Boundless Statistics for Organizations

5.1 CENTRAL TENDENCY

5.1: Central Tendency

5.1.1: Mean: The Average

The term central tendency relates to the way in which quantitative data tend to cluster around some value.

Learning Objectives

Define the average and distinguish between arithmetic, geometric, and harmonic means.

Key Takeaways

Key Points

- An average is a measure of the “middle” or “typical” value of a data set.
- The three most common averages are the Pythagorean means – the arithmetic mean, the geometric mean, and the harmonic mean.
- The arithmetic mean is the sum of a collection of numbers divided by the number of numbers in the collection.
- The geometric mean is a type of mean or average which indicates the central tendency, or typical value, of a set of numbers by using the product of their values. It is defined as the nth

root (where n is the count of numbers) of the product of the numbers.

- The harmonic mean H of the positive real numbers $X_1, X_2, \ldots X_n$ is defined to be the reciprocal of the arithmetic mean of the reciprocals of $X_1, X_2, \ldots X_n$. It is typically appropriate for situations when the average of rates is desired.

Key Terms

average
: any measure of central tendency, especially any mean, the median, or the mode

arithmetic mean
: the measure of central tendency of a set of values computed by dividing the sum of the values by their number; commonly called the mean or the average

central tendency
: a term that relates the way in which quantitative data tend to cluster around some value

Example

The arithmetic mean, often simply called the mean, of two numbers, such as 2 and 8, is obtained by finding a value A such that. One may find that $A = \frac{8+2}{2}$. Switching the order of 2 and 8 to read 8 and 2 does not change the resulting value obtained for A. The mean 5 is not less than the minimum 2 nor greater than the maximum 8. If we increase the number of terms in the list for which we want an average, we get, for example, that the arithmetic mean of 2, 8, and 11 is found by solving for the value of A in the equation $A = \frac{2+8+11}{3}$. One finds that A=7.

The term central tendency relates to the way in which quantitative data tend to cluster around some value. A measure of central tendency is any of a variety of ways of specifying this "central value". Central tendency is contrasted with statistical dispersion (spread), and together these are the most used properties of distributions. Statistics that measure central tendency can be used in descriptive statistics as a summary statistic for a data set, or as estimators of location parameters of a statistical model.

In the simplest cases, the measure of central tendency is an average of a set of measurements, the word average being variously construed as mean, median, or other measure of location, depending on the context. An average is a measure of the "middle" or "typical" value of a data set. In the most common case, the data set

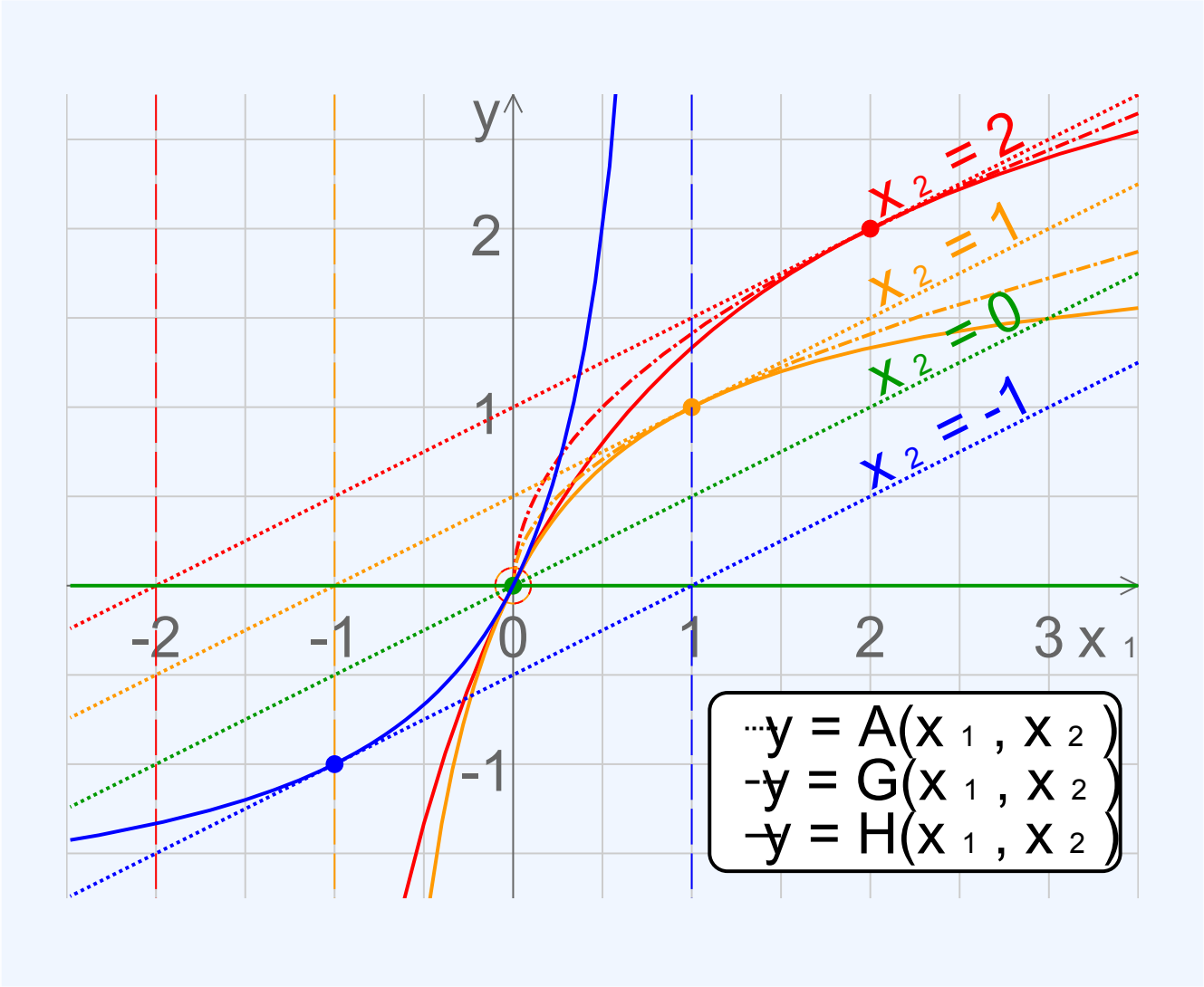
is a list of numbers. The average of a list of numbers is a single number intended to typify the numbers in the list. If all the numbers in the list are the same, then this number should be used. If the numbers are not the same, the average is calculated by combining the numbers from the list in a specific way and computing a single number as being the average of the list.

The term mean has three related meanings:

1. The arithmetic mean of a sample,
2. The expected value of a random variable, or
3. The mean of a probability distribution

The Pythagorean Means

The three most common averages are the Pythagorean means – the arithmetic mean, the geometric mean, and the harmonic mean.



Comparison of Pythagorean Means

Comparison of the arithmetic, geometric and harmonic means of a pair of numbers. The vertical dashed lines are asymptotes for the harmonic means.

The Arithmetic Mean

When we think of means, or averages, we are typically thinking of the arithmetic mean. It is the sum of a collection of numbers divided by the number of numbers in the collection. The collection is often a set of results of an experiment, or a set of results from a survey of a subset of the public. In addition to mathematics and statistics, the arithmetic mean is used frequently in fields such as economics, sociology, and history, and it is used in almost every academic field to some extent. For example, per capita income is the arithmetic average income of a nation's population.

Suppose we have a data set containing the values a_1, ..., a_n. The arithmetic mean is defined via the expression: A=1n∑i=1nai">$A = \frac{1}{n}\sum_{i=1}^{n} a_i$

If the data set is a statistical population (i.e., consists of every possible observation and not just a subset of them), then the mean of that population is called the population mean. If the data set is a statistical sample (a subset of the population) we call the statistic resulting from this calculation a sample mean. If it is required to use a single number as an estimate for the values of numbers, then the arithmetic mean does this best. This is because it minimizes the sum of squared deviations from the estimate.

The Geometric Mean

The geometric mean is a type of mean or average which indicates the central tendency, or typical value, of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). The geometric mean applies only to positive numbers. The geometric mean is defined as the n">nth root (where n">n is the count of numbers) of the product of the numbers.

For instance, the geometric mean of two numbers, say 2 and 8, is just the square root of their product; that is 2⋅8=4">$\sqrt{2 \cdot 8} = 4$. As another example, the geometric mean of the three numbers 4, 1, and 1/32 is the cube root of their product (1/8), which is 1/2; that is $\sqrt[3]{4 \cdot 1 \cdot \frac{1}{32}} = \frac{1}{2}$ .4⋅1⋅1323=12">

A geometric mean is often used when comparing different items – finding a single "figure of merit" for these items – when each item has multiple properties that have different numeric ranges. The use of a geometric

mean "normalizes" the ranges being averaged, so that no range dominates the weighting, and a given percentage change in any of the properties has the same effect on the geometric mean.

For example, the geometric mean can give a meaningful "average" to compare two companies which are each rated at 0 to 5 for their environmental sustainability, and are rated at 0 to 100 for their financial viability. If an arithmetic mean was used instead of a geometric mean, the financial viability is given more weight because its numeric range is larger – so a small percentage change in the financial rating (e.g. going from 80 to 90) makes a much larger difference in the arithmetic mean than a large percentage change in environmental sustainability (e.g. going from 2 to 5).

The Harmonic Mean

The harmonic mean is typically appropriate for situations when the average of rates is desired. It may (compared to the arithmetic mean) mitigate the influence of large outliers and increase the influence of small values.

The harmonic mean H">H of the positive real numbers x1,x2,…,xn">$x_1,x_2,\ldots,x_n$ is defined to be the reciprocal of the arithmetic mean of the reciprocals of x1,x2,…,xn">$x_1,x_2,\ldots,x_n$. For example, the harmonic mean of 1, 2, and 4 is:

311+12+14=113(11+12+14)=127≈1.7143">$\frac{3}{\frac{1}{1}+\frac{1}{2}+\frac{1}{4}} = \frac{1}{\frac{1}{3}\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{4}\right)} = \frac{12}{7} \approx 1.7143$

The harmonic mean is the preferable method for averaging multiples, such as the price/earning ratio in Finance, in which price is in the numerator. If these ratios are averaged using an arithmetic mean (a common error), high data points are given greater weights than low data points. The harmonic mean, on the other hand, gives equal weight to each data point.

5.1.2: The Average and the Histogram

The shape of a histogram can assist with identifying other descriptive statistics, such as which measure of central tendency is appropriate to use.

Learning Objectives

Demonstrate the effect that the shape of a distribution has on measures of central tendency.

Key Takeaways

Key Points

- Histograms tend to form shapes, which when measured can describe the distribution of data within a dataset.
- A key feature of the normal distribution is that the mode, median and mean are the same and are together in the center of the curve.
- A key feature of the skewed distribution is that the mean and median have different values and do not all lie at the center of the curve.
- Skewed distributions with two or more modes are known as bi-modal or multimodal, respectively.

Key Terms

normal distribution
: A family of continuous probability distributions such that the probability density function is the normal (or Gaussian) function.

bell curve
: In mathematics, the bell-shaped curve that is typical of the normal distribution.

histogram
: A representation of tabulated frequencies, shown as adjacent rectangles, erected over discrete

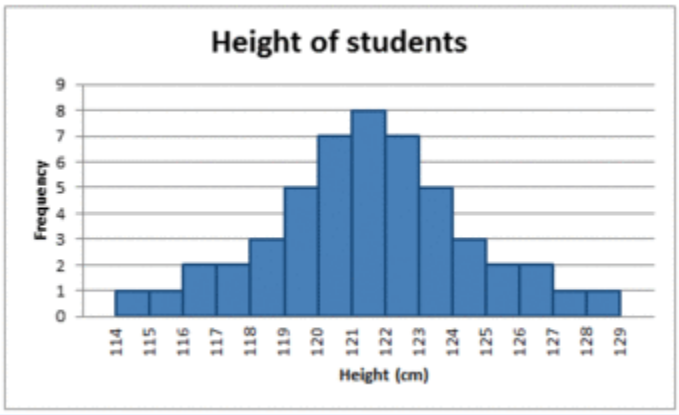
intervals (bins), with an area equal to the frequency of the observations in the interval.

As discussed, a histogram is a bar graph displaying tabulated frequencies. Histograms tend to form shapes, which when measured can describe the distribution of data within a dataset. The shape of the distribution can assist with identifying other descriptive statistics, such as which measure of central tendency is appropriate to use.

The distribution of data item values may be symmetrical or asymmetrical. Two common examples of symmetry and asymmetry are the "normal distribution" and the "skewed distribution. "

Central Tendency and Normal Distributions

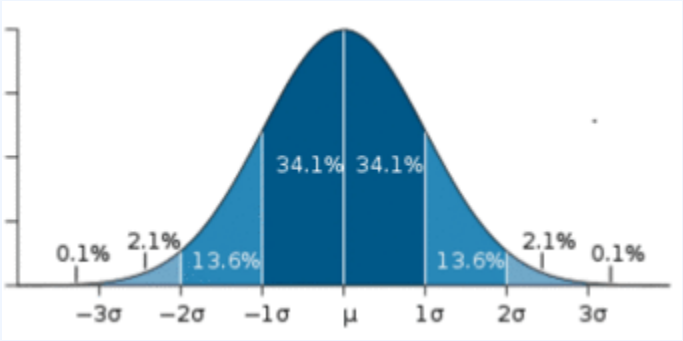
In a symmetrical distribution the two sides of the distribution are a mirror image of each other. A normal distribution is a true symmetric distribution of data item values. When a histogram is constructed on values that are normally distributed, the shape of columns form a symmetrical bell shape. This is why this distribution is also known as a "normal curve" or "bell curve. " is an example of a normal distribution:



The Normal Distribution

A histogram showing a normal distribution, or bell curve.

If represented as a 'normal curve' (or bell curve) the graph would take the following shape (where μ">μ is the mean and σ">σ is the standard deviation):



The Bell Curve

The shape of a normally distributed histogram.

A key feature of the normal distribution is that the mode, median and mean are the same and are together in the center of the curve.

Also, there can only be one mode (i.e. there is only one value which is most frequently observed). Moreover, most of the data are clustered around the center, while the more extreme values on either side of the center become less rare as the distance from the center increases (i.e. about 68% of values lie within one standard deviation (σ">σ) away from the mean; about 95% of the values lie within two standard deviations; and about 99.7% are within three standard deviations. This is known as the empirical rule or the 3-sigma rule).

Central Tendency and Skewed Distributions

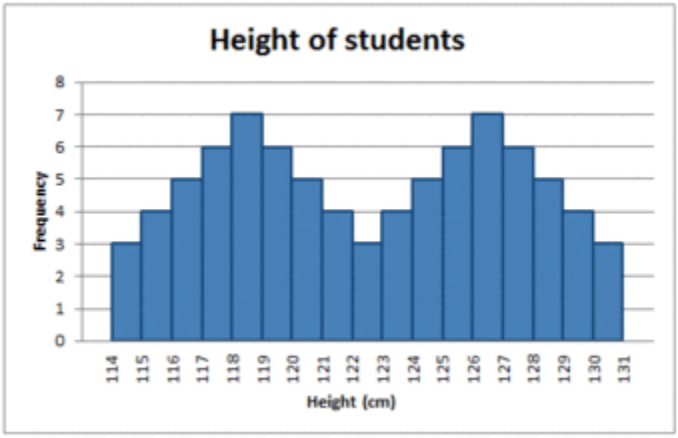
In an asymmetrical distribution the two sides will not be mirror images of each other. Skewness is the tendency for the values to be more frequent around the high or low ends of the x">x-axis. When a histogram is constructed for skewed data it is possible to identify skewness by looking at the shape of the distribution. For example, a distribution is said to be positively skewed when the tail on the right side of the histogram is longer than the left side. Most of the values tend to cluster toward the left side of the x">x-axis (i.e, the smaller values) with increasingly fewer values at the right side of the x">x-axis (i.e. the larger values).

A distribution is said to be negatively skewed when the tail on the left side of the histogram is longer than the right side. Most of the values tend to cluster toward the right side of the x">x-axis (i.e. the larger values), with increasingly less values on the left side of the x">x-axis (i.e. the smaller values).

A key feature of the skewed distribution is that the mean and median have different values and do not all lie at the center of the curve.

