) - {(\overline{\Delta})}\right)^2[/latex] = 760.1	

To compute the variance for the set of difference scores we divide the sum of squares by n-1. In this analysis the variance is computed as follows.

Sum of squares = [latex]\sum \left(\Delta(x_{i}) - {(\overline{\Delta})}\right)^2 = {760.1 \over{n - 1}} = {760.1 \over{10-1}} = {760.1 \over{9}} = 84.46[/latex]

The standard deviation is then calculated by estimating the square root of the variance. Here the calculation is the square root of 84.46, which produces a standard deviation of 9.19. Once we have the standard deviation we can calculate the standard error by dividing the standard deviation by the square root of n. Here we have a standard deviation of 9.19 and n=10 (our total sample). The calculation of the standard error is then 9.19 divided by the square root of 10 which equals 2.91, as shown below.

[latex]s = \sqrt{84.46} = 9.19 [/latex] [latex]\textit{s.e.} = {s\over\sqrt{n}}[/latex] [latex]\rightarrow \textit{s.e.} = {9.19\over\sqrt{10}} = 2.91[/latex]

Once we have the standard error from these calculations then we can calculate the t-test value. Recall that the t-test

score provides an evaluation of the null hypothesis which here is a test that the mean difference between the pre and post blood pressure scores is equal to 0. The calculation for the t score in this scenario is shown here:

[latex]t = \left(\overline{\Delta} - 0 \over{s.e.}\right)[/latex] [latex]\rightarrow t = \left(18.7 - 0 \over{2.91}\right)[/latex] [latex]\rightarrow t = \left(6.43\right)[/latex]

Next use the t observed score to evaluate the null hypothesis by comparing the t observed scores against the t expected score for a research design in which we started with a sample of n=10. The t expected score is identified in a table of critical values – easily retrieved from the internet. The information required to identify the t critical value is the number of cases in our sample and the probability level at which we want to be sure that we are making the right decision in regard to accepting or rejecting the null hypothesis. For most research studies we accept that we want to be 95% confident that we are making the correct decision in regard to accepting or rejecting the null hypothesis, so given that (100% – 95% = 5%) we establish a probability level of 0.05 or 5%; and we call this probability the alpha level. Once we have our alpha level and we know our sample size (n) then we can look up the appropriate t critical value (from a standard table of critical values) against which we will compare our t observed score.

In the evaluation of changes in blood pressure, our t observed score was 6.43, and the corresponding t critical (or t expected) score based on an alpha level of 0.05 and n=10 is given as t critical = 2.228. Therefore, since our t observed score ($t_{observed}$ = 6.43) is greater than the t critical score ($t_{critical}$ = 2.228) we reject the null hypothesis [latex]H_{0}: \left(\overline{\Delta} = 0 \right)[/latex] and determine that the change in blood pressures within our sample was statistically significant.

In this analysis, we can write a SAS program to evaluate the ***pairwise t-test*** using the **PROC MEANS** procedure and evaluate the null hypothesis using both Proc MEANS, as well as the student's single sample t-test. The SAS program is as follows:

```
  DATA PAIRWISE;
INPUT @1 ID SEX $ PREBP POSTBP;
```

Following the input paragraph above, we create a new variable to represent the difference in pre to post systolic blood pressure scores. Here we will call this variable DELTABP. Considering that post bp is expected to rise following exposure to halogen, we include POSTBP as the first term in the equation to compute the difference scores.

```
  DELTABP = (POSTBP – PREBP);
  DATALINES;
01 m 120 140
02 f 132 156
03 m 120 145
04 f 110 130
05 m 115 117
06 m 128 148
07 f 120 137
08 m 112 119
09 f 110 127
10 m 100 135
;
```

Computing the pairwise t-test is similar to computing the student's or single sample t-test. Here we use the measure of difference scores as the variable of interest. The SAS procedural commands are presented here:

```
  PROC SORT DATA=PAIRWISE; BY ID;
PROC UNIVARIATE; VAR DELTABP; RUN;
```

In the statement below, we use PROC MEANS to compute the t-test scores. Notice that the options to produce the statistical output are included within the PROC MEANS statement prior to the semi-colon.

PROC MEANS N MEAN MEDIAN T STDERR PRT; VAR DELTABP; RUN;

In the statement below, we use PROC TTEST to compute the t-test scores.

PROC TTEST; VAR DELTABP; RUN;

As in all analyses we begin with the descriptive statistics. These are reported through the output for the PROC UNIVARIATE procedures. However, since we are comparing a mean for a sample against the population mean of 0, we really don't need to do much more as SAS provides this outcome in the reporting of the student's t-test as part of the PROC UNIVARIATE procedure, shown below.

Table 29.6. Output from PROC UNIVARIATE for dependent variable: DELTABP

N	10	**Sum Weights**	10
Mean	18.7	**Sum Observations**	187
Std Deviation	9.18997038	**Variance**	84.4555556
Skewness	-0.2742618	**Kurtosis**	0.84720499
Coeff Variation	49.1442266	**Std Error Mean**	2.9061238

Test	**Statistic**	**p Value**		
Student's t	**t**	6.43	**Pr > |t|**	0.0001
Sign	**M**	5	**Pr >= |M|**	0.0020
Signed Rank	**S**	27.5	**Pr >= |S|**	0.0020

Next, the PROC MEANS procedure was used to compute the mean and test the comparison of the mean from 0 using the single sample t-test. The output is shown below. Notice this procedure lists the t-value and the probability of the computed t-value.

Table 29.7. Output from PROC MEANS for dependent variable: DELTABP

N	**Mean**	**Median**	**t Value**	**Std Error**	**Pr > |t|**
10	18.7000000	20.0000000	6.43	2.9061238	0.0001

The values computed with the PROC MEANS procedure can be compared to the procedures of the PROC TTEST and are shown below for the pairwise t-test comparison.

Table 29.8. Output from PROC TTEST for dependent variable: DELTABP

N	**Mean**	**Std Dev**	**Std Err**	**Minimum**	**Maximum**
10	18.7000	9.1900	2.9061	2.0000	35.0000

Mean	**95% CL Mean**	
18.7000	12.1259	25.2741

Again, the important bits of information are the t-value and the probability of the t-value. The evaluation of the t-value is to consider the likelihood of the computed value. Recall that there is no fixed demarcation point for this t-test outcome, but what is more important is whether the differences are as expected. Using a probability estimate of $p<0.05$ as a guide we can decide about the difference in means relative to the t-test score. Here we see that the t-test score is 6.43 and the associated probability of observing this t-test value is less than 0.05.

DF	t Value	Pr > |t|
9	6.43	0.0001

However, what we really want to talk about is what this difference means to your research question. In this instance, the results indicate that the mean systolic blood pressure, measured prior to exposure to the halogen lights was significantly lower than the mean systolic blood pressure, measured after exposure to the halogen lights.

Notice, also that the values computed with the SAS program for each of the procedures verify the calculations that we worked through by hand, above. That is: the mean difference score was 18.7 with a variance estimate of 84.45, a standard deviation of 9.19, a standard error score of 2.906 (rounded to 2.91) and a t-test score of 6.43. Well done, SAS!

30. The t-test for Independent Sample Means and Pooled Versus Unpooled Variance

Learner Outcomes

After reading this chapter you should be able to:

1. Compute the significance of the difference between two sample means when the sample variances are different
2. Compute the t-test for independent sample means
3. Compute the t-tests for pooled versus un-pooled variance.
4. Write a SAS program to compute and identify the important elements of the output for the computation

Applications of the t-test under different research scenarios

In statistics, the t-test has a simple approach despite that it uses a variety of error terms in the denominator as shown in Table 30.1, below. Depending on the research design, the error term will differ to ensure that the appropriate variance estimates within each of the samples are included in the analyses. The following equations demonstrate the different error terms related to the types of comparisons.

Table 30.1 A Summary of t-test Formulae

t-test descriptions	Appropriate t-test formula
Evaluation of the single sample mean versus the mean for a population	$t_{students} = \frac{\bar{x}_{observed} - \bar{x}_{expected}}{\frac{s_{observed}}{\sqrt{n_{observed}}}}$
The pairwise t-test uses the average difference in the measure of interest, from the pre-test score to the post-test score, and then divided by the standard error of the average difference. The standard error of the average difference is computed by dividing the standard deviation of the average difference by the square root of the number of cases in the pairwise comparison.	$t_{pairwise} = \frac{\bar{x}_{change} - 0}{\frac{s_{observed}}{\sqrt{n_{observed}}}}$
To evaluate the significance of the difference between two mean scores (regardless of the size of "n" in each level of the independent variable) we might consider using a pooled t-test for independent variables.	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(x_{i1} - \bar{x}_1)^2 + (x_{i2} - \bar{x}_2)^2}{n_1 + n_2 - 2} * \left(\frac{1}{n_1}\right) + \left(\frac{1}{n_2}\right)}}$
To evaluate the significance of the difference between two mean scores (regardless of the sample size "n") when we consider un-pooled or unequal variances	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\frac{(x_{i1} - \bar{x}_1)^2}{n_1 - 1}}{n_1} + \frac{\frac{(x_{i2} - \bar{x}_2)^2}{n_2 - 1}}{n_2}}}$

In Figure 30.1 below we see the two dimensions of variance that can contribute to the differences observed in a t-test calculation. As illustrated, not only does the t-test process the differences between means, but the difference is also influenced by the variability between members within each sample.

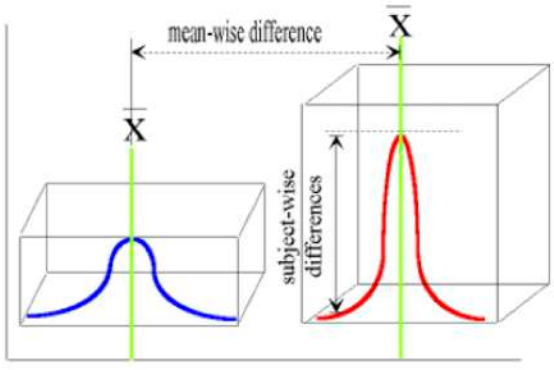


Figure 30.1 Comparison of Two Independent Means

Notice that the decision about which t-test formula to select is dependent on the research design created by the

researcher. For example, as noted in table 30.1, to evaluate the significance of the difference between two mean scores (regardless of the size of "n" in each level of the independent variables) we could consider using a t-test for independent variables, as noted in Table 30.1 formula 3. However, if the number of participants in the samples being compared is equal then it is more appropriate to use Table 30.1 formula 4. In this latter case, given that the sample size between the groups was equal and it is expected that the variance within the two groups is similar. In this case, we can pool or combine the variance estimates as we assume that the two groups have homogenous variance within and between the two groups.

> NB The term homoscedasticity is a term that indicates equality of variance between independent variables. The term is more often used to refer to the variance estimates for each independent variable in a statistical model, as in a multiple linear regression equation. Homoscedasticity suggests that the variables have the same variance.

Scenario 30.1.1 Comparing 2 groups with unknown variance and different sample sizes

Consider the following scenario in which there are two groups with different sample sizes (number of participants in each group) and the variance is unknown within each of the groups. In this situation, the analysis of data uses the t-test for two independent groups. We can use the formula for the t-test for independent groups (with unequal sample sizes) to compute the significance of the differences in the mean scores from group1 in which the sample size is 10 participants and the mean scores from group2 where the sample size is comprised of 8 participants.

Table 30.2 Comparing Responses for two groups with unequal sample sizes and unknown variance

Data for Group 1

Participant ID	001	002	003	004	005	006	007	008	009	010
Score	234	254	260	268	253	270	281	287	265	255

Data for Group 2

Participant ID	001	002	003	004	005	006	007	008
Score	304	235	212	198	273	289	301	209

The degrees of freedom for the two-group t-test is $df=n_1+n_2-2$, and in this case df=10-8-2=16 so that the t critical value based on an *alpha level of* 0.05 is 2.12.

In the following SAS code, we compute the difference between the means for the data in Table 30.2. Here we include a grouping variable so that we can distinguish the data for each group before computing the t-test.

T-test for the difference between means with unequal sample size

```
OPTIONS PAGESIZE=65 LINESIZE=80;
```

```
DATA INDT_TST;
INPUT  ID GROUP SCORE @@;
DATALINES;
001 1 234 002 1 254 003 1 260 004 1 268 005 1 253 006 1 270 007 1 281 008 1 287 009 1 265 010 1 255 011 2 304 012
2 235 013 2 212 014 2 198 015 2 273 016 2 289 017 2 301 018 2 209
;
PROC SORT DATA=INDT_TST; BY GROUP;
PROC TTEST; CLASS GROUP; VAR SCORE;
RUN;
```

The output for the PROC T-TEST procedure for this independent t-test analysis is shown below. The INDEPENDENT t-test Procedure.

GROUP	N	MEAN	STD	STD ERR	MINIMUM	MAXIMUM
1	10	262.7	15.17	4.79	234.0	287.0
2	8	252.6	44.02	15.56	198.0	304.0

GROUP	MEAN	STANDARD DEVIATION
Diff (1-2)	10.08	31.26

group	Method	Mean	Std Dev
1		262.7	15.17
2		252.6	44.
Diff (1-2)	Pooled	10.0750	31,26
Diff (1-2)	Satterthwaite	10.0750	47.37

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	16	0.68	0.51
Satterthwaite	Unequal	8.33	0.62	0.55

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	7	9	8.42	0.0049

To determine if the mean scores in each group were significantly different we typically compare the t-observed to the t critical values, generally available in a reference table. For example, the t critical value for $\alpha = 0.05$, where df=9 is 2.262 for a two-tailed test and 1.833 for a one-tailed test. In the SAS output shown here, a t-critical is not given, however, the probability of achieving the t-value that was computed is reported and this is the indicator of significance. That is to say, when the Pr > |t| is greater than 0.05, as it is in this instance, we would accept the null hypothesis that there is no difference between the mean scores in each group.

Additionally, in the SAS output, we observe a t-value for both a pooled variance estimate and for an un-pooled variance estimate, where the Satterthwaite t Value estimates the unequal/un-pooled variance. As a general rule because the

Folded F stat is a test of unequal variances, when the **folded F statistic** is large and the p-value is <0.05, as shown in the SAS output above, then we refer to the Satterthwaite unequal variances estimate to determine the decision rule regarding the comparison of means via the t-test.

30.2 On the importance of p-values

In the following data set there were 2 groups of 15 individuals. A test was conducted and each individual produced a score. The means were then computed for the scores in each group and a t-test was used to determine if there was a significant difference between the means for each group. The null hypothesis was given as: H_0: mean for group1 = mean for group2

Data in Scenario 1:

The data for each group is shown here using the format (id, group, score):

001 01 12, 002 01 25, 003 01 26, 004 01 23, 005 01 14, 006 01 15, 007 01 17, 008 01 11, 009 01 18, 010 01 14, 021 01 25, 023 01 28, 025 01 26, 027 01 23, 029 01 24 011 02 15, 012 02 34, 013 02 39, 014 02 35, 015 02 34, 016 02 33, 017 02 15, 018 02 31, 019 02 13, 020 02 20, 022 02 16, 024 02 22, 026 02 27, 028 02 26, 030 02 25

The results of the t-test computation using SAS are shown here:

The UNIVARIATE Procedure – Data for the total group for Dependent Variable Score

N	30	**Sum Weights**	30
Mean	22.8666667	**Sum Observations**	686
Std Deviation	7.7135498	**Variance**	59.4988506
Skewness	0.25907362	**Kurtosis**	-0.8503857
Uncorrected SS	17412	**Corrected SS**	1725.46667
Coeff Variation	33.7327251	**Std Error Mean**	1.40829508

The t-TEST Procedure

group	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	15	20.0667	5.8244	1.5039	11.0000	28.0000
2	15	25.6667	8.5161	2.1988	13.0000	39.0000
Diff (1-2)		-5.6000	7.2955	2.6639		

group	Method	Mean	95% CL of the Mean		Std Dev
1		20.0667	16.8412	23.2921	5.8244
2		25.6667	20.9506	30.3827	8.5161
Diff (1-2)	**Pooled**	-5.6000	-11.0568	-0.1432	7.2955
Diff (1-2)	**Satterthwaite**	-5.6000	-11.0893	-0.1107	

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	-2.10	0.0447
Satterthwaite	Unequal	24.746	-2.10	0.0459

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	14	14	2.14	0.1675

So for this output, we would **reject** the null hypothesis and suggest that the mean for *Group 1* was significantly different than the mean for *Group 2*. However, what would happen if in one of the groups we changed one of the scores by 5 points. Notice in the following data set for Scenario 2, all of the scores are exactly the same, except that we changed the data for participant 1 from 12 to 17.

Data in Scenario 2:

The data for each group are shown here using the format (id, group, score):

001 01 17, 002 01 25, 003 01 26, 004 01 23, 005 01 14, 006 01 15, 007 01 17, 008 01 11, 009 01 18, 010 01 14, 021 01 25, 023 01 28, 025 01 26, 027 01 23, 029 01 24, 011 02 15, 012 02 34, 013 02 39, 014 02 35, 015 02 34, 016 02 33, 017 02 15, 018 02 31, 019 02 13, 020 02 20, 022 02 16, 024 02 22, 026 02 27, 028 02 26, 030 02 25

The t-test output for Scenario 2 uses the exact same data set, except that the score for participant 1 in Group 1 was changed from a score of 12 to a score of 17. Notice the highlighted t values and the highlighted confidence intervals.

group	Method	Mean	95% Confidence Limits for the Mean		Std Dev
1		20.4000	17.3755	23.4245	5.4616
2		25.6667	20.9506	30.3827	8.5161
Diff (1-2)	**Pooled**	-5.2667	-10.6175	0.0841	7.1538
Diff (1-2)	**Satterthwaite**	-5.2667	-10.6597	0.1264	

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	-2.02	0.0535
Satterthwaite	Unequal	23.85	-2.02	0.0552

In this second example analysis, we would **accept** the null hypothesis indicating that there was no difference between the means. This decision is based on the comparison of the p-value to the accepted demarcation point of $p<0.05$.

Despite that we have a demarcation point for the probability of the observed t-test, we need to consider the range of scores that we could have seen. The confidence interval provides such information for us. In the first example, the 95% confidence interval tells us that we are 95% confident that the difference between means could be somewhere between -11.09 and -0.11. However, in the second example the 95% confidence interval tells us that we are 95% confident that the difference between means could be somewhere between -10.66 and 0.126.

Sullivan and Feinn[2] provide two quotes to support the need to look beyond the simple comparison of research findings to a p-value. The first quote is by Gene Glass who said, ***"Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude – not just, does a treatment affect people, but how much does it affect them."***.

The second quote is by Jacob Cohen, who said, ***"The primary product of a research inquiry is one or more measures of effect size, not P values."***.

Two important points of interest arise from this comparison. The first being that the intervals are not only similar in size but that they are similar in the bandwidth on the number line[latex]\;\rightarrow\;[/latex] lower limits being (-11.09 and -10.66) and upper limits being (-0.11 and 0.126). However, the second point of interest is that we only had to change one score from the entire set of 30 scores and only by 5 points in order to change from showing a significant difference to a non-significant difference.

If you consider the standard deviation for the scores in *Group 1* in each trial, you will notice that a 5-point change is less than the score by which we expect any score to vary from the mean. That is, in the two examples the standard deviation is 5.82 and 5.46, respectively. Therefore, changing a score by less than the computed standard deviation was sufficient to cause a decision to change from **significant** to **not significant**.

Consider that this was a study in which you invested millions of dollars. If you only relied on the p-value then you would be happy with scenario 1 (a significant difference was found) but you would be tremendously disappointed with scenario 2, and you would unnecessarily throw away valuable information. So a guiding principle may be that despite the reported value of p for any comparison, consider also the standard deviations and the standard errors along with the computed confidence intervals before reporting the findings.

30.3 Estimating the Effect Size

We can compute the effect size – where the effect size is defined as the magnitude of the difference between the two means **when the difference is adjusted by the standard deviation for the mean of interest.** The formula is simply the difference between the two means in a scenario that compares two groups divided by the standard deviation of the group of interest. So how do I establish the group of interest? The confusion in using this formula is often in which standard deviation to select. One way is to simply select one group to be the standard reference group and the other group to be the group of interest.

The *Effect Size* formula: [latex]ES = {\left(\overline{x_{1}} - \overline{x_{2}}\right)\over{s_{1}} }[/latex]

The *effect size* formula is often interpreted using Cohen's criteria where an effect size score of 0.2 is considered as a small but noticeable effect, while an effect size score of 0.5 is considered to be a medium effect size, and an effect size score of 0.8 is considered to be a large effect size.

We can also calculate the confidence interval for the difference between two means in any scenario where we compute the t-observed score. The formula to compute the confidence interval for a mean difference for two independent samples is shown here. The elements for this calculation are produced from the SAS output using PROC UNIVARIATE or PROC MEANS and substituted into the equation. The formula for confidence intervals in a t-test for independent samples is:

[latex]\left(\overline{x_{1}} - \overline{x_{2}}\right) \pm t_{0.05} \times \sqrt{\frac{s^2_{1}}{n_{1}}+ \frac{s^2_{2}}{n_{2}}}[/latex].

Where the $t_{0.05}$ is the critical value for **t** for the degrees of freedom in the study. Considering that we have 50 cases in our sample our degrees of freedom value will be: df = $(n_1 - 1) + (n_2 - 1)$ and the critical value of $t_{0.05} = 2.01$

The decision rule concerning a confidence interval in a t-test for independent samples is to determine if the range of the confidence interval from the lowest value to the highest value includes 0. If 0 is included in the range then we accept the null hypothesis that the mean for group 1 = the mean for group 2.

30.4 The degrees of freedom and critical values

- The term *degrees of freedom*, represents the number of scores within a set of scores that are free to vary, and the number of scores that must be fixed, in order to compute a result.

For example: Consider that the average age for a group of five students is 22. Therefore, the mean of **22** is the outcome or the result. Now consider what each student's age must be in order to calculate an average age of 22 [latex]\rightarrow\;[/latex] the outcome (*aka the result*).

In other words within the set of scores that we observed, what scores are required to make up the set of scores, in order to compute the outcome (result) that we observed?

Let's work through the concept with the following example:

Identify a sample of five students and then decide to ask each student to report their age. The first student tells you that she is **28** years old, but the second student said that he is **12** years old! Student number 3 reports that he is **129** years old and the fourth student suggests that her age is a whopping **-10**! No doubt you are realizing that their ages are totally fictitious but they are what they are, and the outcome for the average age remains at 22. Your challenge is now to determine what the age of the fifth student is in order to ensure that the overall group age is 22. That is, the age of the student cannot be **a free choice** but must be **a fixed age** in order for you to calculate the mean age for the group equal to 22.

In this example, you have some known information. You began with the outcome as the mean age of the group equal to 22, and you also have a set of 4 age values that were reported for the group (28, 12, 129, and -10). Since you know the real average age of the five students is 22. You decide to play along with the group, and you realize that you can use simple arithmetic factoring to solve the unknown value of the age for the 5th student.

[latex]\overline{x} = {\Sigma{\left(x_{1}\; +\; x_{2}\;+\; ... \;+\; x_{n}\right)}\over{n} }[/latex] [latex]= {\Sigma{\left(28\; +\; 12\;+\; 129 \;+\; (-10)\;+\; x_{5}\right)}\over{n} }[/latex]. Work through the computation of the numerator, and then factor out the [latex]x_{5}[/latex] term by multiplying each side by the denominator value of 5, and then subtracting 159 from both sides as shown here.

[latex]22 = {159 + x_{5}\over{5}}[/latex]	[latex]22 \times{5} = {\bcancel{5} \times \left(159 +x_{5}\right)\over{\bcancel{5}}}[/latex]	[latex]110 = \left(159 + x_{5}\right)[/latex] [latex]110 - 159 = \left(x_{5}\right)[/latex] [latex]-49 = \left(x_{5}\right)[/latex]

Since you know the real average age of the five students is 22. You determine that the age for Student #5 MUST BE (-49). The degrees of freedom is a term that represents the number of scores within a set of scores that are free to vary, and the number of scores that must be fixed, in order to compute a result. In your data set, 4 of the ages were free to vary, but the age for Student #5 had to be fixed at (-49) in order to compute the group's known mean age of 22.

A simple formula for degrees of freedom is then to consider that the degrees of freedom equal the number of scores that are free to vary minus the number of scores that are fixed within a set of scores.

Why compute the DEGREES OF FREEDOM?

The degrees of freedom term is used to determine the critical value of a statistic given the research design and the sample size. In other words, the value from the probability distribution function for all possible scores of the statistic of interest under an estimate of the probability for a given research scenario. The statistic's critical score is related to the probability that the null hypothesis is true. In the case of using the t-test, the null hypothesis is a derivative of the **mean observed** being equal to the **mean expected**. The critical value can change for every application of a statistical computation because it depends on the size of each sample and the probability level set by the researcher to establish whether or not to accept or reject the null hypothesis (i.e. the level of significance).

All degrees of freedom computations can be derived from the following formula: **df = (n – 1).** Below is a table of degrees of freedom formulae for different types of t-test designs. Notice that the formula differs to enable freedom of at least 1 measure within each array (set) of data.

Table 30.3 Degrees of freedom computations for different t-test designs

TERM and Null Hypothesis	FORMULA	EXAMPLE
Student's t-test H0: *sample mean* = 0	df = (n − 1)	Given a sample size of 10, degrees of freedom is: df=(10 − 1), df = 9
Pair-wise t-test H0: *preTestmean* = postTestmean	df = (npairs − 1)	Given a sample size of 10 pairs, degrees of freedom is: df=(10 − 1), df = 9
t-test for two group comparisons—equal n in each group[1] H0: *grp1mean* = *grp*2mean	(n1 + n2) − 2	Given that the sample sizes are n1= 10 and n2=10, degrees of freedom is: df=(10 + 10) − 2, df = 18
t-test for two group comparisons—unequal n in each group H0: *grp1mean* = *grp*2mean	(n1− 1) + (n2 − 1)	Given that the sample sizes are n1= 8 and n2=13, degrees of freedom is: df=(8-1) + (13-1), df = 19

[1] assuming equal variances within the two sample distributions

[2] Sullivan, G , and Feinn, R., Using effect size, or why the p Value isn't enough, Journal of Graduate Medical Education, September 2012, 279-282.

31. The One Way Analysis of Variance and Post Hoc Tests

Learner Outcomes

After reading this chapter you should be able to:

- Compute the significance of the difference between three or more sample means using the one way analysis of variance test
- Compute the post-hoc pairwise comparison between sample means when the F statistic is significant
- Write a SAS program to compute and identify the important elements of the output for the computation of the one-way analysis of variance

31.1 Analysis of Variance

We use the analysis of variance (ANOVA) to evaluate tests of hypothesis for differences between two or more treatments. In computing the significant difference between multiple means using the One Way Analysis of Variance – ANOVA along with post hoc tests we compare estimates of variance between and within each of the sample groups. The purpose of the ANOVA is to decide whether the differences in the estimates of variance between the samples are simply due to random error (sampling errors) or whether there are systematic treatment effects that have caused scores in one group to differ from scores in another.

In the one-way analysis of variance (ANOVA), we are only comparing the mean scores from three or more samples on one dimension, such as between each group, as shown in the following diagram.

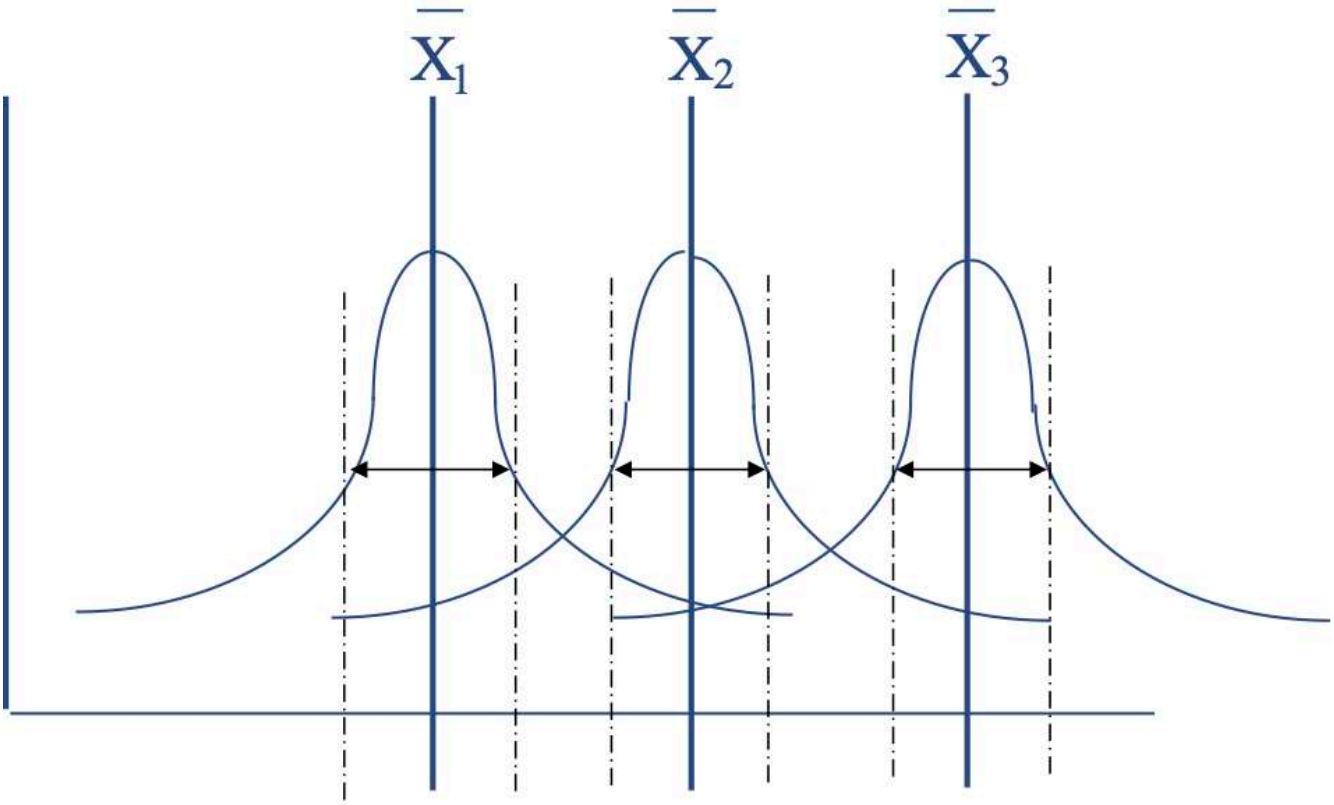


Illustration of the Comparison of Means in the One-way ANOVA

The one-way ANOVA evaluates the variance between samples, and tests the null hypothesis: [latex]H_{0}: \mu_{1} = \mu_{2} = \mu_{k}[/latex], or that [latex]H_{0}: \mu_{1} - \mu_{2} - \mu_{k} = 0[/latex]. The statistic that we use to test this null hypothesis is an F test (producing an F value) and is based on the ratio of the variance between the samples divided by the variance within the samples, as shown here: **F = variance between samples / variance within samples**

Consider an experiment intended to compare the effects of two independent drugs and placebo on an individual's reaction time. Here we can define reaction time as the speed at which an individual demonstrates a response to a given stimulus).

This experimental design would require us to create three groups of participants, in which each group consists of individuals that are randomly selected from a population and randomly distributed to one of the groups (drug group 1, drug group 2, or placebo).

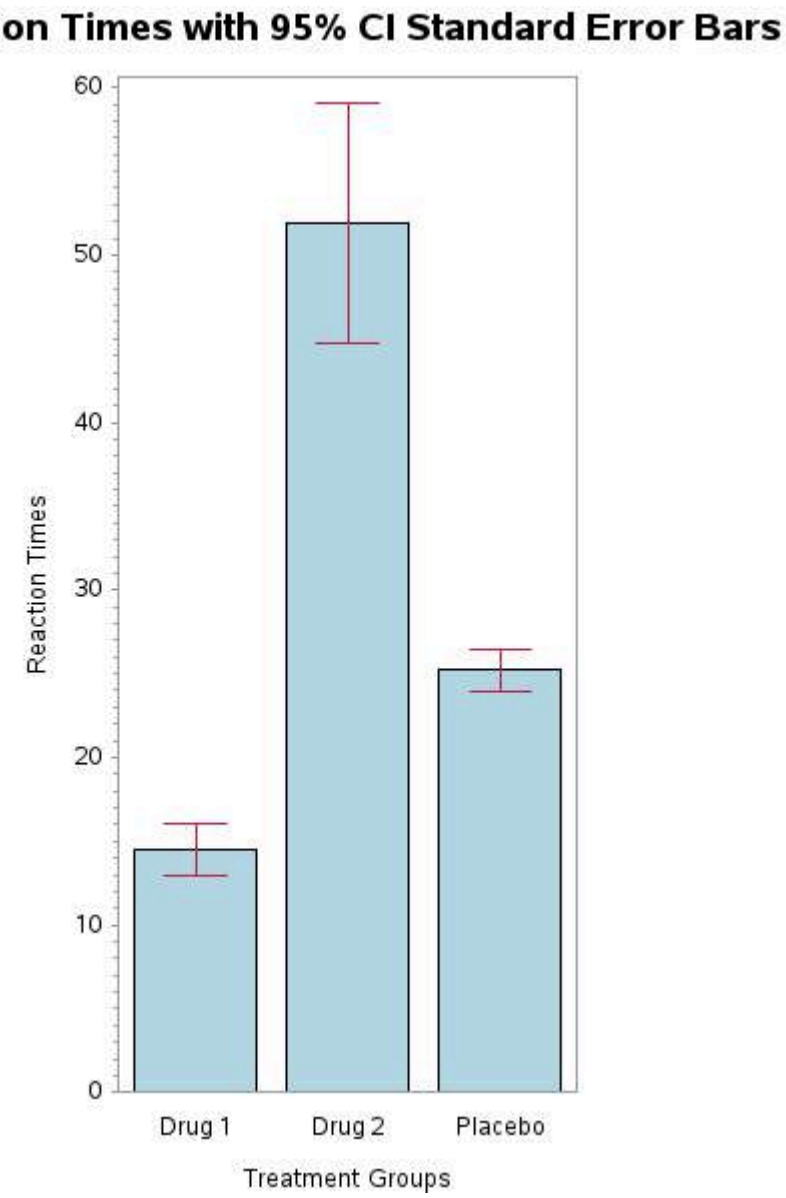
The purpose of the experiment would be to determine if there is a significant difference in the measured reaction times for the individuals in each of the drug groups versus individuals in the placebo group.