There can also be more than one mode in a skewed distribution. Distributions with two or more modes are

known as bi-modal or multimodal, respectively. The distribution shape of the data in is bi-modal because there are two modes (two values that occur more frequently than any other) for the data item (variable).



Bi-modal Distribution

Some skewed distributions have two or more modes.

5.1.3: The Root-Mean-Square

The root-mean-square, also known as the quadratic mean, is a statistical measure of the magnitude of a varying quantity, or set of numbers.

Learning Objective

Compute the root-mean-square and express its usefulness.

Key Takeaways

Key Points

- The root-mean-square is especially useful when a data set includes both positive and negative numbers.
- Its name comes from its definition as the square root of the mean of the squares of the values.
- The process of computing the root mean square is to: 1) Square all of the values 2) Compute the average of the squares 3) Take the square root of the average.
- The root-mean-square is always greater than or equal to the average of the unsigned values.

Key Term

root mean square
: the square root of the arithmetic mean of the squares

The root-mean-square, also known as the quadratic mean, is a statistical measure of the magnitude of a varying quantity, or set of numbers. It can be calculated for a series of discrete values or for a continuously varying function. Its name comes from its definition as the square root of the mean of the squares of the values.

This measure is especially useful when a data set includes both positive and negative numbers. For example, consider the set of numbers [−2,5,−8,9,−4]">[−2,5,−8,9,−4]. Computing the average of this set of numbers wouldn't tell us much because the negative numbers cancel out the positive numbers, resulting in an average of zero. This gives us the "middle value" but not a sense of the average magnitude.

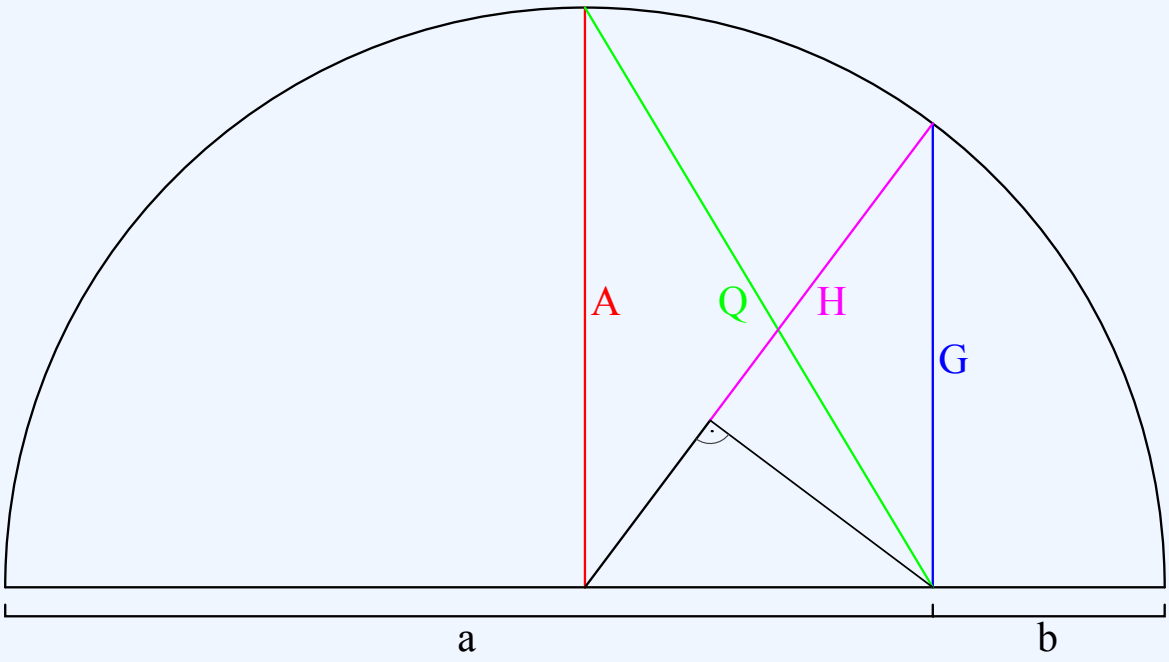
One possible method of assigning an average to this set would be to simply erase all of the negative signs. This would lead us to compute an average of 5.6. However, using the RMS method, we would square every number (making them all positive) and take the square root of the average. Explicitly, the process is to:

1. Square all of the values
2. Compute the average of the squares
3. Take the square root of the average

In our example:

1. (−2)2+52+(−8)2+92+(−4)2">$(-2)^2 + 5^2 + (-8)^2 + 9^2 + (-4)^2$
2. 4+25+64+81+165=38">$\frac{4+25+64+81+16}{5} = 38$
3. $\sqrt{38} \approx 6.16$

The root-mean-square is always greater than or equal to the average of the unsigned values. Physical scientists often use the term "root-mean-square" as a synonym for standard deviation when referring to the square root of the mean squared deviation of a signal from a given baseline or fit. This is useful for electrical engineers in calculating the "AC only" RMS of an electrical signal. Standard deviation being the root-mean-square of a signal's variation about the mean, rather than about 0, the DC component is removed (i.e. the RMS of the signal is the same as the standard deviation of the signal if the mean signal is zero).



Mathematical Means

This is a geometrical representation of common mathematical means. a">a, b">b are scalars. A">A is the arithmetic mean of scalars a">a and b">b. G">G is the geometric mean, H">H is the harmonic mean, Q">Q is the quadratic mean (also known as root-mean-square).

5.1.4: Which Average: Mean, Mode, or Median?

Depending on the characteristic distribution of a data set, the mean, median or mode may be the more appropriate metric for understanding.

Learning Objective

Assess various situations and determine whether the mean, median, or mode would be the appropriate measure of central tendency.

Key Takeaways

Key Points

- In symmetrical, unimodal distributions, such as the normal distribution (the distribution whose density function, when graphed, gives the famous "bell curve"), the mean (if defined), median and mode all coincide.
- If elements in a sample data set increase arithmetically, when placed in some order, then the median and arithmetic mean are equal. For example, consider the data sample {1, 2, 3, 4}. The mean is 2.5, as is the median.
- While the arithmetic mean is often used to report central tendencies, it is not a robust statistic, meaning that it is greatly influenced by outliers (values that are very much larger or smaller than most of the values).
- The median is of central importance in robust statistics, as it is the most resistant statistic, having a breakdown point of 50%: so long as no more than half the data is contaminated, the median will not give an arbitrarily large result.

- Unlike mean and median, the concept of mode also makes sense for "nominal data" (i.e., not consisting of numerical values in the case of mean, or even of ordered values in the case of median).

Key Terms

Mode
: the most frequently occurring value in a distribution

breakdown point
: the number or proportion of arbitrarily large or small extreme values that must be introduced into a batch or sample to cause the estimator to yield an arbitrarily large result

median
: the numerical value separating the higher half of a data sample, a population, or a probability distribution, from the lower half

The Mode

The mode is the value that appears most often in a set of data. For example, the mode of the sample [1,3,6,6,6,6,7,7,12,12,17]">[1,3,6,6,6,6,7,7,12,12,17] is 6. Like the statistical mean and median, the mode is a way of expressing, in a single number, important information about a random variable or a population.

The mode is not necessarily unique, since the same maximum frequency may be attained at different values. Given the list of data [1,1,2,4,4]">[1,1,2,4,4] the mode is not unique – the dataset may be said to be bimodal, while a set with more than two modes may be described as multimodal. The most extreme case occurs in uniform distributions, where all values occur equally frequently.

For a sample from a continuous distribution, the concept is unusable in its raw form. No two values will be exactly the same, so each value will occur precisely once. In order to estimate the mode, the usual practice is to discretize the data by assigning frequency values to intervals of equal distance, as with making a histogram, effectively replacing the values with the midpoints of the intervals they are assigned to. The mode is then the value where the histogram reaches its peak.

The Median

The median is the numerical value separating the higher half of a data sample, a population, or a probability distribution, from the lower half. The median of a finite list of numbers can be found by arranging all

the observations from lowest value to highest value and picking the middle one (e.g., the median of {3,5,9}">{3,5,9} is 5). If there is an even number of observations, then there is no single middle value. In this case, the median is usually defined to be the mean of the two middle values.

The median can be used as a measure of location when a distribution is skewed, when end-values are not known, or when one requires reduced importance to be attached to outliers (e.g., because there may be measurement errors).

Which to Use?

In symmetrical, unimodal distributions, such as the normal distribution (the distribution whose density function, when graphed, gives the famous "bell curve"), the mean (if defined), median and mode all coincide. For samples, if it is known that they are drawn from a symmetric distribution, the sample mean can be used as an estimate of the population mode.

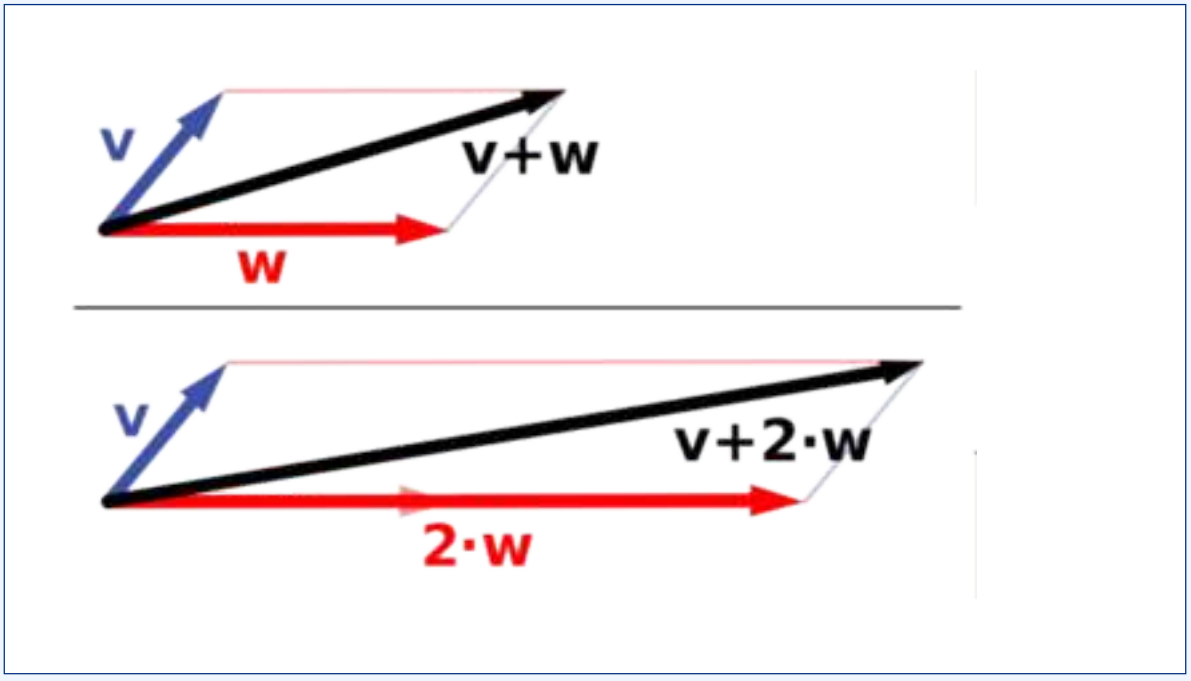
If elements in a sample data set increase arithmetically, when placed in some order, then the median and arithmetic mean are equal. For example, consider the data sample {1,2,3,4}">{1,2,3,4}. The mean is 2.5, as is the median. However, when we consider a sample that cannot be arranged so as to increase arithmetically, such as {1,2,4,8,16}">{1,2,4,8,16}, the median and arithmetic mean can differ significantly. In this case, the arithmetic mean is 6.2 and the median is 4. In general the average value can vary significantly from most values in the sample, and can be larger or smaller than most of them.

While the arithmetic mean is often used to report central tendencies, it is not a robust statistic, meaning that it is greatly influenced by outliers (values that are very much larger or smaller than most of the values). Notably, for skewed distributions, such as the distribution of income for which a few people's incomes are substantially greater than most people's, the arithmetic mean may not be consistent with one's notion of "middle," and robust statistics such as the median may be a better description of central tendency.

The median is of central importance in robust statistics, as it is the most resistant statistic, having a breakdown point of 50%: so long as no more than half the data is contaminated, the median will not give an arbitrarily large result. Robust statistics are statistics with good performance for data drawn from a wide range of probability distributions, especially for distributions that are not normally distributed. One motivation is to produce statistical methods that are not unduly affected by outliers. Another motivation is to provide methods with good performance when there are small departures from parametric distributions.

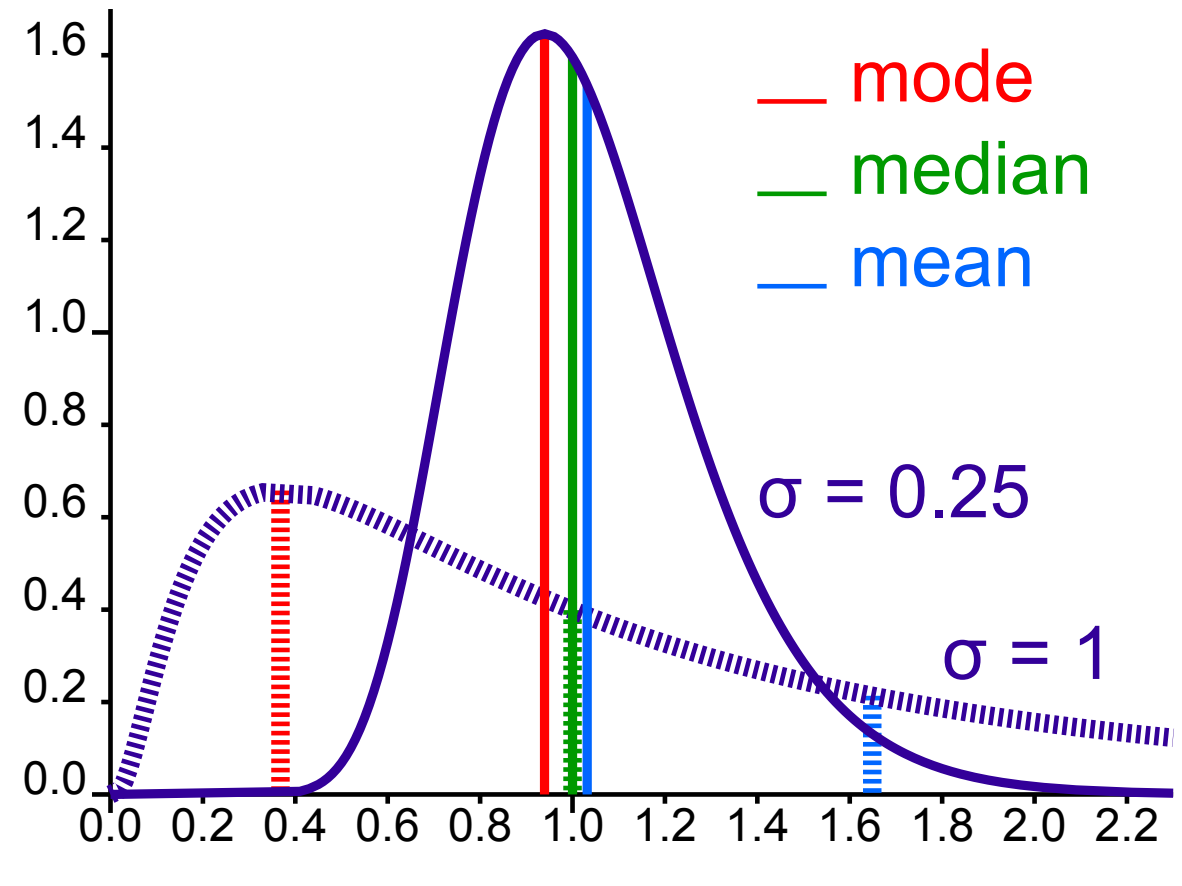
Unlike median, the concept of mean makes sense for any random variable assuming values from a vector space. For example, a distribution of points in the plane will typically have a mean and a mode, but the concept of median does not apply.

Unlike mean and median, the concept of mode also makes sense for "nominal data" (i.e., not consisting of numerical values in the case of mean, or even of ordered values in the case of median). For example, taking a sample of Korean family names, one might find that "Kim" occurs more often than any other name. Then "Kim" would be the mode of the sample. In any voting system where a plurality determines victory, a single modal value determines the victor, while a multi-modal outcome would require some tie-breaking procedure to take place.



Vector Space

Vector addition and scalar multiplication: a vector v">v (blue) is added to another vector w">w (red, upper illustration). Below, w">w is stretched by a factor of 2, yielding the sum v+2w">v+2w.



Comparison of the Mean, Mode & Median

Comparison of mean, median and mode of two log-normal distributions with different skewness.

5.1.5: Averages of Qualitative and Ranked Data

The central tendency for qualitative data can be described via the median or the mode, but not the mean.

Learning Objective

Categorize levels of measurement and identify the appropriate measures of central tendency.

Key Takeaways

Key Points

- Qualitative data can be defined as either nominal or ordinal.
- The nominal scale differentiates between items or subjects based only on their names and/or categories and other qualitative classifications they belong to.
- The mode is allowed as the measure of central tendency for nominal data.
- The ordinal scale allows for rank order by which data can be sorted, but still does not allow for relative degree of difference between them. The median and the mode are allowed as the measure of central tendency; however, the mean as the measure of central tendency is not allowed.
- The median and the mode are allowed as the measure of central tendency for ordinal data; however, the mean as the measure of central tendency is not allowed.

Key Terms

quantitative
: of a measurement based on some quantity or number rather than on some quality

qualitative
: of descriptions or distinctions based on some quality rather than on some quantity

dichotomous
: dividing or branching into two pieces

Levels of Measurement

In order to address the process for finding averages of qualitative data, we must first introduce the concept of levels of measurement. In statistics, levels of measurement, or scales of measure, are types of data that arise in the theory of scale types developed by the psychologist Stanley Smith Stevens. Stevens proposed his typology in a 1946 Science article entitled "On the Theory of Scales of Measurement. " In that article, Stevens claimed that all measurement in science was conducted using four different types of scales that he called "nominal", "ordinal", "interval" and "ratio", unifying both qualitative (which are described by his "nominal" type) and quantitative (to a different degree, all the rest of his scales).

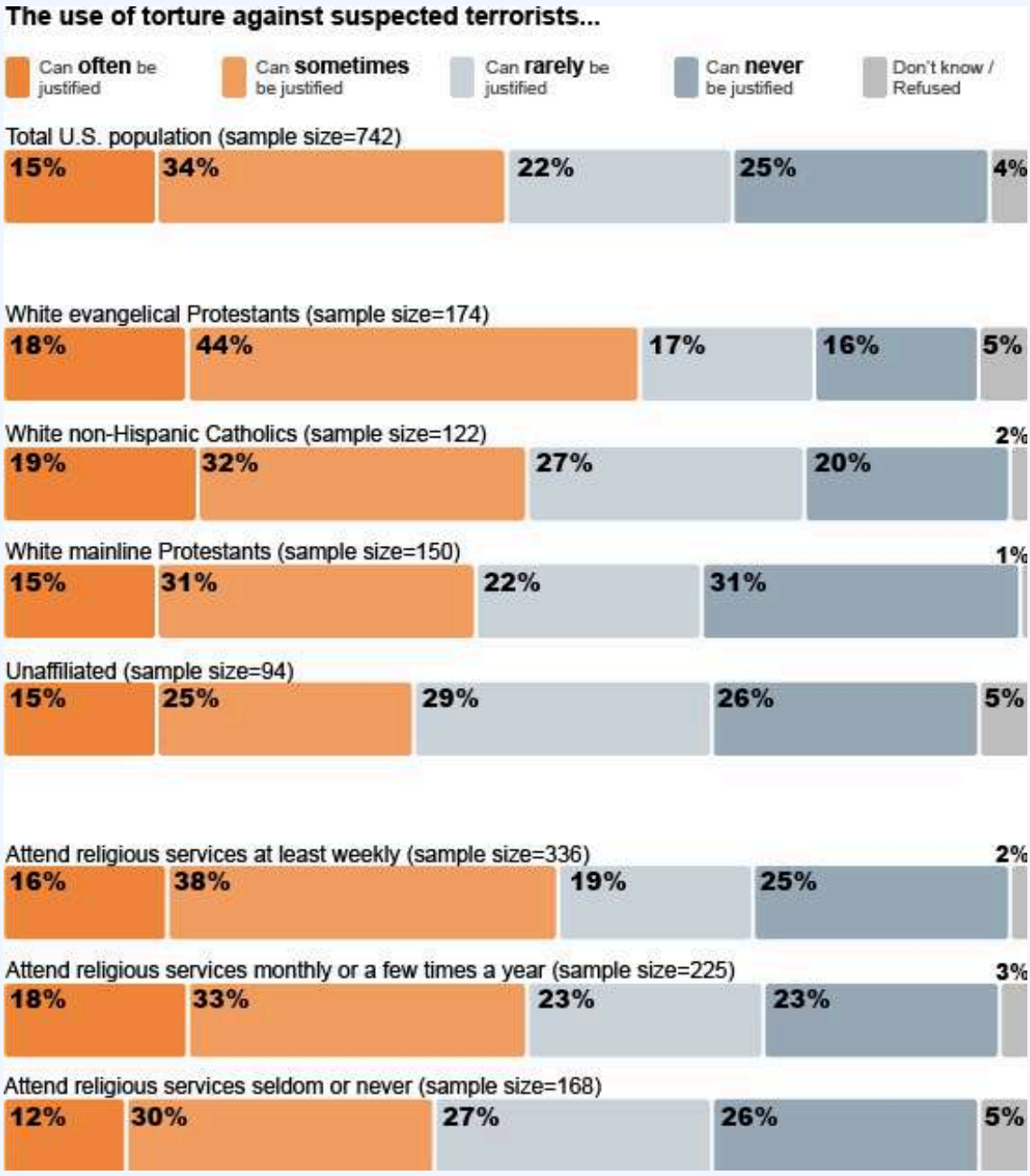
Nominal Scale

The nominal scale differentiates between items or subjects based only on their names and/or categories and other qualitative classifications they belong to. Examples include gender, nationality, ethnicity, language, genre, style, biological species, visual pattern, and form.

The mode, i.e. the most common item, is allowed as the measure of central tendency for the nominal type. On the other hand, the median, i.e. the middle-ranked item, makes no sense for the nominal type of data since ranking is not allowed for the nominal type.

Ordinal Scale

The ordinal scale allows for rank order (1st, 2nd, 3rd, et cetera) by which data can be sorted, but still does not allow for relative degree of difference between them. Examples include, on one hand, dichotomous data with dichotomous (or dichotomized) values such as "sick" versus "healthy" when measuring health, "guilty" versus "innocent" when making judgments in courts, or "wrong/false" versus "right/true" when measuring truth value. On the other hand, non-dichotomous data consisting of a spectrum of values is also included, such as "completely agree," "mostly agree," "mostly disagree," and "completely disagree" when measuring opinion .



Ordinal Scale Surveys

An opinion survey on religiosity and torture. An opinion survey is an example of a non-dichotomous data set on the ordinal scale for which the central tendency can be described by the median or the mode.

The median, i.e. middle-ranked, item is allowed as the measure of central tendency; however, the mean (or average) as the measure of central tendency is not allowed. The mode is also allowed.

In 1946, Stevens observed that psychological measurement, such as measurement of opinions, usually operates on ordinal scales; thus means and standard deviations have no validity, but they can be used to get ideas for how to improve operationalization of variables used in questionnaires.

Attributions

- Mean: The Average
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Geometric mean."
    http://en.wikipedia.org/wiki/Geometric_mean.
    Wikipedia
    CC BY-SA 3.0.
  - "Arithmetic mean."
    https://en.wikipedia.org/wiki/Arithmetic_mean.
    Wikipedia
    CC BY-SA 3.0.
  - "Harmonic mean."
    http://en.wikipedia.org/wiki/Harmonic_mean.
    Wikipedia
    CC BY-SA 3.0.
  - "Average."
    http://en.wikipedia.org/wiki/Average.
    Wikipedia
    CC BY-SA 3.0.
  - "central tendency."
    http://en.wikipedia.org/wiki/central%20tendency.
    Wikipedia
    CC BY-SA 3.0.
  - "Measure of central tendency."
    http://en.wikipedia.org/wiki/Measure_of_central_tendency.
    Wikipedia
    CC BY-SA 3.0.

- "Mean."
  https://en.wikipedia.org/wiki/Mean.
  Wikipedia
  CC BY-SA 3.0.
- "average."
  http://en.wiktionary.org/wiki/average.
  Wiktionary
  CC BY-SA 3.0.
- "arithmetic mean."
  http://en.wiktionary.org/wiki/arithmetic_mean.
  Wiktionary
  CC BY-SA 3.0.
- "Comparison Pythagorean means."
  http://en.wikipedia.org/wiki/File:Comparison_Pythagorean_means.svg.
  Wikipedia
  CC BY-SA.

- The Average and the Histogram
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Error 404."
    http://www.abs.gov.au/websitedbs/a3121120.nsf/89a5f3d8684682b6ca256de4002c809b/81a53a0a10c05d3bca257949001281b5!OpenDocument.
    Austrailian Bureau of Statistics
    CC BY.
  - "bell curve."
    http://en.wiktionary.org/wiki/bell_curve.
    Wiktionary
    CC BY-SA 3.0.
  - "histogram."
    http://en.wikipedia.org/wiki/histogram.
    Wikipedia
    CC BY-SA 3.0.
  - "normal distribution."
    http://en.wiktionary.org/wiki/normal_distribution.
    Wiktionary

CC BY-SA 3.0.

- "Error 404."
  http://www.abs.gov.au/websitedbs/a3121120.nsf/89a5f3d8684682b6ca256de4002c809b/81a53a0a10c05d3bca257949001281b5!OpenDocument.
  Austrailian Bureau of Statistics
  CC BY.
- "Error 404."
  http://www.abs.gov.au/websitedbs/a3121120.nsf/89a5f3d8684682b6ca256de4002c809b/81a53a0a10c05d3bca257949001281b5!OpenDocument.
  Austrailian Bureau of Statistics
  CC BY.
- "Error 404."
  http://www.abs.gov.au/websitedbs/a3121120.nsf/89a5f3d8684682b6ca256de4002c809b/81a53a0a10c05d3bca257949001281b5!OpenDocument.
  Austrailian Bureau of Statistics
  CC BY.

- The Root-Mean-Square
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "root mean square."
    http://en.wiktionary.org/wiki/root_mean_square.
    Wiktionary
    CC BY-SA 3.0.
  - "Root mean square."
    https://en.wikipedia.org/wiki/Root_mean_square.
    Wikipedia
    CC BY-SA 3.0.
  - "MathematicalMeans."
    http://commons.wikimedia.org/wiki/File:MathematicalMeans.svg.
    Wikimedia
    Public domain.
- Which Average: Mean, Mode, or Median?
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning

CC BY-SA 3.0.
- "Median."
  http://en.wikipedia.org/wiki/Median.
  Wikipedia
  CC BY-SA 3.0.
- "median."
  http://en.wikipedia.org/wiki/median.
  Wikipedia
  CC BY-SA 3.0.
- "Mode."
  http://en.wikipedia.org/wiki/Mode.
  Wikipedia
  CC BY-SA 3.0.
- "Robust statistics."
  http://en.wikipedia.org/wiki/Robust_statistics.
  Wikipedia
  CC BY-SA 3.0.
- "Arithmetic mean."
  https://en.wikipedia.org/wiki/Arithmetic_mean.
  Wikipedia
  CC BY-SA 3.0.
- "Arithmetic mean."
  https://en.wikipedia.org/wiki/Arithmetic_mean.
  Wikipedia
  CC BY-SA 3.0.
- "Mode (statistics)."
  http://en.wikipedia.org/wiki/Mode_(statistics).
  Wikipedia
  CC BY-SA 3.0.
- "breakdown point."
  http://en.wiktionary.org/wiki/breakdown_point.
  Wiktionary
  CC BY-SA 3.0.
- "OpenStax College, Prokaryotic Diversity. October 16, 2013."
  http://cnx.org/content/m44603/latest/Figure_22_01_06.jpg.
  OpenStax CNX
  CC BY 3.0.

- "Vector addition ans scaling."
  http://commons.wikimedia.org/wiki/File:Vector_addition_ans_scaling.png.
  Wikimedia
  CC BY-SA.
- "Comparison mean median mode."
  http://commons.wikimedia.org/wiki/File:Comparison_mean_median_mode.svg.
  Wikimedia
  CC BY-SA.

- Averages of Qualitative and Ranked Data
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "qualitative."
    http://en.wiktionary.org/wiki/qualitative.
    Wiktionary
    CC BY-SA 3.0.
  - "dichotomous."
    http://en.wiktionary.org/wiki/dichotomous.
    Wiktionary
    CC BY-SA 3.0.
  - "Level of measurement."
    http://en.wikipedia.org/wiki/Level_of_measurement.
    Wikipedia
    CC BY-SA 3.0.
  - "quantitative."
    http://en.wiktionary.org/wiki/quantitative.
    Wiktionary
    CC BY-SA 3.0.
  - "Religiosity and Torture | Flickr – Photo Sharing!."
    http://www.flickr.com/photos/jurvetson/3492263284/.
    Flickr
    CC BY.

5.2 MEASURES OF RELATIVE STANDING

5.2: Measures of Relative Standing

5.2.1: Measures of Relative Standing

Measures of relative standing can be used to compare values from different data sets, or to compare values within the same data set.

Learning Objective

Outline how percentiles and quartiles measure relative standing within a data set.

Key Takeaways

Key Points

- The common measures of relative standing or location are quartiles and percentiles.
- A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.
- The 25th percentile is also known as the first quartile (Q1), the 50th percentile as the median or second quartile (Q2), and the 75th percentile as the third quartile (Q3).

- To calculate quartiles and percentiles, the data must be ordered from smallest to largest.
- For very large populations following a normal distribution, percentiles may often be represented by reference to a normal curve plot.
- Percentiles represent the area under the normal curve, increasing from left to right.

Key Terms

percentile
: any of the ninety-nine points that divide an ordered distribution into one hundred parts, each containing one per cent of the population

quartile
: any of the three points that divide an ordered distribution into four parts, each containing a quarter of the population

Example

For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race? b. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation. c. A bicyclist in the 90th percentile of a bicycle race between two towns completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation. *SOLUTION* a. For runners in a race it is more desirable to have a low percentile for finish time. A low percentile means a short time, which is faster. b. INTERPRETATION: 20% of runners finished the race in 5.2 minutes or less. 80% of runners finished the race in 5.2 minutes or longer. c. He is among the slowest cyclists (90% of cyclists were faster than him.) *INTERPRETATION*: 90% of cyclists had a finish time of 1 hour, 12 minutes or less.Only 10% of cyclists had a finish time of 1 hour, 12 minutes or longer.

Measures of relative standing, in the statistical sense, can be defined as measures that can be used to compare values from different data sets, or to compare values within the same data set.

Quartiles and Percentiles

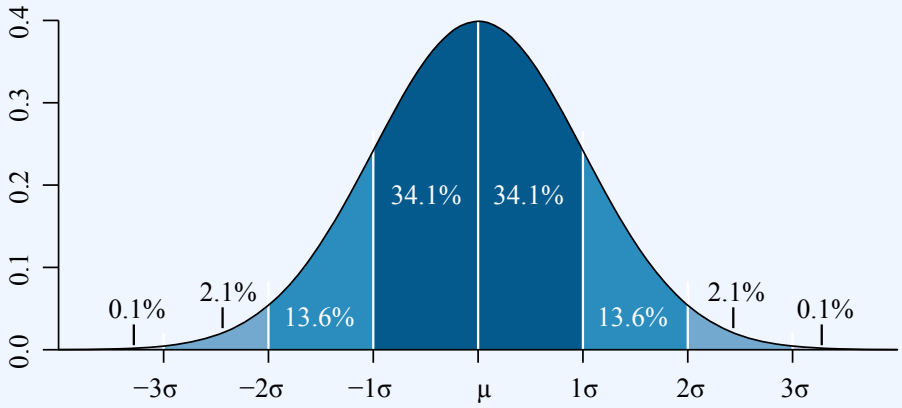
The common measures of relative standing or location are quartiles and percentiles. A *percentile* is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value (or score) below which 20 percent of the observations may be found. The term percentile and the related term, percentile rank, are often used in the reporting of scores from norm-referenced tests. For example, if a score is in the 86th percentile, it is higher than 86% of the other scores. The 25th percentile is also known as the first quartile (Q1), the 50th percentile as the median or second quartile (Q2), and the 75th percentile as the third quartile (Q3).