The data sets are based on the observed reaction times measured for individuals from the two independent drugs and the group that received the placebo. The data are arranged to have 3 groups of 10 individuals per group, wherein each individual within each group provides reaction time measurements. The hypothetical data for this experiment are shown in Table 31.1, below.

Table 31.1 Reaction Time Data for Each Experimental Group

Reaction time scores recorded in seconds	Reaction time scores recorded in seconds	Reaction time scores recorded in seconds
Group 1: Drug A	Group 2: Drug B	Group 3: Placebo
12	45	25
15	54	26
16	39	28
13	65	22
14	34	26
15	63	27
17	55	23
11	51	26
18	53	24
14	60	25
$Sum_1 = 145$	$Sum_2 = 519$	$Sum_3 = 252$

A vertical bar chart of the data in Table 31.1 shows an obvious difference in the mean reaction times for each treatment group ($Drug_1$, $Drug_2$, and Placebo). Given the observed difference between means, it is appropriate to test the statistical significance of this difference using a one-way ANOVA.



31.2 Computing the ANOVA by hand

In this experiment, we can calculate the ANOVA by hand to illustrate the essential calculations that are used to produce the **F statistic**. After we compute the $F_{observed}$ we next compare the F observed value to the $F_{critical}$ value of 3.35 to determine if we accept or reject H_0. The $F_{critical}$ value is derived from a table of critical values which we can retrieve from the Internet, and is based on the number of elements (participants) in each group being compared [latex]\rightarrow[/latex] the degrees of freedom.

We begin our calculations by computing the grand mean. The grand mean is the average for all numbers within the entire set of numbers, and is therefore the sum of all scores divided by N (the number of scores) as shown here:

[latex]\overline{X}= \frac{\Sigma{x}_{i}} {N}[/latex] = ((12 + 15 + 16 + 13 + 14 + 15 + 17 + 11 + 18 + 14 + 45 + 54 + 39 + 65 + 34 + 63 + 55 + 51 + 53 + 60 + 25 + 26 + 28 + 22 + 26 + 27 + 23 + 26 + 24 + 25) / N)

[latex]\overline{X}= \frac{\Sigma{x}_{i}} {N}[/latex] =((916)/30) = 30.53

Once we calculate the grand mean, then we are ready to calculate the specific elements of the one way ANOVA. Our next step is then to calculate the Sum of Squares Total ([latex]SS_{total}[/latex]).

The [latex]SS_{total}[/latex] is calculated by calculating the squared difference of each score from the grand mean and summing the difference scores, as shown here:

[latex]SS_{total}= \Sigma\left(x_{ij} - \overline{X}\right)^2[/latex] = ((12-30.53)^2+(15-30.53)^2+(16-30.53)^2+(13-30.53)^2+(14-30.53)^2+(15-30.53)^2+(17-30.53)^2+(11-30.53)^2+(18-30.53)^2+(14-30.53)^2+(45-30.53)^2+(54-30.53)^2+(39-30.53)^2+(65-30.53)^2+(34-30.53)^2+(63-30.53)^2+(55-30.53)^2+(51-30.53)^2+(53-30.53)^2+(60-30.53)^2+(25-30.53)^2+(26-30.53)^2+(28-30.53)^2+(22-30.53)^2+(26-30.53)^2+(27-30.53)^2+(23-30.53)^2+(26-30.53)^2+(24-30.53)^2+(25-30.53)^2

[latex]SS_{total}= \Sigma\left(x_{ij} - \overline{X}\right)^2[/latex] = 343.36 + 241.18 + 211.12 + 307.3 + 273.24 + 241.18 + 183.06 + 381.42 + 157 + 273.24 + 209.38 + 550.84 + 71.74 + 1188.18 + 12.04 + 1054.3 + 598.78 + 419.02 + 504.9 + 868.48 + 30.58 + 20.52 + 6.4 + 72.76 + 20.52 + 12.46 + 56.7 + 20.52 + 42.64 + 30.58

[latex]SS_{total}= \Sigma\left(x_{ij} - \overline{X}\right)^2[/latex] = 8403.44

Our next step is to compare the [latex]SS_{between}[/latex] terms (sometimes referred to as the sum of squares for the treatment term), which is a comparison of the variance in Group 1 against the variance in Group 2 and concomitantly agaist the variance in Group 3.

This is represented as: [latex]S_{1}^2 + S_{2}^2 +S_{3}^2[/latex]

To calculate the between groups sums of squares we subtract each of the individual means [latex]\left(\overline{x}_{j}\right)[/latex] from the grand mean [latex]\left(\overline{X}\right)[/latex] and square the difference scores [latex]\left(\Delta\right)^2[/latex]. We then multiply the squared difference score by the number of participants in each group,) as shown here:

[latex]SS_{between} = \Sigma{n}_{j} \left(\overline{xj} - \overline{X}\right)^2[/latex]

[latex]SS_{between} = \left(10 \times \left(14.5-30.53\right)^2 + 10 \times \left(51.9-30.53\right)^2 + 10 \times \left(25.2 – 30.53\right)^2\right)[/latex]

[latex]SS_{between}[/latex] = (2569.60 + 4566.80 + 284.10) = 7420.49

The error term is calculated from the squared deviations of each of the scores from the mean score within each group. The term [latex]SS_{error}[/latex] is also referred to as the [latex]SS_{within}[/latex] and is shown here.

[latex]SS_{within}[/latex] = \left(\Sigma{x}_{ij} – \overline{x}_{j} \right)^2[/latex]

[latex]SS_{within} \textit{(Group 1)}[/latex] = (12-14.5)^2 + (15-14.5)^2 + (16-14.5)^2 + (13-14.5)^2 + (14-14.5)^2 + (15-14.5)^2 + (17-14.5)^2 + (11-14.5)^2 + (18-14.5)^2 + (14-14.5)^2

[latex]SS_{within} \textit{(Group 1)}[/latex] = 6.25 + 0.25 + 2.25 + 2.25 + 0.25 + 0.25 + 6.25 + 12.25 + 12.25 + 0.25

[latex]SS_{within} \textit{(Group 1)}[/latex] = 42.25

[latex]SS_{within} \textit{(Group 2)}[/latex] = (45-51.9)^2 + (54-51.9)^2 + (39-51.9)^2 + (65-51.9)^2 + (34-51.9)^2 + (63-51.9)^2 + (55-51.9)^2 + (51-51.9)^2 + (53-51.9)^2 + (60-51.9)^2

[latex]SS_{within} \textit{(Group 2)}[/latex] = 47.61 + 4.41+ 166.41+ 171.61 + 320.41 + 123.21 + 9.61 + 0.81 + 1.21 + 65.61

[latex]SS_{within} \textit{(Group 2)}[/latex] = 910.90

[latex]SS_{within} \textit{(Group 3)}[/latex] = (25-25.2)^2 + (26-25.2)^2 + (28-25.2)^2 + (22-25.2)^2 + (26-25.2)^2 + (27-25.2)^2 + (23-25.2)^2 + (26-25.2)^2 + (24-25.2)^2 + (25-25.2)^2

[latex]SS_{within} \textit{(Group 3)}[/latex] = 0.04 + 0.64 + 7.84 + 10.24 + 0.64 + 3.24 + 4.84 + 0.64 + 1.44 + 0.04

[latex]SS_{within} \textit{(Group 3)}[/latex] = 29.6

[latex]SS_{within}[/latex] = (42.25 + 910.90 + 29.6) = 983

The **mean square term** for the analysis of variance, also known as the ***mean square between*** (**MS_b**), is calculated from the ***sum of squares between*** divided by the degrees of freedom for that term, as shown below. In this example there were 3 groups therefore the degrees of freedom between groups is **(k –1= 3 –1 = 2).**

[latex]MS_{b} = \frac {7420.47} {2}[/latex] = 3710.23

Likewise, the ***mean square for the error term***, also referred to as the **mean square within** (**MSw**) is calculated by dividing the sum of squares from the error statement [latex]SS_{within}[/latex] = 983, shown above, by the degrees of freedom from the error statement (df_{error} = df_w = 30 participants – 3 groups = 27). The degrees of freedom for the error

term is based on the construct that in this example there were 30 participants in total and three groups, therefore we subtract 1 participant per group from the grand total. Therefore, our degrees of freedom for the MS_w term is N-k = 30-3 = 27.

[latex]MS_{w} = \frac {983.00} {27}[/latex] = 36.41

The F- statistic is then calculated by dividing the **mean square between** by the **mean square within**, as shown here.

[latex]F_{observed} = \frac {MS_{b}} {MS{_w}} = \frac {3710.23} {36.41}[/latex] = 101.91

In this experiment we computed the F statistic using a one-way analysis of variance—ANOVA. In this scenario we calculated an observed F statistic, also referred to as the [latex]F_{observed}[/latex] of 101.91. Below we explain how to compare an [latex]F_{observed}[/latex] score to an [latex]F_{expected}[/latex], or an [latex]F_{critical}[/latex] value.

31.3 Creating the SAS PROC ANOVA program

The task of applying the ANOVA to these data is to test the null hypothesis that the means of each group are equal. We create a decision rule to evaluate in which we state that: If F observed is greater than F critical then we will reject H_0; else if F observed is less than or equal to F critical then we will accept H_0 (the null hypothesis).

In the following example, we begin by evaluating the observed or perceived difference between the means using the F statistic from the ANOVA, and the F statistic generated by the PROC GLM procedure of SAS, for the data presented above. The PROC ANOVA and the PROC GLM produce similar output, however, we use the PROC ANOVA term when the number of observations within each group is the same, and we use the PROC GLM procedure when there are a different number of observations in each of the groups being compared. Regardless of whether you choose the PROC ANOVA or the PROC GLM, they both use an F test to evaluate the null hypothesis that there is no difference between means.

SAS for OneWay Anova

```
   OPTIONS PAGESIZE=55 LINESIZE=80 CENTER DATE;
DATA ONEWAY1;
INPUT ID GROUP SCORE @@;
DATALINES;
001 01 12  002 01 15  003 01 16  004 01 13  005 01 14  006 01 15
007 01 17  008 01 11  009 01 18  010 01 14  011 02 45  012 02 54
013 02 39  014 02 65  015 02 34  016 02 63  017 02 55  018 02 51
019 02 53  020 02 60  021 03 25  022 03 26  023 03 28  024 03 22
025 03 26  026 03 27  027 03 23  028 03 26  029 03 24  030 03 25
;
PROC SORT DATA=ONEWAY1; BY GROUP;
PROC ANOVA; CLASS GROUP; MODEL SCORE = GROUP;
MEANS GROUP / tukey scheffe;
TITLE 'ONE WAY ANOVA FOR SCORE BY GROUP WITH POST HOC TESTS';
RUN;
PROC GLM; CLASS GROUP;MODEL SCORE= GROUP;
LSMEANS GROUP /tdiff adjust=scheffe ;
TITLE 'GLM for Dep Var = SCORE by Group with Post Hoc';
```

```
run;
```

31.4 Annotated output from PROC ANOVA

In this scenario, we used the PROC ANOVA procedure to generate the following output. Here we had three groups of 10 individuals wherein we suggested that the members of the groups were randomly selected and randomly assigned to each group. In this way, we were attempting to eliminate any apriori bias that may have influenced the results.

TABLE 31.2 SAS Output for the Analysis of Variance Procedure

Class	Levels	Values
group	3	1 2 3

Number of Observations Used	30

The next table summarizes the stepwise calculations that generated the F statistic. The model statement, presented in the ANOVA SUMMARY table below refers to the comparison of the distributed variance within each of the groups. In this table, we see that the comparison of the means across the three groups produced a MEAN SQUARE score of **3710.23** which matches our calculations done by hand and presented above.

The MEAN SQUARE score is calculated by dividing the sum of squares **(7420.46)** by the degrees of freedom, which in this instance is **2**.

The **error statement** in the summary table refers to the comparison of the variance within each group. The MEAN SQUARE from the between means term is divided by the MEANS SQUARE from the error term **36.41** to produce the F VALUE *a.k.a.* $F_{observed}$ or *the F-statistic* of **101.91**.

Table 31.3 SAS Output for the Analysis of Variance Procedure Summary Table

Dependent Variable: score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model (between)	2	7420.466667	3710.233333	101.91	< 0.0001
Error (within)	27	983.000000	36.407407		
Corrected Total	29	8403.466667			

R-Square	Coeff Var	Root MSE	score Mean
0.883024	19.76153	6.033855	30.53333

Source	DF	Anova SS	Mean Square	F Value	Pr > F
group	2	7420.466667	3710.233333	101.91	< 0.0001

31.5 Evaluating $F_{observed}$ Against the Null Hypothesis

Notice in the output presented above – the probability of producing this F value is less than 0.0001. Typically, by conven-

tion in evaluating the null hypothesis, we evaluate a probability or p-value for the statistic observed against the probability of the statistic expected. In most cases, we expect that the probability associated with the statistic expected has a value of 1/20 or ($p < 0.05$). Further, in order to establish our decision rule, we consider that if the p-value associated with the statistic that we calculate is less than the p-value of the statistic that we expect (which has a probability of 0.05) then we say that there is a significant difference between the distributed variances in each group.

Using our decision rule we can compare the F statistic that we observed against the F critical that we expect for the degrees of freedom of df=(k-1, N-3), at $p<0.05$. Here we see that the F statistic was F = 101.91 and it had a probability of ($p<0.0001$). The F critical value for this experiment with k=3 groups and a total sample of N=30 was 3.354 [latex]\rightarrow[/latex] F critical = 3.354 df=(2,27) p=0.05. Therefore, if we compare the F statistic to the F critical then we see that the F statistic is greater than the F critical and therefore we would reject the null hypothesis. Likewise, we can compare the probability of the F statistic and the probability of the F critical and we see that the probability of the F statistic ($p<0.0001$) is less than the probability of the F critical ($p<0.05$) and therefore we would reject the null hypothesis.

31.5.1 The R-Square Estimate

The output of the ANOVA summary table provides several bits of information that help us to understand the measures within the experiment. One measure that is provided is the **R-square value.** The R^2 or R-square value is an estimate of the amount of variance in the dependent variable that is explained by the independent variable. Recall that every score we measure is comprised of true score plus error. When we combine the error of every score in a distribution we approximate the variance of the scores in the distribution. When we calculate the ANOVA we are comparing the means for several groups within a distribution after correcting for the amount of variance within each of the groups in the distribution. The r-square value is calculated by dividing the SS_B term (the sum of squares between groups) by the SS_T term (the sum of square total). Applying this computation to this data set we see the following: R^2 = SSB : SST = 7420.49 : 8403.44 = 0.88[latex]\rightarrow[/latex] 88%. Therefore, we can say that 88% of the variance in the dependent variable is explained by the independent variable, which in this case is a function of the grouping variable, and therefore 88% of the variance within the dependent variable is a result of the dependent variable being calculated from three different groups. Another way to say this is that the independent variable (here being groups) predicts 88% of the variance in the dependent variable.

31.6 Determining the Location of the Difference in Means Using Post Hoc Tests or Confidence Intervals

In the calculation of the ANOVA, we simply measure that there is a significant difference between the means of the groups representing the total sample. Yet, the ANOVA cannot tell us which means are different or are causing the F statistic to be significantly different than the F expected. In order to determine which means were significantly different in the computation of the F statistic, we have several options. For example, one way is to calculate the confidence intervals for each of the means from each of the comparison groups. Over the years, several statisticians have also worked to develop equations to identify which group means within the ANOVA are driving the difference that is measured by the F test. We refer to these computations as the post hoc tests. There are several different types of post hoc tests, typically named for the author of the method, and while they are each different in their own way, they share a common approach. The commonality between post hoc tests is that they typically use a pairwise approach to compare the difference between the means that are used in the F statistic. However, unlike the computation of repeated t-tests, which

also use a pairwise comparison approach for any two means, the post hoc tests use a selective component of the shared variance for all of the groups taken from the overall F statistic.

31.6.1 Tukey and Scheffe post hoc tests

Two commonly used approaches are the Scheffe post hoc F test and the Tukey HSD[1] post hoc t-Test. As you can see, while these are two types of post hoc procedures, they are each based on different underlying distributions (i.e. F and t). In Computing the ANOVA using PROC ANOVA in SAS we can also compute the post hoc analysis using the comparison of the means from either Tukey or Scheffe or both. The computation of the Tukey HSD test is used when the groups in the sample have the same number of participants (n is the same for all groups). The statistic is based on a t-distribution and uses the following formula.

The formula for Tukey's HSD post hoc test: [latex]{\overline{x}_{1} -\overline{x}_{2} } \over{\sqrt{MS_{W}\left(\frac{1}{n_{1}}\right)}}[/latex]

Where the numerator refers to any two means in the F Statistic computation, and the MS_W term in the denominator refers to the mean square within score that is taken from the output of the one-way ANOVA summary table. Notice in the SAS output in the one-way ANOVA procedure, the MS_W term is replaced by the MS_{error} term. Although the SAS output for the Tukey HSD post hoc test was calculated using SAS and is shown below, the following table demonstrates the hand computation for the Tukey HSD post hoc test with our data set.

The important data to compute the Tukey HSD post hoc test include the mean scores in each group and the sample size within each group. Recall that the Tukey HSD post hoc test is based on equal samples in each group. In our data the number of participants per group was n=10, and the means were: group 1 = 14.5, group 2 = 51.9, group 3 = 25.2. The other bit of information required for this calculation is the MSW term, which was 36.41.

Group 1 vs Group 2	Group 1 vs Group 3	Group 2 vs Group 3
HSD = [latex]{\overline{14.5} -\overline{51.9} } \over{\sqrt{36.41\ left(\frac{1}{10}\right)}}[/latex]	HSD = [latex]{\overline{14.5} -\overline{25.2} } \over{\sqrt{36.41\ left(\frac{1}{10}\right)}}[/latex]	HSD = [latex]{\overline{51.9} -\overline{25.2} } \over{\sqrt{36.41\ left(\frac{1}{10}\right)}}[/latex]
HSD = [latex]{ \lvert-37.4\rvert\ over{1.91}} = 19.08[/latex]	HSD = [latex]{ \lvert-10.7\rvert\ over{1.91}} = 5.06[/latex]	HSD = [latex]{ 26.7\over{1.91}} = 13.97[/latex]

Computed HSD values are compared to the Tukey Studentized Range Statistic Critical value based on the number of groups (k= 3) and the degrees of freedom (n-k)= (30-3)=27. In our computation the critical value used for comparison is given in the SAS table as: Critical Value of Studentized Range = 3.51. Comparing each HSD from the rows above with the Critical value we determine which pairs of means are significantly different.

Grp1 vs Grp2	Grp1 vs Grp3	Grp3 vs Grp2
HSD: 19.08 > 3.51	HSD: 19.08 > 5.06	HSD: 13.97 > 5.06
[latex]\therefore \textit{reject } H_{1}\left({\overline{x_{1}} =\overline{x_{2}} }\right)[/latex]	[latex]\therefore \textit{reject } H_{2}\left({\overline{x_{1}} =\overline{x_{3}} }\right)[/latex]	[latex]\therefore \textit{reject } H_{3}\left({\overline{x_{3}} =\overline{x_{2}} }\right)[/latex]

SAS computations for the post hoc Tukey HSD test are shown here.

Table 31.4 SAS Output for Tukey's Studentized Range (HSD) Test with DEPENDENT VARIABLE: score

Alpha (the p value)	0.05
Error Degrees of Freedom (N-k)	27
Error Mean Square (MS_E)	36.40741
Critical Value of Studentized Range	3.50633
Minimum Significant Difference	6.6903

Tukey Grouping*	**Mean**	**N**	**group**
A	51.900	10	2
B	25.200	10	3
C	14.500	10	1

Means with the same letter are not significantly different.

Note: These results indicate that each of the means were significantly different from each other according to the Tukey Test.

The Scheffe Test

The comparison of means proposed in the Scheffe post hoc test are based on an F-distribution and uses the following two-step approach. In the first step, the absolute difference between the pairs of means is calculated. This is similar to the procedure we used above with the Tukey HSD comparison.

Group 1 vs Group 2	Group 1 vs Group 3	Group 2 vs Group 3
[latex]{\overline{14.5} -\overline{51.9} } \over{\sqrt{36.41\ left(\frac{1}{10}\right)}}[/latex]	[latex]{\overline{14.5} -\overline{25.2} } \over{\sqrt{36.41\ left(\frac{1}{10}\right)}}[/latex]	[latex]{\overline{51.9} -\overline{25.2} } \over{\sqrt{36.41\ left(\frac{1}{10}\right)}}[/latex]
[latex]{ \lvert-37.4\rvert\over{1.91}} = 19.08[/latex]	[latex]{ \lvert-10.7\rvert\over{1.91}} = 5.06[/latex]	[latex]{ 26.7\over{1.91}} = 13.97[/latex]

In the second step of the Scheffe test the minimum significant difference is computed for each pair of means using the following formula where F critical = 3.35 and the MS_{error} = 36.41 are taken directly from the SAS output. Note also that since n_k is 10 in each comparison then the Scheffe minimum significance score will be constant for all comparisons.

Scheffe minimum significant difference:[latex]\sqrt{\left(k - 1\right) \times \textit{F critical} \times \textit{MSE} \times \left(\frac{1}{n_{1}} + \frac{1}{n_{2}} \right) }[/latex]

Scheffe$_{MSD}$: [latex]\sqrt{\left(3 - 1\right) \times 3.35 \times 36.41 \times \left(\frac{1}{10} + \frac{1}{10} \right) }[/latex]

Scheffe$_{MSD}$: **6.98 [latex]\leftarrow[/latex]** matches value given in SAS output

The researcher then compares the absolute difference between means against the Scheffe minimum significant difference term for that pair of means and if the absolute difference term is larger than the Scheffe minimum significant difference term then the null hypothesis is rejected..

[latex]\left({\lvert\overline{14.5} -\overline{51.9} }\rvert\right)[/latex]	[latex]\left({\lvert\overline{14.5} -\overline{25.2} }\rvert\right)[/latex]	[latex]\left({\lvert\overline{51.9} -\overline{25.2} }\rvert\right)[/latex]
= 37.4 > 6.98	10.7 > 6.98	26.7 > 6.98
[latex]\therefore \textit{reject } H_{1}\left({\overline{x_{1}} =\overline{x_{2}} }\right)[/latex]	[latex]\therefore \textit{reject } H_{2}\left({\overline{x_{1}} =\overline{x_{3}} }\right)[/latex]	[latex]\therefore \textit{reject } H_{3}\left({\overline{x_{3}} =\overline{x_{2}} }\right)[/latex]

Table 31.5 SAS Output for Scheffe's Test for the DEPENDENT VARIABLE: score

Alpha	0.05
Error Degrees of Freedom	27
Error Mean Square	36.40741
Critical Value of F	3.35413
Minimum Significant Difference	6.989

Scheffe Grouping	Mean	N	group
A	51.900	10	2
B	25.200	10	3
C	14.500	10	1

Means with the same letter are not significantly different.

Notice in the results produced for both the Tukey test and the Scheffe test there is a statement referring to the control of the Type I and Type II errors. These errors refer to the statistical errors – and are associated with the researcher's decision related to accepting or rejecting the null hypothesis. In other words, despite that you have used a powerful program like SAS to compute the statistical tests and produce the necessary output, there is a chance that the statistics that you produce may not provide the information that you require to make a decision about your null hypothesis.

> Consider a simple null hypothesis like: [latex]H_{0}: \left({\overline{x_{1}} =\overline{x_{2}} }\right)[/latex]. In statistics, a Type I error occurs when the researcher fails to accept the null hypothesis when in fact it is true. That is, the researcher rejected the null hypothesis because although the statistic may be correct the information upon which the statistic is computed may be misleading. Similarly, a Type II error occurs when the researcher fails to reject the null hypothesis even though the evidence supports that the null hypothesis should be rejected. As noted in the post hoc tables, each of the pairwise tests presented by SAS indicates the extent to which the test is controlling either the Type I or Type II error, or both.

Using Confidence Intervals to Identify the Significance of the Difference in the Group Means

Once a significant F statistic is calculated then the next step is to determine which means are causing the overall difference. **The confidence interval** can be used with the ANOVA to determine which means are different when comparing more than two means. In computing the **confidence interval** for the ANOVA we first compute the standard error for the **mean square within (MS_W) or mean square error (MS_E)** using the following formula:

[latex]s.e. = \sqrt{MS_{w} \over{n_{k}}}[/latex]

Here **(MS_W)** which is denoted as the mean square within groups, was 36.41, and again in our example, each group had an equal sample size of 10 individuals so that $n_k = 10$. Therefore, the calculation of the standard error is:

[latex]s.e. = \sqrt{36.41 \over{10}} = 1.91[/latex]

The 95% confidence interval is then determined individually for the mean in each group. So that for group 1 the mean is (14.5) and the critical value for the confidence interval is based on an approximation to the t distribution using the following formula: [latex]\left(\alpha, n - 1\right) = \left(0.05, 10 - 1\right) = \left(0.05, 9\right) = 2.262[/latex]

So that the confidence interval for group 1 is: 14.5 ± 2.262 * 1.908 which provides a range from a lower limit of: 14.5 – 4.32 to an upper limit of 14.5 + 4.32. The interval is then: 10.18 [latex]\leftarrow\;\rightarrow[/latex] 18.82.

We can then compare the means from the other groups to this interval and determine if they fall within or beyond the interval range. In this experiment, we have means of 51.9 and 25.2 in the other two groups, both of which fall outside the range of 10.18 to 18.82. The confidence interval for group 2 is: 51.9 ± 2.262 * 1.908 = 51.9 ± 4.32 and the range is: 47.58 [latex]\leftarrow\;\rightarrow[/latex] 56.22; and the confidence interval for group 3 is: 25.2± 2.262 * 1.908 = 25.2 ± 4.32 and the range is: 20.88 [latex]\leftarrow\;\rightarrow[/latex] 29.52.

Calculating the confidence intervals for each group provides the opportunity to compare the means of the other groups and determine if the values fall outside the range of the confidence interval. When the mean for a group falls outside a confidence interval range of a comparison group then we can say that there is a significant difference between the groups, and again use this information to support our decision to reject the null hypothesis.

31.7 A word about Bonferroni Correction Factor

As noted previously, the F statistic observed from the ANOVA provides a ratio of the relationship between the variance between groups divided by the variance within groups. The resulting value is then compared to an expected F statistic – referred to as the F critical score. The F critical score is derived from the sample size and the probability level () that is associated with decisions related to the null hypothesis. Generally, the decision rule for the ANOVA considers that if the F statistic is greater than the F critical score, then we can say that there is an overall significant difference between the groups upon which the analysis is based.

Post Hoc tests are used as a follow-up to the ANOVA to determine which pairwise comparison of means contributes to the overall significant difference that is observed in the computation of the F statistic.

One difficulty that arises in evaluating multiple pairwise means is that as we increase the number of comparisons, so too do we increase the chance that a significant difference will be found. Realizing that in any research design there could be any number of relevant pairwise comparisons Carlo Bonferroni (b. 1892 – d. 1960) suggested that in order to control for spurious significant differences associated with multiple pairwise comparisons, one should adjust the probability level () by dividing the probability by the number of comparisons made.

An example of a Bonferroni application is shown here:

In the examples shown above for both the Tukey and the Scheffe post hoc tests, there were 3 hypotheses evaluated in each test. An [latex]\alpha\;[/latex]= 0.05 was selected for the decision rule regarding the null hypothesis. The Bonferroni correction factor is represented as: [latex]\left(\alpha \over{\textit{n tests}} \right)[/latex]. In this situation, we may be very conservative in our evaluation of the null hypothesis by considering that the [latex]\alpha[/latex]= 0.05 should be changed from 0.05 to [latex]\left(0.05 \over{3} \right) = 0.017[/latex].

[1] The term HSD refers to the Honestly Significant Difference

[2] Note that the $F_{critical}$ = 3.35 and the MS error = 36.41 are taken directly from the SAS output

32. Research Design Applications with PROC GLM

Learner Outcomes

After reading this chapter you should be able to:

- Compute the significance of the difference between three or more sample means using PROC GLM for the one-way analysis of variance test
- Compute the significance of the association between an outcome and one or several predictors using PROC GLM as a linear regression model
- Compute the post hoc comparison between sample means when the F statistic is significant using posthoc analysis procedures (in either ANOVA applications or linear regression applications)
-

INTRODUCTION TO GENERAL LINEAR MODELS IN SAS

A univariate general linear model is defined as a statistical model in which a dependent variable is modeled in relation to a set of predictor variables. The predictor variables can be categorical independent variables with multiple levels, or they can be a continuous variable, or the predictor variables can be a combination of categorical and continuous independent variables. In the application of statistical processing for research designs, where the dependent variable is a continuous scaled score, and the independent variables are categorically scored, the researcher can use either the analysis of variance or a general linear model.

In SAS, the F statistic can be computed with either the PROC ANOVA procedures described previously or with the PROC GLM procedure with similar post-analytic processes to establish not only the significance of the main effects but also of the characteristics of the distribution, like measures of normality and equality of variance, there are limitations to the application of the PROC ANOVA which suggest that the use of PROC GLM is more appropriate. For example, the PROC GLM procedure is preferable to PROC ANOVA when using unbalanced comparison groups, when combining categorical and continuous predictors as in an analysis of covariance, and when attempting to evaluate the dependent measure using complex interactions as in nested designs.

In this chapter, we will explore the SAS application of the PROC GLM procedures to evaluate the F statistic represented by the statement: F = variance between samples divided by the variance within samples. Next, we will explore the relationship between the outcome and predictor variables based on the concept that the dependent variable = independent variable ± error, which we can represent algebraically as: [latex]Y_{ij} = \beta_{0} \pm \beta_{i}X_{i} + \epsilon[/latex]

Extending from this General Linear Model (GLM) approach, we will introduce the General Linear **Mixed** Model, which we will analyze with the **PROC MIXED** application, which adds the following parameter [latex]U_{i}[/latex] into the General Linear Model Equation. This parameter represents the random effect in the model. [latex]Y_{ij} = \beta_{0} \pm \beta_{i}X_{i} \pm U_{i} + \epsilon[/latex]

Applying PROC GLM to evaluate a one-way ANOVA design.

The following describes a 12 week experiment in which researchers were interested in the effects of coffee consumption on resting systolic blood pressure for a sample of healthy male participants. The study participants were randomly selected from the total sample of volunteers and randomly allocated into three groups. Group 1 was comprised of 20

individuals that were asked to consume a total of 2000 ml of coffee each morning of the 12-week program between the hours of 6 and 8 am. Group 2 was comprised of 20 individuals that were asked to consume a total of 2000 ml of de-caffeinated coffee each morning of the 12-week program between the hours of 6 and 8 am, and Group 3 was comprised of 20 individuals that were asked to consume a total of 2000 ml of hot water with no additive each morning of the 12-week program between the hours of 6 and 8 am. Resting systolic blood pressure measures were taken on day 84 and recorded in the following table. The dependent variable was then determined to be the systolic resting blood pressure on day 84. The raw data and SAS code are shown below:

Group 1 – caffeinated coffee	Group 2 – de-caffeinated coffee	Group 3 – Placebo
Systolic Blood Pressure (mmHg)	Systolic Blood Pressure (mmHg)	Systolic Blood Pressure (mmHg)
134	115	125
152	114	126
161	119	128
139	115	122
149	114	126
158	113	117
167	115	113
151	111	116
148	123	114
144	110	115
124	115	129
122	116	116
121	113	118
129	119	112
129	111	116
128	112	127
127	110	123
131	115	126
128	111	124
124	114	125

```
   options pagesize=55 linesize=120 center date;
data glm1;
Title 'GLM analysis of Systolic Blood Pressure Data';
input id 1-2 @4 grp sysbp;
datalines;
134 115 125
152 114 126
161 119 128
139 115 122
149 114 126
```

```
158 113 117
167 115 113
151 111 116
148 123 114
144 110 115
124 115 129
122 116 116
121 113 118
129 119 112
129 111 116
128 112 127
127 110 123
131 115 126
128 111 124
124 114 125
;
proc sort data=glm1; by id;
proc glm;
class grp; model sysbp = grp;
run;
```

The output from this SAS Program is explained below.