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Recall that quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

For very large populations following a normal distribution, percentiles may often be represented by reference to a normal curve plot. The normal distribution is plotted along an axis scaled to standard deviations, or sigma units. Percentiles represent the area under the normal curve, increasing from left to right. Each standard deviation represents a fixed percentile. Thus, rounding to two decimal places, −3″>−3

is the 0.13th percentile, −2″>−2 the 2.28th percentile, −1″>−1 the 15.87th percentile, 0 the 50th percentile (both the mean and median of the distribution), +1″>+1 the 84.13th percentile, +2″>+2 the 97.72nd percentile, and +3″>+3 the 99.87th percentile. This is known as the 68–95–99.7 rule or the three-sigma rule.



Percentile Diagram

Representation of the 68–95–99.7 rule. The dark blue zone represents observations within one standard deviation (σ">σ) to either side of the mean (μ">μ), which accounts for about 68.2% of the population. Two standard deviations from the mean (dark and medium blue) account for about 95.4%, and three standard deviations (dark, medium, and light blue) for about 99.7%.

Note that in theory the 0^{th} percentile falls at negative infinity and the 100^{th} percentile at positive infinity; although, in many practical applications, such as test results, natural lower and/or upper limits are enforced.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order, from smallest to largest. p">p% of data values are less than or equal to the p">p^{th} percentile. For example, 15% of data values are less than or equal to the 15^{th} percentile. Low percentiles always correspond to lower data values. High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad". The interpretation of whether a certain percentile is good or bad depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good'; in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to properly interpret percentiles is important not only when describing data, but is also important when calculating probabilities.

Guideline:

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information:

- information about the context of the situation being considered,
- the data value (value of the variable) that represents the percentile,
- the percent of individuals or items with data values below the percentile.
- Additionally, you may also choose to state the percent of individuals or items with data values above the percentile.

5.2.2: Median

The median is the middle value in distribution when the values are arranged in ascending or descending order.

Learning Objective

Identify the median in a data set and distinguish it's properties from other measures of central tendency.

Key Takeaways

Key Points

- The median divides the distribution in half (there are 50% of observations on either side of the median value). In a distribution with an odd number of observations, the median value is

the middle value.

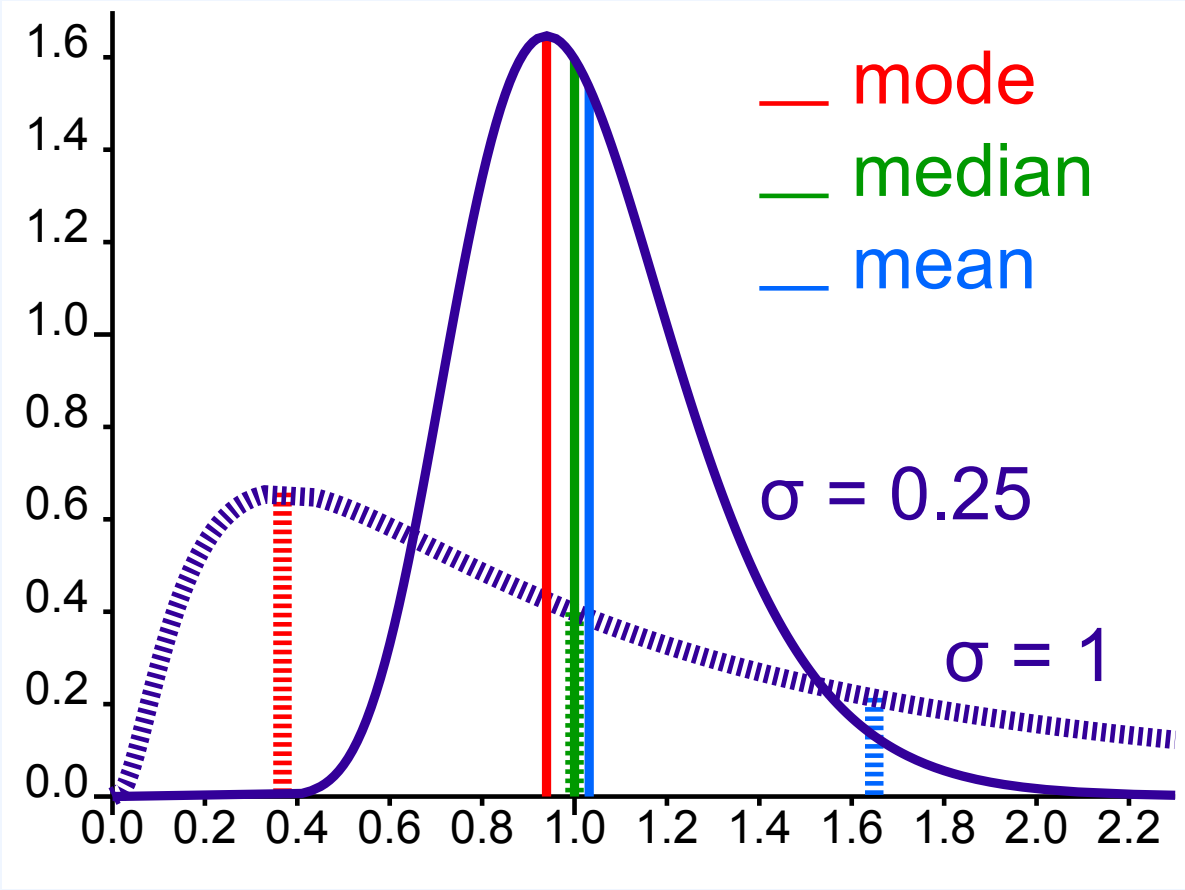
- When the distribution has an even number of observations, the median value is the mean of the two middle values.
- The median is less affected by outliers and skewed data than the mean, and is usually the preferred measure of central tendency when the distribution is not symmetrical.
- he median cannot be identified for categorical nominal data, as it cannot be logically ordered.

Key Terms

outlier
: a value in a statistical sample which does not fit a pattern that describes most other data points; specifically, a value that lies 1.5 IQR beyond the upper or lower quartile

median
: the numerical value separating the higher half of a data sample, a population, or a probability distribution, from the lower half

A measure of central tendency (also referred to as measures of center or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or center of its distribution. There are three main measures of central tendency: the mode, the median and the mean . Each of these measures describes a different indication of the typical or central value in the distribution.



Central tendency

Comparison of mean, median and mode of two log-normal distributions with different skewness.

The median is the middle value in distribution when the values are arranged in ascending or descending order. The median divides the distribution in half (there are 50% of observations on either side of the median value). In a distribution with an odd number of observations, the median value is the middle value.

Looking at the retirement age distribution (which has 11 observations), the median is the middle value, which is 57 years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

When the distribution has an even number of observations, the median value is the mean of the two middle values. In the following distribution, the two middle values are 56 and 57, therefore the median equals 56.5 years:

52, 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

The median is less affected by outliers and skewed data than the mean, and is usually the preferred measure of central tendency when the distribution is not symmetrical. The median cannot be identified for categorical nominal data, as it cannot be logically ordered.

5.2.3: Mode

The mode is the most commonly occurring value in a distribution.

Learning Objectives

Define the mode and explain its limitations.

Key Takeaways

Key Points

- There are some limitations to using the mode. In some distributions, the mode may not reflect the center of the distribution very well.
- It is possible for there to be more than one mode for the same distribution of data, (eg bi-modal). The presence of more than one mode can limit the ability of the mode in describing the center or typical value of the distribution because a single value to describe the center cannot be identified.
- In some cases, particularly where the data are continuous, the distribution may have no mode at all (i.e. if all values are different). In cases such as these, it may be better to consider using the median or mean, or group the data in to appropriate intervals, and find the modal class.

Key Term

skewness

A measure of the asymmetry of the probability distribution of a real-valued random variable; is the third standardized moment, defined as where is the third moment about the mean and is the standard deviation.

A measure of central tendency (also referred to as measures of center or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or center of its distribution. There are three main measures of central tendency: the mode, the median and the mean . Each of these measures describes a different indication of the typical skewness in the distribution.

The mode is the most commonly occurring value in a distribution. Consider this dataset showing the retirement age of 11 people, in whole years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

The most commonly occurring value is 54, therefore the mode of this distribution is 54 years. The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.

There are some limitations to using the mode. In some distributions, the mode may not reflect the center of the distribution very well. When the distribution of retirement age is ordered from lowest to highest value, it is easy to see that the center of the distribution is 57 years, but the mode is lower, at 54 years. It is also possible for there to be more than one mode for the same distribution of data, (bi-modal, or multi-modal). The presence of more than one mode can limit the ability of the mode in describing the center or typical value of the distribution because a single value to describe the center cannot be identified. In some cases, particularly where the data are continuous, the distribution may have no mode at all (i.e. if all values are different). In cases such as these, it may be better to consider using the median or mean, or group the data in to appropriate intervals, and find the modal class.

Attributions

http://en.wikipedia.org/wiki/Standard_score.
Wikipedia
CC BY-SA 3.0.

- "percentile."
  http://en.wiktionary.org/wiki/percentile.
  Wiktionary
  CC BY-SA 3.0.
- "quartile."
  http://en.wiktionary.org/wiki/quartile.
  Wiktionary
  CC BY-SA 3.0.
- "Susan Dean and Barbara Illowsky, Descriptive Statistics: Measuring the Location of the Data. September 19, 2013."
  http://cnx.org/content/m16314/latest/.
  OpenStax CNX
  CC BY 3.0.
- "David Lane, Percentiles. October 12, 2013."
  http://cnx.org/content/m10805/latest/.
  OpenStax CNX
  CC BY 3.0.
- "Percentile."
  http://en.wikipedia.org/wiki/Percentile.
  Wikipedia
  CC BY-SA 3.0.
- "Standard deviation diagram."
  http://en.wikipedia.org/wiki/File:Standard_deviation_diagram.svg.
  Wikipedia
  CC BY.

- Median
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "median."
    http://en.wikipedia.org/wiki/median.
    Wikipedia
    CC BY-SA 3.0.

- “outlier.”
  http://en.wiktionary.org/wiki/outlier.
  Wiktionary
  CC BY-SA 3.0.
- “Error 404.”
  http://www.abs.gov.au/websitedbs/a3121120.nsf/89a5f3d8684682b6ca256de4002c809b/3a5b4029ba31b1cbca257949001281f8!OpenDocument.
  Austrailian Bureau of Statistics
  CC BY-SA.
- “Comparison mean median mode.”
  http://en.wikipedia.org/wiki/File:Comparison_mean_median_mode.svg.
  Wikipedia
  Public domain.

- Mode
  - “Boundless.”
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - “Error 400.”
    http://www.abs.gov.au/websitedbs/D3310114.nsf/Home/%C2%A9+Copyright?OpenDocument.
    Austrailian Bureau of Statistics
    CC BY.
  - “skewness.”
    http://en.wiktionary.org/wiki/skewness.
    Wiktionary
    CC BY-SA 3.0.
  - “Comparison mean median mode.”
    http://en.wikipedia.org/wiki/File:Comparison_mean_median_mode.svg.
    Wikipedia
    CC BY-SA.

5.3 THE LAW OF AVERAGES

5.3: The Law of Averages

5.3.1: What Does the Law of Averages Say?

The law of averages is a lay term used to express a belief that outcomes of a random event will "even out" within a small sample.

Learning Objectives

Evaluate the law of averages and distinguish it from the law of large numbers.

Key Takeaways

Key Points

- The law of averages typically assumes that unnatural short-term "balance" must occur. This can also be known as "Gambler's Fallacy" and is not a real mathematical principle.
- Some people mix up the law of averages with the law of large numbers, which is a real theorem that states that the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are

performed.

- The law of large numbers is important because it "guarantees" stable long-term results for the averages of random events. It does not guarantee what will happen with a small number of events.

Key Term

expected value
: of a discrete random variable, the sum of the probability of each possible outcome of the experiment multiplied by the value itself

The Law of Averages

The law of averages is a lay term used to express a belief that outcomes of a random event will "even out" within a small sample. As invoked in everyday life, the "law" usually reflects bad statistics or wishful thinking rather than any mathematical principle. While there is a real theorem that a random variable will reflect its underlying probability over a very large sample (the law of large numbers), the law of averages typically assumes that unnatural short-term "balance" must occur.

The law of averages is sometimes known as "Gambler's Fallacy. " It evokes the idea that an event is "due" to happen. For example, "The roulette wheel has landed on red in three consecutive spins. The law of averages says it's due to land on black! " Of course, the wheel has no memory and its probabilities do not change according to past results. So even if the wheel has landed on red in ten consecutive spins, the probability that the next spin will be black is still 48.6% (assuming a fair European wheel with only one green zero: it would be exactly 50% if there were no green zero and the wheel were fair, and 47.4% for a fair American wheel with one green "0" and one green "00"). (In fact, if the wheel has landed on red in ten consecutive spins, that is strong evidence that the wheel is not fair – that it is biased toward red. Thus, the wise course on the eleventh spin would be to bet on red, not on black: exactly the opposite of the layman's analysis.) Similarly, there is no statistical basis for the belief that lottery numbers which haven't appeared recently are due to appear soon.

The Law of Large Numbers

Some people interchange the law of averages with the law of large numbers, but they are different. The law of averages is not a mathematical principle, whereas the law of large numbers is. In probability theory, the law of large numbers is a theorem that describes the result of performing the same experiment a large number of

times. According to the law, the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

The law of large numbers is important because it "guarantees" stable long-term results for the averages of random events. For example, while a casino may lose money in a single spin of the roulette wheel, its earnings will tend towards a predictable percentage over a large number of spins. Any winning streak by a player will eventually be overcome by the parameters of the game. It is important to remember that the law of large numbers only applies (as the name indicates) when a large number of observations are considered. There is no principle that a small number of observations will coincide with the expected value or that a streak of one value will immediately be "balanced" by the others.

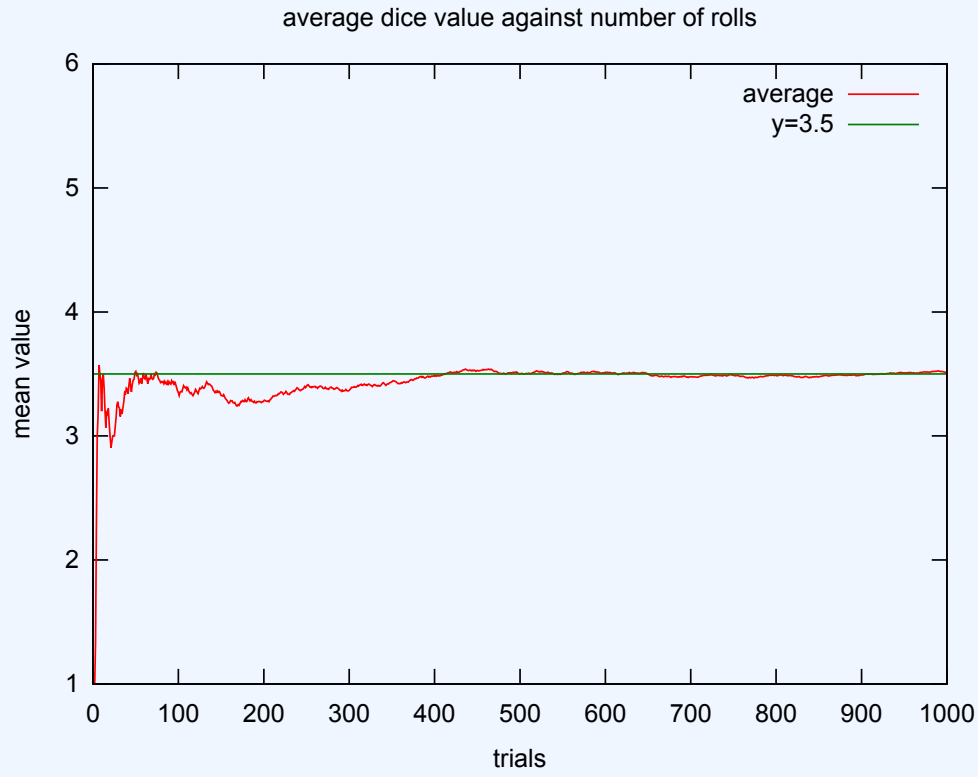
Another good example comes from the expected value of rolling a six-sided die. A single roll produces one of the numbers 1, 2, 3, 4, 5, or 6, each with an equal probability (16″>16

) of occurring. The expected value of a roll is 3.5, which comes from the following equation:

1+2+3+4+5+66=3.5″>(1+2+3+4+5+6)/6=3.5

According to the law of large numbers, if a large number of six-sided dice are rolled, the average of their values (sometimes called the sample mean) is likely to be close to 3.5, with the accuracy increasing as more dice are rolled. However, in a small number of rolls, just because ten 6's are rolled in a row, it doesn't mean a 1 is more likely the next roll. Each individual outcome still has a probability of 16″>16

.



The Law of Large Numbers: This shows a graph illustrating the law of large numbers using a particular run of rolls of a single die. As the number of rolls in this run increases, the average of the values of all the results approaches 3.5. While different runs would show a different shape over a small number of throws (at the left), over a large number of rolls (to the right) they would be extremely similar.

5.3.2: Chance Processes

A stochastic process is a collection of random variables that is often used to represent the evolution of some random value over time.

Learning Objective

Summarize the stochastic process and state its relationship to random walks.

Key Takeaways

Key Points

- One approach to stochastic processes treats them as functions of one or several deterministic arguments (inputs, in most cases regarded as time) whose values (outputs) are random variables.
- Random variables are non-deterministic (single) quantities which have certain probability distributions.
- Although the random values of a stochastic process at different times may be independent random variables, in most commonly considered situations they exhibit complicated statistical correlations.
- The law of a stochastic process is the measure that the process induces on the collection of functions from the index set into the state space.
- A random walk is a mathematical formalization of a path that consists of a succession of random steps.

Key Terms

random variable
: a quantity whose value is random and to which a probability distribution is assigned, such as the possible outcome of a roll of a die

random walk
: a stochastic path consisting of a series of sequential movements, the direction (and sometime length) of which is chosen at random

stochastic
: random; randomly determined

Example

Familiar examples of processes modeled as stochastic time series include stock market and exchange rate fluctuations; signals such as speech, audio and video; medical data such as a patient's EKG, EEG, blood pressure or temperature; and random movement such as Brownian motion or random walks.

Chance = Stochastic

In probability theory, a stochastic process–sometimes called a random process– is a collection of random variables that is often used to represent the evolution of some random value, or system, over time. It is the probabilistic counterpart to a deterministic process (or deterministic system). Instead of describing a process which can only evolve in one way (as in the case, for example, of solutions of an ordinary differential equation), in a stochastic or random process there is some indeterminacy. Even if the initial condition (or starting point) is known, there are several (often infinitely many) directions in which the process may evolve.

In the simple case of discrete time, a stochastic process amounts to a sequence of random variables known as a time series–for example, a Markov chain. Another basic type of a stochastic process is a random field, whose domain is a region of space. In other words, a stochastic process is a random function whose arguments are drawn from a range of continuously changing values.

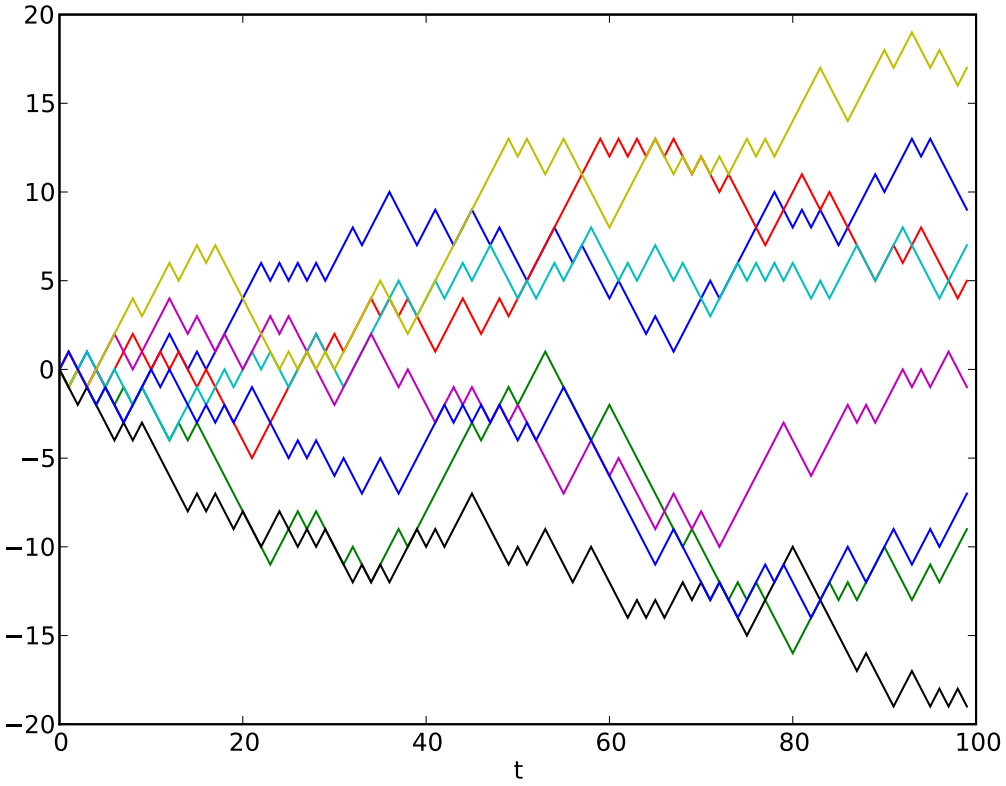
One approach to stochastic processes treats them as functions of one or several deterministic arguments (inputs, in most cases regarded as time) whose values (outputs) are random variables. Random variables are non-deterministic (single) quantities which have certain probability distributions. Random variables corresponding to various times (or points, in the case of random fields) may be completely different. Although the random values of a stochastic process at different times may be independent random variables, in most commonly considered situations they exhibit complicated statistical correlations.

Familiar examples of processes modeled as stochastic time series include stock market and exchange rate fluctuations; signals such as speech, audio, and video; medical data such as a patient's EKG, EEG, blood pressure, or temperature; and random movement such as Brownian motion or random walks.

Law of a Stochastic Process

The law of a stochastic process is the measure that the process induces on the collection of functions from the index set into the state space. The law encodes a lot of information about the process. In the case of a random walk, for example, the law is the probability distribution of the possible trajectories of the walk.

A random walk is a mathematical formalization of a path that consists of a succession of random steps. For example, the path traced by a molecule as it travels in a liquid or a gas, the search path of a foraging animal, the price of a fluctuating stock, and the financial status of a gambler can all be modeled as random walks, although they may not be truly random in reality. Random walks explain the observed behaviors of processes in such fields as ecology, economics, psychology, computer science, physics, chemistry, biology and, of course, statistics. Thus, the random walk serves as a fundamental model for recorded stochastic activity.



Random Walk

Example of eight random walks in one dimension starting at 0. The plot shows the current position on the line (vertical axis) versus the time steps (horizontal axis).

5.3.3: The Sum of Draws

The sum of draws is the process of drawing randomly, with replacement, from a set of data and adding up the results.

Learning Objective

Describe how chance variation affects sums of draws.

Key Takeaways

Key Points

- By drawing from a set of data with replacement, we are able to draw over and over again under the same conditions.
- The sum of draws is subject to a force known as chance variation.
- The sum of draws can be illustrated in practice through a game of Monopoly. A player rolls a pair of dice, adds the two numbers on the die, and moves his or her piece that many squares.

Key Term

chance variation
: the presence of chance in determining the variation in experimental results

The sum of draws can be illustrated by the following process. Imagine there is a box of tickets, each having a number 1, 2, 3, 4, 5, or 6 written on it.

The sum of draws can be represented by a process in which tickets are drawn at random from the box, with the ticket being replaced to the box after each draw. Then, the numbers on these tickets are added up. By replacing the tickets after each draw, you are able to draw over and over under the same conditions.

Say you draw twice from the box at random with replacement. To find the sum of draws, you simply add the first number you drew to the second number you drew. For instance, if first you draw a 4 and second you draw a 6, your sum of draws would be 4+6=10″>4+6=10. You could also first draw a 4 and then draw 4 again.

In this case your sum of draws would be 4+4=8">4+4=8. Your sum of draws is, therefore, subject to a force known as chance variation.

This example can be seen in practical terms when imagining a turn of Monopoly. A player rolls a pair of dice, adds the two numbers on the die, and moves his or her piece that many squares. Rolling a die is the same as drawing a ticket from a box containing six options.

Sum of Draws In Practice

Rolling a die is the same as drawing a ticket from a box containing six options.

To better see the affects of chance variation, let us take 25 draws from the box. These draws result in the following values:

3 2 4 6 3 3 5 4 4 1 3 6 4 1 3 4 1 5 5 5 2 2 2 5 6

The sum of these 25 draws is 89. Obviously this sum would have been different had the draws been different.

5.3.4: Making a Box Model

A box plot (also called a box-and-whisker diagram) is a simple visual representation of key features of a univariate sample.

Learning Objectives

Produce a box plot that is representative of a data set.

Key Takeaways

Key Points

- Our ultimate goal in statistics is not to summarize the data, it is to fully understand their complex relationships.
- A well designed statistical graphic helps us explore, and perhaps understand, these relationships.
- A common extension of the box model is the 'box-and-whisker' plot, which adds vertical lines extending from the top and bottom of the plot to, for example, the maximum and minimum values.

Key Terms

regression
: An analytic method to measure the association of one or more independent variables with a dependent variable.

box-and-whisker plot
: a convenient way of graphically depicting groups of numerical data through their quartiles

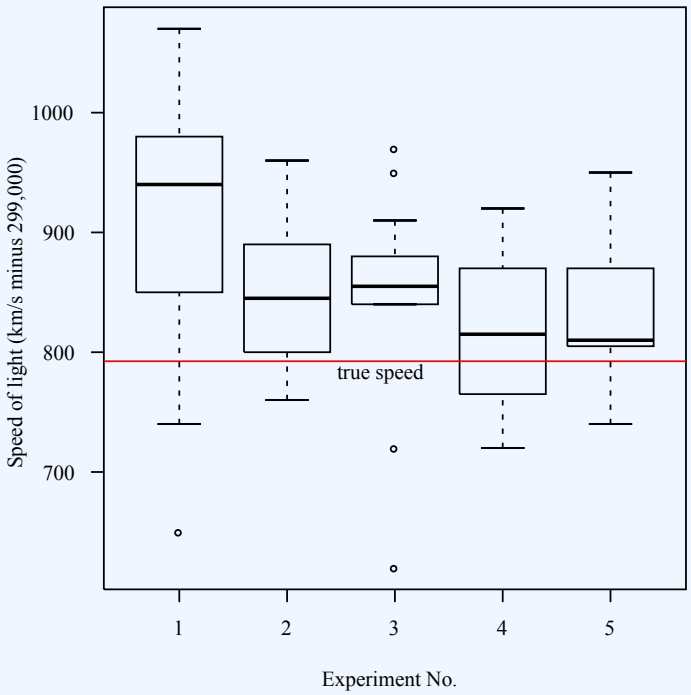
A single statistic tells only part of a dataset's story. The mean is one perspective; the median yet another. When we explore relationships between multiple variables, even more statistics arise, such as the coefficient estimates

in a regression model or the Cochran-Maentel-Haenszel test statistic in partial contingency tables. A multitude of statistics are available to summarize and test data.

Our ultimate goal in statistics is not to summarize the data, it is to fully understand their complex relationships. A well designed statistical graphic helps us explore, and perhaps understand, these relationships. A box plot (also called a box and whisker diagram) is a simple visual representation of key features of a univariate sample.

The box lies on a vertical axis in the range of the sample. Typically, a top to the box is placed at the first quartile, the bottom at the third quartile. The width of the box is arbitrary, as there is no x-axis. In between the top and bottom of the box is some representation of central tendency. A common version is to place a horizontal line at the median, dividing the box into two. Additionally, a star or asterisk is placed at the mean value, centered in the box in the horizontal direction.

Another common extension of the box model is the 'box-and-whisker' plot , which adds vertical lines extending from the top and bottom of the plot to, for example, the maximum and minimum values. Alternatively, the whiskers could extend to the 2.5 and 97.5 percentiles. Finally, it is common in the box-and-whisker plot to show outliers (however defined) with asterisks at the individual values beyond the ends of the whiskers.



Box-and-Whisker Plot

Box plot of data from the Michelson-Morley Experiment, which attempted to detect the relative motion of matter through the stationary luminiferous aether.

Attributions

- What Does the Law of Averages Say?
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Law of averages."
    http://en.wikipedia.org/wiki/Law_of_averages.
    Wikipedia
    CC BY-SA 3.0.

- "expected value."
  http://en.wiktionary.org/wiki/expected_value.
  Wiktionary
  CC BY-SA 3.0.
- "Law of large numbers."
  http://en.wikipedia.org/wiki/Law_of_large_numbers.
  Wikipedia
  CC BY-SA 3.0.
- "Law of large numbers."
  http://en.wikipedia.org/wiki/Law_of_large_numbers.
  Wikipedia
  GNU FDL.

- Chance Processes
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Law (stochastic processes)."
    http://en.wikipedia.org/wiki/Law_(stochastic_processes).
    Wikipedia
    CC BY-SA 3.0.
  - "Stochastic process."
    http://en.wikipedia.org/wiki/Stochastic_process.
    Wikipedia
    CC BY-SA 3.0.
  - "stochastic."
    http://en.wiktionary.org/wiki/stochastic.
    Wiktionary
    CC BY-SA 3.0.
  - "random variable."
    http://en.wiktionary.org/wiki/random_variable.
    Wiktionary
    CC BY-SA 3.0.
  - "Random walk."
    http://en.wikipedia.org/wiki/Random_walk.
    Wikipedia
    CC BY-SA 3.0.

- "random walk."
  http://en.wiktionary.org/wiki/random_walk.
  Wiktionary
  CC BY-SA 3.0.
- "Random Walk example."
  http://commons.wikimedia.org/wiki/File:Random_Walk_example.svg.
  Wikimedia
  GNU FDL 1.2.

- The Sum of Draws
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Sampling (statistics)."
    http://en.wikipedia.org/wiki/Sampling_(statistics).
    Wikipedia
    CC BY-SA 3.0.
  - "All sizes | Monopoly | Flickr – Photo Sharing!."
    http://www.flickr.com/photos/elpadawan/8480394254/sizes/z/in/photostream/.
    Flickr
    CC BY-SA.
- Making a Box Model
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "box-and-whisker plot."
    http://en.wikipedia.org/wiki/box-and-whisker%20plot.
    Wikipedia
    CC BY-SA 3.0.
  - "regression."
    http://en.wiktionary.org/wiki/regression.
    Wiktionary
    CC BY-SA 3.0.
  - "Statistics/Displaying Data/Box Plots."
    http://en.wikibooks.org/wiki/Statistics/Displaying_Data/Box_Plots.
    Wikibooks

CC BY-SA 3.0.

- "Statistics/Displaying Data." http://en.wikibooks.org/wiki/Statistics/Displaying_Data. Wikibooks CC BY-SA 3.0.
- "Michelsonmorley-boxplot." http://commons.wikimedia.org/wiki/File:Michelsonmorley-boxplot.svg. Wikimedia Public domain.
- "Michelsonmorley-boxplot." http://en.wikipedia.org/wiki/File:Michelsonmorley-boxplot.svg. Wikipedia Public domain.

5.4 FURTHER CONSIDERATIONS FOR DATA

5.4: Further Considerations for Data

5.4.1: The Sample Average

The sample average/mean can be calculated taking the sum of every piece of data and dividing that sum by the total number of data points.

Learning Objectives

Distinguish the sample mean from the population mean.

Key Takeaways

Key Points

- The sample mean makes a good estimator of the population mean, as its expected value is equal to the population mean. The law of large numbers dictates that the larger the size of the sample, the more likely it is that the sample mean will be close to the population mean.
- The sample mean of a population is a random variable, not a constant, and consequently it will have its own distribution.

- The mean is the arithmetic average of a set of values, or distribution; however, for skewed distributions, the mean is not necessarily the same as the middle value (median), or the most likely (mode).