GLM analysis of Systolic Blood Pressure Data using Systolic Blood Pressure (SYSBP) as the Dependent Variable

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6169.23333	3084.61667	37.57	<.0001
Error	57	4679.75000	82.10088		
Corrected Total	59	10848.98333			

R-Square	Coeff Var	Root MSE	sysbp Mean
0.568646	7.278849	9.060953	124.4833

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grp	2	6169.233333	3084.616667	37.57	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grp	2	6169.233333	3084.616667	37.57	<.0001

The comparison of means across groups was analyzed using the SAS code **lsmeans grp/ adjust= scheffe;** as shown here.

GLM analysis of Systolic Blood Pressure Data

The GLM Procedure using Least Squares Means Adjustment for Multiple Comparisons: Scheffe

grp	sysbp LSMEAN	LSMEAN Number
1	138.300000	1
2	114.250000	2
3	120.900000	3

Least Squares Means for effect grp
Pr > |t| for H0: LSMean(i)=LSMean(j) Dependent Variable: sysbp

i/j	1	2	3
1		<.0001	<.0001
2	<.0001		0.0763
3	<.0001	0.0763	

means grp /hovtest welch tukey scheffe;

GLM analysis of Systolic Blood Pressure Data- Main Effects Analysis

Levene's Test for Homogeneity of sysbp Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
grp	2	406309	203155	15.59	<.0001
Error	57	742833	13032.2		

Welch's ANOVA for sysbp

Source	DF	F Value	Pr > F
grp	2.0000	33.35	<.0001
Error	32.1316		

GLM analysis of Systolic Blood Pressure Data with the Post Hoc t Tests (LSD) for sysbp

Note: This test controls the Type I comparison wise error rate, not the experiment wise error rate.

Alpha	0.05
Error Degrees of Freedom	57
Error Mean Square	82.10088
Critical Value of t	2.00247
Least Significant Difference	5.7377

Means with the same letter are not significantly different.

t Grouping	Mean	N	grp
A	138.300	20	1
B	120.900	20	3
C	114.250	20	2

GLM analysis of Systolic Blood Pressure Data with the Tukey's Studentized Range (HSD) Test for sysbp

Note: This test controls the Type I experiment-wise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	57
Error Mean Square	82.10088
Critical Value of Studentized Range	3.40311
Minimum Significant Difference	6.895

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	grp
A	138.300	20	1
B	120.900	20	3
B	114.250	20	2

GLM analysis of Systolic Blood Pressure Data with the Scheffe's Test for sysbp

Note: This test controls the Type I experiment-wise error rate.

Alpha	0.05
Error Degrees of Freedom	57
Error Mean Square	82.10088
Critical Value of F	3.15884
Minimum Significant Difference	7.202

Means with the same letter are not significantly different.

Scheffe Grouping	Mean	N	grp
A	138.300	20	1
B	120.900	20	3
B	114.250	20	2

If we rerun the analysis with the class statement removed we can generate the coefficients for the independent variables.

```
  proc glm ;
model sysbp = grp;
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	141.8833333	3.96644269	35.77	<.0001
grp	-8.7000000	1.83610618	-4.74	<.0001

Adding A Second Grouping Factor To a GLM Model

Consider the analysis we used in the PROC ANOVA computations used in Chapter 9, where we were interested in evaluating the effects of a one-hour activity break into the workday, believing that such an opportunity could reduce the resting heart rates of the participants and thereby lead to a healthier workforce.

You will recall that the research design began with 66 participants that were randomly selected from a sample of

employees within the company, and randomly allocated to one of three treatment groups. In the following analysis, we used PROC GLM and the post hoc procedure LSMEANS to evaluate the cell-wise interaction component to evaluate the individual cell means between the treatment levels (walking versus dancing versus book reading), for each level of sex (males versus females).

```
PROC glm data=anova2x3;
title 'Using PROCGLM to determine interaction effect ';
class sex group ;
model hrchange =sex group sex*group;
lsmeans sex*group/ diff;
run;
```

The results from the LSMEANS analysis are shown here Using PROC GLM to determine interaction effect

The GLM Procedure: Least Squares Means

sex	group	hrchange LSMEAN	LSMEAN Number
F	**1**	-4.5454545	1
F	**2**	-10.3181818	2
F	**3**	5.8181818	3
M	**1**	-4.2727273	4
M	**2**	-2.0000000	5
M	**3**	6.5454545	6

Least Squares Means for effect sex*group
Pr > |t| for H0: LSMean(i)=LSMean(j) Dependent Variable: hrchange

i/j	1	2	3	4	5	6
1		<.0001	<.0001	0.8183	0.0336	<.0001
2	<.0001		<.0001	<.0001	<.0001	<.0001
3	<.0001	<.0001		<.0001	<.0001	0.5404
4	0.8183	<.0001	<.0001		0.0573	<.0001
5	0.0336	<.0001	<.0001	0.0573		<.0001
6	<.0001	<.0001	0.5404	<.0001	<.0001	

Notice the matrix indicates the probability level at which the pairwise comparisons between cell means are different. Sine most comparisons were significantly different, only the comparisons that showed a probability level of p >0.05, are highlighted in red. These results support the notion that being physically active, whether it be dancing or walking as planned exercise, has a positive effect on reducing resting heart rates, and more so for females than males.

33. Statistical applications with linear regression analyses

Learner Ootcomes

After reading this chapter you should be able to:

1. Define and describe simple linear regression
2. Create a SAS program to compute the outcome for a linear regression application including the slope of a line
3. Create a line of best fit
4. Identify the critical components in the output generated from a linear regression application

Calculating the slope of a line

The first step in understanding linear regression is to review the calculation for the slope of a line.

Although presented early in your introduction to mathematics and algebra, most likely in secondary school, learning about the slope of a line may have been one of those topics that you missed, or forgot, or decided that you would never need in the future. Of course, since your intended vocation was not going to require statistical analyses and only essential math, why bother listening. However, now we are reviewing statistical applications and so understanding the calculation for the slope of a line is actually meaningful.

Providing an estimate for the slope of a line can be one way to calculate the **rate of change** in a variable of interest on the vertical axis – usually denoted as y, in relation to an independent variable, such as time plotted on the horizontal axis – denoted as x. The slope of a line provides a measure of the steepness of the line as a function of the change in the variable (y) in relation to the change in the variable (x). Consider for example the data plotted in the following graph (Figure 33.1).

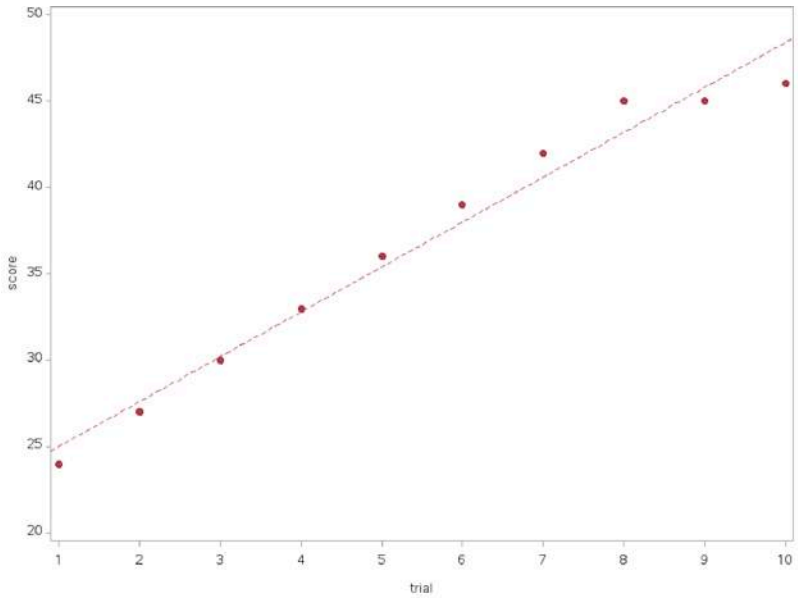


Figure 33.1 Line of Best fit or Y (Score) over X (Time)

In the graph above, we observe a distinct positive relationship between the scores for the variable on the y-axis and the scores for the variable on the x-axis. That is, as we move from left to right on the x-axis we observe an increase in the scores on the y-axis. The slope of the line is an estimate or what we refer to as a coefficient, a single number that represents all of the points on the line of the relationship between the variables: x and y.

The basic calculation to determine this estimate (i.e. the slope) of this relationship is given here as [latex]slope = {\Delta{y}\over\Delta{x}}[/latex] which is read as the change in the y variable divided by the change in the x variable.

Key Takeaways

The slope of a straight line is constant for the entire line and therefore any two points, chosen at random on the line will provide the estimate of the slope (this is not the case for non-linear lines). You might recall that the slope of a line is often represented by the letter m, and the formula for slope is read as: slope = rise over run.

Using SAS programming as shown below, we see that The slope of the line for the graph above is **2.60**, and the y-intercept (the point where the line crosses the y-axis) is **22.40.** Both estimates are shown in the table below.

Parameter Estimates for Figure 33.1

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	**1**	22.40	0.916	24.44	<.0001	20.29	24.51
trial	**1**	2.60	0.147	17.60	<.0001	2.26	2.94

Determining the Slope of a line

The SAS code for the graph above is shownhere. The program uses two variables: the independent variable (x) called **TRIAL** and the outcome variable (y) called **SCORE**. Lines 1 to 3 in the code below set up the program environment. Line 4 identifies the program workflow. Line 5 explains the arrangement of the columns of data. Line 6 cues the program that the data will follow. Ten data points are included in this dataset. Lines 7 to 16 are the SAS code to produce the regression estimates and create the graph with both a line and with dots to represent each point.

```
1.  options pagesize=55 linesize=120 center date;
2.  goptions reset=all cback=white border htitle=12pt htext=10pt;
3.  LIBNAME txtbook '/home/username/textbookExamples/regression';
4.  data txtbook.slope1;
5.  input trial 1-2 score 4-5;
6.  datalines;
   01 24
   02 27
   03 30
   04 33
   05 36
   06 39
   07 42
   08 45
   09 45
   10 46
   ;
7.  Title 'Estimating the slope';
8.  axis1 label=("Trial"); axis2 label=(angle=90 "Score") minor=(n=4);

   /* Define the symbol characteristics for the plot groups */
9.  symbol1 interpol=none value=dot color=depk;
10.  symbol2 interpol=none value=dot color=vibg;

   /* Define the symbol characteristics for the regression line */
11.  symbol3 interpol=rl value=none color=black;

   /* proc gplot data=txtbook.slope1;
12.  plot score*trial / haxis=axis1 vaxis=axis2 ;
13.  plot2 score*trial / noaxis; */
```

```
14.   proc reg; model score=trial / clb; /* command to produce estimates */
15.   plot score*trial ="*";
16.   run;
```

Let's consider a research application of simple linear regression. In the following research study, we are interested in the change in pain estimates for horses that undergo castration. Here we collected 4 pain estimates at 30-minute intervals following equine castration surgery. The data for the estimates of pain are represented in the following table. Notice for each 30-minute time point the estimate of pain on a 10 point scale diminishes by one point.

Time in reference to surgery	30 minutes after	60 minutes after	90 minutes after	120 minutes after
10-point pain scale	9	8	7	6

We can graph these data using the following SAS code.

Example 33.2 Application of Simple Linear Regression

```
DATA LINE;
INPUT ID TIME PAIN;
DATALINES;
01  30 9
02  60 8
03  90 7
04 120 6
;
PROC SGPLOT NOBORDER NOAUTOLEGEND;
REG Y=PAIN X=TIME;
RUN;
```

Figure 33.2, below shows the line of best fit that illustrates the relationship between pain ratings over time since surgery.

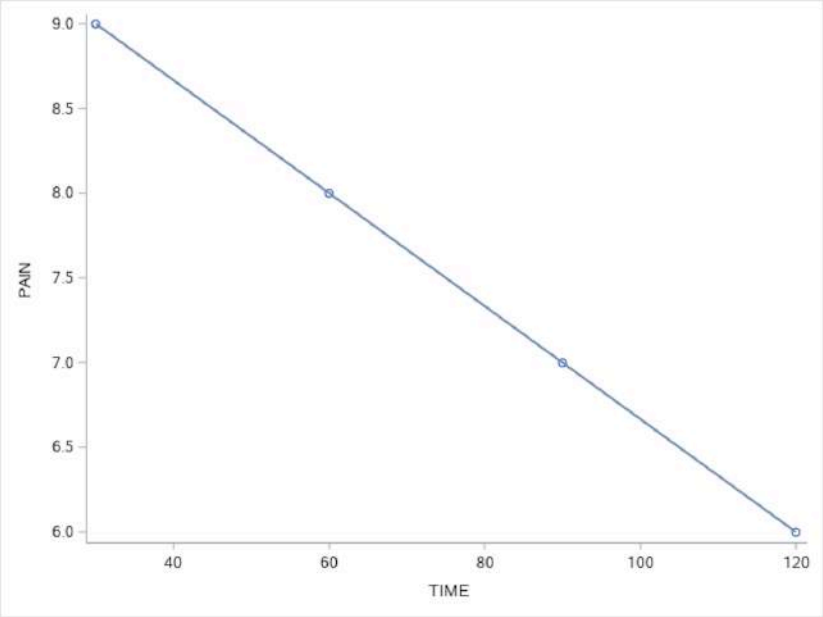


Figure 33.2. Pain Ratings Following Surgery

The calculation of the slope of the line for the estimate of pain over time is represented by the following equation.

[latex]slope = {\Delta{y}\over\Delta{x}}[/latex]

[latex]slope = {(y_2 - y_1)\over (x_2 - x_1)} =[/latex]

[latex]slope = {(8 - 9)\over (60 - 30)} = {(-1)\over (30)} ={0.033}[/latex]

Notice that the slope has a value of m -0.033 which when multiplied by the time value in the equation [latex]y = {mx + b}[/latex] indicates that as the value of TIME increases from left to right on the x-axis, the value for PAIN decreases on the y-axis.

With the following lines of SAS code added to our program we can compute the slope and the y-intercept term for the data above and cofirm that which we calculated by hand.

```
PROC REG; MODEL PAIN = TIME;
RUN;
```

This SAS code produced the following output.

The REG Procedure – Model: MODEL1 for Dependent Variable (Y) : pain

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	10.00000	0	Infty	<.0001
time (x variable)	1	-0.03333	0	-Infty	<.0001

Notice both the slope score of -0.033 and the y-intercept (10) are included in this table. Likewise, because there are only four scores in the data set the standard error is 0 and the estimates for slope and y-intercept are significantly greater than 0.

Using Regression to Compute a Laboratory Standard Curve

In many laboratory experiments, researchers will create what is known as a curve of standards or a standard curve to establish the relationship between substrates and products.. One way that we can use regression and the calculation of the line of best fit is to test the linearity of a relationship between the concentrations of a substrate (x) and a product (y). In the following example we were measuring the presence of a gene of interest at 5 different concentrations.

The Gene of interest was Bactin and the concentrations were: EXPAND..

The following SAS program was used to calculate the relationship using regression, and the graphical representation of the relationship.

SAS progran to compute line of best fit with PROC REG & PROC SGPLOT

```
   PROC FORMAT;
VALUE TRFMT 1='CONC30'
2='CONC60'
3='CONC120'
4='CONC240'
5='CONC480';
options pagesize=55 linesize=80 center;
LIBNAME LINES20 '/home/username/bioInfomatics/lnBstFt';
DATA LINES20.BACTIN01;
input TRIAL CONC1 CONC2 CONC3;
DATALINES;
1 19.63 18.63 20.12
2 18.06 17.23 17.05
3 15.63 15.80 15.59
4 14.58 15.00 14.57
5 13.64 13.70 13.56
;
proc reg;
model CONC1=TRIAL;
model CONC2=TRIAL;
model CONC3=TRIAL;
run;
   proc sgplot data=LINES20.BACTIN01
noautolegend;
reg x=TRIAL y=CONC1 / CLM
CLMATTRS=(CLMLINEATTRS=
(COLOR=Green PATTERN= ShortDash));
FORMAT TRIAL TRFMT. ;
TITLE ' CONFIDENCE LIMITS FOR TRIAL 1 OF BACTIN';
run;
```

```
proc sgplot data=LINES20.BACTIN01
noautolegend;
reg x=TRIAL y=CONC2 / CLM
CLMATTRS=(CLMLINEATTRS=
(COLOR=Green PATTERN= ShortDash));
FORMAT TRIAL TRFMT. ;
TITLE ' CONFIDENCE LIMITS FOR TRIAL 2 OF BACTIN';
run;

  proc sgplot data=LINES20.BACTIN01
noautolegend;
reg x=TRIAL y=CONC3 / CLM
CLMATTRS=(CLMLINEATTRS=
(COLOR=Green PATTERN= ShortDash));
FORMAT TRIAL TRFMT. ;
TITLE ' CONFIDENCE LIMITS FOR TRIAL 3 OF BACTIN';
run;

  proc sgplot data=LINES20.BACTIN01;
xaxis type=discrete;
series x=TRIAL y=CONC1;
series x=TRIAL y=CONC2;
series x=TRIAL y=CONC3;
;
FORMAT TRIAL TRFMT. ;
TITLE ' Simple plot for 3 trials OF BACTIN using overlay';

  run;
```

Output from PROC REG

The REG Procedure

Model: MODEL3

Dependent Variable: CONC3

Number of Observations Read	5
Number of Observations Used	5

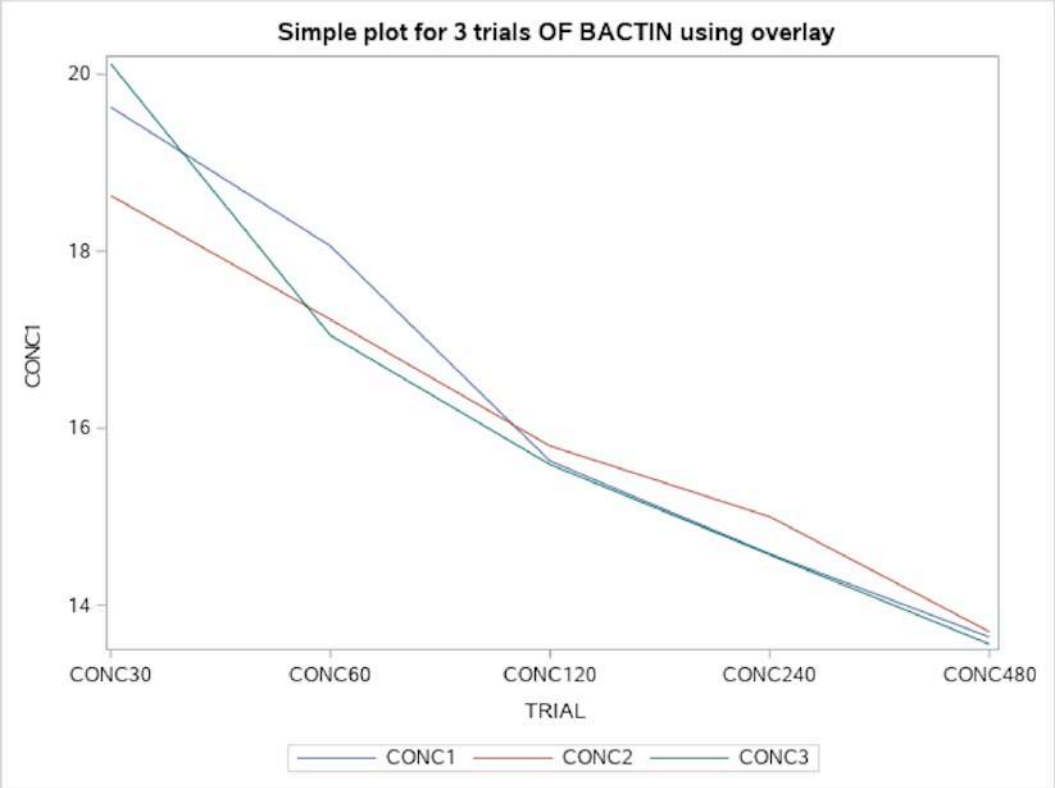
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	24.33600	24.33600	41.74	0.0075
Error	3	1.74908	0.58303		
Corrected Total	4	26.08508			

Root MSE	0.76356	**R-Square**	0.9329
Dependent Mean	16.17800	**Adj R-Sq**	0.9106
Coeff Var	4.71975		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	20.85800	0.80083	26.05	0.0001
TRIAL	**1**	-1.56000	0.24146	-6.46	0.0075



34. Logistic Regression Analysis using PROC LOGISTIC

Learner Outcomes

After reading this chapter you should be able to:

- Define and compute a logit
- Define and describe simple logistic regression
- Create a SAS program to compute the outcome for a logistic regression application
- Describe the use of logistic regression in evaluating the null hypothesis
- Identify the critical components in the output generated from a logistic regression application
- Determine if your study design is multi-level (hierarchical design) and how to use PROC GLIMMIX to account for this design in your analyses

Introduction to Logistic Regression

Consider the application of logistic regression to be synonymous with the computation of ordered least squares regression (OLS) which we studied previously using Proc Reg and Proc GLM applications. However, the difference between these general linear model applications is that the dependent variable was a continuous variable. Conversely, in the use of logistic regression we are interested in evaluating a dependent variable that is binary and has outcome values limited to two possibilities (e.g. 0 or 1).

Most often we apply the logistic regression approach when the dependent variable is binary or dichotomous. We can call this approach binary logistic regression. The dependent variable can take on one of two outcome values like yes or no, 0 or 1, success or failure.

However, we can also use logistic regression to analyze data when the dependent variable has multiple categories, which we call multinomial logistic regression. In the case of multinomial logistic regression the dependent variable is categorical – presenting a discrete value in which there are more than two possible responses, as is the case in a multiple response categorical scale. The outcome measure can be a subjective value produced by a respondent, or it can result from arranging participants into specific groups or categories.

Figure 15.1 presents an example of a binary logistic regression model in which the dependent variable has one of two outcome values: cancer positive or cancer negative, and an exposure variable: exposure to tobacco smoke, as shown here.

Figure 15.1 Example of a binary logistic regression model

Consider the following data set for the model shown above.

Table of cancer outcomes related to smoking			
Smoker status	**Disease status: lung cancer**		
Cancer positive	**Cancer negative**	**Total**	
Smoker	13	20	33
Non-smoker	6	41	47
Total	19	61	80

Table 15.1 Distribution of data for the nominal dependent variable cancer and the independent variable smoking status.

In studying linear regression analyses we discussed the computation of the coefficients that are used to adjust the x variables (independent measure(s) – here the measure is smoking status) as they influence the y variable (dependent measure). That is, we used simple linear regression and the PROC REG procedure to produce a slope score (the regression coefficient or parameter estimate) which acts on to produce () the outcome. However, as we stated previously, in simple linear regression the dependent or outcome variable can take on any value from the real number line.

In logistic regression the measure of interest is a binary value (one of two possible outcomes) which is converted mathematically to a value ranging from 0 to 1 that we call a logit. The mathematical transformation of the binary outcome score to a logit value is computed using the following process.

Logit =

The logit is used in the logistic regression procedure where the logit represents the dependent variable and is forecast by a linear combination of the predictor variables.

PART VI

MEASURING CORRELATION, ASSOCIATION, RELIABILITY AND VALIDITY

Learning Objectives

After reading the chapters in this section you should be able to:

- Compute correlation coefficients using the Pearson Product Moment Correlation Coefficient for continuous data with SAS programming.
- Compute correlation coefficients using the Spearman Non-Parametric Correlation Coefficient for data based on rankswith SAS programming.
- Compute the Bland Altman measures of association using specific SAS programming code
- Evaluate the null hypothesis for a correlation coefficient at $p<0.05$
- Compute the Contingency Coefficient based on data from a Chi-square with SAS programming and with the webulators
- Write SAS programs for each method and review the output that is produced from the computations

The calculation of a correlation coefficient is the method by which a researcher can show a relationship between two measures of interest This estimate **DOES NOT** imply cause or causality between the two variables. Rather, the measure is merely an estimate of how closely two variables describe independent responses for a sample. In the following chapters, we will explore the relationship and association between variables using different approaches that include: calculating a Pearson product-moment correlation coefficient, calculating the correlation coefficient with the Non-Parametric Spearman approach based on ranks, calculating measures of association with the methods of Bland and Altman, and calculating the measures of association with chi-square based techniques such as contingency tables and estimates of kappa.

35. Computing the Pearson Product Moment Correlation Coefficient

Defining the Correlation Coefficient for Continuous Data: the Pearson Product Moment Correlation Coefficient

Pearson's Product Moment Correlation Coefficient is aptly named for the mathematical computational steps that are used to produce the outcome. First, the computation is based on the multiplication of variables, which result in generating a product. Second, the variables, labelled x and y, are computed as independent statistical moments, whereby each score is evaluated against the variable's algebraic measure of centrality of the group scores, a.k.a. the mean. In this way, the correlation coefficient provides the researcher with an estimate of a relationship between two dependent variables.

The outcome estimate of the Pearson Product Moment Correlation Coefficient does NOT imply cause or causality between the two variables. Rather, the outcome estimate is merely an estimate of how closely two variables describe independent responses for a sample. The correlation coefficient is calculated by combining the estimates of variance on each of the separately measured variables of interest.

Specifically, the correlation coefficient is a ratio score that ranges from -1.00 to +1.00 and is created from a cross product:

$$\sum(x_i - \bar{x})(y_i - \bar{y})$$

as the numerator term and then adjusted through division by the error term in the denominator, which is the square root of the product of the sum of squares for the x variable and the sum of squares for the y variable, as shown here:

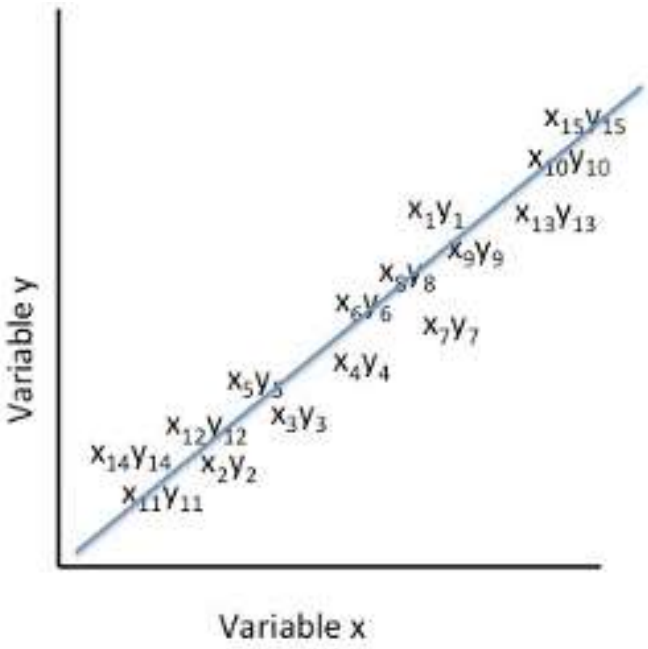
$$\sqrt{\left[\sum(x_i - \bar{x})^2\right]\left[\sum(y_i - \bar{y})^2\right]}$$

The formula to compute the Pearson product-moment correlation coefficient is shown here:

$$r = \frac{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}}{\sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}\frac{\sum(y_i - \bar{y})^2}{n-1}}} \rightarrow r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum(x_i - \bar{x})^2][\sum(y_i - \bar{y})^2]}}$$

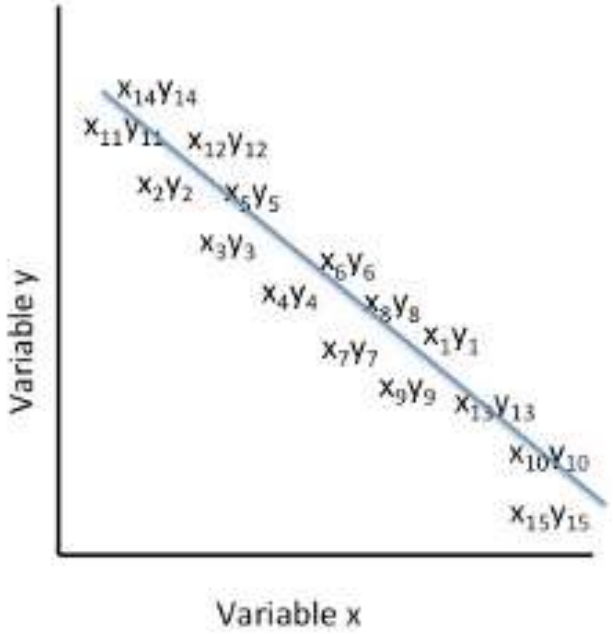
When presented graphically, the paired measures of the x and y variables for the set of scores are represented as a single point within the graphing space. The calculation of the correlation coefficient describes the mathematical relationship between the x and y variable pairs are associated within a geometric space. The following three graphs illustrate the extremes of the Pearson Product Moment Correlation Coefficients for the relationships between x and y variable pairs. In Figure 1, the relationship illustrates a Pearson Product Moment Correlation Coefficient as extreme positive with an r value of 1.00. In Figure 2, the relationship illustrates an extreme negative r value of -1.00. In Figure 3, the relationship is shown as a Pearson Product Moment Correlation Coefficient of zero, or no mathematical relationship between each participant's paired x and y scores.

Figure 1. Representation of the Pearson Product Moment Correlation Coefficient as r=1.00



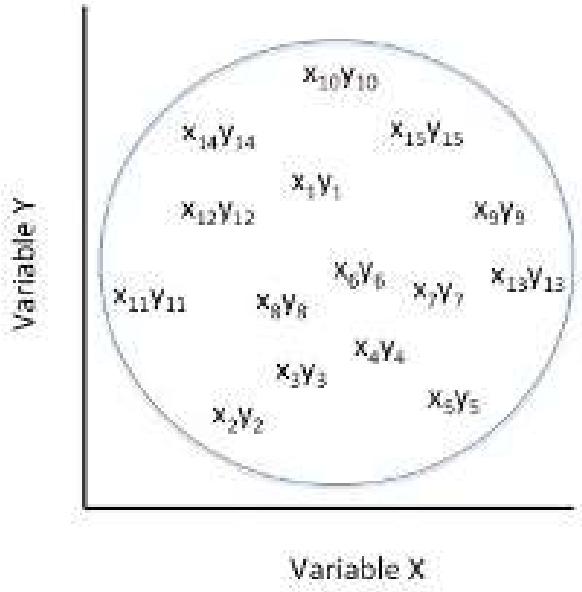
In the figure above, the cluster of paired scores on the variable x and the variable y are presented in a positive direction within the graphing space. For each individual's score on the x variable there is an equal score on the y variable relative to the scale of scores. This means that a correspondingly high score on the y variable matches a high score on the x variable; and similarly a correspondingly low score on the y variable matches a low score on the x variable. When the x and y variables as represented as a single point in the graphing space the resulting image suggests a positive correlation between the variables x and y for the set of participants' scores.

Figure 2. Representation of the Pearson Product Moment Correlation Coefficient as r= -1.00



In the figure above, the cluster of paired scores on the variable x and the variable y are presented in a negative direction within the graphing space. For each individual's score on the x variable, there is an inverse score on the y variable relative to the scale of scores. This means that a correspondingly high score on the y variable matches a low score on the x variable, and similarly, a correspondingly low score on the y variable matches a high score on the x variable. When the x and y variables as represented as a single point in the graphing space the resulting image suggests a negative correlation between the variables x and y for the set of participants' scores.

Figure 3. Representation of the Pearson Product Moment Correlation Coefficient as r= 0.00



In Figure 3, above, the cluster of paired scores on the variable x and the variable y are presented as having no definitive direction within the graphing space. For each individual's score on the x variable, there is no corresponding score on the y variable relative to the scale of scores. In this instance, the x and y variables are simply represented as pairs of scores within the graphing space. No relationship exists when there is complete randomness in the scoring by an individual on the x and on the y variables and thereby do not depict a relationship between the two variables for the set of participants' scores.

The correlation coefficient provides the researcher with an estimate of a relationship between two variables. The Pearson Product Moment Correlation Coefficient does **NOT** imply cause or causality between the two variables. Rather, the measure is merely an estimate of how closely two variables describe independent responses for a sample. The correlation coefficient is calculated by combining the measures of variance on each of the separately measured variables, as shown in the following formula:

$$r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\Sigma (x_i - \bar{x})^2\right]\left[\Sigma (y_i - \bar{y})^2\right]}}$$

A SAS Example

Consider the following data set, with 6 participants measured on two separate variables. Your intention is to determine if the group measures on VAR_1 are similar to the group measures of VAR_2. For those of you that need more tangible labels, let's consider that VAR_1 represents IQ scores and that VAR_2 represents SHOE SIZE.

So that our query is simply, is there a relationship between IQ scores and SHOE SIZE. In order to analyze these data, you create the following SAS program which produces the output shown below.

SAS Code to produce Pearson Correlation Coefficient and A Line of Best Fit in a Scatterplot