Key Terms

random variable
: a quantity whose value is random and to which a probability distribution is assigned, such as the possible outcome of a roll of a die

finite
: limited, constrained by bounds, having an end

Sample Average vs. Population Average

The sample average (also called the sample mean) is often referred to as the arithmetic mean of a sample, or simply, x¯">¯x (pronounced “x bar”). The mean of a population is denoted μ">μ, known as the population mean. The sample mean makes a good estimator of the population mean, as its expected value is equal to the population mean. The sample mean of a population is a random variable, not a constant, and consequently it will have its own distribution. For a random sample of n">n observations from a normally distributed population, the sample mean distribution is:

$\bar{x} \sim N\left\{\mu, \frac{\sigma^2}{n}\right\}$

For a finite population, the population mean of a property is equal to the arithmetic mean of the given property while considering every member of the population. For example, the population mean height is equal to the sum of the heights of every individual divided by the total number of individuals.The sample mean may differ from the population mean, especially for small samples. The law of large numbers dictates that the larger the size of the sample, the more likely it is that the sample mean will be close to the population mean.

Calculation of the Sample Mean

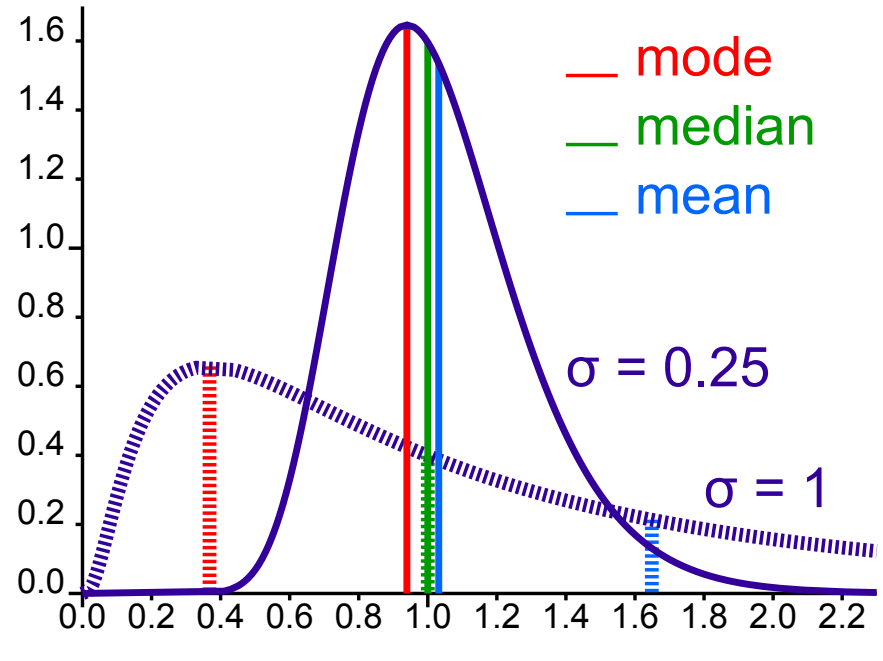
The arithmetic mean is the “standard” average, often simply called the “mean”. It can be calculated taking the sum of every piece of data and dividing that sum by the total number of data points:

x¯=1n⋅∑i=1nxi">$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$

For example, the arithmetic mean of five values: 4, 36, 45, 50, 75 is:

4+36+45+50+755=2105=42">$\frac{4+36+45+50+75}{5} = \frac{210}{5} = 42$

The mean may often be confused with the median, mode or range. The mean is the arithmetic average of a set of values, or distribution; however, for skewed distributions, the mean is not necessarily the same as the middle value (median), or the most likely (mode). For example, mean income is skewed upwards by a small number of people with very large incomes, so that the majority have an income lower than the mean. By contrast, the median income is the level at which half the population is below and half is above. The mode income is the most likely income, and favors the larger number of people with lower incomes. The median or mode are often more intuitive measures of such data.



Measures of Central Tendency: This graph shows where the mean, median, and mode fall in two different distributions (one is slightly skewed left and one is highly skewed right).

5.4.2: Which Standard Deviation (SE)?

Although they are often used interchangeably, the standard deviation and the standard error are slightly different.

Learning Objective

Differentiate between standard deviation and standard error.

Key Takeaways

Key Points

- Standard error is an estimate of how close to the population mean your sample mean is likely to be, whereas standard deviation is the degree to which individuals within the sample differ from the sample mean.
- Standard deviation (represented by the symbol sigma, **σ**) shows how much variation or dispersion exists from the average (mean), or expected value.
- The standard error is the standard deviation of the sampling distribution of a statistic, such as the mean.
- Standard error should decrease with larger sample sizes, as the estimate of the population mean improves. Standard deviation will be unaffected by sample size.

Key Terms

standard error
: A measure of how spread out data values are around the mean, defined as the square root of the variance.

central limit theorem
: The theorem that states: If the sum of independent identically distributed random variables has a finite variance, then it will be (approximately) normally distributed.

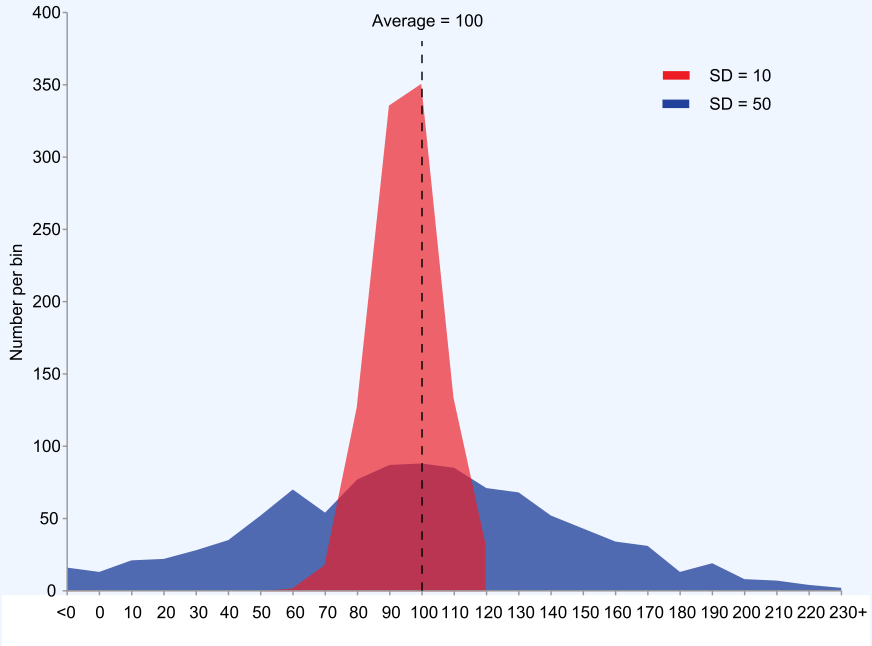
sample mean

the mean of a sample of random variables taken from the entire population of those variables

The standard error is the standard deviation of the sampling distribution of a statistic. The term may also be used to refer to an estimate of that standard deviation, derived from a particular sample used to compute the estimate.

For example, the sample mean is the usual estimator of a population mean. However, different samples drawn from that same population would in general have different values of the sample mean. The standard error of the mean (i.e., of using the sample mean as a method of estimating the population mean) is the standard deviation of those sample means over all possible samples (of a given size) drawn from the population. Secondly, the standard error of the mean can refer to an estimate of that standard deviation, computed from the sample of data being analyzed at the time.

In scientific and technical literature, experimental data is often summarized either using the mean and standard deviation or the mean with the standard error. This often leads to confusion about their interchangeability. However, the mean and standard deviation are descriptive statistics, whereas the mean and standard error describes bounds on a random sampling process. Despite the small difference in equations for the standard deviation and the standard error, this small difference changes the meaning of what is being reported from a description of the variation in measurements to a probabilistic statement about how the number of samples will provide a better bound on estimates of the population mean, in light of the central limit theorem. Put simply, standard error is an estimate of how close to the population mean your sample mean is likely to be, whereas standard deviation is the degree to which individuals within the sample differ from the sample mean. Standard error should decrease with larger sample sizes, as the estimate of the population mean improves. Standard deviation will be unaffected by sample size.



Standard Deviation

This is an example of two sample populations with the same mean and different standard deviations. The red population has mean 100 and SD 10; the blue population has mean 100 and SD 50.

5.4.3: Estimating the Accuracy of an Average

The standard error of the mean is the standard deviation of the sample mean's estimate of a population mean.

Learning Objective

Evaluate the accuracy of an average by finding the standard error of the mean.

Key Takeaways

Key Points

- Any measurement is subject to error by chance, which means that if the measurement was taken again it could possibly show a different value.
- In general terms, the standard error is the standard deviation of the sampling distribution of a statistic.
- The standard error of the mean is usually estimated by the sample estimate of the population standard deviation (sample standard deviation) divided by the square root of the sample size.
- The standard error and standard deviation of small samples tend to systematically underestimate the population standard error and deviations because the standard error of the mean is a biased estimator of the population standard error.
- The standard error is an estimate of how close the population mean will be to the sample mean, whereas standard deviation is the degree to which individuals within the sample differ from the sample mean.

Key Terms

standard error
: A measure of how spread out data values are around the mean, defined as the square root of the variance.

confidence interval
: A type of interval estimate of a population parameter used to indicate the reliability of an estimate.

central limit theorem
: The theorem that states: If the sum of independent identically distributed random variables has a finite variance, then it will be (approximately) normally distributed.

Any measurement is subject to error by chance, meaning that if the measurement was taken again, it could possibly show a different value. We calculate the standard deviation in order to estimate the chance error for

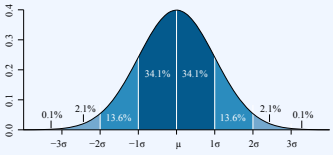
a single measurement. Taken further, we can calculate the chance error of the sample mean to estimate its accuracy in relation to the overall population mean.

Standard Error

In general terms, the standard error is the standard deviation of the sampling distribution of a statistic. The term may also be used to refer to an estimate of that standard deviation, derived from a particular sample used to compute the estimate. For example, the sample mean is the standard estimator of a population mean. However, different samples drawn from that same population would, in general, have different values of the sample mean.

The standard error of the mean (i.e., standard error of using the sample mean as a method of estimating the population mean) is the standard deviation of those sample means over all possible samples (of a given size) drawn from the population. Secondly, the standard error of the mean can refer to an estimate of that standard deviation, computed from the sample of data being analyzed at the time.

In practical applications, the true value of the standard deviation (of the error) is usually unknown. As a result, the term standard error is often used to refer to an estimate of this unknown quantity. In such cases, it is important to clarify one's calculations, and take proper account of the fact that the standard error is only an estimate.



Standard Deviation as Standard Error

For a value that is sampled with an unbiased normally distributed error, the graph depicts the proportion of samples that would fall between 0, 1, 2, and 3 standard deviations above and below the actual value.

Standard Error of the Mean

As mentioned, the standard error of the mean (SEM) is the standard deviation of the sample-mean's estimate of a population mean. It can also be viewed as the standard deviation of the error in the sample mean relative to the true mean, since the sample mean is an unbiased estimator. Generally, the SEM is the sample estimate of the population standard deviation (sample standard deviation) divided by the square root of the sample size:

SEx¯=sn">$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

Where s is the sample standard deviation (i.e., the sample-based estimate of the standard deviation of the population), and n">n is the size (number of observations) of the sample. This estimate may be compared with the formula for the true standard deviation of the sample mean:

SDx¯=σn">$SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Where σ">σ is the standard deviation of the population. Note that the standard error and the standard deviation of small samples tend to systematically underestimate the population standard error and

deviations because the standard error of the mean is a biased estimator of the population standard error. For example, with n=2″>n=2, the underestimate is about 25%, but for n=6″>n=6, the underestimate is only 5%. As a practical result, decreasing the uncertainty in a mean value estimate by a factor of two requires acquiring four times as many observations in the sample. Decreasing standard error by a factor of ten requires a hundred times as many observations.

Assumptions and Usage

If the data are assumed to be normally distributed, quantiles of the normal distribution and the sample mean and standard error can be used to calculate approximate confidence intervals for the mean. In particular, the standard error of a sample statistic (such as sample mean) is the estimated standard deviation of the error in the process by which it was generated. In other words, it is the standard deviation of the sampling distribution of the sample statistic.

Standard errors provide simple measures of uncertainty in a value and are often used for the following reasons:

- If the standard error of several individual quantities is known, then the standard error of some function of the quantities can be easily calculated in many cases.
- Where the probability distribution of the value is known, it can be used to calculate a good approximation to an exact confidence interval.
- Where the probability distribution is unknown, relationships of inequality can be used to calculate a conservative confidence interval.
- As the sample size tends to infinity, the central limit theorem guarantees that the sampling distribution of the mean is asymptotically normal.

5.4.4: Chance Models

A stochastic model is used to estimate probability distributions of potential outcomes by allowing for random variation in one or more inputs over time.

Learning Objective

Support the idea that stochastic modeling provides a better representation of real life by building randomness into a simulation.

Key Takeaways

Key Points

- Accurately determining the standard error of the mean depends on the presence of chance.
- Stochastic modeling builds volatility and variability (randomness) into a simulation and, therefore, provides a better representation of real life from more angles.
- Stochastic models help to assess the interactions between variables and are useful tools to numerically evaluate quantities.

Key Terms

Monte Carlo simulation
: a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results–i.e., by running simulations many times over in order to calculate those same probabilities

stochastic
: random; randomly determined

The calculation of the standard error of the mean for repeated measurements is easily carried out on a data set; however, this method for determining error is only viable when the data varies as if drawing a name out of a

hat. In other words, the data should be completely random, and should not show a trend or pattern over time. Therefore, accurately determining the standard error of the mean depends on the presence of chance.

Stochastic Modeling

"Stochastic" means being or having a random variable. A stochastic model is a tool for estimating probability distributions of potential outcomes by allowing for random variation in one or more inputs over time. The random variation is usually based on fluctuations observed in historical data for a selected period using standard time-series techniques. Distributions of potential outcomes are derived from a large number of simulations (stochastic projections) which reflect the random variation in the input(s).

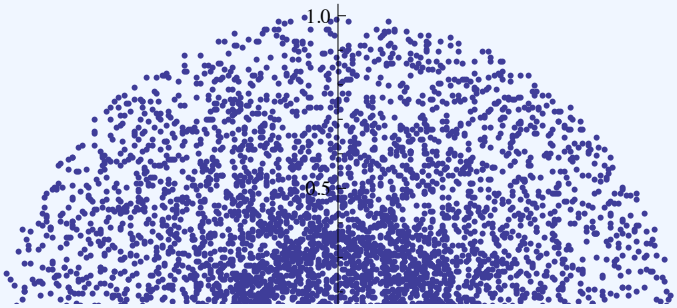
In order to understand stochastic modeling, consider the example of an insurance company projecting potential claims. Like any other company, an insurer has to show that its assets exceed its liabilities to be solvent. In the insurance industry, however, assets and liabilities are not known entities. They depend on how many policies result in claims, inflation from now until the claim, investment returns during that period, and so on. So the valuation of an insurer involves a set of projections, looking at what is expected to happen, and thus coming up with the best estimate for assets and liabilities.

A stochastic model, in the case of the insurance company, would be to set up a projection model which looks at a single policy, an entire portfolio, or an entire company. But rather than setting investment returns according to their most likely estimate, for example, the model uses random variations to look at what investment conditions might be like. Based on a set of random outcomes, the experience of the policy/portfolio/company is projected, and the outcome is noted. This is done again with a new set of random variables. In fact, this process is repeated thousands of times.

At the end, a distribution of outcomes is available which shows not only the most likely estimate but what ranges are reasonable, too. The most likely estimate is given by the center of mass of the distribution curve (formally known as the probability density function), which is typically also the mode of the curve. Stochastic modeling builds volatility and variability (randomness) into a simulation and, therefore, provides a better representation of real life from more angles.

Numerical Evaluations of Quantities

Stochastic models help to assess the interactions between variables and are useful tools to numerically evaluate quantities, as they are usually implemented using Monte Carlo simulation techniques .



Monte Carlo Simulation

Monte Carlo simulation (10,000 points) of the distribution of the sample mean of a circular normal distribution for 3 measurements.

While there is an advantage here, in estimating quantities that would otherwise be difficult to obtain using analytical methods, a disadvantage is that such methods are limited by computing resources as well as simulation error. Below are some examples:

Means

Using statistical notation, it is a well-known result that the mean of a function, f">f, of a random variable, x">x, is not necessarily the function of the mean of x">x. For example, in finance, applying the best estimate (defined as the mean) of investment returns to discount a set of cash flows will not necessarily give the same result as assessing the best estimate to the discounted cash flows. A stochastic model would be able to assess this latter quantity with simulations.

Percentiles

This idea is seen again when one considers percentiles. When assessing risks at specific percentiles, the factors that contribute to these levels are rarely at these percentiles themselves. Stochastic models can be simulated to assess the percentiles of the aggregated distributions.

Truncations and Censors

Truncating and censoring of data can also be estimated using stochastic models. For instance, applying a non-

proportional reinsurance layer to the best estimate losses will not necessarily give us the best estimate of the losses after the reinsurance layer. In a simulated stochastic model, the simulated losses can be made to "pass through" the layer and the resulting losses are assessed appropriately.

5.4.5: The Gauss Model

The normal (Gaussian) distribution is a commonly used distribution that can be used to display the data in many real life scenarios.

Learning Objective

Explain the importance of the Gauss model in terms of the central limit theorem.

Key Takeaways

Key Points

- If μ=0">μ=0 and σ=1">σ=1, the distribution is called the standard normal distribution or the unit normal distribution, and a random variable with that distribution is a standard normal deviate.
- It is symmetric around the point x=μ">x=μ, which is at the same time the mode, the median and the mean of the distribution.
- The Gaussian distribution is sometimes informally called the bell curve. However, there are many other distributions that are bell-shaped as well.
- About 68% of values drawn from a normal distribution are within one standard deviation σ away from the mean; about 95% of the values lie within two standard deviations; and about

99.7% are within three standard deviations. This fact is known as the 68-95-99.7 (empirical) rule, or the 3-sigma rule.

Key Term

central limit theorem

The theorem that states: If the sum of independent identically distributed random variables has a finite variance, then it will be (approximately) normally distributed.

The Normal (Gaussian) Distribution

In probability theory, the normal (or Gaussian) distribution is a continuous probability distribution, defined by the formula:

f(x)=1σ2πe−(x−μ)22σ2">$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

The parameter μ">μ in this formula is the mean or expectation of the distribution (and also its median and mode). The parameter σ">σ is its standard deviation; its variance is therefore σ2">σ^2. A random variable with a Gaussian distribution is said to be normally distributed and is called a normal deviate.

If μ=0">μ=0 and σ=1">σ=1, the distribution is called the standard normal distribution or the unit normal distribution, and a random variable with that distribution is a standard normal deviate.

Importance of the Normal Distribution

Normal distributions are extremely important in statistics, and are often used in the natural and social sciences for real-valued random variables whose distributions are not known. One reason for their popularity is the central limit theorem, which states that, under mild conditions, the mean of a large number of random variables independently drawn from the same distribution is distributed approximately normally, irrespective of the form of the original distribution. Thus, physical quantities that are expected to be the sum of many independent processes (such as measurement errors) often have a distribution very close to normal. Another reason is that a large number of results and methods (such as propagation of uncertainty and least squares parameter fitting) can be derived analytically, in explicit form, when the relevant variables are normally distributed.

The normal distribution is symmetric about its mean, and is non-zero over the entire real line. As such it may not be a suitable model for variables that are inherently positive or strongly skewed, such as the weight of

a person or the price of a share. Such variables may be better described by other distributions, such as the log-normal distribution or the Pareto distribution.

The normal distribution is also practically zero once the value x">x lies more than a few standard deviations away from the mean. Therefore, it may not be appropriate when one expects a significant fraction of outliers, values that lie many standard deviations away from the mean. Least-squares and other statistical inference methods which are optimal for normally distributed variables often become highly unreliable. In those cases, one assumes a more heavy-tailed distribution, and the appropriate robust statistical inference methods.

The Gaussian distribution is sometimes informally called the bell curve. However, there are many other distributions that are bell-shaped (such as Cauchy's, Student's, and logistic). The terms Gaussian function and Gaussian bell curve are also ambiguous since they sometimes refer to multiples of the normal distribution whose integral is not 1; that is, for arbitrary positive constants a">a, b">b and c">c.

Properties of the Normal Distribution

The normal distribution f(x)">f(x), with any mean μ">μ and any positive deviation σ">σ, has the following properties:

- It is symmetric around the point x=μ">x=μ, which is at the same time the mode, the median and the mean of the distribution.
- It is unimodal: its first derivative is positive for x<μ">x<μ, negative for x>μ">x>μ, and zero only at x=μ">x=μ.
- It has two inflection points (where the second derivative of f">f is zero), located one standard deviation away from the mean, namely at x=μ−σ">x=μ−σ and x=μ+σ">x=μ+σ.
- About 68% of values drawn from a normal distribution are within one standard deviation σ">σ away from the mean; about 95% of the values lie within two standard deviations; and about 99.7% are within three standard deviations. This fact is known as the 68-95-99.7 (empirical) rule, or the 3-sigma rule.

Notation

The normal distribution is also often denoted by N(μ,σ2)">$N(\mu,\sigma^2)$. Thus when a random variable x">x is distributed normally with mean μ">μ and variance σ2">σ^2, we write X∼N(μ,σ2)">$X\sim N(\mu,\sigma^2)$.

5.4.6: Comparing Two Sample Averages

Student's t-test is used in order to compare two independent sample means.

Learning Objective

Contrast two sample means by standardizing their difference to find a t-score test statistic.

Key Takeaways

Key Points

- Very different sample means can occur by chance if there is great variation among the individual samples.
- In order to account for the variation, we take the difference of the sample means and divide by the standard error in order to standardize the difference, resulting in a t-score test statistic.
- The independent samples t-test is used when two separate sets of independent and identically distributed samples are obtained, one from each of the two populations being compared.
- Paired samples t-tests typically consist of a sample of matched pairs of similar units or one group of units that has been tested twice (a "repeated measures" t-test).
- An overlapping samples t-test is used when there are paired samples with data missing in one or the other samples.

Key Terms

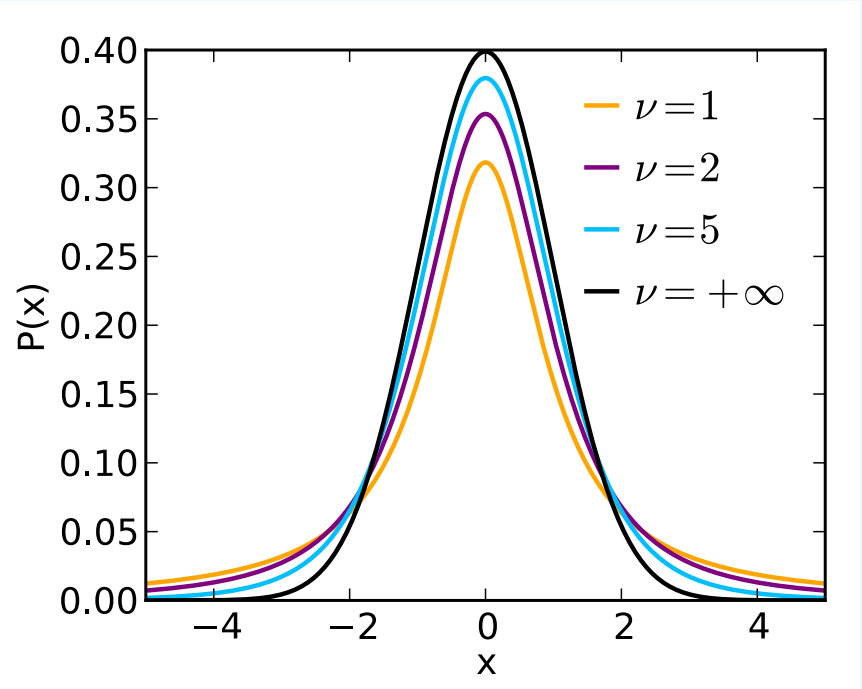
null hypothesis
: A hypothesis set up to be refuted in order to support an alternative hypothesis; presumed true until statistical evidence in the form of a hypothesis test indicates otherwise.

Student's t-distribution
: A distribution that arises when the population standard deviation is unknown and has to be estimated from the data; originally derived by William Sealy Gosset (who wrote under the pseudonym "Student").

The comparison of two sample means is very common. The difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means, X1¯−X2¯">$\bar{X}_1 - \bar{X}_2$,
and divide by the standard error in order to standardize the difference. The result is a t-score test statistic.

t-Test for Two Means

Although the t-test will be explained in great detail later in this textbook, it is important for the reader to have a basic understanding of its function in regard to comparing two sample means. A t-test is any statistical hypothesis test in which the test statistic follows Student's t distribution, as shown in , if the null hypothesis is supported. It can be used to determine if two sets of data are significantly different from each other.



Student t Distribution

This is a plot of the Student t Distribution for various degrees of freedom.

In the t-test comparing the means of two independent samples, the following assumptions should be met:

1. Each of the two populations being compared should follow a normal distribution.
2. If using Student's original definition of the t-test, the two populations being compared should have the same variance. If the sample sizes in the two groups being compared are equal, Student's original t-test is highly robust to the presence of unequal variances.
3. The data used to carry out the test should be sampled independently from the populations being compared. This is, in general, not testable from the data, but if the data are known to be dependently sampled (i.e., if they were sampled in clusters), then the classical t-tests discussed here may give misleading results.

Two-sample t-tests for a difference in mean involve independent samples, paired samples and overlapping samples. The independent samples t-test is used when two separate sets of independent and identically distributed samples are obtained, one from each of the two populations being compared. For example, suppose we are evaluating the effects of a medical treatment. We enroll 100 subjects into our study, then randomize 50

subjects to the treatment group and 50 subjects to the control group. In this case, we have two independent samples and would use the unpaired form of the t-test.

Paired sample t-tests typically consist of a sample of matched pairs of similar units or one group of units that has been tested twice (a "repeated measures" t-test). A typical example of the repeated measures t-test would be where subjects are tested prior to a treatment (say, for high blood pressure) and the same subjects are tested again after treatment with a blood-pressure lowering medication. By comparing the same patient's numbers before and after treatment, we are effectively using each patient as their own control.

An overlapping sample t-test is used when there are paired samples with data missing in one or the other samples. These tests are widely used in commercial survey research (e.g., by polling companies) and are available in many standard crosstab software packages.

5.4.7: Odds Ratios

The odds of an outcome is the ratio of the expected number of times the event will occur to the expected number of times the event will not occur.

Learning Objective

Define the odds ratio and demonstrate its computation.

Key Takeaways

Key Points

- The odds ratio is one way to quantify how strongly having or not having the property A">A is associated with having or not having the property B">B in a population.

- The odds ratio is a measure of effect size, describing the strength of association or non-independence between two binary data values.
- To compute the odds ratio, we 1) compute the odds that an individual in the population has A">A given that he or she has B">B, 2) compute the odds that an individual in the population has A">A given that he or she does not have B">B and 3) divide the first odds by the second odds.
- If the odds ratio is greater than one, then having A">A is associated with having B">B in the sense that having B">B raises the odds of having A">A.

Key Terms

logarithm
: for a number x, the power to which a given base number must be raised in order to obtain x

odds
: the ratio of the probabilities of an event happening to that of it not happening

The odds of an outcome is the ratio of the expected number of times the event will occur to the expected number of times the event will not occur. Put simply, the odds are the ratio of the probability of an event occurring to the probability of no event.

An odds ratio is the ratio of two odds. Imagine each individual in a population either does or does not have a property A">A, and also either does or does not have a property B">B. For example, A">A might be "has high blood pressure," and B">B might be "drinks more than one alcoholic drink a day." The odds ratio is one way to quantify how strongly having or not having the property A">A is associated with having or not having the property B">B in a population. In order to compute the odds ratio, one follows three steps:

1. Compute the odds that an individual in the population has A">A given that he or she has B">B (probability of A">A given B">B divided by the probability of not-A">A given B">B).
2. Compute the odds that an individual in the population has A">A given that he or she does not have B">B.
3. Divide the first odds by the second odds to obtain the odds ratio.

If the odds ratio is greater than one, then having A">A

is associated with having B">B in the sense that having B">B raises (relative to not having B">B) the odds of

having A″>A. Note that this is not enough to establish that B″>B is a contributing cause of A″>A. It could be that the association is due to a third property, C″>C, which is a contributing cause of both A″>A and B″>B.

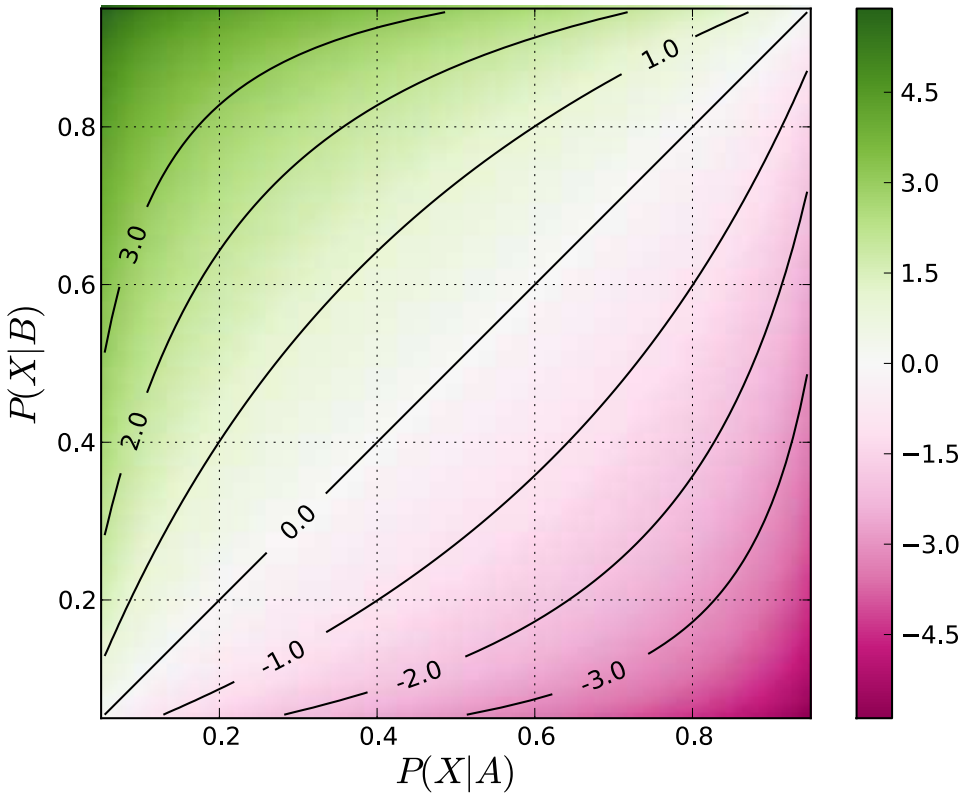
In more technical language, the odds ratio is a measure of effect size, describing the strength of association or non-independence between two binary data values. It is used as a descriptive statistic and plays an important role in logistic regression.

Example

Suppose that in a sample of 100">100 men 90">90 drank wine in the previous week, while in a sample of 100">100 women only 20">20 drank wine in the same period. The odds of a man drinking wine are 90">90 to 10">10 (or 9:1">9:1) while the odds of a woman drinking wine are only 20">20 to 80">80 (or 1:4=0.25:1">1:4=0.25:1). The odds ratio is thus 90.25">90.25 (or 36">36) showing that men are much more likely to drink wine than women. The detailed calculation is:

0.9/0.10.2/0.8=0.9⋅0.80.1⋅0.2=0.720.02=36">$\frac{0.9/0.1}{0.2/0.8} = \frac{0.9 \cdot 0.8}{0.1 \cdot 0.2} = \frac{0.72}{0.02} = 36$

This example also shows how odds ratios are sometimes sensitive in stating relative positions. In this sample men are 9020=4.5">9020=4.5 times more likely to have drunk wine than women, but have 36">36 times the odds. The logarithm of the odds ratio—the difference of the logits of the probabilities—tempers this effect and also makes the measure symmetric with respect to the ordering of groups. For example, using natural logarithms, an odds ratio of 361">361 maps to 3.584">3.584, and an odds ratio of 136">136 maps to −3.584">−3.584.