```
DATA CORREX1;
  TITLE 'CORRELATION BETWEEN 2 MEASURES IN SAME PERSON';
INPUT ID VAR1 VAR2;
LABEL VAR1='MEASURE FOR VARIABLE 1'
     VAR2='MEASURE FOR VARIABLE 2';
DATALINES;
001 55 7
002 77 9
003 88 10
004 92 10.5
005 100 11
006 105 12
;
RUN;
PROC UNIVARIATE; VAR VAR1 VAR2;
PROC CORR PEARSON; VAR VAR1 VAR2;
PROC SGPLOT NOAUTOLEGEND DATA= CORREX1;
TITLE1  "SCATTER PLOT OF CORRELATION BETWEEN VAR1 AND VAR2";
TITLE2  "LINE OF BEST FIT ADDED TO SCATTERPLOT";
REG Y=VAR1 X=VAR2;
RUN;
```

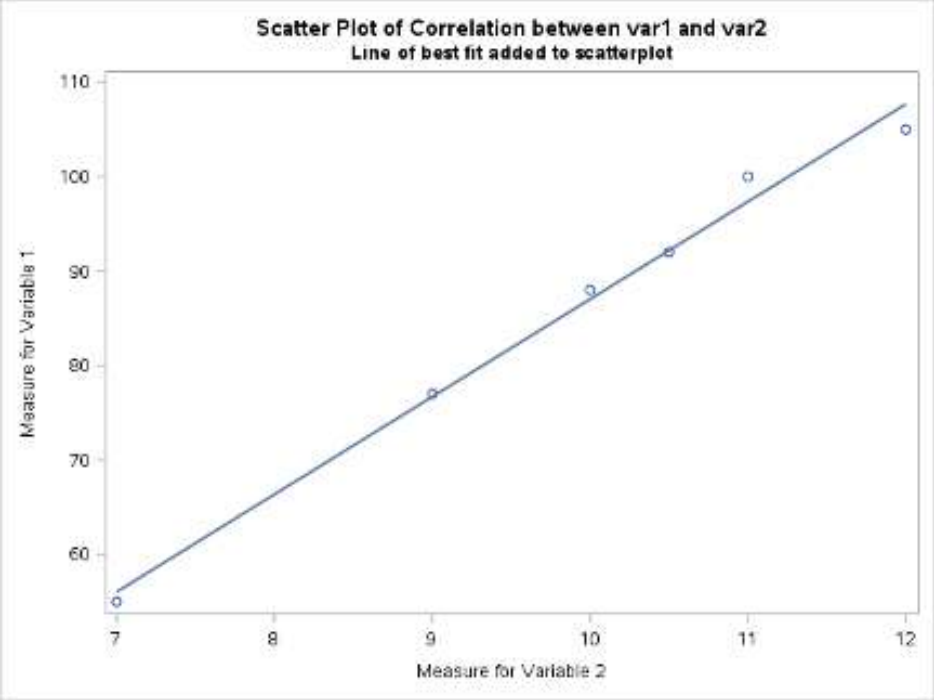
Output from PROC CORR with Corresponding Line of Best Fit in a Scatterplot

Correlation between 2 measures in the same person using the SAS PROC CORR Procedure. Table 1. Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Measure for Var_1 (IQ scores)	6	86.16667	18.10433	517.00000	55.00000	105.00000
Measure for Var_2 Shoe Size	6	9.91667	1.74404	59.50000	7.00000	12.00000

Table 2. Matrix of Pearson Correlation Coefficients, N = 6, Prob > |r| under H0: Rho=0

	IQ scores (Var_1)	Shoe Size (Var_2)
IQ scores (Var_1)	1.00000	r= 0.99500 p <.0001
Shoe Size (Var_2)	r= 0.99500 p <.0001	1.00000



The results indicate that there is a strong correlation between IQ scores and shoe size as indicated by the correlation coefficient value in Table 2 of the SAS output in the scatterplot, above. The reported correlation coefficient (r value) was 0.995. which is considered statistically significant ($p<0.0001$)[1]. The r value indicates that as a group the data on each variable are traveling in the same direction. In other words, low IQ scores were associated with small shoe sizes and high IQ scores were associated with larger shoe sizes.

Another SAS Working Example

Consider the following comparison of VO_2 maximum tests to illustrate the computations for the Pearson product-moment correlation coefficient.

1. Note: just because the p value extends beyond 0.01 you need only to report p<0.01

An individual's ability to perform maximal re-synthesis of ATP is often determined from their estimated VO_2 maximum. Yet the true estimate of VO_2 maximum requires an individual to undergo an extensive performance test in a "clinical or laboratory" setting.

Further, because physical fitness is often established from the VO_2 maximum, researchers strive to develop approaches that can provide a suitable estimate for VO_2 maximum without requiring the laboratory test. To this end, several "field tests" have been presented as reliable and valid proxies for the laboratory estimate of VO_2 maximum.

The validity of field-tests for VO_2 maximum as effective measures of true physiological functioning within this system are based on the linear relationship between heart rate and oxygen consumption. Since the changes in heart rate match the changes in aerobic metabolism, an individual's ability to respire can be predicted indirectly by heart rate rather than the direct determination of oxygen utilization from a controlled laboratory setting.

Laboratory tests for VO_2 maximum typically require that the individual runs or walks on a treadmill, or cycles on a standard bicycle ergometer while expired air is collected and the oxygen concentration in the expired air, is measured. The field tests, however, can use several stimulus modalities (e.g. walking, running, stepping, and cycling) and rather than measure the oxygen concentration in expired air, the technician simply records heart rate at specific stages of the exercise stress test. The relationship between field-tests and a declared "gold standard" is determined by comparing the statistical probability of the association between the predictive nature of the field test and a baseline reference standard measured in the laboratory.

In determining the accuracy of a clinical or diagnostic test, researchers compare the performance of subjects on the selected diagnostic (field) test against the "gold standard". In the following example, this clinical epidemiological approach was used to determine the accuracy of field tests for predicting maximal oxygen consumption against a laboratory standard treadmill test (Bruce protocol). The field tests included the Cooper's 12-minute run, the Astrand-Rhyming bicycle test, the Canadian Aerobic Fitness Test, and the one-mile walking test. A sample of the results for each test is presented in the table, below.

Table of VO_2 Maximum Estimates For A Sample Of Ten Individuals On 5 Different Procedures

Subject	treadmill VO_2 max ml/kg/min	12-minute run VO_2 max ml/kg/min	Bicycle ergometer VO_2 max ml/kg/min	Step test VO_2 max ml/kg/min	1-mile walk test VO_2 max ml/kg/min
01	47	33	44	57	45
02	36	19	53	53	43
03	55	25	49	61	47
04	39	41	23	56	44
05	57	27	50	58	45
06	46	39	52	61	53
07	43	36	55	57	44
08	29	31	42	53	42
09	54	28	44	56	44
10	37	44	54	57	45

The following computations illustrate the application of the Pearson product-moment correlation coefficient to the oxygen consumption data shown above.

STEP 1: Select two variables to compareà, for example, we will begin with VO2 measured with the treadmill test versus VO2 measured with the 1-mile walk test. The mean for the scores on the treadmill, which in this example we designate as the x variable is 44.3 (± 9.23), and the scores on the walk test, which in this example we designate as the y variable and is 45.2 (± 3.04).

$$\bar{x} = \frac{\sum x_i}{n} \rightarrow \bar{x} = 44.3 \text{ and } \bar{y} = 45.2$$

The mean score for the variable x and the mean score for the variable y are used independently in Step 2: computing variance for each selected variable.

STEP 2: Compute the sum of squares for the variance elements of each selected variable, separately, as shown below:

$$Sum\ of\ squares = \sum (x_i - \bar{x})^2$$

Scores on "x" treadmill		
47	47 – 44.3 = 2.7	(47 – 44.3)2= 7.29
36	36 – 44.3 = -8.3	(36 – 44.3)2= 68.89
55	55 – 44.3 = 10.7	(55 – 44.3)2= 114.49
39	39 – 44.3 = -5.3	(39 – 44.3)2= 28.09
57	57 – 44.3 = 12.7	(57 – 44.3)2= 161.29
46	46 – 44.3 = 1.7	(46 – 44.3)2= 2.89
43	43 – 44.3 = -1.3	(43 – 44.3)2= 1.69
29	29 – 44.3 = -15.3	(29 – 44.3)2= 234.09
54	54 – 44.3 = 9.7	(54 – 44.3)2= 94.09
37	37 – 44.3 = -7.3	(37 – 44.3)2= 53.29
	= 0	= 766.01

Scores on "y" the walk test		
45	45 – 45.2= -0.2	(45 – 45.2)2= 0.04
43	43 – 45.2 = -2.2	(43 – 45.2)2= 4.84
47	47 – 45.2 = 1.8	(47 – 45.2)2= 3.24
44	44 – 45.2 = -1.2	(44 – 45.2)2= 1.44
45	45 – 45.2 = -0.2	(45 – 45.2)2= 0.04
53	53 – 45.2 = 7.8	(53 – 45.2)2= 60.84
44	44 – 45.2 = -1.2	(44 – 45.2)2= 1.44
42	42 – 45.2 = -3.2	(42 – 45.2)2= 10.24
44	44 – 45.2 = -1.2	(44 – 45.2)2= 1.44
45	45 – 45.2 = -0.2	(45 – 45.2)2= 0.04
	= 0	= 83.6

STEP 3: Compute the cross products of x and y using: [latex](x_i - \overline{x})(y_i - \overline{y})[/latex]

participant	[latex](x_i - \overline{x})(y_i - \overline{y})[/latex]
01	2.7 * -0.2 = -0.54
02	-8.3 * -2.2 = 18.26
03	10.7 * 1.8 = 19.26
04	-5.3 * -1.2 = 6.36
05	12.7 * -0.2 = -2.54
06	1.7 * 7.8 = 13.26
07	-1.3 * -1.2 = 1.56
08	-15.3 * -3.2 = 48.96
09	9.7 * -1.2 = -11.64
10	-7.3 * -0.2 =1.46
$\sum[(x_i - \bar{x}) \times (y_i - \bar{y})]$	= 94.4

STEP 4: Compute the denominator of the correlation coefficient using the square root of the product of the sum of squares for the x and the y variables.

$$\sqrt{\left[\sum (x_i - \bar{x})^2\right]\left[\sum (y_i - \bar{y})^2\right]}$$

The specific values can be taken from the computations above where the $\sum (x_i - \bar{x})^2$ = 766.01 and the $\sum (y_i - \bar{y})^2$ = 83.6 The denominator term is thus the square root of: SQRT**(766.01 x 83.6) = SQRT(64038.44) = 253**

STEP 5: Compute the correlation coefficient by substituting the values from that which was calculated, above.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum (x_i - \bar{x})^2][\sum (y_i - \bar{y})^2]}} = \frac{94.4}{\sqrt{64038.44}} = \frac{94.4}{253} = 0.37$$

In our example, the correlation coefficient computed for the relationship between the VO2 on the treadmill and the VO2 on the 1 mile walk test = 0.37

Verifying Our Computations with SAS

The data set used in the example above included two columns of data: the VO2 max scores recorded for each participant on the treadmill test and the 1-mile walk test. The columns were separated by a tab character which we denote in our INFILE statement with the command: **delimiter='09'x** . The data set has three variables: id, treadmill scores and 1-mile walk test scores.

raw data set for three variables

```
01   47   45
02   36   43
03   55   47
04   39   44
05   57   45
06   46   53
07   43   44
08   29   42
09   54   44
10   37   45
```

The SAS program written for SAS Studio to compute the correlation coefficient for the relationship between the scores on the treadmill and the 1-mile walk test is shown here:

SAS Code to produce Pearson Correlation Coefficient for treadmill and the 1-mile walk test

```
DATA CORR1(LABEL= 'CORRELATION COEFFICIENTS');
TITLE 'PRACTICE DATA SET TO COMPUTE CORRELATION COEFFICIENTS';
INFILE '/FOLDERS/DATASETS/CORR1.DAT' DELIMITER='09'X ;
* THE DATA IN THE FILE CORR1.DAT ARE TAB DELIMITED;
INPUT ID 1-3 @4 TRDMILL ONEMLWLK;
PROC CORR; VAR TRDMILL ONEMLWLK;
RUN;
```

Output from the SAS PROC CORR PROCEDURE

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
trdmill	10	44.30000	9.22617	443.00000	29.00000	57.00000
Onemlwlk	10	45.20000	3.04777	452.00000	42.00000	53.00000

	trdmill	Onemlwlk
Trdmill	r= 1.00000	r = 0.37301 p < 0.2884
Onemlwlk	r = 0.37301 p < 0.2884	r = 1.00000

Testing the significance of the sample correlation coefficient

When we compute the r statistic (correlation coefficient) the relationship that we observe is specific to the sample from which the data were drawn. If we wish to compute the population parameter of a correlation coefficient, that is the extension of the statistic for the sample to the parameter in the population, then our next step is to compute the rho statistic – rho is the population parameter estimate of the correlation coefficient and it is determined using a t test with (n-2 degrees of freedom).

To determine if the computed correlation coefficient is equal to 0 as in the null hypothesis test: H_0: r = 0 at p<0.05 we use the following equation:

[latex]t = r \sqrt\frac{n-2}{1-r^2}[/latex]

with the degrees of freedom (npairs – 2).

Therefore, in the example presented above, the correlation coefficient was r=0.37, and the sample size was n=10. The t value to evaluate the population parameter is shown below with degrees of freedom = (npairs – 2) = (10 – 2) = 8.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.37\sqrt{\frac{10-2}{1-0.37^2}} = 0.37\sqrt{\frac{8}{0.86}} = 0.37 \times 3.04 = 1.126$$

The *t* critical value for t=1.126 with df=8 is 2.31 for p<0.05. Therefore, because our t statistic is less than the critical value, we would accept the null hypothesis at p<0.05 and suggest that in this sample, the estimate of the relationship between treadmill VO2 maximum scores with 1-mile walk test scores is not an estimate of the true population parameter (i.e. the rho score). Finally, in addition to computing the significance of the correlation coefficient, we can also determine the strength of the correlation coefficient. Typically, the decision rule concerning the strength of a correlation coefficient is as follows: if the absolute value of r is low (<0.7) then we say that is no correlation, but if the absolute value is (>0.7) then we say that there is a correlation between the scores on the variable x and the scores on the variable y.

Strength of the Correlation

The strength of the correlation is determined by squaring the value of r (the correlation coefficient) and multiplying the result by 100. Taken in this way, we are able to express the bivariate correlation (i.e. the relationship between two variables) as a percent of variance that is explained between the two variables.

For example, if we square the correlation coefficient computed for the relationship between the VO_2 on the treadmill and the VO_2 on the 1-mile walk test, as follows:

r = 0.37 and r^2 = 0.139 x 100 = approximately 14%

We can then say that the variable x and the variable y have a very low relationship and when this relationship is

expressed in terms of variance, then the variables x and y in this example share only about 14% of the variance between these two variables, leaving more than 86% of the variance between these two variables unexplained.

Conversely, if we had computed a correlation coefficient of 0.9 between the variable x and the variable y then we would say that there is a strong relationship and could express the variance shared by these two variables as follows:

r = 0.9 and r^2 = 0.81 x 100 = approximately 81% explained variance.

Here is what we covered in this chapter

In this chapter you were introduced to:

- The Pearson Product Moment Correlation Coefficient which we defined as an estimate of a relationship between two dependent variables.
- We also stated, explicitly, that the outcome estimate of the Pearson Product Moment Correlation Coefficient does NOT imply cause or causality between the two variables.
- The correlation coefficient, which is represented by the letter 'r' expresses a positive relationship between two variables (a bivariate relationship) when the r-value is greater than 0 and less than 1, a negative relationship between two variables when the r value is less than 0 and greater than -1, and no relationship when the r-value is in close proximity to 0 and is deemed to be not significant based on the formula: [latex]t = r \sqrt\ frac{n-2}{1-r^2}[/latex] with degrees of freedom (n_pairs – 2).
- In the examples presented here we observed how SAS code enables us to not only produce the estimates of the relationship but also produce a visual representation of the data from the two variables being compared and include a representative line of best fit.

Practice Question for This Chapter

So, you've decided to take on a positive lifestyle that, among other behaviours, includes changes in diet and exercise. You are really interested in eating smarter and becoming more active to enhance your cardiovascular efficiency. You want to become physically fit. But, how does one know that they are achieving a higher level of physical fitness? For some individuals, they may perceive that they are achieving a state of physical fitness by simply looking in the mirror after bathing and observing a change in body shape. While this approach invokes a response, it is not a good indicator of the intrinsic cardiovascular changes that you may actually be gaining. An alternative to mirror gazing is to measure heart rate response after performing a physically demanding task. For example, for many individuals measuring heart rate after climbing stairs is a good indication that their involvement in physical fitness pursuits is having a positive change in cardiovascular dynamics. While this approach seems extremely simple it is certainly easily measured, and changes are easily recognized.

In the Canadian Physical Activity Test novice fire fighters are evaluated on a stepping test in which they are required to walk on a stepping ergometer (a treadmill like device that simulates stepping) at a stepping rate of 60 steps per minute for three minutes while wearing two 5.67 kg weights on their shoulders. The weights are used to represent wearing fully charged air packs during an actual fire-fighting event. In the following fictitious experiment, data were collected from a sample of 20 individuals ranging in ages from 40 to 60 years that climbed 180 steps at a rate of 60 steps per minute. Heart rates were measured in each participant within the first minute following activity and 3 minutes following the

activity. The number of minutes of physical activity in which the individual was engaged per week was also recorded for each participant.

In this exercise compute the correlation coefficient of the relationship between the reported number of minutes of physical activity in which the individual reportedly averaged per week with both the immediate heart rate response to stair climbing (1-minute post-exercise heart rate) and the 3-minute recovery heart rate following the stair climbing exercise. Produce the correlation coefficient in table form and a graphical representation of each bivariate relationship. Likewise, determine the significance of the relationship using the formulae shown above.

Table of Raw Data from the Mock Stair Climbing Experiment

Participant	Age	Amount of exercise per week (minutes)	1-minute post-exercise heart rate (bpm)	3-minute post-exercise heart rate (bpm)
01	56	420	144	75
02	56	315	163	84
03	45	210	178	97
04	49	210	147	66
05	47	140	144	88
06	56	105	142	101
07	53	70	180	117
08	59	350	144	53
09	44	90	190	106
10	47	315	144	57
11	46	420	124	55
12	56	315	133	54
13	55	210	158	77
14	59	210	147	86
15	60	140	164	98
16	40	105	162	101
17	53	70	176	117
18	54	350	144	59
19	40	90	184	126
20	60	315	134	77

Some hints to get you started

```
DATA hrCorr01;
TITLE 'CORRELATION BETWEEN REPORTED WEEKLY PHYSICAL ACTIVITY AND HR RESPONSES TO STAIR CLIMBING';
```

```
INFILE ‘/HOME/USERNAME/YOURFOLDERS/HRCORR.DAT’ DELIMITER= ’09’X ;
/* NOTES – IF YOU USE THE INFILE STATEMENT AS SHOWN ABOVE THEN BE SURE TO
CHANGE THE USERNAME AND FOLDERS WHERE THE DATA RESIDE, ALSO SINCE THIS EXAMPLE IS TAKEN
DIRECTLY FROM THE TABLE BE SURE TO SET DELIMITER TO TAB BETWEEN VALUES */
INPUT ID AGE EX_MIN HR_1MIN HR_3MIN;
PROC CORR PEARSON; VAR EX_MIN HR_1MIN HR_3MIN; RUN;
PROC SGPLOT NOAUTOLEGEND DATA= hrCorr01;
TITLE1 “SCATTERPLOT OF CORR BETWEEN REPORTED WEEKLY PHYSICAL ACTIVITY AND IMMEDIATE POST
EXERCISE HEART RATE RESPONSE TO STAIR CLIMBING”;
TITLE2 “LINE OF BEST FIT ADDED TO SCATTERPLOT”;
LABEL  EX_MIN  =’REPORTED EXERCISE IN MINUTES PER WEEK’
HR_1MIN =’IMMEDIATE POST EXERCISE HEART RATE’;
REG Y= EX_MIN   X= HR_1MIN; RUN;

   PROC SGPLOT NOAUTOLEGEND DATA= hrCorr01;

   TITLE1  “SCATTERPLOT OF CORR BETWEEN REPORTED WEEKLY PHYSICAL ACTIVITY AND 3 MINUTE
POST EXERCISE HEART RATE RECOVDERY TO STAIR CLIMBING”;

   TITLE2  “LINE OF BEST FIT ADDED TO SCATTERPLOT”;

   LABEL  EX_MIN  =’REPORTED EXERCISE IN MINUTES PER WEEK’

   HR_3MIN =’3 MINUTE RECOVERY HEART RATE’;

   REG Y= EX_MIN   X= HR_3MIN;
```

* Explain the outcome of your calculation in terms related to the original research question.

36. Computing Correlations Based on Ranks

Computing Spearman's Rank Order Statistic and Kendall's Tau B

Using Spearman Rank

Kenadll's tau is a method by which to compute the consistency of pairs of ranks – concordance and discordance

37. Demonstrating the Bland-Altman Tests for Agreement

At the University of Prince Edward Island, in Canada, we have an Atlantic Veterinary College where we train future clinicians (DVM), researchers (M.Sc and PhD), along with Internists and Residents in Veterinary Medicine. Among the many areas of interest, one, in particular, is equine-based research. In the following example, we demonstrate the application of Bland-Altman measures of association for equine pain behaviours.

38. Measures of Association - Part I: The McNemar Chi-Square

Part 1: The McNemar Test of Symmetry[1]

This chapter is presented in 2 parts. In the first part, we explore the McNemar Chi-Square Test of Symmetry, where the notion of symmetry is based on the 2 x 2 chi-square comparing elements on the diagonal axis from top left to bottom right, against the off-diagonal elements from bottom left to top-right. In the second part, we explore the question, if we do not observe a significant difference between two measures based on the calculations of the McNemar Chi-Square test, then should we expect that there is an implicit association between the two measures?

Building on the concept of association between independently measured outcome variables, the 2 x 2 table can also be used to organize our data so that we can test the association between the outcome on one variable against the outcome on another variable.

Field testing, whereby we collect data in real-world research applications to evaluate constructs that we would normally see in laboratory tests are often used in health research to evaluate community-based samples. In health research, there are several opportunities to compare field test variables against suggested gold standard variables to determine how closely the field test outcomes match that which we would expect to observe from more controlled laboratory measures.

For example, in measuring the fitness levels within a community, it may not be practical to have every person in the community perform a laboratory-based treadmill exercise test that would evaluate the precise level of maximal oxygen consumption – a measure of cardiorespiratory fitness capacity (VO_2 max ml/kg min). Rather, we can simply recruit a sample of individuals and ask them to perform a one-mile walk test using a planned procedure in which we confirm the distance travelled and the time taken to walk the distance. Combining this information with their height, weight and heart rate response while performing the walk will provide reliable evidence to accurately predict the individual's oxygen consumption capacity – their VO_2 max (ml/kg min).

The one-mile walk test is considered to be an efficient field test to provide accurate predictions of an individual's oxygen consumption capacity because the outcomes on such a test have been compared to laboratory estimates of VO_2 max (ml/kg min) and verified by several independent researchers. A simple way to test the notion of an association between a field test and a laboratory test is to have a group of individuals complete both tests and then measure the relationship between the combined outcome measures. That is, to organize the data as pairs of outcomes and then determine the ratio of the number of concordant or expected pairs against the number of discordant pairs.

In the application of the McNemar Chi-Square test, use the following diagram to organize the association between a given field test response and a corresponding laboratory test response. The organization of these outcomes provides the data that are used to calculate the chi-square statistic (based on a 2 x 2 design).

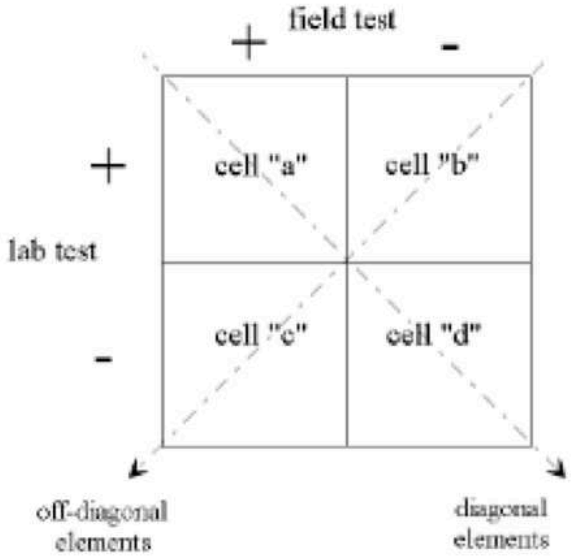
The diagram presents a box with four cells labelled "a", "b", "c", and "d". The top margin of the box is labelled field test (+, -), while the left margin is labelled laboratory test (+,-). In order for a participant's response to be placed in the "a" cell, the paired outcome of their scores on the field and laboratory test would be (+) on BOTH tests. Likewise, for a participant's response to be placed in the "d" cell, the paired outcome of their scores on the field and laboratory test would be (-) on BOTH tests. The "a" and "d" cells represent the concordant pairs.

Similarly, for a participant's response to be placed in the "b" cell, the paired outcome of their scores would be (+) on

the laboratory test and (-) on the field test. Finally, for a participant's response to be placed in the "c" cell, the paired outcome of their scores would be (-) on the laboratory test and (+) on the field test. The "b" and "c" cells represent the concordant pairs.

Using this design and the McNemar Chi-Square statistic, the researcher can evaluate the concordant pairs (the diagonal elements) while adjusting for the discordant pairs (the off-diagonal elements).

Figure 1: Design for the Application of the McNemar Chi-square Statistic



It is important to note that there are two specific preparatory steps to be considered before applying the McNemar Chi-square to test the association between two independent tests.

1. First, the two tests, which in this example will produce a field test response and a laboratory test response, must be organized to demonstrate pair-wise data. That is, the same individual (or a matched pair of individuals) performs both the laboratory test and the field test.
2. Second, regardless of the initial variable type, the results for each test are transformed into a binary score. This can be done by splitting the array of data for each test at the median (middle) score and establishing the polarity of the top and bottom halves of the array. In order to arrange the data in this way, simply list the scores from highest to lowest (or lowest to highest) and split the list (the array) at the mid-point (median score) on the list. All scores above the median score are labelled (+) and all scores below the median (-) are labelled negative. when we arrange the outcomes of scores as matched pairs we can distribute the outcomes in the 2 x 2 design relative to the combined response on the two tests.

The following abbreviated data table presents the cell outcomes for a set of data based on a sample size of 86 participants organized to meet the criteria stated above. Notice that each individual completed both the laboratory (Column 2) and field-test (Column 4) and therefore has a VO_2 max score on each test. The median scores for each variable (median score on the laboratory test = 44 (Column 3) , and median score on the field test = 37 (Column 35)) were identified and the individual's response in relation to the respective median scores was noted.

Next, the 2 x 2 cell membership was established for each participant in relation to the design shown in Figure 1, above. That is, the cell assignment in the 2 x 2 table for each participant based on whether they scored above or below the median score on both variables is indicated with "+" for scores at or above median score, and "-" for scores below the median score (Column 6).

Table of Raw data used in the McNemar Chi-Square Test of Symmetry

Participant ID (Column 1)	Score on laboratory test: The Treadmill test (Column 2)	Score in relation to treadmill median score => 44 (Column 3)	Score on field test: One-mile walk test (Column 4)	Score in relation to one mile walk median score => 37 (Column 5)	2 x 2 cell membership (Column 6)
001	54	+	45	+	+ + (cell a)
002	25	–	23	–	– – (cell d)
003	46	+	33	–	+ – (cell b)
004	42	–	38	+	– + (cell c)
005	40	–	40	+	– + (cell c)
…	…	…	…	…	…
084	29	–	30	–	– – (cell d)
085	47	+	28	–	+ – (cell b)
086	71	+	72	+	+ + (cell a)

The membership for pairs of outcomes is presented in (Column 6) of the chart above. Using the median score within each variable enabled the separation of the group into one of four outcomes, based on the variable's median reference value. The data used to identify membership is binary for each variable.

Further, while cells a and d are important in that they provide the number of concordant pairs, the McNemar Chi-square is actually a test of symmetry used to determine the significance of the number of discordant pairs – the off-diagonal elements.

Simply put, The McNemar procedure tests the equality of frequencies in pairs of cells that are symmetric around the diagonal of a 2 by 2 design (the diagonal elements are the paired data in the upper-left cell: cell "a", and lower right cell: cell "d"). In the computation of the McNemar equivalence estimates, the frequencies in the major diagonal (upper left cell to lower right cell) are ignored. The null hypothesis **(Ho: p1. = p.1)** which is that the row 1 probability = the column 1 probability, and implies that the proportion of individuals who score high on the laboratory test and low on the field test will match the proportion of individuals who score high on the field test and low on the laboratory test.

The outcome data for a sample of 86 individuals that completed both tests are presented in Table 2 below.

Table 2. Observed pairwise counts for laboratory and field test measures arranged for the McNemar Chi-square test

N=86	Field test (+)	Field test (-)	Row Probabilities
Laboratory test (+)	Cell a =23	cell b =12	[latex]p_{1.} = {(a+b)\over{N}}[/latex] [latex]p_{1.} = 0.41[/latex]
Laboratory test (-)	cell c = 19	cell d = 32	[latex]p_{2.} = {(c+d)\over{N}}[/latex] [latex]p_{2.} = 0..59[/latex]

Once we organize the observed data according to the appropriate cell membership we can write the following SAS program to estimate the chi-square observed value

SAS Program to Compute the McNemar Chi-Square Statistic

```
DATA MCNKAP;
TITLE 'MCNEMAR AND KAPPA STATISTICS';
INPUT ROW COL OUTCOME;
DATALINES;
1 1 23
1 2 12
2 1 19
2 2 32
;
PROC SORT DATA=MCNKAP; BY ROW COL;
PROC FREQ;
TABLES ROW*COL /AGREE;
WEIGHT OUTCOME;
RUN;
```

Again, as in the previous applications of the 2 x 2 test, we can evaluate the chi-square observed score against the chi-square critical score of 3.84. The chi-square critical value is the expected chi-square statistic for a 2 x 2 table with a degrees of freedom (row-1) x (column -1) = 1 and an alpha level or probability level of $p < 0.05$.

The McNemar Test is used to test the relationship between the matched pairs of data on the two variables in the 2 x 2 table. That is we are interested in the proportion of responses in the cells related to the marginal variables (i.e. the row variable – the lab test with the column variable – the field test). Specifically, the null hypothesis is testing the following comparison of proportions s that the row 1 probability = the column 1 probability l. Which can also be written as: [latex]H_0 : p_{1.} =p_{.1}[/latex]

The results of the SAS computation of the McNemar Chi-Square are shown in the table below.

McNemar Chi-square test for pairwise counts of laboratory and field test measures

McNemar's Test	
Statistic (S)	1.5806
DF	1
Pr > S	0.2087

Since the chi-square observed value in this sample computation is **1.59** with a probability of **0.21** then we accept the null hypothesis of no difference between the field test and the laboratory test.

The Stepwise formula to compute the McNemar Chi-Square statistic is shown here:

1. [latex]z_{1}={(n_{12} - n_{21})\over{\sqrt(n_{12} + n_{21})}}[/latex]

where: **[latex]n_{12} = \textit{cell b as in } row_{1}, column_{2}[/latex]** and **[latex]n_{21} = \textit{cell c as in } row_{2}, column_{1}[/latex]**

2. **[latex]z_{1}={(12- 19)\over{\sqrt(12 + 19)}}[/latex]**

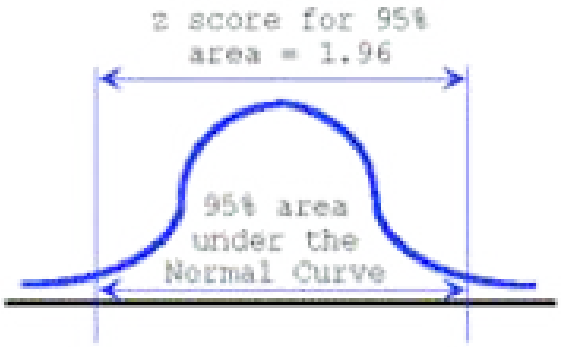
3. [latex]z_{1}={(-7)\over{\sqrt(31)}}[/latex]

4. [latex]z_{1}= -1.26[/latex]

To calculate the McNemar chi-square statistic from a z-score, simply square the z-score.

The McNema Statistic can be shown as a Z score =-1.26 or as a [latex]{\chi}^2 = 1.58[/latex]

If the McNemar estimate is presented as a Z score then we compare the value against -1.96 to +1.96, as the region of accepting the null hypothesis. Likewise, if the McNemar estimate is presented as a [latex]{\chi^2}[/latex] then we compare the value against 3.84, where [latex]{\chi^2}[/latex] scores < 3.84 are included in the region of accepting the null hypothesis (area under the normal curve) shown below.



NOTE: A webulator to calculate the McNemar and Kappa Statistics is presented in the chapter after next, and is currently available at: https://health.ahs.upei.ca/webulators/test_mcnKap.php

[1] This section is based on the following published work: Montelpare, W.J., and McPherson M., (2000) Client-side processing on the InterNet: Computing the McNemar test of symmetry and the kappa statistic for paired response data. *The International Electronic Journal for Health Education*, 3(3): 253-271.

39. Measures of Association -- Part II: The Kappa Statistic

Part II: The Kappa Statistic to Measure Agreement

Given that the results of the McNemar Chi-Square statistic, calculated in the previous chapter, were not significant, then the question becomes, "if the outcome variables representing the results of a participant's performance on each test are not statistically significant in their difference, does that necessarily mean that the outcome scores are in agreement?"

Since the Kappa statistic is a measure of agreement we can test this notion using the Kappa statistic applied to the fourfold or 2 x 2 table. Converse to the McNemar Chi-square which processes the data in the off-diagonal elements (cell "b" and cell "c"), the Kappa computations focus on the data in the major diagonal from upper left to lower right (cell "a" and cell "d"), examining whether counts along this diagonal differ significantly from what is expected to occur by chance. If no agreement exists between the counts on the major diagonal then we would expect the proportion of individuals scoring high or low on the lab and field tests to be similar.

Similar to the computation of the McNemar Chi-square, the Kappa statistic uses the data from the row and column probabilities of the 2 x 2 table. The exact computations for Kappa are specifically shown as follows:

1. COMPUTE ROW PROPORTIONS

Row 1 Proportion: p1. = (a+b) : N; p1. = (23+12) : 86 = 0.41
Row 2 Proportion: p2. = (c+d) : N; p2. = (19+32) : 86 = 0.59
Column 1 Proportion: p.1 = (a+c) : N; p.1 = (23+19) : 86 = 0.49
Column 1 Proportion: p.2 = (b+d) : N; p.2 = (12+32) : 86 = 0.51

2. COMPUTE THE [latex]P_{i}[/latex] TERMS

OBSERVED: [latex]({\pi}) \textit{obs}[/latex]
[latex]({\pi}) \textit{obs}[/latex]: *the observed term of the main diagonal elements*
[latex]({\pi}) \textit{obs}[/latex] = ((cell a) : N) + ((cell d) : N);
[latex]({\pi}) \textit{obs}[/latex] = ((23:86) + (32 : 86));
[latex]({\pi}) \textit{obs}[/latex] = (.27+.37);
[latex]({\pi}) \textit{obs}[/latex] = 0.64
EXPECTED: [latex]({\pi}) \textit{exp}[/latex]
[latex]({\pi}) \textit{exp}[/latex]: the expected term of the main diagonal elements
[latex]({\pi}) \textit{exp}[/latex] = ((p1. * p.1) + (p2. * p.2);
[latex]({\pi}) \textit{exp}[/latex] = ((0.41 * 0.49) + (0.59 * 0.51));
[latex]({\pi}) \textit{exp}[/latex] = (0.20 + 0.30);
[latex]({\pi}) \textit{exp}[/latex] = (0.50);

3. COMPUTE KAPPA [latex]({\kappa})[/latex]

Kappa = (([latex]({\pi}) \textit{obs} - ({\pi}) \textit{exp}) : (1- ({\pi}) \textit{exp}[/latex]))
Kappa = ((0.64 – 0.50) : (1- 0.50))
Kappa = (0.14 : 0.50)

Kappa = 0.28

The computed Kappa value is **κ** = 0.28. Our next task is then to determine if this is a true measure of agreement or an agreement that can happen by chance. Therefore, in order to evaluate this Kappa statistic we need to determine if the computed value is significantly different than 0.

We can do this by first computing the standard error of the Kappa statistic and then using this value to determine the z statistic for Kappa and comparing the value to the normal curve. Recall that 95% of scores on the normal curve are < ±1.96. Therefore, if our Z**κ** score is between -1.96 and +1.96 then we would accept the null hypothesis that **κ**=0.

To compute the standard error for our computed **KAPPA SCORE** we use the following procedure under the null hypothesis that Ho: **k**=0

4. COMPUTE THE SUM OF PROPORTIONS

p1. = 0.41; p.1 = 0.49; p2. = 0.59; p.2 = 0.51

sumP = (p1. * p.1 * (p1. + p.1)) + (p2. * p.2 * (p2. + p.2));

sumP = (0.41 * 0.49 * (0.41 + 0.49)) + (0.59 * 0.51 * (0.59 + 0.51));

sumP = (0.20 * (0.90)) + (0.30 * (1.10));

sumP = (0.18) + (0.33);

sumP = (0.51);

5. COMPUTE THE STANDARD ERROR

std error = 1/((1-) *)*

std error = 1/((1- 0.5) *) *

std error = 1/((0.5) *) * 0.49

std error = 0.22 * 0.49

std error = 0.106

Use the following formula to compute Zκ which is the z score for Kappa, under the null hypothesis of Ho: k=0:

zKappa = (kappa/ stderr1) zKappa = (0.28/ 0.106) zKappa = 2.65

Considering that 2.65 is greater that 1.96 we can say that the zKappa is within the region of rejection in regard to the null hypothesis stated as Ho:k=0 and therefore we can say that there is agreement between the lab and field test.

Finally, we can also determine the significant difference of our Kappa estimate from 0 by using the standard error to compute the 95% confidence intervals for the Kappa statistic as follows:

6. COMPUTE THE STANDARD ERROR AND 95% CONFIDENCE INTERVAL

Use the following measurement terms taken from the McNemar Chi-square table:

p1. = 0.41	P11 = 23/86=0.27
p.1 = 0.49	p12 = 12/86=0.14
p2. = 0.59	p21 = 19/86 = 0.22
p.2 = 0.51	p22 = 32/86 =0.37

Aterm = (p11*(1-(p1. + p.1)*(1-kappa))**2 + p22*(1-(p2. + p.2)*(1-kappa))**2);

Aterm = (0.27*(1-(0.41 + 0.49)*(1-0.28))**2 + 0.37*(1-(0.59 + 0.51)*(1-0.28))**2);

Aterm = (0.27*(1-(0.9)*(1-0.28))**2 + 0.37*(1-(1.10)*(1-0.28))**2);
Aterm = (0.27*(1-(0.9)*(0.72))**2 + 0.37*(1-(1.10)*(0.72))**2);
Aterm = (0.27*(1-(0.648))**2 + 0.37*(1-(0.792))**2);
Aterm = (0.27*(0.352)**2 + 0.37*(0.208)**2);
Aterm = (0.27*(0.124) + 0.37*(0.043));
Aterm = (0.033 + 0.02);
Aterm = (0.049);
Bterm=((p12*(p.1 + p2.)2 + p21*(p.2 + p1.)2)*(1-kappa)2);
Bterm=((0.14*(0.49 + 0.59)2 + 0.22*(0.51 + 0.41)2)*(1-0.28)2);
Bterm=((0.14*(1.08)2 + 0.22*(0.92)2)*(0.72)2);
Bterm=((0.14*(1.17) + 0.22*(0.84))*(0.52));
Bterm=((0.16 + 0.185)*(0.52));
Bterm=(0.179);
Cterm=((kappa − *(1-kappa))**2);
Cterm=((0.28 − 0.5*(0.72))2);
Cterm=(0.0064)
A + B + C= (Aterm + Bterm + Cterm);
A + B + C= (0.049 + 0.179 + 0.0064);
A + B + C= 0.23
Compute the standard error used in the computation of the confidence interval:
stderr = = = = 0.01
ci95LL = (kappa − 1.96*(stderr));
ci95LL = (0.28 − 1.96 * 0.01);
ci95LL = (0.28 − 0.022);
ci95LL = (0.258)
ci95UL = (kappa + 1.96*(stderr2));
ci95UL = (0.28 + 1.96 * 0.01);
ci95UL = (0.28 + 0.022);
ci95UL = (0.302);

If the upper and lower limits of the 95% confidence interval do not include 0 then we can say that the Kappa value is significantly different from 0.

The SAS program to produce KAPPA in the 2 x 2 matrix was handled by the McNemar Chi-Square, where a=23, b=12, c= 19, d=32. Since the data were entered as cell summary data and not strings of raw data, the weight <dependent variable> format is used to read each cell value. The essential option is /AGREE which produces the Kappa measure of agreement.

```
PROC FREQ;
TABLES ROW*COL /AGREE;
WEIGHT OUTCOME;
RUN;
```

Statistics for Table of ROW BY COL

Simple Kappa Coefficient	
Kappa	0.2759
ASE	0.1024
95% Lower Conf Limit	0.0752
95% Upper Conf Limit	0.4767

PART VII

ADVANCED CONCEPTS FOR APPLIED STATISTICS IN HEALTHCARE

In this section, the following topics are included:

Calculating Sample Size and power under different Scenarios

Learner Outcomes

- Describe the importance in establishing a sample to represent the population
- Identify the difference between probabilistic and non-probabilistic sampling strategies.
- Compute sample size under different scenarios using SAS code
- Understand when a given sample size calculation is most appropriate
- Apply the appropriate sampling strategy to a given research design

Mixed model analysis

Learner Outcomes

Understanding repeated measures designs, split-plot factorial models, nested designs and mixed model anovas that incorporate fixed and random effects

Survival analysis

Learner Outcomes

Type your learning objectives here.

- First
- Second

Computer Simulation and Random Number Generation

Learner Outcomes

In this chapter we will create new data sets using computer generated random numbers. In this way we can simulate research outcomes without actually performing the research.

Some basic rules of the exercise are that we must begin by understanding our variables and the parameters that the variables represent. Min max estimates variance, N

Through computer simulation approaches we can combine logic with combinations and permutations in factorial models to explore wicked problems.

40. Computing Sample Size and Power

Introducing sampling

In applied health research, we deal primarily with humans; therefore, the term *sampling* refers to how we select individuals to form a sample (small group) from the larger group of all individuals (the population). For example, we might be interested in knowing if a mindfulness intervention reduces stress among emergency room physicians. Although we could try to include every single emergency room physician in the world, that isn't necessary because we can actually get a very close estimate of the true effects for the population by studying some of the people in the group (i.e., a sample).

The foundation of sampling is based on the following three essential elements: samples, inferences, and the population.

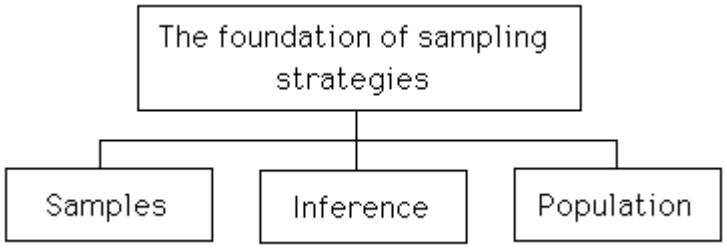


Figure 40.1 The Foundation of Sampling

The term *population* refers to the complete set of all people being studied.

The sample is then denoted as the subset of people that are being studied, and enable us to describe the elements of the larger group: the population through estimation and inference. The term *inference* refers to a deduction or a conclusion and is used in research to describe the process of relating information derived from a sample onto a population.

One of the most important concepts in sampling is that because we can rarely evaluate the population directly, it is expected that the sample is a true representation of the population from which it was drawn. This latter point – *that the sample is a true representation of the population from which it was drawn*, is imperative if we are going to make inferences about the population based on estimates from the sample.

Figure xxx The sample is a true representation of the population

Once we have defined our population, we need to decide how many people we need to include in our research study in order to be confident in our results. There are several ways to compute sample size and the calculations differ depending on the research design and analysis that you are planning to do. In quantitative research we can simplify sampling strategies into two basic categories: probabilistic sampling and non-probabilistic. A list of the more common types of probabilistic sampling strategies includes: simple random sampling, systematic sampling, stratified sampling, and cluster sampling. While these are elementary sampling approaches there are several other types of more complex probabilistic sampling strategies that are derived from these four basic types. Similarly, with regard to non-probabilistic sampling, convenience sampling is the most common, but one may also consider consecutive sampling, and judgemental sampling as approaches that provide samples, but where the members of the sample are not drawn at random and therefore, not all members of the population had a chance to be selected.

Concept 40.2 Types of sampling

40.2.1 Probability Sampling

The term *probability sampling* or *probabilistic sampling* refers to sampling procedures that are based on random selection from a population or a predefined set. The component of randomness ensures that each unit within the larger group (the population) has an equal chance of being selected. If the researcher uses a true random selection approach, then the process of sample selection will be more likely to reduce the influence of selection bias in the research process. The goal of probability sampling is to end up with a sample that is representative of the population.

Random and independent sampling:

A probability sampling approach that uses random and independent sampling implies that each member of the population has the same chance of being selected for the sample. The term independent suggests that the selection of any one member to the sample in no way influences the selection of any other member to the sample from the population. A sample that is comprised of individuals that were selected using a random and independent approach will enable the researcher can generalize the results to a larger population.

Systematic Sampling:

In some circumstances the researcher will have a complete list of all of the members of a group from which the sample is to be derived. For example, the list of all patients within a medical practice, the list of constituents within a voting area, the list of all members in a club. Systematic sampling is an effective approach to draw the sample from the list of the members of the larger group (i.e., the population) because the researcher can select individuals using a method that can have an apriori plan, that can be replicated. For example, a researcher may decide to select every other name or every "nth" name on a list. Likewise, using the entire list, the researcher may decide to create a strategy for dividing the list into specific groups to represent the total population.

Stratified random sample (common in public opinion polls):

The stratified random sample is a probabilistic sampling approach that maintains the elements of randomness and independence but is established within the constraints of a user defined subgrouping system referred to as strata. In the stratified random sampling approach each strata is defined by a particular characteristic of interest, for example an age range, or a level of household income, or a grade level. The characteristic of interest is fixed within the strata so that there is only one characteristic contained in any one strata, as shown in the following example in Table 20.1.

Table 40.1 Example of stratification by income

Here an individual can be selected from one of the three possible income strata. Any individual cannot belong to more than one strata as the income level is a unique identifier for an individual.

The stratified random sample approach is most effective when the researcher's interest is related to the variables upon which the strata are based. For example, if a researcher is interested on health service utilization within a cohort based on income, then household income may be an appropriate characteristic upon which to base the sampling strata. When a researcher applies the stratified random sampling approach, each stratum is sampled randomly and the various sub-samples collected from the strata are combined to form the representative sample. However, when there are

noticeable imbalances in the total number of individuals within a strata, it is suggested that the researcher create proportions within the strata to preserve the natural concentrations of sections within the population.

Cluster Sampling

The cluster sampling approach is a form of random sampling that is used to reduce the large numbers of individuals needed for stratified and random methods. For example, consider a study of university students that are separated by academic discipline. Each discipline is a pseudo-intact group that forms a group which can then be further processed through stratification and random selection. The cluster may be considered as the first level filter of a population that is either impossible or impractical to sample all members because of size. The effect of clustering enables organization of the population based on arbitrary grouping criteria established by the researcher. The clusters are generally easy to define and often individuals will self-identify within the cluster.

A form of fixed clustering is to use postal area codes as the criteria for cluster membership. The postal area code establishes fixed boundaries to a geographic region. Individuals representing households within the postal code area are then selected at random to represent all of the households in the entire area. One caveat to consider in using postal code area however is the minimum sample size. This is especially true when considering selecting individuals based on postal codes for rural and remote areas whereby only a few households are included in the entire postal code area. As a general rule a minimum sample of n=15 is used to ensure anonymity of selected individuals when using cluster sampling approaches.

20.2.2 Non-Probability Sampling

The term non-probability sampling or non-probabilistic sampling refers to sampling procedures in which randomness of selection is less important than meeting apriori characteristics that are specifically related to the research question. Typically the resulting sample is small in comparison to a larger population and therefore may not necessarily be generalized to describe the larger population from which the sample was drawn.

Often in health-based studies, non-probability samples refer to individuals that are best suited to the conditions of the research question. That is, in non-probability sampling, such individuals may be more likely to be accessible for the study or comply with the specific regimen of the researcher's interest.

For example, in studies of children with complex and chronic conditions we selected individuals that met the specific apriori planned criteria for the definition of chronic and complex. In this way we were able to select from all children that met the criteria established by the definition. This approach allowed is to infer our results to the specific target "group" but not the general population.

Convenience samples

Among the most common non-probabilistic sampling procedures is the convenience sample. This procedure is used most often because it is inexpensive, it takes advantage of the availability of subjects, and it is functional when other methods are less practical.

While convenience samples are efficient in including willing patients where informed consent is required and may inhibit participants, convenience samples include only those individuals who are willing to participate. Therefore, a challenge in selecting a convenience sample is participant bias because they want to be involved. In using convenience sampling the researcher should always check for geographic proximity of samples, the likelihood that subjects may refrain from participation and the intrinsic bias of sampling.

Consecutive samples

Consecutive sampling is a form of non-probabilistic sampling whereby individual are selected if they meet the study criteria, and if they are available during the duration of the study. Likewise, consecutive sampling refers connecting with an accessible group of individuals but may not be inferential to a larger population. Finally, consecutive sampling is problematic when the duration of the selection process is inappropriate; for example measuring an outcome over a month when a year's worth of data is suggested.

Judgmental samples

Judgmental sampling is a form of non-probabilistic sampling whereby individuals are selected by "handpicking" individuals for the study. The judgmental sampling approach resembles convenience sampling in its disregard for the effects of bias, but can produce results which are related to an accessible group of individuals. Participants selected using a judgmental sample may not be generally inferential to a larger population.

40.3 Applications Of Sample Size Formulas For Various Designs

40.3.1 Statistical Power

Statistical power is defined as the likelihood to find an effect when in fact the effect really does exist. In other words, *statistical power* refers to the probability of correctly rejecting the null hypothesis.

The terms *power*, *alpha* (α) and *beta* (β) are all related to statistical decisions about accepting and rejecting the null hypothesis. Recall that the null hypothesis always proposes there is no statistically significant effect or difference between groups. Therefore, if you reject the null hypothesis, that means there IS a significant effect or difference. In contrast if you fail to reject the null hypothesis, it stands.

Consider a simple comparison of average heart rate between two groups. Here, the null hypothesis would be that the two groups have the same mean heart rate, given as follows:

Ideally, you want to plan your research study so that you have a large enough sample to be able to accurately discern whether or not there is a meaningful difference between the two groups and avoid making a false conclusion.

A **Type I statistical error** would be made if the researcher found a significant difference between the two group means (thus, rejecting the null hypothesis that there was no difference) when one did not actually exist. In this situation we would say you have a false positive.

The probability of making a Type I error is also referred to as "alpha", which uses the Greek letter, α. By convention, most researchers decide that a 5% chance of making a Type I error is acceptable. Therefore α is often set at .05 (5%).

A **Type II statistical error** would be made if the researcher did not find a significant difference between the two group means (thus, failing to reject the null hypothesis), when in fact a true difference exists. This would be a false negative.

The probability of making a Type II error is inversely related to your statistical power:

β = 1- Power

As power increases, the probability of making a Type II error (missing a significant finding) decreases. However, the sample size required for your study also increases as power increases, sometimes making it impractical. Generally, power is set to .80-.90, resulting in a β value from .20-.10, meaning that even with a power of .90 there is a 10% chance of missing a significant effect.

Statistical errors are based on the interconnection between the size of the sample, the effect size, and the power of the test. Statistical power is computed as:

where:

therefore, if we establish that then and . In which case we would say that the statistical test has 80% power.

Effect size tells us about the magnitude or strength of an effect, difference, or relationship and is specific to a given statistical test. Smaller effect sizes require larger sample sizes, while medium to large effect sizes require smaller sample sizes. Many sample size calculations require effect size estimates which can be tricky if there is not a lot of past research on the variables you are interested in. Generally researchers use the best available estimates and calculate the actual effect size in their study once they have collected their data.

In this section we will work through four different approaches to sample size calculations using probabilistic formulae. To begin, the data in Table 40.2 list the Z-scores associated with common levels of Alpha and Beta estimates. These are essential values for the formula used below.

Table 40.2 Z- Scores associated with common Alpha and Beta Probabilities

Alpha Terms	**Beta Terms**		
Alpha Probability	Z Scores	Beta Probability	Z-scores
.10	1.64	.10	1.64
.09	1.70	.12	1.55
.08	1.75	.14	1.48
.07	1.81	.16	1.41
.06	1.88	.18	1.34
.05	1.96	.20	1.28
.04	2.05		
.03	2.17		
.02	2.33		
.01	2.58		

40.3.2 Size of a Simple Random Sample to Estimate a Population Proportion

Let's say you want to create a representative sample for a population using simple random sampling but you're not sure how big of a sample you need. The following formula can be used to calculate your sample size, "n".

Unpacking the formula for simple random sampling we see that N represents the population from which the sample will be drawn; p refers to the proportion of individuals displaying the characteristic of interest, while $1-p$ or q refers to the proportion of individuals in the population not displaying the characteristic of interest.

To determine the proportion of cases in a population, p is the ratio of all individuals displaying the characteristic of interest divided by the set of all cases from which the sample was drawn. In real life you might determine p based on past research or population statistical data such as the census.

The error term in the denominator refers to the expected accuracy or the allowable difference between the estimate of the proportion in the selected sample, as a result of the sample size calculation, and the true population proportion. A typical value here would range between 0.02 and 0.10.

The term refers to the two-tailed standardized score of researcher confidence. If =0.05 then z= 1- = 0.95 or a 95% confidence value.

Figure 40.6. Explaining the parts of the formula

Applying the formula

Consider the following application of the formula. You are asked by the Medical Officer of Health (MOH) to determine the sample size required to represent a proportion of persons who inject drugs (PWID) in a starting population of 157,000 individuals. From previous reports, the proportion of PWID within a sample (i.e. the proportion of the population that display the characteristic of interest) was p = 0.13 and therefore q = (1-p) = 0.87. The MOH wants the estimate to be within 3% of the true population proportion with 95% confidence.

Based on the application of the formula for simple random sampling for the estimate of a Population Proportion you report that the MOH will require a sample size of at least 481 individuals in order to be 95% confident that the proportion of PWID in her sample will represent the true population proportion of PWID individuals within 3%.

40.3.3 Sample Size when the Original Population is Unknown

In some situations, you may not know the original size of the population from which to draw the sample. In these circumstances the more appropriate sample size formula is shown here as:

Applying this formula to compute sample size where the original "N" is unknown for the unbiased proportional estimate given that p = 0.5; error= 0.05; and =0.05.

Working through the formula with the values given above produces a sample size of n = 384.

40.3.4 Sample Size for a Comparison Study

In some research designs we are interested in comparing results between two or more groups. In this case we may not know the original population size, but we may know about variability with respect to the dependent variables. To determine sample size for a comparison study we need to know **the measure of central tendency** for the dependent variable and the **amount of variability** we could expect for the measures of the dependent variable.

Typically, the amount of variability is based on information from previous studies. In some situations this information may come from pilot work or from an actual study. The sample size formula for comparison studies is shown here as:

The terms of the formula are explained as follows. The term is based on the variance reported in the literature or from pilot studies. Similarly, the expected mean is based on the mean from the literature or from pilot studies, while the *expected % accuracy is the proximity of the estimated mean score to the true population score and* is set by the researcher to be about 10%. *The terms* Zalpha or Zand Zbeta or Z use the standard scores for the alpha level and for the beta value (power) levels. Common values for and are =0.05 and =1.28, respectively.

Application of the sample size formula for a comparison study

In a recent healthy heart study, researchers measured the effects of red wine consumption on blood cholesterol concentrations of males over the age of 35 years. The researchers showed that males (n1= 133) who consumed on average one serving of red wine per day, for a minimum of six days per week, had a lower concentration of the athero-genic low-density lipoprotein cholesterol than a group of age matched control subjects (n2= 143) who abstained from any alcohol consumption. The investigation followed the total group of 276 males for a 24 week period.

You believe that the data are valuable and therefore you wish to conduct the study with a group of males in your local community. In the reference study the average cholesterol concentration for the sample of interest (red wine consumers) was 4.6 mmol/L with a standard deviation of 0.32 millimoles per liter (mmol/L). Considering that you wish to use an alpha level of 0.05 with a corresponding beta level of 4 x alpha (where beta = 4 x 0.05 = 0.20) and a power level of "1 – beta = 0.80". Further you expect that the difference between your estimate of the mean and the TRUE estimate of the mean is within 3 percent.

The data you need in order to compute the appropriate sample size are:

1. The estimated mean from previous studies = 4.6 mmol/L
2. The standard deviation from previous studies = 0.32
3. The z scores for the alpha probability (.05), $Z\alpha=1.96$
4. The z score for the beta probability (.20), $Z\beta=1.28$
5. The allowable percent difference between your estimate for the dependent variable and the expected estimate for the dependent variable from the true population = 3%

= 16

The results of this computation indicate that in order to be 95% confident that the estimates for the proposed sample will be within 3 percent of the true population value, you will need to have a minimum of 16 participants in the test group and 16 participants in the control group.

40.3.5 Sample Size for a Case-Control Study

In the case-control study design, individuals with a specific measurable condition are "compared" to individuals that do not demonstrate the condition of interest. The case-control design is a retrospective study design type that evaluates, by comparison, the differences in outcome measures between groups of individuals with and without a disease, or the signs/symptoms of a condition.

Case-control studies are useful in demonstrating associations but may not show causation. The temporal characteristics (elements of time) are important to demonstrating the relationship. An essential consideration in a case-control study is the clear definition of the cases and of the controls.

In a case control design we may also consider that the cases are more likely to occur given exposure to the stimulus, to which we say this is a directional hypothesis or a one tailed hypothesis. If we consider a one tailed decision rule (cases are more likely than controls, given the characteristics of the scenario) then we see that a power of 80% has a beta term of 0.20 and a $Z\beta= 0.084$. Estimates of statistical power for the one-tailed (directional hypothesis) and corresponding $Z\beta$ values are shown in the following table.

Figure 40.7 Power estimates and corresponding z scores

The sample size formula to determine the number of cases (or the number of controls) in each group of a case-control study is shown here as:

The formula differs slightly from that published more recently by Kasiulevicius, Sapoka, and Filipaviciute (2006)[1] but produces similar estimates. Where the elements of the formula include : the proportion of cases among those individuals suspected to have been exposed. : the proportion of cases among those individuals suspected to not have been

exposed. : is the Z score for the term, where 1.96, and : is the Z score for the term for a one-tailed directional hypothesis (0.84).

Application of the sample size formula for a case-control study

In order to determine the effects of cannabis smoking on lung cancer you decide to conduct a retrospective case control study in which your sample size estimate is based on the consideration that the relative risk of lung cancer among frequent cannabis smokers is about 5.7 times that of non-smokers. You decide to use p1 = 0.285 to represent the proportion lung cancer patients who were frequent cannabis smokers and p0 = 0.05 to represent the proportion lung cancer patients as controls who never smoked cannabis.

The data needed to compute the appropriate sample size are shown here as:

1. Proportion of individuals that had lung cancer among individuals that were considered frequent smokers of cannabis, P1 = 0.285
2. Proportion of individuals that had lung cancer among individuals that never smoked cannabis, P0 = 0.05
3. The z scores for the alpha term (α=0.05), Zα=1.96
4. The z scores for the beta term based on a one-tailed hypothesis, Zβ=0.84

The results of your computation showed that the appropriate sample size needed to conduct your study will require at least 36 individuals in the case group and at least 36 individuals in the control group.

40.3.6 Sample Size for a Cohort Comparison Study

As described previously, the cohort comparison study design is a type of observational study in which the researcher simply observes an outcome without intervening. As a longitudinal study design the cohort study design follows a group of individuals with similar characteristics either forward in time (prospectively) or backward in time (retrospectively).

In the cohort comparison study design a group demonstrating the characteristic(s) of interest are followed for a period of time while being compared to a similar group or multiple similar comparison groups (the cohorts) that do not demonstrate the characteristic(s) of interest. The researcher is intending to measure specific variables within the designated cohort of interest and to compare such measures to those reported for the comparison cohort(s). Throughout the monitoring stage, the selected measures are recorded at the onset of the monitoring activity, at pre-designated time points throughout the study, and at the completion of the study.

The formula to compute the sample size for the group of interest in a cohort comparison study where the data are normally distributed is shown here:

(1)

However, if the data are based on a chi-square distribution the recommended approach by Fleiss (1981) is to use the following continuity correction formula, after computing the initial sample size with formula (1) above.

(2)

The elements required for formula (1) and formula (2) are shown here.

i). The alpha level – also referred to as the level of statistical significance…the value against which the estimated "test statistic" will be compared to determine if there is something happening in the research question under investigation (i.e. the drug worked, the neighbourhoods differed, more symptoms were reported, the light is brighter, the sound was louder, etc.). The alpha level is also referred to as the probability of committing a Type I error (failure to accept the null hypothesis when in fact it is true). The typical value for the alpha level is 0.05 (also written as = 0.05).

ii). The beta level – also associated with statistical power as in power = 1- beta. The beta value is an estimate of the probability associated with making a Type II statistical error (i.e. failure to reject the null hypothesis when in fact it is false).

According to Cohen (described in Fleiss, 1981), given that committing a Type I error is four times as serious as committing a Type II error, a researcher should set the beta value to 4 x alpha. That is, when a researcher states an alpha value of 0.05, the corresponding beta value should be set to 4 x 0.05 = 0.20.

A beta value of 0.20 is therefore an indication of the researcher's willingness to accept a 20% chance of missing an event (i.e. the effect) that actually occurred. Considering the concept of power, a beta value of 0.20 represents a statistical power level of 0.80 or 80%. The typical value for the beta level is 0.20.

iii). The ratio estimate represented by the letter m in the formula refers to the ratio of the number of control (comparison) participants to the number of participants of interest. In the computation of sample size for a prospective multiple cohort design the researcher may be faced with cohorts of different sizes. The ratio term refers to the fraction of difference between the cohort of interest and the control or comparison cohorts. In computing sample size for the prospective multiple cohort comparison, the researcher may wish to consider that the group of interest is half as large as the control group, in which case the ratio value is presented as 0.5:1. Similarly, the researcher may consider a ratio of 2:1 or 3:1 for the group of interest and the control group. Likewise, in some situations the researcher may consider ratios as high as 5:1, 10:1, or even 20:1. The value entered for m is simply the value of the ratio.

1. iv) The term P1 represents the expected proportion in the group of interest. In the computation of sample size for a prospective multiple cohort design the researcher may have access to previous research that indicates the expected proportion of outcome for individuals within a given cohort, or the expected proportion of individuals that are considered exposed or present with a given characteristic in a study.

For example, in previous research measuring the epidemiology of injuries in ice hockey, Montelpare, Pelletier and Stark (1996) reported injury rates that ranged from 17% to 68%, with an average proportion of injured among individuals that body check of about 43% (a proportion value of 0.43).

In computing the sample size for a prospective multiple cohort design the researcher should enter a decimal value to represent the expected proportion (i.e. outcome proportion) in their study.

1. v) The term P0 represents the expected proportion in the CONTROL group. In the computation of sample size for a prospective multiple cohort design the researcher may also have access to previous research that indicates the expected proportion of outcome for individuals within the control group or not exposed or not at risk cohort. If unsure about the true value of the expected proportion for the control group then enter 0.50 as this would be considered as an unbiased level of expected exposure or risk. However, the formula take any value for P0 between 0 and 1.
2. vi) The term refers to the average expected proportion and is computed from the formula shown here, which includes the estimate of the ratio score and the estimate for the expected proportion in the control group.

Application of the sample size formula for a cohort comparison study

In order to determine the effects of cannabis smoking on driving related events, you decide to develop a prospective cohort comparison study, by following a group of individuals that self-report driving after using cannabis containing the active ingredient delta-9-tetrahydracannabinol (. The comparison control group will be composed of a similar number of individuals (i.e. 1:1 ratio) that self-report not using cannabis and also self-report that they do not drive after using alcohol. The expected proportion of traffic related events associated with cannabis use is based on previous work by Kelly, Darke, and Ross ([2]004)[3] which reported that approximately 4% of the general population drive while under the influence of drugs but that 25% of traffic related events involve drivers who tested positive for using drugs.

Based on the formula above, the data used to determine the size of each cohort (observed and control) are as follows:

n =

n = =

n 57

Ratio (m) 1:1

= 0.145

The results of this computation suggest that both the cohort of interest, individuals that self-report driving after using cannabis and the control cohort, individuals that abstain from cannabis as well as alcohol when driving should include approximately 57 individuals each.

Verifying the computations shown above with the following SAS code.

```
DATA COHORT2;

  /*
  Begin by establishing the alpha and beta terms as probabilities and as z-scores.
  */

  ALPHA = 0.05;  ZALPHA = 1.96; BETA = 0.20; ZBETA = 1.28;
  M = 1;  P_1 = 0.25;  P_0 = 0.04; PBAR = 0.145;

  /*
  Next unpack the elements of the formula above by creating variables and working through the mechanics of the formula.
  For example: create a variable called NUMRTR1A to represent the first element of the numerator in the formula. Here the element is to compute just the value under the square root sign
  */
```

```
NUMRTR1A = SQRT((1+(1/M))* (PBAR*(1-PBAR)));

/*
Multiply the value computed above by the ZALPHA term, which you declared in the second line of the program as ZALPHA = 1.96
*/

NUMRTR1B = (ZALPHA*NUMRTR1A);

/*
Work through each element of the formula as you have above to simplify the stepwise calculations.
Here we create the third numerator element NUMRTR1C by calculating the value under the square root sign
*/

NUMRTR1C = SQRT(((P_0*(1-P_0))/M)+ (P_1*(1-P_1)));

/*
We multiply the value of NUMRTR1C by the ZBETA term that we declared in line 2 of the program as ZBETA = 1.28
*/
NUMRTR1D = (ZBETA*NUMRTR1C);

/* Strictly adhere to the rules of BEDMAS to ensure that the formula is deconstructed and reconstructed in the proper order.
For example, after computing the two elements of the numerator that are contained within the square brackets, add these values together, and then raise them to the exponent 2.
*/

NUMRTR1E = (NUMRTR1B + NUMRTR1D)**2;

/* Next work through the elements of the denominator in the same way as you did with the numerator.
*/

DENOM=(P_0 − P_1)**2;

N_APPRX = (NUMRTR1E/DENOM);

PROC PRINT;
VAR NUMRTR1A NUMRTR1B NUMRTR1C NUMRTR1D NUMRTR1E
DENOM N_APPRX;
RUN;
```

The SAS output shown here was generated from the processing of the formula. The elements produced by this formula are a result of the values declared by the user. For example, in this calculation for sample size, you set an alpha level of $p<0.05$ to control the Type I error. This translates to a Z-alpha value of 1.96. Likewise you set a beta level of 0.20 to control the Type II error, and therefore the corresponding Z-beta term is 0.84. You also knew apriori that the probability of the event of interest was 25%, so you set a P1 value to 0.25, and the probability related to the control was 4% so you

set a P0 value to 0.04. The calculations worked through by hand matched the calculations from the SAS program. That is in both approaches you determined that the sample size for this study should be approximately 57 individuals.

NUMRTR1A	NUMRTR1B	NUMRTR1C	NUMRTR1D	NUMRTR1E	DENOM	N_APPRX
0.5	0.98	0.48	0.61	2.51	0.44	56.92

20.4 Here is what we covered in this chapter.

In this chapter you were introduced to:

- Describing the importance in establishing a sample to represent the population
- Identifying the difference between probabilistic and non-probabilistic sampling strategies.
- Computing sample size under different scenarios using SAS code
- Understanding when a given sample size calculation is most appropriate
- Applying the appropriate sampling strategy to a given research design

[3] Kelly, E., Darke , S, & Ross, J., A review of drug use and driving: epidemiology, impairment, risk factors and risk perceptions, Drug and Alcohol Review(September 2004), 23, 319–344

41. Using Webulator Applications to Compute Sample Size

In this chapter, we will work through four different approaches to sample size calculations using bespoke Webulators based on probabilistic formulae.

1. Determining Sample Size for a Simple Random Sample to Estimate a Population Proportion

Elements of the Webulator and Sample Size for a Population Proportion

The sample size formula for a population proportion is based on the formula shown here.

$$n = \frac{(N \times p \times q) \times [Z_{\alpha}]^2}{(p \times q) \times [Z_{\alpha}]^2 + [N - 1] \times (error)^2}$$

Figure 41.1. Formula to Determine Sample Size for a Population Proportion

Unpacking the formula for simple random sampling we see that N represents the population from which the sample will be drawn; the proportion (*p*) refers to the proportion of individuals displaying the characteristic of interest, while (1-*p*) or (*q*) refers to the proportion of individuals in the population not displaying the characteristic of interest.

To determine the proportion of cases in a population, *p* is the ratio of all individuals displaying the characteristic of interest divided by the set of all cases from which the sample was drawn. In real life, you might determine *p* based on past research or population statistical data such as the census.

The error term in the denominator refers to the expected accuracy or the allowable difference between the estimate of the proportion in the selected sample, as a result of the sample size calculation, and the true population proportion. A typical value here would range between 0.02 and 0.10.

The term refers to the two-tailed standardized score of researcher confidence. If =0.05 then z= 1- = 0.95 or a 95% confidence value.

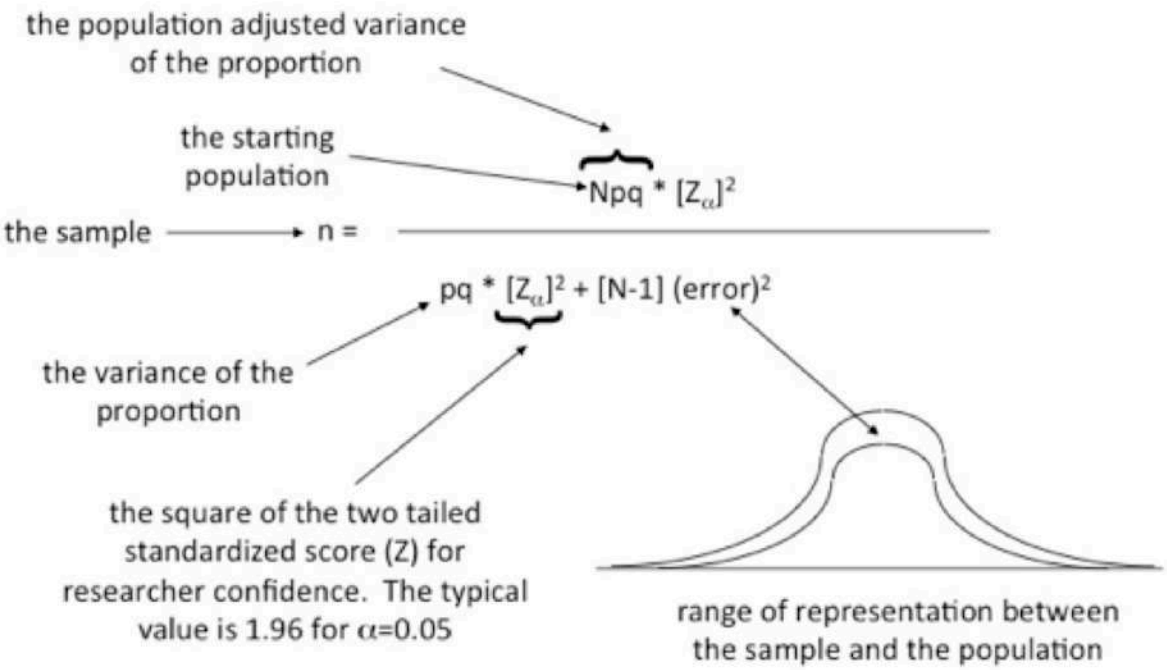


Figure 41.2. Explaining the parts of the formula

Application of the Sample Size Formula for a Population Proportion

Using the Webulator for a Population Proportion you can produce a representative sample for the population based on a simple random sampling approach. Consider that you are asked by the Chief Public Health Officer (CPHO) to determine the sample size required to represent a proportion of persons who inject drugs (PWID) based on a starting population of 157,000 individuals. From previous reports, the proportion of PWID within a sample (i.e. the proportion of the population that displays the characteristic of interest) was about 13% or p = 0.13 and therefore q = (1-p) = 0.87. The CPHO wants the estimate to be within 3% of the true population proportion with 95% confidence. Below is a Webulator to calculate the sample size for this scenario – i.e. a population proportion.

An interactive or media element has been excluded from this version of the text. You can view it online here:
https://pressbooks.library.upei.ca/montelpare/?p=1098

Follow these basic steps to use this webulator to compute the sample size for this scenario, or any situation where you need to determine the sample size for a population proportion.

- Enter 157000 into the Webulator in the field labelled Initial Population
- Next, enter 0.13 of 13 into the Webulator in the field labelled Expected Proportion
- Enter 1.96 into the field labelled Z_{alpha} for the percent confidence, as this is the corresponding z-score for 95% confidence
- Finally, enter the number 3 into the Webulator in the field labelled Percent Error
- Click the button labelled calculate.

Based on the application of the formula for simple random sampling for the estimate of a Population Proportion you report that the CPHO will require a sample size of at least **481** individuals in order to be **95% confident** that the proportion of PWID in her sample will represent the true population proportion of PWID individuals **within 3%**.

---

2. Determining Sample Size for a Comparison Study

In some research designs, we are interested in comparing results between two or more groups. In this case, we may not know the original population size, but we may know about variability with respect to the dependent variables. To determine the sample size for a comparison study we need to know **the measure of central tendency** for the dependent variable and the **amount of variability** we could expect for the measures of the dependent variable.

Typically, the amount of variability is based on information from previous studies. In some situations, this information may come from pilot work or from an actual study. The sample size formula for comparison studies is shown here as:

$$n \geq \frac{2(s)^2 \times [Z_\alpha + Z_\beta]^2}{(\text{expected mean}) \times (\%\ \text{change})}$$

Figure 41.2 Formula to Determine Sample Size for a Population Proportion

The $[Z_{alpha}]$ and $[Z_{beta}]$ terms are provided in the following tables. The area under the normal curve for the $[Z_{beta}]$ term is illustrated in the Figure 41.3

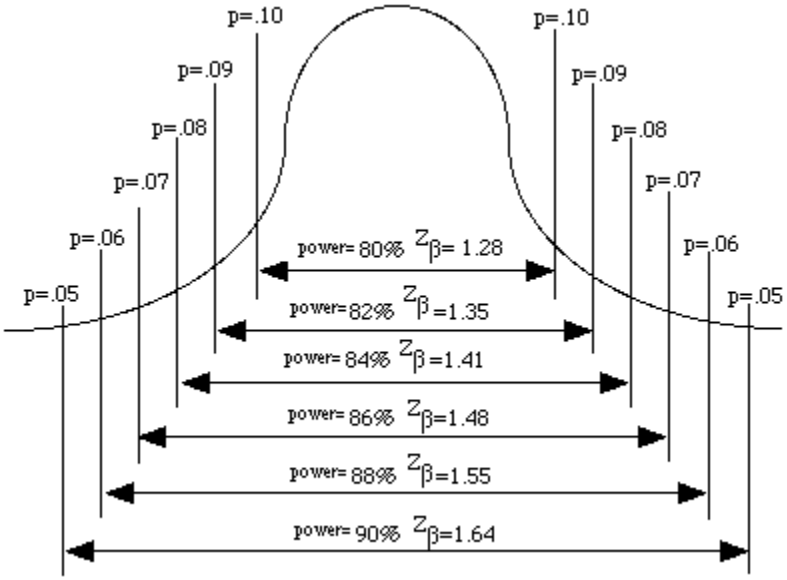


Figure 41.3 Areas under the Normal Curve representing [latex]z_{beta}[/latex] scores

The terms of the formula are explained as follows. The term is based on the variance reported in the literature or from pilot studies. Similarly, the expected mean is based on the mean from the literature or from pilot studies, while the *expected % accuracy is the proximity of the estimated mean score to the true population score and* is set by the researcher to be about 10%. *The terms* Zalpha or Zand Zbeta or Z use the standard scores for the alpha level and for the beta value (power) levels. Common values for and are =0.05 and =1.28, respectively. A more comprehensive table of conversions

of areas under the normal curve representing [latex]z_{alpha}[/latex] and [latex]z_{beta}[/latex] scores are shown in Tables 41.2 and 41.3, below.

Table 41.2 Conversion table to create the Z beta term

(one tailed probability estimate=0.05); beta = 0.10 (power = 90%) ; $[Z_{beta}]$ = 1.64	(one tailed probability estimate=0.06); beta = 0.12 (power = 88%) ; $[Z_{beta}]$ = 1.55
(one tailed probability estimate=0.07); beta = 0.14 (power = 86%) ; $[Z_{beta}]$ = 1.48	(one tailed probability estimate=0.08); beta = 0.16 (power = 84%) ; $[Z_{beta}]$ = 1.41
(one tailed probability estimate=0.09); beta = 0.18 (power = 82%) ; $[Z_{beta}]$ = 1.34	(one tailed probability estimate=0.10); beta = 0.20 (power = 80%) ; $[Z_{beta}]$ = 1.28

Table 41.3 Conversion table for alpha and percent confidence to Z

(alpha probability estimate of 0.10)=90% $[Z_{alpha}]$ = 1.64	(alpha probability estimate of 0.09)=91% $[Z_{alpha}]$ = 1.70
(alpha probability estimate of 0.08)=92% $[Z_{alpha}]$ = 1.75	(alpha probability estimate of 0.07)=93% $[Z_{alpha}]$ = 1.81
(alpha probability estimate of 0.06)=94% $[Z_{alpha}]$ = 1.88	(alpha probability estimate of 0.05)=95% $[Z_{alpha}]$ = 1.96
(alpha probability estimate of 0.04)=96% $[Z_{alpha}]$ = 2.05	(alpha probability estimate of 0.03)=97% $[Z_{alpha}]$ = 2.17
(alpha probability estimate of 0.02)=98% $[Z_{alpha}]$ = 2.33	(alpha probability estimate of 0.01)=99% $[Z_{alpha}]$ = 2.58

Application of the Sample Size Formula for a Comparison Study

In a recent healthy heart study, researchers measured the effects of red wine consumption on blood cholesterol concentrations of males over the age of 35 years. The researchers showed that males (n1= 133) who consumed on average one serving of red wine per day, for a minimum of six days per week, had a lower concentration of the athero-genic low-density lipoprotein cholesterol than a group of age-matched control subjects (n2= 143) who abstained from any alcohol consumption. The investigation followed the total group of 276 males for a 24 week period.

You believe that the data are valuable and therefore you wish to conduct the study with a group of males in your local community. In the reference study, the average cholesterol concentration for the sample of interest (red wine consumers) was 4.6 mmol/L with a standard deviation of 0.32 millimoles per litre (mmol/L). Considering that you wish to use an alpha level of 0.05 with a corresponding beta level of 4 x alpha (where beta = 4 x 0.05 = 0.20) and a power level of "1 – beta = 0.80". Further, you expect that the difference between your estimate of the mean and the TRUE estimate of the mean is within 3 percent.

Using the Webulator for a Comparison Study you can produce a representative sample for the population based on a simple random sampling approach. The data you need in order to compute the appropriate sample size is:

1. The estimated mean from previous studies = 4.6 mmol/L
2. The standard deviation from previous studies = 0.32

3. The z scores for the alpha probability (.05), Zα=1.96
4. The z score for the beta probability (.20), Zβ=1.28
5. The allowable percent difference between your estimate for the dependent variable and the expected estimate for the dependent variable from the true population = 3%

An interactive or media element has been excluded from this version of the text. You can view it online here: https://pressbooks.library.upei.ca/montelpare/?p=1098

The results of this computation indicate that in order to be 95% confident that the estimates for the proposed sample will be within 3 percent of the true population value, you will need to have a minimum of 16 participants in the test group and 16 participants in the control group.

---

3. Determining Sample Size for a Case-Control Study

In the case-control study design, individuals with a specific measurable condition are "compared" to individuals that do not demonstrate the condition of interest. The case-control design is a retrospective study design type that evaluates, by comparison, the differences in outcome measures between groups of individuals with and without a disease, or the signs/symptoms of a condition.

Case-control studies are useful in demonstrating associations but may not show causation. The temporal characteristics (elements of time) are important in demonstrating the relationship. An essential consideration in a case-control study is the clear definition of the cases and of the controls.

In a case-control design we may also consider that the cases are more likely to occur given exposure to the stimulus, to which we say this is a directional hypothesis or a one-tailed hypothesis. If we consider a one-tailed decision rule (cases are more likely than controls, given the characteristics of the scenario) then we see that a power of 80% has a beta term of 0.20 and a Zβ= 0.084. Estimates of statistical power for the one-tailed (directional hypothesis) and corresponding Zβ values are shown in the following table.

Table 41.5 Power estimates from the [latex]z_{beta}[/latex] terms in a one-tailed hypothesis

POWER	[latex]z_{beta}[/latex]
80%	0.84
85%	1.04
90%	1.28
95%	1.64

The sample size formula to determine the number of cases (or the number of controls) in each group of a case-control study is shown here as:

$$\text{“n” (each group)} = \frac{(p_0 q_0 + p_1 q_1)\,(z_{1-\alpha/2} + z_{1-\beta})^2}{(p_1 - p_0)^2}$$

Figure 41.4. Formula to Determine Sample Size for a Case-Control Application

The formula differs slightly from that published more recently by Kasiulevicius, Sapoka, and Filipaviciute (2006)[1] but produces similar estimates. Where the elements of the formula include: the proportion of cases among those individuals suspected to have been exposed. : the proportion of cases among those individuals suspected to not have been exposed. : is the Z score for the term, where 1.96, and: is the Z score for the term for a one-tailed directional hypothesis (0.84).

Application of the sample size formula for a case-control study

In order to determine the effects of cannabis smoking on lung cancer you decide to conduct a retrospective case control study in which your sample size estimate is based on the consideration that the relative risk of lung cancer among frequent cannabis smokers is about 5.7 times that of non-smokers. You decide to use p1 = 0.285 to represent the proportion lung cancer patients who were frequent cannabis smokers and p0 = 0.05 to represent the proportion lung cancer patients as controls who never smoked cannabis.

The data needed to compute the appropriate sample size are shown here as:

1. Proportion of individuals that had lung cancer among individuals that were considered frequent smokers of cannabis, P1 = 0.285
2. Proportion of individuals that had lung cancer among individuals that never smoked cannabis, P0 = 0.05
3. The z scores for the alpha term (α=0.05), Zα=1.96
4. The z scores for the beta term based on a one-tailed hypothesis, [latex]z_{beta}[/latex] =0.84

> *An interactive or media element has been excluded from this version of the text. You can view it online here: https://pressbooks.library.upei.ca/montelpare/?p=1098*

As you can see in the results of the computation using the Webulator the appropriate sample size needed to conduct your study will require at least 36 individuals in the case group and at least 36 individuals in the control group.

---

4. Determining Sample Size for a Cohort Comparison Study

As described previously, the cohort comparison study design is a type of observational study in which the researcher simply observes an outcome without intervening. As a longitudinal study design, the cohort study design follows a group of individuals with similar characteristics either forward in time (prospectively) or backward in time (retrospectively).

In the cohort comparison study design a group demonstrating the characteristic(s) of interest are followed for a period of time while being compared to a similar group or multiple similar comparison groups (the cohorts) that do not demonstrate the characteristic(s) of interest. The researcher is intending to measure specific variables within the designated cohort of interest and to compare such measures to those reported for the comparison cohort(s). Throughout the monitoring stage, the selected measures are recorded at the onset of the monitoring activity, at pre-designated time points throughout the study, and at the completion of the study.

The formula to compute the sample size for the group of interest in a cohort comparison study where the data are normally distributed is shown here:

[latex]n ={ \left[Z_\alpha \sqrt{ (1+{\left(1\over{m}\right)}) \times{\left(\overline{p} (1-{\overline{p}})\right) }} + Z_\beta\sqrt{{p_0(1-p_0)\over{m}} + p_1 (1-p_1)}\right]^2 \over{(p_0 - p_1)^2} }[/latex]

The essential elements required for the computation of sample sizeinclude:

The sample size computations are based on Fleiss (1981). The following SAS program was originally written by Dr. P.N. Corey (University of Toronto) and includes a continuity correction that was not included in Fleiss' original calculations. This SAS program uses a probit function to enable the input of any alpha and beta values.

The SAS program to determine a sample size for a cohort comparison based on Fleiss (1981) is shown here.

```
* SAMPLE SIZES FOR THE COMPARISON OF TWO INDEPENDENT SAMPLES;

* PROGRAM NAME IS SScohort.SAS ;

   OPTIONS PS = 65 LS = 80 NODATE NONUMBER ;
   TITLE1 'SAMPLE SIZE DETERMINATION USING THE FORMULAE FROM Fleiss 1981';
   ************************************************************
** Program SAMPSIZEFLEISSA.SAS has been modified to allow **
** DO LOOPS to be defined in %LET statements outside the **
** body of the program and calculates sample sizes N1 = m **
** and N2 = rm for a cohort or cross-sectional study that **
** involves the comparison of two independent samples **
** using a correction factor for continuity. The program **
** makes a small modification to the program SAMPSIZE2.FLS**
** by defining PBAR differently. **
************************************************************;
********* PARAMETER DEFINITION USING %LET STATEMENT ********;
%LET ALPHA = 0.05 ;
%LET BETA = 0.20 ;
%LET RATIOLIST = 0.50, 1.0, 5.0, 20.0;
%LET P1LIST = 0.01, 0.05, 0.10, 0.2, 0.43;
%LET RRLIST = 1.5, 2, 3 ;
   ************************************************************
* If we use the DO LOOP DO RATIO = (1/2),1,2 we would be *
* asking the program to estimate the sample sizes for the *
* following situations *
```

```
* (a) N1 is one half the number N2 *
* (b) N1 is equal to N2 *
* (c) N1 is twice the size of N2 *
************************************************************;
************************************************************
* If we use the DO LOOP DO P1 = 0.10, 0.20 we would be *
* asking the program to estimate the sample sizes for the *
* following situations *
* (a) P1 = 0.10 *
* (b) P1 = 0.20 *
************************************************************;
************************************************************
* If we use the Do LOOP DO RELRISK = 2 to 4 we would be *
* asking the program to estimate the sample sizes for the *
* following situations *
* (a) P2 is twice the size of P1 *
* (b) P2 is three times the size of P1 *
* (c) P2 is four times the size of P1 *
************************************************************;
********** NO CHANGES BEYOND THIS POINT IN PROGRAM *********;
DATA TEMP1 ;
ALPHA = 0 + &ALPHA ;
BETA = 0 + &BETA ;
DO RATIO = &RATIOLIST ; ***<<<– See explanation above ***;
DO P1 = &P1LIST ; ***<<<– See explanation above ***;
DO RELRISK = &RRLIST ; ***<<<– See explanation above ***;
POWER = 1 – &BETA ;
ZALPHA = PROBIT((1 – &ALPHA/2));
ZBETA = PROBIT((1 – &BETA)) ;
P2 = RELRISK * P1 ; Q1 = 1 – P1 ; Q2 = 1 – P2 ; DELTA = P2 – P1 ;
PBAR = (P1 + RATIO * P2 ) / (RATIO + 1) ; QBAR = 1 – PBAR ;
   NUMERAT1 = ZALPHA * SQRT((RATIO + 1) * PBAR*QBAR) +
ZBETA * SQRT(RATIO*P1*Q1 + P2*Q2) ;
M1 = ((NUMERAT1/DELTA)**2) / RATIO ;
M2 = M1 + (RATIO + 1)/(RATIO * ABS(P2 – P1)) ; M = INT(M2 + 1) ;
TOTALN = M + RATIO * M ;
OUTPUT ; END ; END ; END ;
PROC PRINT DOUBLE NOOBS ;
VAR RATIO P1 P2 RELRISK M1 M2 M TOTALN ;
TITLE1 'SAMPLE SIZE DETERMINATION USING THE FORMULAE FROM THE BOOK';
TITLE2 'STATISTICAL METHODS FOR RATES AND PROPORTIONS' ;
TITLE3 'BY JOSEPH L. FLEISS SECOND EDITION JOHN WILEY AND SONS' ;
TITLE4 "FOR TYPE I ERROR RATE &ALPHA AND TYPE II ERROR RATE &BETA";
RUN ;
```

The output generated by the SScohort.sas program is shown below.

Sample size determination using the formulae from “Statistical Methods for Rates & Proportions” by Joseph L. Fleiss, 2nd Edition, John Wiley & Sons

RATIO	P1	P2	RELRISK	M1	M2	M	TOTALN
0.5	0.01	0.015	1.5	11307.92	11907.92	11908	17862.0
0.5	0.01	0.020	2.0	3316.17	3616.17	3617	5425.5
0.5	0.01	0.030	3.0	1070.02	1220.02	1221	1831.5
0.5	0.05	0.075	1.5	2148.71	2268.71	2269	3403.5
0.5	0.05	0.100	2.0	623.23	683.23	684	1026.0
0.5	0.05	0.150	3.0	196.32	226.32	227	340.5
0.5	0.10	0.150	1.5	1003.79	1063.79	1064	1596.0
0.5	0.10	0.200	2.0	286.59	316.59	317	475.5
0.5	0.10	0.300	3.0	87.07	102.07	103	154.5
0.5	0.20	0.300	1.5	431.30	461.30	462	693.0
0.5	0.20	0.400	2.0	118.21	133.21	134	201.0
0.5	0.20	0.600	3.0	32.31	39.81	40	60.0
0.5	0.43	0.645	1.5	124.93	138.88	139	208.5
0.5	0.43	0.860	2.0	27.76	34.74	35	52.5
0.5	0.43	1.290	3.0	.	.	.	.
1.0	0.01	0.015	1.5	7749.59	8149.59	8150	16300.0
1.0	0.01	0.020	2.0	2318.16	2518.16	2519	5038.0
1.0	0.01	0.030	3.0	768.01	868.01	869	1738.0
1.0	0.05	0.075	1.5	1470.49	1550.49	1551	3102.0
1.0	0.05	0.100	2.0	434.43	474.43	475	950.0
1.0	0.05	0.150	3.0	140.10	160.10	161	322.0
1.0	0.10	0.150	1.5	685.60	725.60	726	1452.0
1.0	0.10	0.200	2.0	198.96	218.96	219	438.0
1.0	0.10	0.300	3.0	61.60	71.60	72	144.0
1.0	0.20	0.300	1.5	293.15	313.15	314	628.0
1.0	0.20	0.400	2.0	81.22	91.22	92	184.0
1.0	0.20	0.600	3.0	22.33	27.33	28	56.0
1.0	0.43	0.645	1.5	83.23	92.53	93	186.0
1.0	0.43	0.860	2.0	18.21	22.87	23	46.0
1.0	0.43	1.290	3.0	.	.	.	.
5.0	0.01	0.015	1.5	4876.35	5116.35	5117	30702.0
5.0	0.01	0.020	2.0	1497.47	1617.47	1618	9708.0
5.0	0.01	0.030	3.0	509.46	569.46	570	3420.0
5.0	0.05	0.075	1.5	922.50	970.50	971	5826.0
5.0	0.05	0.100	2.0	278.84	302.84	303	1818.0
5.0	0.05	0.150	3.0	91.63	103.63	104	624.0
5.0	0.10	0.150	1.5	428.25	452.25	453	2718.0
5.0	0.10	0.200	2.0	126.50	138.50	139	834.0
5.0	0.10	0.300	3.0	39.39	45.39	46	276.0
5.0	0.20	0.300	1.5	181.10	193.10	194	1164.0

RATIO	P1	P2	RELRISK	M1	M2	M	TOTALN
5.0	0.20	0.400	2.0	50.30	56.30	57	342.0
5.0	0.20	0.600	3.0	13.23	16.23	17	102.0
5.0	0.43	0.645	1.5	48.78	54.36	55	330.0
5.0	0.43	0.860	2.0	9.33	12.12	13	78.0
5.0	0.43	1.290	3.0	.	.	.	.
20.0	0.01	0.015	1.5	4330.01	4540.01	4541	95361.0
20.0	0.01	0.020	2.0	1337.06	1442.06	1443	30303.0
20.0	0.01	0.030	3.0	455.75	508.25	509	10689.0
20.0	0.05	0.075	1.5	818.22	860.22	861	18081.0
20.0	0.05	0.100	2.0	248.38	269.38	270	5670.0
20.0	0.05	0.150	3.0	81.54	92.04	93	1953.0
20.0	0.10	0.150	1.5	379.23	400.23	401	8421.0
20.0	0.10	0.200	2.0	112.27	122.77	123	2583.0
20.0	0.10	0.300	3.0	34.73	39.98	40	840.0
20.0	0.20	0.300	1.5	159.68	170.18	171	3591.0
20.0	0.20	0.400	2.0	44.15	49.40	50	1050.0
20.0	0.20	0.600	3.0	11.22	13.85	14	294.0
20.0	0.43	0.645	1.5	42.00	46.88	47	987.0
20.0	0.43	0.860	2.0	7.26	9.71	10	210.0
20.0	0.43	1.290	3.0	.	.	.	.

*TYPE I ERROR RATE 0.05 AND TYPE II ERROR RATE 0.20

42. Computer Simulation and Random Number Generators

Research using simulated data is often done to predict future events based on real-world data. For example, health researchers can use demographic information about a cohort within the population or about the health care workforce to predict how many new health care professionals we will need in the years to come. This information can then be used to make decisions about the number of students that universities and colleges should accept into their programs in order to meet the predicted needs. Likewise, using computer simulation tools, administrators can create financial forecasting models based on selected expenditure statements and health sector administrative data to estimate future costs and establish budgetary guidelines appropriately.

Creating and using a simulated dataset is also an excellent way to practice the application of statistical methods without having to collect real-world data. That is, we can create a simulated dataset by first establishing the set of independent and dependent variables that are of interest to us in our research project. Next we provide a range of possible responses for each variable. Then we use a random number generator to create a dataset that is estimated from the set of values we provide to the computer.

This is an amazing learning opportunity as it enables you to create, albeit artificially, a complete dataset with the variables in which you are interested. The experience is invaluable as it provides you with the opportunity to critically evaluate both the strategies for input as well as the interpretation for output. Although not required, working through a computer-simulated dataset during the development of your research proposal can help you develop your data analysis plan, and enable you to become familiar with the ranges and nuances of the important variables.

In the following program we will generate a single value based on a SAS random number generator. Here we will control the function of the random number generator by controlling the parameters of the processor to ensure that our output falls within a specific range.

Annotated Example Using SAS to produce data for a random variable.

```
DATA SASRNG_01;
```

The following SAS program creates a continuous-discrete variable that we will call AGE. We begin by initializing the variable with the SAS RAND function and we use *1000000 to create the range of the distribution (0 to 1 million) from which the SAS RAND function will draw a number

```
AGE=RAND("NORMAL")*1000000;
```

Next we set the range for the number so that it is in a logical age range for our use. The statements below ensure that the SAS RAND function will produce a value that is above 40 but less than 72

```
AGE=40+ABS((MOD(AGE,62)));
IF AGE>72 THEN AGE=25+ABS((MOD(AGE,50)));
IF AGE<40 THEN AGE=AGE+ABS((MOD(AGE,50)));
AGE=ROUND(AGE);
RUN;
```

We then print the value that we produced with the random number generator.

```
PROC PRINT; VAR AGE;
RUN;
```

The program above produced one age value within a predesignated age range.

OBS	AGE
1	52

Using a random number generator to produce ICD-9 codes

In this next example, we will produce a randomly generated dataset consisting of ICD-9 codes. In this SAS program we first create the data, then we organize the output into categories, and finally we produce a horizontal bar chart of the relative percentile values for each category of ICD-9 codes.

Consider the following program to evaluate the primary diagnosis for a group of patients visiting a healthcare clinic. The data are generated using a customized random number generator that generates data in the form of ICD-9[1] codes. Since the codes are based on a continuous number line several unique values can be generated to represent the various sub-conditions of that which a patient may present to a healthcare provider. Here we simplify the organization of the codes by creating categories and using the SAS PROC FORMAT command to assign the categories to the output.

SAS Code To Organize Categories Of ICD-9 Codes

```
PROC FORMAT;
  VALUE CATFMT 1='Infectious/parasitic'
  2='Neoplasms'
  3=' Endo/nutri/metabolic'
  4=' Blood/blood-forming organs'
  5=' Mental disorders'
  6=' Nervous system'
  7=' Sense organs'
  8=' Circulatory system'
  9=' Respiratory system'
  10=' Digestive system'
  11=' Genitourinary system'
  12=' Pregnancy/childbirth'
  13=' Skin & subcutaneous tissue'
  14=' MSK & connective tissue'
  15=' Congenital anomalies'
  16=' Perinatal period Conditions'
  17=' Injury and poisoning'
  20=' Diagnosis not reported'
  18=' Supplementary classification';
```

In this example, the random number generator produces ICD-9 scores as the dependent variable which we assign with the label (PRDIAG). The SAS code uses a DO loop to create a set of 500 scores, representing ICD-9 score for each patient. The data are drawn from a normal distribution at random using the command: PRDIAG=RAND("NORMAL")*10000; We seed the random number generator for k=500 times with the CALL STREAMINIT(K); command. We also set a maximum absolute value for the dependent variable using the modulus math function MOD().

```
DO K=1 TO 500;
   CALL STREAMINIT(K); /* SEED THE RNG ON EACH LOOP FOR K TIMES */
   PRDIAG=RAND("NORMAL")*10000;
   /* SET MAX RANDOM NUMBER TO 1500 */
   PRDIAG=0+ABS((MOD(PRDIAG,1500)));
   /* ROUND THE RANDOM NUMBERS TO 2 DECIMAL PLACES */
   PRDIAG=ROUND(PRDIAG,.01);
```

Next we use if-then logic statements to organize the randomly generated numbers into specific categories based on specific cutpoints. Notice these commands are included within the DO loop. The loop is closed with the commands OUTPUT; followed by END;

```
IF PRDIAG = 95 OR PRDIAG = 99 THEN CATEGORY=20;
   IF PRDIAG >=001 AND PRDIAG<94 THEN CATEGORY=1;
   IF PRDIAG >=96 AND PRDIAG<99 THEN CATEGORY=1;
   IF PRDIAG >99 AND PRDIAG<140 THEN CATEGORY=1;
   IF PRDIAG >=140 AND PRDIAG<240 THEN CATEGORY=2;
   IF PRDIAG >=240 AND PRDIAG<280 THEN CATEGORY=3;
   IF PRDIAG >=280 AND PRDIAG<290 THEN CATEGORY=4;
   IF PRDIAG >=290 AND PRDIAG<320 THEN CATEGORY=5;
   IF PRDIAG >=320 AND PRDIAG<390 THEN CATEGORY=6;
   IF PRDIAG >=390 AND PRDIAG<460 THEN CATEGORY=7;
   IF PRDIAG >=460 AND PRDIAG<520 THEN CATEGORY=8;
   IF PRDIAG >=520 AND PRDIAG<580 THEN CATEGORY=9;
   IF PRDIAG >=580 AND PRDIAG<630 THEN CATEGORY=10;
   IF PRDIAG >=630 AND PRDIAG<677 THEN CATEGORY=11;
   IF PRDIAG >=680 AND PRDIAG<710 THEN CATEGORY=12;
   IF PRDIAG >=710 AND PRDIAG<740 THEN CATEGORY=13;
   IF PRDIAG >=740 AND PRDIAG<760 THEN CATEGORY=14;
   IF PRDIAG >=760 AND PRDIAG<780 THEN CATEGORY=15;
   IF PRDIAG >=780 AND PRDIAG<800 THEN CATEGORY=16;
   IF PRDIAG >=800 AND PRDIAG<1000 THEN CATEGORY=17;
   IF PRDIAG >=1000 THEN CATEGORY=18;
   OUTPUT;
   END;
```

The SAS commands to create a frequency distribution table are shown below. By using a frequency distribution table the author can provide a standard presentation of important summry statistics within the data set. For example, here we show the organization of the randomly generated numbers within each of the designated categories while also presenting the relative percentages that the categories represent within this data set (see Cumulative Percent column). The frequency distribution table is followed by the horizontal bar chart of the percentage of diagnoses within each category. In this figure we included the data values at the end of each horizontal bar.

```
PROC FREQ; TABLES CATEGORY;
   TITLE1 'FREQUENCY DISTRIBUTION FOR RNG ICD-9 CODES';

   PROC SGPLOT DATA=PRDIAG; HBAR CATEGORY/ GROUPDISPLAY = CLUSTER
```

STAT=PERCENT DATALABELFITPOLICY=NONE DATALABEL;
XAXIS LABEL=”PERCENT OF CASES”;
YAXIS LABEL=”DISEASE/DIAGNOSIS CATEGORIES”;
FORMAT CATEGORY CATFMT. ;
TITLE1 ‘PERCENT OF REPORTED DIAGNOSIS CATEGORY’; RUN;

Frequency distribution for RNG ICD-9 codes
The FREQ Procedure

CATEGORY	FREQUENCY	PERCENT	CUMULATIVE FREQUENCY	CUMULATIVE PERCENT
1	61	12.20	61	12.20
2	41	8.20	102	20.40
3	10	2.00	112	22.40
4	2	0.40	114	22.80
5	7	1.40	121	24.20
6	24	4.80	145	29.00
7	18	3.60	163	32.60
8	22	4.40	185	37.00
9	18	3.60	203	40.60
10	11	2.20	214	42.80
11	20	4.00	234	46.80
12	16	3.20	250	50.00
13	8	1.60	258	51.60
14	13	2.60	271	54.20
15	7	1.40	278	55.60
16	6	1.20	284	56.80
17	57	11.40	341	68.20
18	159	31.80	500	100.00

7

[1] ICD-9 codes refer to the International Classification of Disease Codes – version 9.

Consider an example using the Lotto 649

combination of six numbers from 1 to 49 is extremely low:
1/(49C6)

Which expands to 1 chance in 13,983,816 combinations.

Considering the low probability of winning the grand prize (i.e. all 6 numbers the player chooses will be selected), it is expected that the Lotto 649 lottery should be strategy free. If however, the selection process is not random, but rather follows a specific pattern, then the chance of winning will not remain constant and a strategy to predict outcome could be developed.

Here we will generate data for one draw. That is, using SAS code we will create a random number generator to produce a unique set of 6 numbers that simulates the data that could be generated by the Lotto 649.

The program to generate 6 numbers at random from a set of 49 numbers is shown here. In this instance we have a few constraints. First, we need to be sure that once the first number is drawn, it is not placed back into the set of 49 to be redrawn on a subsequent step. This is because the lotto uses a strategy of **sampling without replacement** and therefore each draw selects only 6 unique numbers. Likewise, in presenting the output from the random number gnerators we need to be sure that the data are reported as discrete scores and not as decimal based continuous scores; and finally, in filtering the numbers produced we need to be sure that the numbers range from 1 to 49 inclusive.

Copy the following program to your SAS workspace and run the program to see which lucky lottery numbers you can produce. This program has several important features that are noted by the comments /* comment */ within the code.

```
/* NOTE THE CALL STREAMINIT(13); Command

To create reproducible random numbers then seed the system with the streaminit command. If RAND() is used without an initial streaminit the program will use the value of the system clock and the random numbers will change each time the program is run.
*/

DATA LOTTO1;

* CALL STREAMINIT(13); /* CREATES REPRODUCIBLE NUMBERS */
DO UNTIL (CHOICE1 NE 0);
CHOICE1 = RAND("NORMAL")*1000000000000;
CHOICE1 = ROUND(CHOICE1);
CHOICE1 = 1+(MOD(CHOICE1,49));
CHOICE1 = ABS(CHOICE1);
END;
* CALL STREAMINIT(999);
DO UNTIL (CHOICE2 NE CHOICE1 AND CHOICE2 NE 0);
CHOICE2 = RAND("NORMAL")*1000000000000;
CHOICE2 = ROUND(CHOICE2);
CHOICE2 = 1+(MOD(CHOICE2,49));
CHOICE2 = ABS(CHOICE2);
END;
* CALL STREAMINIT(28);
DO UNTIL (CHOICE3 NE CHOICE2 AND CHOICE3 NE CHOICE1 AND CHOICE3 NE 0);
CHOICE3 = RAND("NORMAL")*1000000000000;
CHOICE3 = ROUND(CHOICE3);
CHOICE3 = 1+(MOD(CHOICE3,49));
CHOICE3 = ABS(CHOICE3);
```

```
END;
* CALL STREAMINIT(218);
DO UNTIL (CHOICE4 NE CHOICE3 AND CHOICE4 NE CHOICE2 AND CHOICE4 NE CHOICE1 AND CHOICE4 NE 0);
CHOICE4 = RAND("NORMAL")*1000000000000;
CHOICE4 = ROUND(CHOICE4);
CHOICE4 = 1+(MOD(CHOICE4,49));
CHOICE4 = ABS(CHOICE4);
END;

* CALL STREAMINIT(28);
DO UNTIL (CHOICE5 NE CHOICE4 AND CHOICE5 NE CHOICE3 AND CHOICE5 NE CHOICE2 AND CHOICE5 NE
CHOICE1 AND CHOICE5 NE 0);
CHOICE5 = RAND("NORMAL")*1000000000000;
CHOICE5 = ROUND(CHOICE5);
CHOICE5 = 1+(MOD(CHOICE5,49));
CHOICE5 = ABS(CHOICE5);
END;

* CALL STREAMINIT(68);
DO UNTIL (CHOICE6 NE CHOICE5 AND CHOICE6 NE CHOICE4 AND CHOICE6 NE CHOICE3 AND CHOICE6 NE
CHOICE2 AND CHOICE6 NE CHOICE1 AND CHOICE6 NE 0);
CHOICE6 = RAND("NORMAL")*1000000000000;
CHOICE6 = ROUND(CHOICE6);
CHOICE6 = 1+(MOD(CHOICE6,49));
CHOICE6 = ABS(CHOICE6);
END;
RUN;
PROC PRINT; VAR CHOICE1 CHOICE2 CHOICE3 CHOICE4 CHOICE5 CHOICE6;
RUN;
```

Obs	CHOICE1	CHOICE2	CHOICE3	CHOICE4	CHOICE5	CHOICE6
1	37	32	48	11	26	30

So then how many combinations of six numbers are we really talking about?

To compute the number of possible combinations of 6 numbers from the 49 numbers, we need to use the following combinatorial (or factorial) formula. We have 49 numbers choose 6. The number 49 represents the population from which the sample of 6 numbers will be chosen. We write the formula for determining the combinations using the following combinatorial equation:

or we may wish to write the formula using a factorial format as:

Therefore the number of all possible combinations of 6 numbers from a set of 49 consecutive numbers is:

=

Yet you won't be happy unless all of your numbers were chosen, but REALLY what is the chance that all six of your numbers will be selected by the lottery machine. Well since you only bought one ticket, then your chance of winning the lottery is 1 in 13,983,816 chances, or

The value 0.000000071 represents the probability associated with your set of scores.

While this example is fairly straight-forward it is somewhat abstract and is not guaranteed to make you a winner. It does however present the basic concepts in presenting a value for a variable that is generated randomly from the set of all possible outcomes. Let's now turn our attention to an applied health example and see how we can use the utility of the random number generators and computer simulation to create a dataset that exemplifies a real world example.

An Applied Health Example using Simulated Data

An Applied Health Example using Simulated Data

Consider for example that you are asked to assess the benefits of a 12-week pulmonary rehabilitation program, consisting of exercise and education, for a cohort of individuals with varying classifications of chronic obstructive pulmonary disease (COPD). The intake data include demographic variables such as the individual's age, sex, height, and weight; and performance data such as the distance walked in 6 minutes, a physician based rating of COPD, the program participant's self reported smoking status, years smoked; and physiological measures such as forced expiratory volume in 1 second, and resting heart rate.

In the following example we will generate data artificially using random number generators written with SAS code. In this way we can produce a simulated dataset that we can then use to observe what might happen if we were to actually conduct a research study with the same parameters and considerations.

Using random number generators we create the data set to produce a set of values representing 20 individuals (a random selection of males and females). The variables used in the table along with the variable types and the possible minimum and maximum range for each variable are presented in Table 6.1 below.

Table 6.1 Variables Used To Produce A Sample Of Raw Data For The COPD Clinic

Variable Name & Variable label	**Variable Type**	**Range of Values**
Patient identification – Px id	discrete	1 to 20
Age in years. – *age*	discrete	45 to 75
Sex – *sex*	discrete	m: male; f: female;
Height – *ht*	continuous	1.5 m to 2.0 m
Weight – *wt*	continuous	50 kg to 150 kg
– Distance walked in 6 minutes – *walkdist*	continuous	54 metres to 150 meters
Rating of COPD severity – *severity*	discrete	MI: mild; MO: moderate; S: severe
Smoking status – *smoke*	discrete	S: smoker; EX: ex-smoker; NON: never smoked
Years as a smoker – *yrsmoke*	continuous	<1 to max years smoked
Forced expiratory vol in 1 sec – *FEV1*	continuous	1.5 – 4.0
Resting heart rate – *rhr*	continuous	50 to 100

6.3.1 Creating your dataset with a random number generator

Here we will use SAS code to produce the table of random numbers for each of the variables listed above. Recent developments in high speed computing and the creation of the Mersenne-Twister Random Number Generator which is now used by SAS, have led to the creation of the RAND() function. As stated in the SAS Knowledge Base (SAS(R) 9.3 Functions and CALL Routines), the RAND function can generate random numbers for a distribution specified by the user.

In the following example the random number generator was seeded with the statement: call streaminit(n);

/* where n refers to any number you wish to use */

Here we specify that the data we generate will be drawn from the normal distribution.

... RAND("normal")...

code snippet:

```
data sasrng;
call streaminit(13);
/* here we use n=13 to seed the RNG */
```

SAS User Notes provide an explanation of the RAND() function as follows:

where is an observation from the normal distribution with a mean of θ and a standard deviation of λ that has the following probability density function:

Range:

θ: is the mean parameter ? Default:0

: is the standard deviation parameter ? Default:1

Range: > 0.

Once we have established the parameters for random number selection we begin writing the SAS program to create random number generators as we would for any SAS program. Start by stating the options that you would like included in the output and then name the workspace using normal SAS code.

```
OPTIONS PAGESIZE=63 LINESIZE=90 DATE;
DATA SASRNG;
```

Our next statement is to create an array. An array is a set of variables that generally have some commonality and that you wish to process together. In our example, we will start by creating an array that we name **SCORES**, and which has three elements or variables.

```
ARRAY SCORES SEX SEVERITY SMOKING;
```

By naming the array, as we have here (**SCORES**) we can refer to the array **SCORES** later to reference the specific elements that are contained within. For example, since the array has three elements, then **SCORES**(1) refers to the first

element–the participant's *SEX*, while **SCORES**(2) refers to the second element–the *SEVERITY* of the COPD condition, and **SCORES**(3) refers to the third element–the patient's *SMOKING* status.

Once we create the workspace in SAS, we next use the do-loop statements to generate a data set consisting of 20 cases. The first do-loop (**DO** K=**1** TO **20**) tells SAS to execute the statements within the loop 20 times.

The second do-loop (**DO** K=**1** TO **3**) is contained within the first loop and is designed to provide data specifically for the variables *SEX*, *SEVERITY, and SMOKING*

```
DO K=1 TO 20;
  DO I=1 TO 3;
```

Figure 6.1 Functions of The Do-Loop To Generate Random Numbers For The Array SCORES

Finally, we end the do loops with the following statement sequence.

```
END;
OUTPUT;
END;
```

The first END; statement closes the inside loop that begins with DO I=1 TO 3; likewise, the OUTPUT; statement is needed to assign the RNG values to each variable for each participant, the outside loop (DO K=**1** TO **20**? is closed with the second END; statement.

Figure 6.2 Closing the do-loops and producing output

The actual statement sequence to generate a random number for each of the variables in the array is shown here as a three step process beginning by seeding the Random Number Generator (RNG) with CALL STREAMINIT(N); where the (N) can be any number you wish to use. In this first example here we used the number 13 (only because 13 is MY lucky number!).

```
CALL STREAMINIT(13);
```

The call statement initiates or seeds the random number generator.

```
OPTIONS PAGESIZE=63 LINESIZE=90 DATE;
  DATA SASRNG;

  ARRAY SCORES SEX SEVERITY SMOKING;
  DO K=1 TO 20;

  DO I=1 TO 3;

  CALL STREAMINIT(13);
  SCORES(I)=RAND("NORMAL")*100000000;
  SCORES(I)=ROUND(SCORES(I));
  SCORES(I)=1+ABS((MOD(SCORES(I),333)));
```

This sequence of statemente invokes the random number generator and places a value in each element of the array (i.e., the list of variables).

After generating the random numbers for each variable in the array **SCORE** (SEX, SEVERITY, SMOKING) we then

process the number with a logic filter so that it makes sense in relation to the range of scores that we would expect to see for each given variable.

For example, if the RNG produces a value of 75 for the variable sex, then what does that mean?

Well actually it is meaningless until you assign the meaning.

We assign meaning to the values within a variable using logic statements. For each of the elements (variables) within the array we process the RNG value with logic statements that will make the data relevant to our variables.

For example, the logic statements to convert the RNG values for sex are shown here. In this situation we convert the numeric variable for sex to a text variable that we call sex. Since we have text labels that extend beyond 8 characters we use the length statement with the $ to ensure that the full length of the text label is used.

```
/* LOGIC STATEMENTS FOR THE VARIABLE: SEX */

  LENGTH SEX $12;
  IF SEX > 175 THEN SEX = 'NOT STATED';
  IF SEX >54 AND SEX<175 THEN SEX = 'FEMALE';
  IF SEX >0 AND SEX<55 THEN SEX = 'MALE';
```

In SAS, the logic statements use the if-then conventional approach. That is, for every IF statement we use a corresponding THEN statement. In this way we process the RNG values to be within the range of logical outcomes for the variable that we are creating.

```
/* LOGIC STATEMENTS TO CREATE CATEGORIES FOR THE VARIABLE: COPD SEVERITY TYPE */
  IF SEVERITY >55  THEN SEVERITY = 3;
  IF SEVERITY >27 AND SEVERITY<56 THEN SEVERITY = 2;
  IF SEVERITY >4 AND SEVERITY<28 THEN SEVERITY = 1;

  /* LOGIC STATEMENTS TO CREATE CATEGORIES FOR THE VARIABLE: SMOKING STATUS */
  IF SMOKING >55  THEN SMOKING = 3;
  IF SMOKING >27 AND SMOKING<56 THEN SMOKING = 2;
  IF SMOKING >4 AND SMOKING<28 THEN SMOKING = 1;
```

Next we create RNGs for the remaining seven variables that we plan to include in the analysis. We do not need to include these in the array and can simply generate the values when SAS walks through the outer do loop. The independent execution of the rand("normal") function can run with a new seed and a new maximum score. Notice that these follow the array processing statements.

The continuous discrete variables were age, years smoked, resting heart rate, weight, and distance walked in 6 minutes (measured in metres).

```
  /* CONTINUOUS DISCRETE VARIABLE AGE */
  CALL STREAMINIT(13); AGE=RAND("NORMAL")*1000000000000;
  AGE=40+ABS((MOD(AGE,62))); AGE=ROUND(AGE);
  IF AGE>72 THEN AGE=35+ABS((MOD(AGE,50)));

  /* CONTINUOUS DISCRETE VARIABLE YRSMOKE */
  CALL STREAMINIT(13); YRSMOKE=RAND("NORMAL")*1000000000000;
  YRSMOKE=1+ABS((MOD(YRSMOKE,12))); YRSMOKE=ROUND(YRSMOKE);

  /* CONTINUOUS DISCRETE VARIABLE RHR */
```

```
CALL STREAMINIT(69); RHR=RAND("NORMAL")*1000000000000;
RHR=54+ABS((MOD(RHR,80))); RHR=ROUND(RHR);

/* CONTINUOUS DISCRETE VARIABLE WT */
CALL STREAMINIT(45); WT=RAND("NORMAL")*1000000000000;
WT=45+ABS((MOD(WT,65)));WT=ROUND(WT,0.01);
IF WT>85 THEN WT=55+ABS((MOD(WT,12)));

/* CONTINUOUS DISCRETE VARIABLE WALKDIST */
CALL STREAMINIT(69);WALKDIST=RAND("NORMAL")*1000000000000;
WALKDIST=54+ABS((MOD(WALKDIST,80))); WALKDIST=ROUND(WALKDIST);
```

Next we created the continuous decimal variables. Again these statements are placed within the do loops to produce a full set of 20 outputs.

```
/* CONTINUOUS DECIMAL VARIABLE FEV1 */
CALL STREAMINIT(99); FEV1=RAND("NORMAL")*1000000000000;
FEV1=1+ABS((MOD(FEV1,3)));FEV1=ROUND(FEV1,0.01);

/* CONTINUOUS DECIMAL VARIABLE HT */
CALL STREAMINIT(21); HT=RAND("NORMAL")*1000000000000;
HT=1+ABS((MOD(HT,1.1)));HT=ROUND(HT,0.01);
IF HT<1.5 THEN HT=1.2+ABS((MOD(HT,1.1)));
```

43. Survival Analysis

Essential Background in Survival Analysis

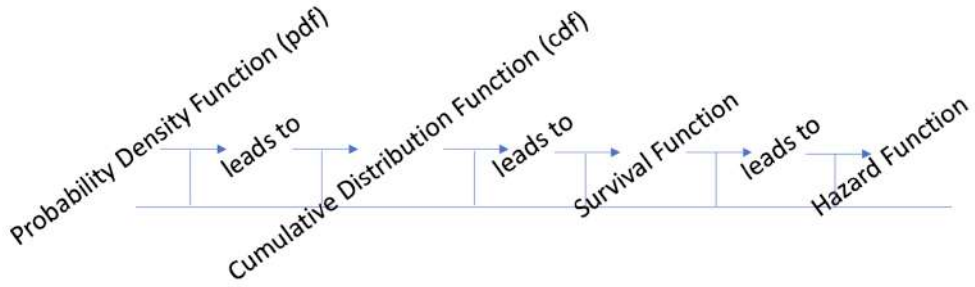
Survival analysis can be considered in its simplest form as a method to analyze longitudinal data for a cohort, or for a comparison of cohorts with a specific interest in the proportion of individuals that reached or exceeded a definite point on a time scale.

In survival analysis, the demarcation point for the event of interest on a time scale is referred to in a variety of ways but is dependent upon the perspective of the researcher. For example, if the researcher is interested in the application of survival analysis to estimate mortality as a result of a given treatment regimen then the demarcation point may be used to count the number of individuals that died within the interval up to a specific time, versus the number of individuals that lived beyond the selected time (i.e. survived). However, given the intention of the research, the mathematics of survival analysis need not be limited to only counting deaths (or survival), rather, the approaches of survival analyses may be thought of as a set of mathematical functions that enable statistical techniques which can be applied to the evaluation of any selected event at a specific period of time. Hence, there are several methods that can be used to perform survival analysis, however, in this chapter, the focus will be on the application of SAS for survival analysis using life tables, the calculation of the log-rank test, and the application of the Cox Proportional Hazard Model.

Important Functions Used in Survival Analysis

The progression of information about functions used in the computation of survival analyses is presented in Figure 19.1. In the following section, we will review the important concepts of the probability density function for a random discrete variable and a random continuous variable, the cumulative distribution function, the survival function, and the hazard function.

The flow of function processing in survival analysis



There are several ways to demonstrate survival analysis, but we will begin here by reviewing the basic terminology and the elements of the different functions used in the calculation of survival analysis so that we can measure the risk of an event happening at a specific period of time.

The probability density function represents a value that describes the probability of an outcome or a combination of outcomes occurring within a known outcome space – such as an interval.

The probability density function (pdf) can refer to either the associated probability value from a discrete random variable or from a continuous random variable. When the pdf refers to a discrete random variable then it is also referred to as the probability mass function (pmf) for a positive discrete random variable. In this case, we define a positive discrete random variable as a variable that holds numbers from the whole number line, meaning that the scores are whole numbers (ranging from 0 to + ∞) and may resemble (0,1,2,3, ..., ∞) without decimal values.

Probability Density Function (pdf) Related to Tossing a Single die

Possible outcome expressed as [latex]P(X = x)[/latex]	The probability associated with the outcome
[latex]P(X = 1)[/latex]	1/6
[latex]P(X = 2)[/latex]	1/6
[latex]P(X = 3)[/latex]	1/6
[latex]P(X = 4)[/latex]	1/6
[latex]P(X = 5)[/latex]	1/6
[latex]P(X = 6)[/latex]	1/6

A graph of the frequency distribution for these data would produce a platykurtic (flat) distribution profile since each outcome value has a frequency of 1.

However, we could create a graph to demonstrate the cumulative outcomes for the probabilities of the random discrete variable (X) ranging from 1 to 6; which would be to consider the discrete outcome ranging as follows: P(X=1) ≤ P(X=6).

The Cumulative Distribution Function commonly referred to as the c.d.f. and written as F(x)=P(X≤x) represents the set of values associated with the probabilities of the random variable (X) occurring equal to or less than a given value (x) in an outcome space.

In the example of the toss of a fair six-sided die, the outcome space is based only on the discrete numbers 1 through 6, as shown in the following outcome chart.

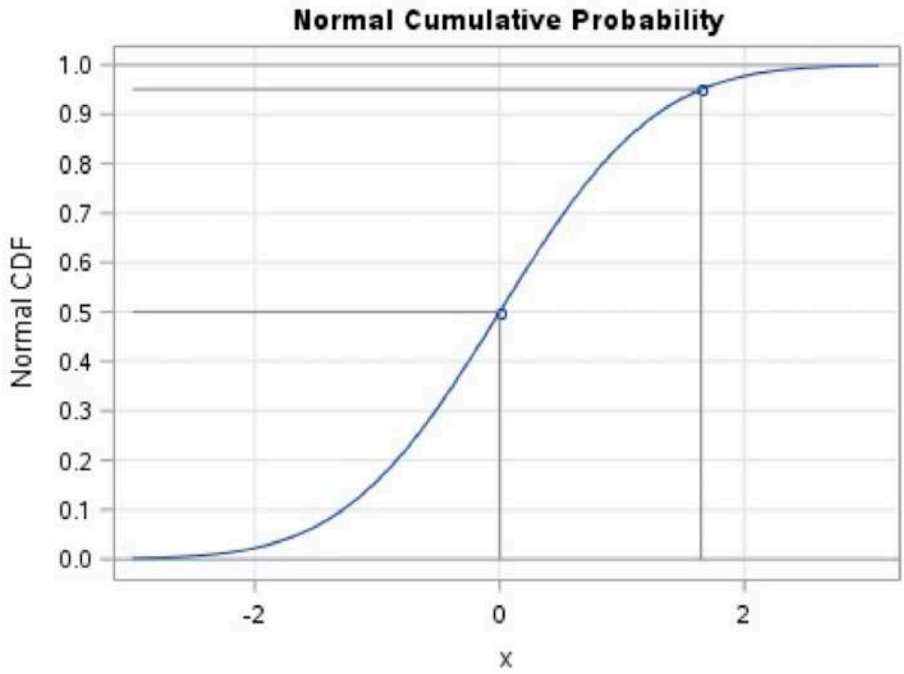
Cumulative Distribution Function (c.d.f) Related to Tossing a Single die

Possible outcome expressed as [latex]P(X \le x)[/latex]	Probability associated with the outcome
[latex]P(X \le 1)[/latex]	1/6 = 0.17
[latex]P(X \le 2)[/latex]	2/6 = 0.33
[latex]P(X \le 3)[/latex]	3/6 = 0.50
[latex]P(X \le 4)[/latex]	4/6 = 0.67
[latex]P(X \le 5)[/latex]	5/6 = 0.83
[latex]P(X \le 6)[/latex]	6/6 = 1.00

While the example presented here describes the c.d.f. for discrete random variable outcomes (and their associated probabilities based on the probability mass function (pmf) or probability density function (pdf)), the c.d.f. is also relevant for continuous variable values and the pdf is based on the outcomes (X) in an interval (a, b) represented by $P(a < X < b)$, where all numbers from the real number line are eligible within the interval of the distribution, typically ranging from 0 to 1.

If the data for the c.d.f. were attributed to a continuous random variable such as time, then the graph of the set of

probabilities for all possible outcomes of the c.d.f. is presented as a positive *S-shaped* curve ranging from 0 to 1, as shown in the figure below.



Schematic of a c.d.f. for a Continuous Random Variable

The SAS code to generate this image was written by Wicklin (2011)[1] and was processed unedited in SAS Studio shown here.

```
data cdf;
do x = -3 to 3 by 0.1;
y = cdf("Normal", x);
output; end;
x0 = 0;
cdf0 = cdf("Normal", x0);
output;
x0 = 1.645; cdf0 = cdf("Normal", x0); output;
run;
ods graphics / height=500;
proc sgplot data=cdf noautolegend;
```

```
title “Normal Cumulative Probability”;
series x=x y=y;
scatter x=x0 y=cdf0;
vector x=x0 y=cdf0 /xorigin=x0 yorigin=0 noarrowheads lineattrs=(color=gray);
vector x=x0 y=cdf0 /xorigin=-3 yorigin=cdf0 noarrowheads lineattrs=(color=gray);
xaxis grid label=”x”;
yaxis grid label=”Normal CDF” values=(0 to 1 by 0.05);
refline 0 1/ axis=y;
run;
```

The c.d.f. is an important step in the computation of the survival analysis because it is part of the computation of the survival function. In a time relevant model as is typical in a biostatistics application, the cumulative distribution function can be represented as [latex]F(t)=P(T \le{t}) \textit{where t}[/latex] is the value of the random variable representing a measured time and [latex]{t}[/latex] is the value of the intended time at the event.

The survival function [latex]S{(t)}[/latex] provides the estimate of the duration of time to an event, be it a failure, death, or a specified incident. The survival function begins at 1, the point where an individual enters the dataset and ends at 0 the point where data monitoring stops, usually because the event of interest has occurred.

In simple terms, the Survival Function is the complement of the c.d.f. and is computed as [latex]S{(t)}= 1- F(t)\textit{, where t >0}[/latex]. More important, the survival function is the denominator in the computation of the Hazard Function, which is a main element in one approach to the computation of the survival analysis. The survival function can show the probability of surviving up to a designated event, based on units of time.

For example, consider the following data set in which a measure of time to an event is recorded.

The cutoff time is set at 48 ***(totally arbitrary units)*** so that any value above 48 is assigned a censor score of 1 and any value less than 48 is a value of 0.

Table depicting number of individuals that exceeded the time to event

Patient ID	Time to Event: The measure of the length of time to the event happening	Event Counter variable (0=event has not happened, 1=event has happened)
01	40	0
02	38	0
03	54	1
04	56	1
05	28	0
06	36	0
07	42	0
08	51	1
09	45	0
10	49	1

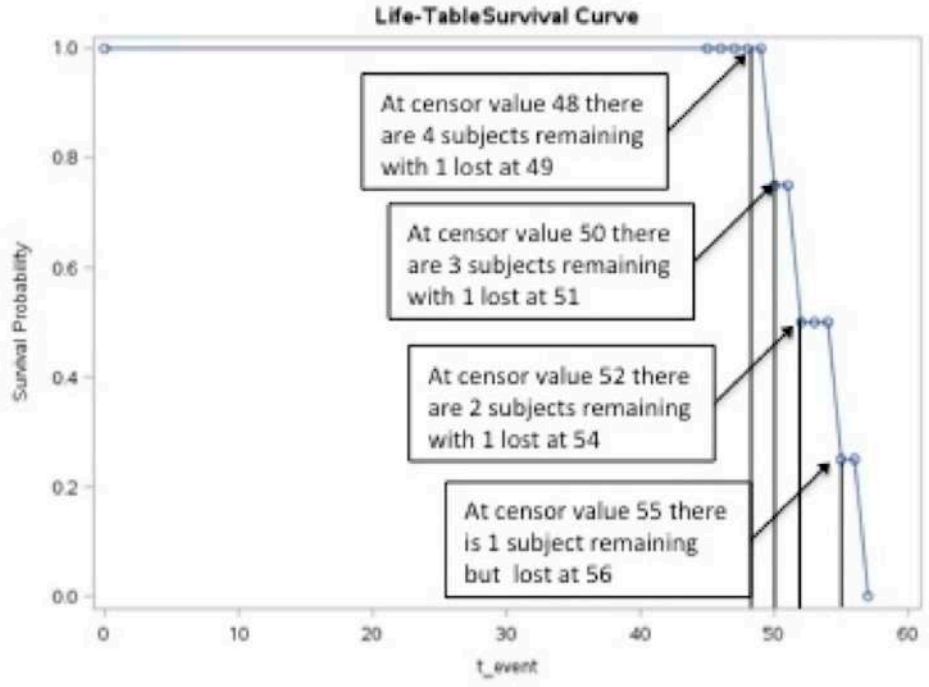
The data are processed with the following SAS code[2] to produce a graph of the survival curve shown below.

```
title 'program to show a survival curve';
data survcurv;
input id t_event censor;
datalines;
01 40 0
02 38 0
03 54 1
04 56 1
05 28 0
06 36 0
07 42 0
08 51 1
09 45 0
10 49 1
;
proc lifetest data=survcurv(where=(censor=1)) method=lt
intervals=(45 to 60 by 1) plots=survival; time t_event*censor(0); run;
```

The results of this analysis include the table of the survival estimates and the survival curve below – note that the failure point was set at 48. The curve shows the probability of surviving to 48 and then beyond 48.

Notice that the entire group begins at probability = 1 and ends at probability = 0.

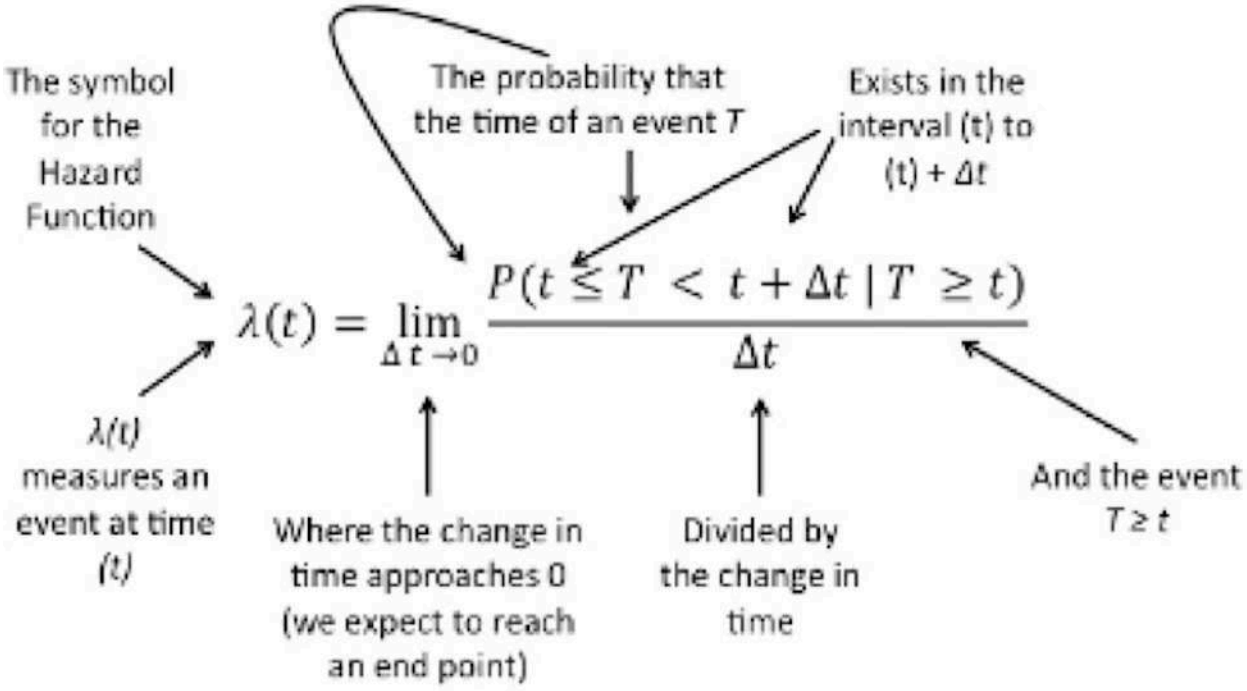
Figure of SAS representation of the survival function for n=10 with censoring at x=48.



The Hazard Function is determined by the ratio of the probability density function (pdf) to the survival function [latex]S{(t)}[/latex] and can be written as:[latex]\lambda = {(p.d.f.) \over{S(t)}}[/latex]

The following explanation may help to describe the elements of the **hazard function** in greater detail. In this annotated formula the hazard function is shown to represent the likelihood of an event such as death or survival occurring within an interval at time [latex]{(t)}[/latex].

[latex]\lambda(t) = {\lim\limits_{\Delta{t}\to {o}}} {P(t \le{T} \lt{t} + \Delta{t} \mid{T} \ge{t})\over{\Delta(t)}}[/latex]



Annotated image of the Hazard Function Equation

- The hazard function [latex]\lambda{(t)}[/latex] measures a specific event with respect to time [latex]{(t)}[/latex]
- The hazard function [latex]\lambda{(t)}[/latex] is based on the probability that the observed event occurring at time [latex]{T}[/latex] will happen within the interval beginning at time point [latex]{(t)}[/latex] and ranging to the end of the interval [latex]{(t + \Delta{t})}[/latex], so that we say [latex]{(t \le T \lt t + \Delta{t})}[/latex]
- Since the hazard function [latex]\lambda{(t)}[/latex] is not a probability estimate but is a ratio, the hazard function [latex]\lambda{(t)}[/latex] can exceed 1.

The following table shows the output from a *life table* approach to evaluating the set of data that were used in the SAS program above to produce the survival function. The *hazard function* is included in the tabled output when the method=LT command is included in the proc lifetest procedure. An abbreviated form of the table is shown here.

Life Table Survival Estimates			
Interval (sum of failed)	Number failed after censoring	PDF	Hazard
47-48	(0)	0	0
49-50	(1)	0.25	0.29
51-52	(1)	0.25	0.40
54-55	(1)	0.25	0.67
56-57	(1)	0.25	2.00

Recall from the program listed above, that the important SAS code to produce the hazard function using the proc lifetest procedure is:

```
 proc lifetest data=survcurv(where=(censor=1)) method=lt
intervals=(45 to 60 by 1) plots=survival;
 time t_event*censor(0);
run;
```

Censoring Data

In the computation of survival analyses, not all participants will fail (or die) at the demarcation point set by the researcher. As shown in the data set analyzed above, the demarcation point for the event of interest was set at an arbitrary value of 48 and therefore 4 individuals extended beyond the value 48.

In a survival analysis, where the time to an event is noted, any cases that "survive" beyond the point stated will be considered censored. Censoring does not mean that the participants are dropped from the analysis. Rather, when censored, the individuals that have not demonstrated the event of interest prior to the pre-designated demarcation point are not calculated as part of the group measured with the event of interest (i.e. dying, failing).

When we plot the survival curves for a cohort in SAS, we can specify the censoring point and thereby produce survival probability curves that represent both the cases – those individuals that have demonstrated the event of interest by the end of the interval measured; or we can plot the non-cases – those individuals that have not demonstrated the event of interest by the end of the interval measured. In the following example, survival probability curves are used to demonstrate the influence of censoring and the Kaplan-Meier estimates used to develop the survival probability curves.

Annotated SAS application for a Survival Analysis

As noted, survival analysis is a time-based evaluation. That is, in survival analysis, we are interested in evaluating the time point at which an event occurs within a cohort. Survival analysis helps researchers evaluate the proportion of individuals at a time to reach a demarcation point, *and therefore the number of individuals within a cohort that extends beyond an event (a time point of interest).*

In the following scenario, we will use a random number generator to create a SAS dataset and simulate the scenario of the ZIKA Virus at the Summer Olympics (2017). Next, we will apply the different tools of the SAS Survival Analysis suite to evaluate the data set, with examples that include a comparison of outcomes across athlete cohorts.

Background:

In August 2016, Brazil hosted the Olympic Summer Games. However, several athletes decided to boycott the games because of the risk of exposure to the ZIKA virus. ZIKA is a virus that can be transmitted through the bite from an infected Aedes mosquito. The ZIKA virus is extremely dangerous for young women as it can reside in the blood for up to 3 months and if the woman becomes pregnant, the virus can have negative consequences for the developing fetus. In particular, the ZIKA virus has been implicated in the development of microcephaly in newborn children.

Generating the dataset with a random number generator:

In this example, we will use a series of random number generating commands to create a data set with four variables and 100 cases.

Three discrete variables are: sex, sport and case and we will use the following format: sex (1=m, 2=f), sport (1=golf, 2=equestrian, 3=swimming, 4=gymnastics, 5=track & field), and case (1=yes, 2=no).

A continuous variable, labelled days will represent the number of days prior to the individual contracting the ZIKA virus from the Aedes mosquito.

The program to generate the simulated SAS data set is shown here

```
options pagesize=60 linesize=80 center date;
LIBNAME sample '/home/Username/your directory/';
proc format; value sexfmt  1 ='male' 2 ='female';
value sprtfmt  1 ='golf' 2 ='equestrian' 3 ='swimming' 4 ='gymnastics' 5 ='track & field';
value casefmt 1='present' 0='absent';
data sample.zika;
/* create 3 new variables set as score1 score2 score3 */
array scores score1-score3;
/* set 100 cases per variable */
do k=1 to 100;
/* set days to 100 days of exposure */
days=ranuni(13)*100; days=round(days, 0.02);
/* Loop through each variable to establish 100 randomly generated scores */
do i=1 to 3;
call streaminit(23);
scores(i)=RAND("normal")*1000000000000;
scores(i)=ROUND(scores(i));
```

```
scores(i)=1+ABS((mod(scores(i),150)));
/*  the variable sex will relate to score1, we can create a filter to establish the binary score for sex based on the randomly generated output */
if score1 > 55 then sex = 2;
if score1 >2 and score1<56 then sex = 1;
/* the variable sport type will relate to score2, we can create a filter to establish the determination of an athletes sport based on the randomly generated output */
if score2 >90 then sport = 5;
if score2 >80 and score2<91 then sport = 4;
if score2 >60 and score2<81 then sport = 3;
if score2 >30 and score2<61 then sport = 2;
if score2 >5 and score2<31 then sport=1;
/* the determination of a case will relate to score3, we can create a filter to establish the determination of a case based on the randomly generated output */
if score3 > 48 then case = 1;else case = 0;
/* a case=1 is a case present, and a case=0 is a case absent */
if days<=15 then daygrp=1;
if days>15 and days<=30 then daygrp=2;
if days>30 and days<=45 then daygrp=3;
if days>45 and days<=60 then daygrp=4;
if days>60 and days<=75 then daygrp=5;
if days>75 and days<=90 then daygrp=6;
if days>90 and days<=105 then daygrp=7;
if days>105 and days<=120 then daygrp=8;
if days>120 and days<=135 then daygrp=9;
if days>135 then daygrp=10;
/* create an interaction term for sex and sport to be used later in the Cox regression analysis */
sex_sport=sex*sport;
end; output; end;
```

Describing the Output

Part 1: Descriptive Statistics

Prior to computing the survival analysis, descriptive statistics are produced for each of the four variables generated by

the computer simulation. Initially a grouping variable called ***daygrp*** was created to summarize the continuous variable (counting number of days) into a discrete variable for use in later presentations.

Next, the data were sorted and the proc freq command was applied.

```
proc sort data=sample.zika; by sex;
proc freq; tables sex sport daygrp case;
format  case casefmt. ;
```

The FREQ Procedure

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	29	29.00	29	29.00
2	71	71.00	100	100.00

case	Frequency	Percent	Cumulative Frequency	Cumulative Percent
absent	30	30.00	30	30.00
present	70	70.00	100	100.00

sport	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	23	23.00	23	23.00
2	22	22.00	45	45.00
3	17	17.00	62	62.00
4	24	24.00	86	86.00
5	14	14.00	100	100.00

daygrp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1	1.00	1	1.00
2	5	5.00	6	6.00
3	7	7.00	13	13.00
4	5	5.00	18	18.00
5	9	9.00	27	27.00
6	6	6.00	33	33.00
7	4	4.00	37	37.00
8	13	13.00	50	50.00
9	7	7.00	57	57.00
10	43	43.00	100	100.00

The demarcation point for a case was set at a value of 100 for the random variable days from the array:

```
do i=1 to 2;
call streaminit(23);
scores(i)=RAND("normal")*1000000000000;
scores(i)=ROUND(scores(i));
scores(i)=1+ABS((mod(scores(i),150)));
```

The variable ***days*** was given a range of 1 to 150 and 100 days was used as a demarcation point to censor individuals as non-cases.

```
if days < 101 then case = 1;
if days>100 then case = 0;
```

The labelling of individuals in this way was used to generate a random assignment of the individual as a case (1) or as a non-case (0). The proc univariate procedure was used to present descriptive statistics for individuals that were considered cases ((where=(case=1))and individuals that were censored (where=(case=0));

```
proc univariate data=sample.zika(where=(case=1));
var days;
histogram days/normal;
```

```
title 'Survivor function for zika virus plot of pdf';
label days ='days to infection';
```

The results from the random number generator produced a mean days among cases of 60.49, for a sample of 70 individuals. These data also produced a 95% confidence interval for the mean of 60.49 ± 6.45 which ranged from 54.03 to 66.93.

The UNIVARIATE Procedure – Variable: days (days since exposure)

Moments			
N	70	**Sum Weights**	70
Mean	60.4874286	**Sum Observations**	4234.12
Std Deviation	27.0551932	**Variance**	731.98348
Skewness	-0.2644189	**Kurtosis**	-1.1242295
Uncorrected SS	306617.891	**Corrected SS**	50506.8601
Coeff Variation	44.7286219	**Std Error Mean**	3.2337141

Basic Confidence Limits Assuming Normality			
Parameter	**Estimate**	**95% Confidence Limits**	
Mean	60.48743	54.03635	66.93851

The following code produced a set of percentiles from the data set for cases. These data show the percentage of the group being affected by a certain day. For example, 25% of the group were affected within 39.4 days of the start of the games. By day 96 some 90% of the cohort were infected with the Zika Virus. Note, these are not real data but were generated with a random number generator.

```
  output out=Pctls pctlpts  = 25 40 50 60 75 90
pctlpre  = days_
pctlname = pct25 pct40 pct50 pct60 pct75 pct90;
proc print data= Pctls;
run;
```

Percentiles for days

Obs	**days_pct25**	**days_pct40**	**days_pct50**	**days_pct60**	**days_pct75**	**days_pct90**
1	39.4	51.84	66.34	72.44	84.08	96.44

```
proc univariate data=sample.zika(where=(case=0)) cibasic;
```

```
var days;
histogram days;
title 'Survivor function for zika virus plot of pdf ';
label days ='days to infection';
```

The results from the random number generator produced a mean days among cases of 60.49, for a sample of 70 individuals. These data also produced a 95% confidence interval for the mean of 127.50 ± 5.36 which ranged from 122.14 to 132.86.

The UNIVARIATE Procedure – Variable: days (days since exposure)

Moments			
N	30	**Sum Weights**	30
Mean	127.503333	**Sum Observations**	3825.1
Std Deviation	14.3410047	**Variance**	205.664416
Skewness	-0.3026793	**Kurtosis**	-1.0739046
Uncorrected SS	493677.268	**Corrected SS**	5964.26807
Coeff Variation	11.2475528	**Std Error Mean**	2.61829726

Basic Confidence Limits Assuming Normality			
Parameter	**Estimate**	**95% Confidence Limits**	
Mean	127.50333	122.14831	132.85835

The following code produced a set of percentiles from the data set for non-cases. As shown in the example above, these data show the percentage of the group being affected by a certain day. All individuals in this data set were censored as they had passed the 100 days demarcation point before being infected. This is the reason that individuals in a survival analysis are not dropped from the study but rather censored. The data show that even though an individual exceeded the time to an event, they were continued to be at risk for the event of interest.

```
   output out=Pctls pctlpts  = 25 30 40 50 60 75 80 90 100
pctlpre  = days_
pctlname = pct25 pct30 pct40 pct50 pct60 pct75 pct80 pct90 pct100;
proc print data= Pctls;
run;
```

Percentiles for days

Obs	days_pct25	days_pct40	days_pct50	days_pct60	days_pct75	days_pct90
1	115.18	125.24	129.69	133.47	139	146.24

In each of the proc univariate statements there was a call for a histogram to illustrate the distribution of the data for the variable days. The graphs of the histogram for each distribution for days in each of the cohorts (cases versus non-cases) are shown in Figure 19.4 below. Notice that in each distribution the number of days shows a slight negative skewness with more cases appearing after the mean days.

Figure 19.4 Comparison of the distribution days in each cohort

Part 2: Creating Life Tables

The survival analysis applications using METHOD=LIFE in the PROC LIFETEST procedure are presented in this section:

In this first stage of survival processing we can observe the influence of censoring the data. Recall that initially the data are censored at 100 days. Censoring was accomplished by creating the variable days, described above and then combined with the binary variable case. If an individual had a days score of less than 100 then they were assigned to the cohort of cases. Conversely, if the individual had a days score exceeding 100 then they were censored and assigned to the non-cases cohort.

The SAS code to compute the survival curve for the entire data set is given here:

```
proc sort data=sample.zika; by case;
PROC LIFETEST METHOD=LIFE plots=(s) data=sample.zika notable;
time days ;
format  case casefmt. ;
title 'Survivor function for zika virus – implicit right censoring of cases';
label days ='days to infection';
```

This SAS code produced the image shown in Figure 19.5, below, which is the survival probability curve for the entire sample of N=100 cases monitored over 150 days. Notice that there is an inflection point in the curve at 100 days. This inflection point corresponds to the censoring limit of 100 days and is shown more explicitly in Figure 19.6 where we change the command time days; to the command: time days * case(0);

Figure 19.5 Life Table Survival curve for all individuals in the data set

Figure 19.6 Life Table Survival curve with explicit right censoring at 100 days

Figure 19.6 above shows the survival probability for each event among the cases and holds the non-cases constant at a probability level of 0.3. Further, when we include the censoring criteria using the command: time days * case(0);

a summary table indicating the number of cases that fail prior to the demarcation point (100 days) and the number of cases that exceed the demarcation point is also included, as shown here.

Summary of the Number of Censored and Uncensored Values			
Total	**Failed**	**Censored**	**Percent Censored**
100	70	30	30.00

Next we include a command to show the differences in time to event with a grouping variable. Here we use the strata command to group the data by sex, while maintaining the influence of censoring at 100 days.

```
PROC LIFETEST METHOD=LIFE plots=(s)data=sample.zika notable;
time days * case(0) ;
strata sex;
format case casefmt. sex sexfmt. ;
```

The code produces a summary table of the number of males and females that failed or exceeded the demarcation point of 100 days and a graph of the survival probability curves for male and females.

Summary of the Number of Censored and Uncensored Values					
Stratum	**sex**	**Total**	**Failed**	**Censored**	**Percent Censored**
1	female	71	52	19	26.76
2	male	29	18	11	37.93
Total		100	70	30	30.00

Figure 19.7 Life Table Survival Curves With Explicit Right Censoring at 100 Days for Males and Females

In this next analysis we separate the data using strata=sport, while maintaining the right censoring of the data at 100 days. As shown in the approach used to separate the data by sex, this code produces a summary table of the number of individuals in each of the sport groups that failed or exceeded the demarcation point of 100 days as well as a graph of the survival probability curves for each sport.

Summary of the Number of Censored and Uncensored Values					
Stratum	**sport**	**Total**	**Failed**	**Censored**	**Percent Censored**
1	equestrian	22	15	7	31.82
2	golf	23	19	4	17.39
3	gymnastics	24	14	10	41.67
4	swimming	17	13	4	23.53
5	track & field	14	9	5	35.71
Total		100	70	30	30.00

Figure 19.8 Life Table Survival Curves With Explicit Right Censoring at 100 Days for Sport Groups

Part 3: The Kaplan-Meier Approach

The Kaplan-Meier approach to survival analysis differs slightly from the applications using METHOD=LIFE in the PROC LIFETEST procedure. When we use the METHOD=KM in the PROC LIFETEST procedure we generate a series of survival probability estimates referred to as the Kaplan-Meier estimates (heretofore referred to as the KM estimates), and corresponding survival probability curves for the KM estimates.

In the KM estimates values are given for the probability change each time an individual becomes a case up to the demarcation point of 100 days. This approach is more precise in reporting the time at event and does not summarize the data across an interval as is done with the METHOD=LIFE in the PROC LIFETEST procedure.

A comparison of the output from the METHOD=LIFE and the METHOD=KM is shown in the comparison of the tables up to the first 12 cases that became infected. Notice that the METHOD=LIFE approach summarizes the estimates within a set of intervals, while the METHOD=KM approach provides the continuous probability values for each individual within the cohort of interest.

Table 19.5 Survivor function for Zika virus using METHOD = LIFE in Proc Lifetest

Days Interval	**Abbreviated table showing results for The LIFETEST Procedure**								
Lower interval	**Upper interval**	**Number failed**	**Number censored**	**Effective sample size**	**Conditional probability of failure**	**Conditional probability of failure Standard error**	**Survival**	**Failure**	**Survival Standard error**
0	20	6	0	100.0	0.0600	0.0237	1.0000	0	0
20	40	12	0	94.0	0.1277	0.0344	0.9400	0.06	0.023

When we use the METHOD=KM approach in the PROC LIFETEST procedure the following estimates are generated. Note these estimates only refer to the first 12 cases designated as infected within the original data set of n=100 cases.

Table 19.6 Survivor function for Zika virus using METHOD = KM in Proc Lifetest

Abbreviated table showing results for The LIFETEST Procedure

Days	Survival	Failure	Survival Standard Error	Number Failed	Number Remaining
0.000	1.0000	0	0	0	100
8.800	0.9900	0.0100	0.00995	1	99
11.780	0.9800	0.0200	0.0140	2	98
12.540	0.9700	0.0300	0.0171	3	97
12.800	0.9600	0.0400	0.0196	4	96
14.120	0.9500	0.0500	0.0218	5	95
15.860	0.9400	0.0600	0.0237	6	94
21.240	0.9300	0.0700	0.0255	7	93
22.560	0.9200	0.0800	0.0271	8	92
24.720	0.9100	0.0900	0.0286	9	91
27.280	0.9000	0.1000	0.0300	10	90
27.780	.	.	.	11	89
27.780	0.8800	0.1200	0.0325	12	88

The difference in the two methods is further exemplified in the comparison of the two survival curves shown in Figure 19.9. The survival curve for the METHOD=LIFE approach is a summary curve while the survival curve for the METHOD=KM approach shows more precise estimates of failures (individuals reporting infection) over the entire time interval. In both curves the data are right censored at days=100, and as such no survival probabilities are reported for individuals that have not become a case as of the 100 days demarcation point, in the data set.

Survival probability curve using Method=LIFE in SAS Proc lifetest	Survival probability curve using Method=KM in SAS Proc lifetest

Figure 19.9 Comparison of Survival Curves With Explicit Right Censoring for life-table analysis versus Kaplan-Meier estimation

Part 4: Comparing Kaplan-Meier Survival Estimates with Log Rank and Wilcoxon Tests

In the PROC LIFETEST procedure we can evaluate the difference between survival probability curves by computing two non-parametric tests: i) the Log Rank Test and ii) the Wilcoxon test. The tests are computed with the PROC LIFETEST procedure when including the strata command, as shown here:

```
PROC LIFETEST plots=(s) data=sample.zika2 ;
time days * case(0);
strata sex;
format case casefmt. sex sexfmt. ;
```

```
title 'Kaplan Meier Estimates with log rank and Wilcoxon tests';
label days ='days to infection';
```

The strata command separates the computation of survival probabilities by different subgroups of the variable used in the strata command. In our Zika data set, survival probabilities are estimated for the males and females in the observed sample. The graphical illustration of the survival probability curves is shown in Figure 19.10 below and the statistical comparison of the survival curves is shown in the following two tests.

The Log-Rank test and the Wilcoxon test are two non-parametric tests that enable users to compare the survival probability curves based on Kaplan-Meier Survival Estimates for each subgroup within designated strata. The results for the comparison of the Survival Probability Curves for males versus females are shown here.

Table 19.7 Test to evaluate the survival curves

Test	Chi-Square	DF	Pr > Chi square
Log-Rank	2.8240	1	0.0929
Wilcoxon	4.2191	1	0.0400

The p value indicates that the difference in survival curves for males versus females was found to be significantly different at $p<0.04$ for the Wilcoxon test, while the difference was significant at $p<0.09$ when tested using the log-rank test. The overall conclusion from this test is that the curves for the two survival probabilities were different. However, it should be noted that the Log-Rank test is the more powerful of the two tests because it is based on the assumption that the proportional hazard rate is constant at each time point. This means that the likelihood for an individual to be infected (i.e. become a case) is constant across all time points for all individuals[3].

Figure 19.10 illustrates the survival probability curves for males versus females in our Zika dataset. These curves are based on the product-limit estimates (aka Kaplan-Meier estimates) for the survival probability series within each level of the strata. Notice that the two survival curves cross early in the recording. This cross over of KM curves corresponds to the p value identified with the Wilcoxon analysis. In the statistical comparison of survival curves a stronger Wilcoxon outcome is likely to occur when one of the comparison groups has a higher risk of demonstrating the time to the event (becoming a case) earlier in the recording, versus a higher risk of being infected later. The higher risk of being infected (i.e. failing, dying, becoming a case) corresponds with a higher number of days to the event which increases the likelihood of a significant log-rank test outcome if this is demonstrated by one group more than another.

Figure 19.10 Comparison of Survival Curves With Explicit Right Censoring for Kaplan-Meier estimation of males versus females

Part 5: Computing the Cox Proportional Hazard Regression Analysis

The data in a survival analysis can be used in a special type of regression procedure known as the proportional hazard model. This approach to using regression modeling was developed by Cox[4] and builds on the regression approaches that we have discussed earlier in this text.

In simple linear regression we can create equations in which a predictor variable, or set of predictor variables are used

to explain the variance in an outcome variable (the dependent variable), as shown in the following simple linear regression and multiple regression equations.

A simple straight-line or linear regression equation:

where: is the dependent variable, is the slope element by which we adjust the predictor () variable, is the independent or predictor variable, and is the

– intercept (i.e. the point where the response graph crosses the vertical axis).

The simple linear regression equation in its most basic form helps us to understand the relationship between two variables, one designated as the and the other designated as the . Together these variables help us to predictor or explain an outcome, while adjusting for the variance between the two measures.

A multiple regression equation:

where: is the dependent variable, is the slope element by which we adjust the predictor () variable, is the independent or predictor variable, and is the

– intercept. In this equation, the subscript (*i*) is a counter for each of the predictor variables used in the equation.

The multiple linear regression equation is an expansion of the simple linear regression, and under a univariate model has one but two or more Again, the regression procedure helps us to predict or explain the outcome while adjusting for the variance in the predictor () variables. In multiple regression we can determine the slope of a predictor variable – the coefficient by which the variable is multiplied, while holding all other variables in the model constant. In this way we are able to determine the significance of each variable in the equation with respect to all of the variables in the equation.

In the Cox proportional hazard regression, also referred to as the Cox regression, the concepts of simple and multiple regression equations are the same, however the dependent variable is comprised not of a single scalar score, but rather of the hazard function representing the relationship between survival probability and time to an event.

As stated earlier, the hazard function provides an estimate of an event happening by a given time or within a given interval of time. The hazard function does not provide a probability estimate; therefore the estimate can exceed 1. Rather the hazard function indicates how likely an event is expected to occur by a given time.

In the computation of the Cox regression we develop a statistical regression model comprised of a dependent variable which consists of a hazard function and a set of independent variables which consist of predictors of the dependent variable, all based on a time based distribution referred to as the Weibull distribution. The Weibull distribution is familiar to the field of engineering because it is helpful in describing reliability and failure of a measured device over time. The applicable characteristic of the Weibull distribution for survival analysis is that it provides a mathematical foundation for failure rate throughout the lifetime of a measurement period. In the Weibull distribution the failure rate is shown to decrease with time reaching a plateau that is relatively constant[5]. The Weibull distribution fits applications for survival analysis since higher failure rates (i.e. time to an event) occur more often prior to the censoring demarcation point as shown in Figure 19.11.

Figure 19.11 Schematic of a Weibull distribution

As in the application of simple and multiple linear regression procedures, in the application of the Cox regression the user can establish regression coefficients for each of the predictors of the dependent variable to determine the magnitude and direction of the predictor acting on the dependent variable.

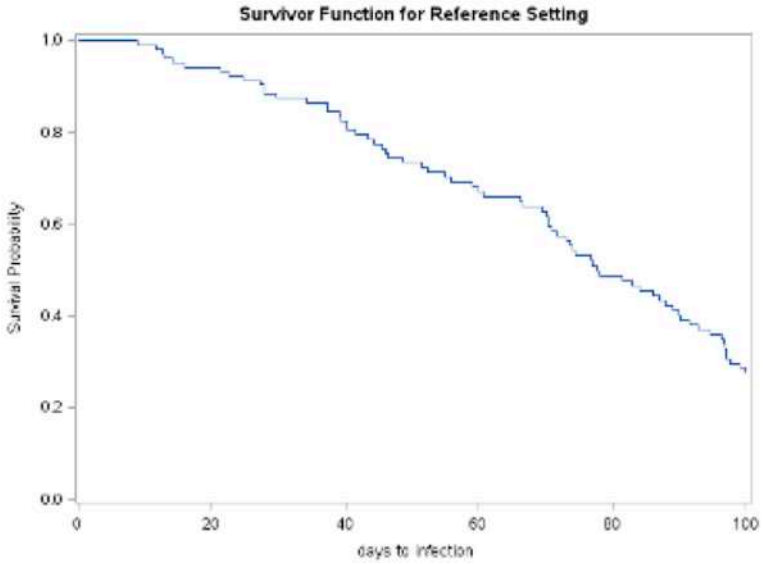
In our Zika virus example, we use Cox regression to determine the risk of infection based on the ratio of the probability density function and survival probabilities for time to infection as the dependent variable, and individual's sex and sport as predictor variables.

In other words, using the simulated dataset for the Olympic athletes and Cox regression we can evaluate the likelihood of being infected with Zika virus based on whether the individual was male or female, and the type of Olympic sport in which they were participating.

In the following sample code we use the proc phreg; procedure to produce output for the Cox Proportional Hazard Function. However, it is good practice to explain the overall model that we are testing. Here our hazard function is based on the number of days to infection, and the covariates are sex and sport type, along with the interaction of sex by sport type.

```
proc phreg plots=survival;
class sex sport;
model days*case(0) = sex sport sex_sport;
title 'Cox Proportional Hazard Analysis for Zika Virus by sex and sport';
label days ='days to infection';
```

The output shown below provides a graphic image of the survival curve and associated tables representing the statistical analyses.



Plot of the survival probability curve from proc phreg

The summary table of the number of cases that exceeded the censoring demarcation point is presented in Table 19.8 below. The results indicate that 30 of the 100 simulated cases.

Table of Proportion of Censored Observations from the Survival Curves

Summary of the Number of Event and Censored Values			
Total	**Event**	**Censored**	**Percent Censored**
100	70	30	30.00

Next, the model fit statistics are presented followed by the test of the null hypothesis that the predictor variables as greater than 0. The model fit statistics are most often used when comparing more than one model, in which case we

evaluate the AIC criteria to select the lowest value as suggesting a more appropriate fitting model. In the example shown here, this output is less relevant as we on have one model to consider. The column representing ***With Covariates*** is important to consider as it indicates that as we add predictor variables to the equation we decrease the criteria value, whereby lower values are considered to represent a better fit.

Table of a Model Fit Statistics for the Application of the Cox PHREG

Model Fit Statistics		
Criterion	**Without Covariates**	**With Covariates**
-2 LOG L	578.185	569.236
AIC	578.185	581.236
SBC	578.185	594.727

The main outputs for us to consider from the application of the **proc phreg;** procedure for this example are the tables of test for Global Null Hypothesis: Beta=0 and the Analysis of the Maximum Likelihood, shown below. The test of the Global Null Hypothesis: Beta=0 is suggesting that the predictor variables do not have an effect on the calculated value of the hazard function.

Table of Tests of Beta=0 for the Application of the Cox PHREG

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	8.9485	6	0.1765
Score	9.8384	6	0.1316
Wald	9.3857	6	0.1530

The results presented in the table above for the test of the Global Null Hypothesis: Beta=0 illustrate the results of three tests of the null hypothesis: i) the likelihood ratio test, ii) the Score test, and iii) the Wald test. Notice that the probability estimates for each Chi-square test are similar in that none of the p values supported a significant difference between the predictor variables and 0.

Since the predictor variables included the example were discrete class variables (no continuous covariates were included in the model), we also included the class sex sport; statement in the proc phreg; procedure. The output generated a table of the Type 3 tests (also referred to as Joint tests) to determine if each of the categorical discrete variables were significantly different than 0. The results of the Wald Chi-square statistic indicate that there was no significant effect of any of the categorical variables on the computed hazard function for the days to infection from the Zika virus.

Table of Type 3 Tests from Proc PHREG

Type 3 Tests			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
sex	1	0.5229	0.4696
sport	4	1.3533	0.8523
sex_sport	1	0.0311	0.8601

The maximum likelihood estimates produced by the SAS ***proc phreg*** enable us to provide the parameter estimates that correspond to the predictor variables included in the regression equation. The underlying algebraic regression equation[6] for the Cox Proportional Hazard Model is given as:

[latex]h(t) = h_0 (t)exp(x\beta_{x})[/latex]

Therefore, the parameter estimates refer to the coefficients for each predictor variable in the equation.

Maximum Likelihood Estimates from PROC PHREG

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
sex	**1**	1	-0.49114	0.67917	0.5229	0.4696	0.612	sex 1
sport	**1**	1	0.47592	1.51352	0.0989	0.7532	1.609	sport 1
sport	**2**	1	0.00554	1.11315	0.0000	0.9960	1.006	sport 2
sport	**3**	1	0.13256	0.80301	0.0273	0.8689	1.142	sport 3
sport	**4**	1	-0.13920	0.52854	0.0694	0.7923	0.870	sport 4
sex_sport		1	-0.03689	0.20926	0.0311	0.8601	0.964	

The results presented in the table above indicate that none of the predictor variables produced a significant parameter estimate, therefore we can conclude that the days to infection were not different by gender nor the sport in which the athlete participated.

---

[1] Wicklin, R. (2011) http://blogs.sas.com/content/iml/2011/10/19/four-essential-functions-for-statistical-programmers.html

[2] Introduction to Survival Analysis in SAS.UCLA: Statistical Consulting Group.From http://www.ats.ucla.edu/stat/sas/seminars/sas_survival/ (accessed Feb 20, 2017)

[3] Bewick, V., Cheek, L., Ball, J., Statistics review 12: Survival analysis, Critical Care 2004, 8:389-394.

[4] The Cox Proportional Hazard regression is based on Sir David Cox 1972 paper: Regression Models and Life-Tables (1972), J. R. Stat. Soc. B, 34:187–220).

[5] The weibull.com reliability engineering resource website is a service of ReliaSoft Corporation.
Copyright © 1992 – 2017 ReliaSoft Corporation. All Rights Reserved.

[6] Introduction to Survival Analysis in SAS.UCLA: Statistical Consulting Group.From http://www.ats.ucla.edu/stat/sas/seminars/sas_survival/ (accessed Feb 20, 2017)

44. Repeated Measures, Split Plots, and Mixed Model ANOVAS

Exploring the Analysis of Variance through various research designs

In this chapter, we will explore the various applications of the analysis of variance through different research designs that include Random Block, Repeated Measures, Split Plot Factorial, 3 way models with interaction, mediating and moderating issues, and the fixed and random effects (Mixed Model) ANOVA.

Let's begin with an in depth exploration of the Analysis of Variance.