Odds Ratios: A graph showing how the log odds ratio relates to the underlying probabilities of the outcome X">X occurring in two groups, denoted A">A and B">B. The log odds ratio shown here is based on the odds for the event occurring in group B">B relative to the odds for the event occurring in group A">A. Thus, when the probability of X">X occurring in group B">B is greater than the probability of X">X occurring in group A">A, the odds ratio is greater than 1">1, and the log odds ratio is greater than 0">0.

5.4.8: When Does the Z-Test Apply?

A z-test is a test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution.

Learning Objective

Identify how sample size contributes to the appropriateness and accuracy of a z-test

Key Takeaways

Key Points

- The term z">z-test is often used to refer specifically to the one- sample location test comparing the mean of a set of measurements to a given constant.
- To calculate the standardized statistic $Z = \frac{X - \mu_0}{s}$Z=X−μ0s">, we need to either know or have an approximate value for σ2">σ^2 z">z σ^2, from which we can calculate s2=σ2n">$s^2 = \frac{\sigma^2}{n}$.
- For a z">z-test to be applicable, nuisance parameters should be known, or estimated with high accuracy.
- For a z">z-test to be applicable, the test statistic should follow a normal distribution.

Key Terms

null hypothesis
: A hypothesis set up to be refuted in order to support an alternative hypothesis; presumed true until statistical evidence in the form of a hypothesis test indicates otherwise.

nuisance parameters
: any parameter that is not of immediate interest but which must be accounted for in the analysis of those parameters which are of interest; the classic example of a nuisance parameter is the variance σ^2, of a normal distribution, when the mean, μ, is of primary interest

Z-test

A Z">Z-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Because of the central limit theorem, many test statistics are approximately normally distributed for large samples. For each significance level, the Z">Z-test has a single critical value (for example, 1.96">1.96 for 5% two tailed) which makes it more convenient than the Student's t-test which has separate critical values for each sample size. Therefore, many statistical tests can be conveniently performed as approximate Z">Z-tests if the sample size is large or the population variance known. If the population variance is unknown (and therefore has to be estimated from the sample itself) and the sample size is not large (n<30">n<30), the Student t">t-test may be more appropriate.

If T">T is a statistic that is approximately normally distributed under the null hypothesis, the next step in performing a Z">Z-test is to estimate the expected value θ">θ of T">T under the null hypothesis, and then obtain an estimate s">s of the standard deviation of T">T. We then calculate the standard score Z=(T−θ)s">$Z = \frac{(T-\Theta)}{s}$, from which one-tailed and two-tailed p">p-values can be calculated as φ(−Z)">φ(−Z) (for upper-tailed tests), φ(Z)">φ(Z) (for lower-tailed tests) and 2φ(|−Z|)">2φ(|−Z|) (for two-tailed tests) where φ">φ is the standard normal cumulative distribution function.

Use in Location Testing

The term Z">Z-test is often used to refer specifically to the one-sample location test comparing the mean of a set of measurements to a given constant. If the observed data X1,⋯,Xn">$X_1,\cdots,X_n$ are uncorrelated, have a common mean μ">μ, and have a common variance σ2">σ^2, then the sample average $\bar{X}$X¯"> has mean μ">μ and variance σ2n">$\frac{\sigma^2}{n}$. If our null hypothesis is that the mean value of the population is a given number μ0">μ_0, we can use X¯−μ0">$\bar{X} - \mu_0$ as a test-statistic, rejecting the null hypothesis if X¯−μ0">X¯−μ0">

$X - \mu_0[/latex < /span >< /span >< /span >] < /span >< /span >< /span >< /span >< /span >< /span >< /span >< /span > islarge.$

To calculate the standardized statistic ¡span id=”MathJax-Element-148-Frame” class=”mjx-chtml MathJax$_C HTML” tabindex = ”0” role = ”presentation” data-mathml = ” < mathxmlns = ”http : //www.w3.org/1998/Math/MathML” >< mtext > Z < /mtext >< mo >=< /mo >< mfrac >< mrow >< mostretchy = ”false” > (< /mo >< mtext > X < /mtext >< mo > x2212; < /mo >< msub >< mrowclass = ”MJX-TeXAtom-ORD” >< mo > x3BC; < /mo >< /mrow >< mn > 0 < /mn >< /msub >< mostretchy = ”false” >) < /mo >< /mrow >< mtext > s < /mtext >< /mfrac >< /math > ” >< spanid = ”MJXc - Node - 798” class = ”mjx - math” aria - hidden = ”true” >< spanid = ”MJXc - Node - 799” class = ”mjx - mrow” >< spanid = ”MJXc - Node - 800” class = ”mjx - mtext” >< spanclass = ”mjx - charMJXc - TeX - main - R” > [latex]Z = \frac{(X-\mu_0)}{s}$, we need to either know or have an approximate value for σ2”>σ^2, from which we can calculate s2=σ2n”>$s^2 = \frac{a^2}{n}$. In some applications, σ2”>σ^2 is known, but this is uncommon. If the sample size is moderate or large, we can substitute the sample variance for σ2”>σ2, giving a plug-in test. The resulting test will not be an exact Z”>Z-test since the uncertainty in the sample variance is not accounted for—however, it will be a good approximation unless the sample size is small. A t”>t-test can be used to account for the uncertainty in the sample variance when the sample size is small and the data are exactly normal. There is no universal constant at which the sample size is generally considered large enough to justify use of the plug-in test. Typical rules of thumb range from 20 to 50 samples. For larger sample sizes, the t”>t-test procedure gives almost identical p”>p-values as the Z”>Z-test procedure. The following formula converts a random variable X”>X to the standard Z”>Z:

Z=X−μσ”>$Z = \frac{X-\mu}{\sigma}$

Conditions

For the Z-test to be applicable, certain conditions must be met:

- Nuisance parameters should be known, or estimated with high accuracy (an example of a nuisance parameter would be the standard deviation in a one-sample location test). Z-tests focus on a single parameter, and treat all other unknown parameters as being fixed at their true values. In practice, due to Slutsky’s theorem, “plugging in” consistent estimates of nuisance parameters can be justified. However if the sample size is not large enough for these estimates to be reasonably accurate, the Z-test may not perform well.
- The test statistic should follow a normal distribution. Generally, one appeals to the central limit theorem to justify assuming that a test statistic varies normally. There is a great deal of statistical research on the question of when a test statistic varies approximately normally. If the variation of the test statistic is strongly non-normal, a Z-test should not be used.

Attributions

- The Sample Average
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "random variable."
    http://en.wiktionary.org/wiki/random_variable.
    Wiktionary
    CC BY-SA 3.0.
  - "Sample mean and sample covariance."
    http://en.wikipedia.org/wiki/Sample_mean_and_sample_covariance.
    Wikipedia
    CC BY-SA 3.0.
  - "Mean."
    http://en.wikipedia.org/wiki/Mean%23Population_and_sample_means.
    Wikipedia
    CC BY-SA 3.0.
  - "finite."
    http://en.wiktionary.org/wiki/finite.
    Wiktionary
    CC BY-SA 3.0.
  - "Mean."
    http://en.wikipedia.org/wiki/Mean%23Population_and_sample_means.
    Wikipedia
    GNU FDL.
- Which Standard Deviation (SE)?
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Standard deviation."
    http://en.wikipedia.org/wiki/Standard_deviation.
    Wikipedia
    CC BY-SA 3.0.
  - "Standard error."

http://en.wikipedia.org/wiki/Standard_error.
Wikipedia
CC BY-SA 3.0.

- "central limit theorem."
  http://en.wiktionary.org/wiki/central_limit_theorem.
  Wiktionary
  CC BY-SA 3.0.
- "standard error."
  http://en.wiktionary.org/wiki/standard_error.
  Wiktionary
  CC BY-SA 3.0.
- "Standard deviation."
  http://en.wikipedia.org/wiki/Standard_deviation.
  Wikipedia
  GNU FDL.

- Estimating the Accuracy of an Average
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "confidence interval."
    http://en.wiktionary.org/wiki/confidence_interval.
    Wiktionary
    CC BY-SA 3.0.
  - "Standard error."
    http://en.wikipedia.org/wiki/Standard_error.
    Wikipedia
    CC BY-SA 3.0.
  - "central limit theorem."
    http://en.wiktionary.org/wiki/central_limit_theorem.
    Wiktionary
    CC BY-SA 3.0.
  - "standard error."
    http://en.wiktionary.org/wiki/standard_error.
    Wiktionary
    CC BY-SA 3.0.
  - "Standard deviation diagram."

http://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg.
Wikimedia
CC BY.

- Chance Models
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Monte Carlo simulation."
    http://en.wikipedia.org/wiki/Monte%20Carlo%20simulation.
    Wikipedia
    CC BY-SA 3.0.
  - "Stochastic modelling (insurance)."
    http://en.wikipedia.org/wiki/Stochastic_modelling_(insurance).
    Wikipedia
    CC BY-SA 3.0.
  - "stochastic."
    http://en.wiktionary.org/wiki/stochastic.
    Wiktionary
    CC BY-SA 3.0.
  - "CircUniformDistOfMean."
    http://commons.wikimedia.org/wiki/File:CircUniformDistOfMean.svg.
    Wikimedia
    Public domain.
- The Gauss Model
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "central limit theorem."
    http://en.wiktionary.org/wiki/central_limit_theorem.
    Wiktionary
    CC BY-SA 3.0.
  - "Gaussian distribution."
    http://en.wikipedia.org/wiki/Gaussian_distribution.
    Wikipedia
    CC BY-SA 3.0.

- Comparing Two Sample Averages
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "null hypothesis."
    http://en.wiktionary.org/wiki/null_hypothesis.
    Wiktionary
    CC BY-SA 3.0.
  - "Student's t-test."
    http://en.wikipedia.org/wiki/Student's_t-test.
    Wikipedia
    CC BY-SA 3.0.
  - "Barbara Illowsky and Susan Dean, Collaborative Statistics. September 17, 2013."
    http://cnx.org/content/m17025/latest/?collection=col10522/latest.
    OpenStax CNX
    CC BY 3.0.
  - "Student's t-distribution."
    http://en.wikipedia.org/wiki/Student's%20t-distribution.
    Wikipedia
    CC BY-SA 3.0.
  - "Student t pdf."
    http://commons.wikimedia.org/wiki/File:Student_t_pdf.svg.
    Wikimedia
    CC BY.
- Odds Ratios
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "odds."
    http://en.wiktionary.org/wiki/odds.
    Wiktionary
    CC BY-SA 3.0.
  - "4441.0.55.002 – A Comparison of Volunteering Rates from the 2006 Census of Population and Housing and the 2006 General Social Survey, Jun 2012."
    http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/

4441.0.55.002Explanatory+Notes5Jun+2012.
Austrailian Bureau of Statistics
CC BY.

- "logarithm."
  http://en.wikipedia.org/wiki/logarithm.
  Wikipedia
  CC BY-SA 3.0.
- "Odds ratio."
  http://en.wikipedia.org/wiki/Odds_ratio.
  Wikipedia
  CC BY-SA 3.0.
- "Odds ratio map."
  http://commons.wikimedia.org/wiki/File:Odds_ratio_map.svg.
  Wikimedia
  CC BY-SA.

- When Does the Z-Test Apply?
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "null hypothesis."
    http://en.wiktionary.org/wiki/null_hypothesis.
    Wiktionary
    CC BY-SA 3.0.
  - "Z-test."
    http://en.wikipedia.org/wiki/Z-test.
    Wikipedia
    CC BY-SA 3.0.
  - "Standard score."
    http://en.wikipedia.org/wiki/Standard_score.
    Wikipedia
    CC BY-SA 3.0.
  - "Statistics/Testing Data/z-tests."
    http://en.wikibooks.org/wiki/Statistics/Testing_Data/z-tests.
    Wikibooks
    CC BY-SA 3.0.

SECTION IX

5.XLSX – EXCEL CHALLENGE - FORMULAS, FUNCTIONS, LOGICAL AND LOOKUP FUNCTIONS

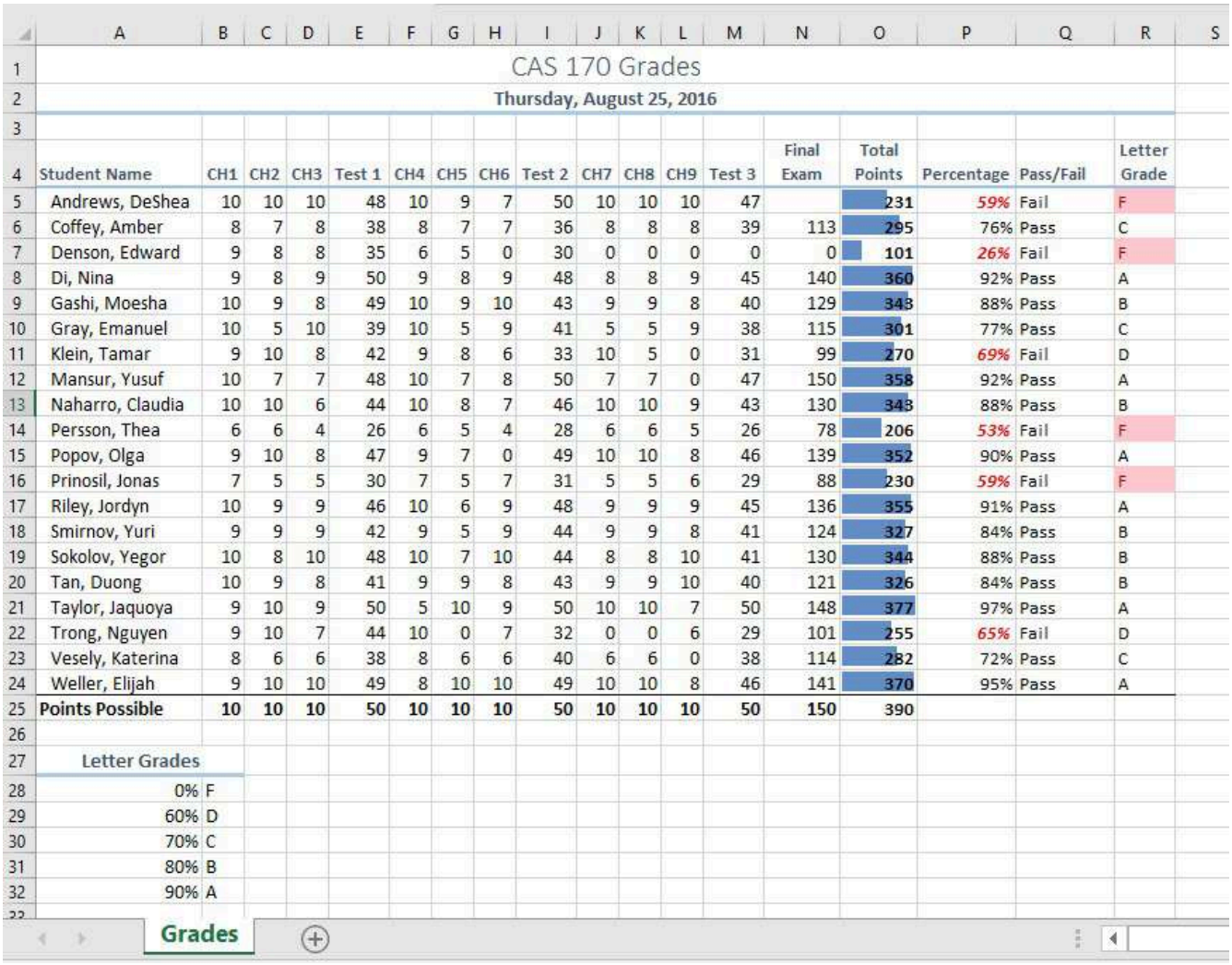
Excel workbooks are designed to allow you to create useful and complex calculations. In addition to doing arithmetic, you can use Excel to look up data, and to display results based on logical conditions. We will also look at ways to highlight specific results. These skills will be demonstrated in the context of a typical gradebook spreadsheet that contains the results for an imaginary Excel class.

In this chapter, we will:

- Use the **Quick Analysis Tool** to find the Total Points for all students and Points Possible.
  (**Note for Mac Users**: the Quick Analysis Tool is not available with Excel for Mac. We have alternate steps for Mac Users)
- Write a division formula to find the Percentage for each student, using an absolute reference to the Total Points Possible.
- Write an **IF Function** to determine Pass/Fail – passing is 70% or higher.
- Write a **VLOOKUP** to determine the Letter Grade using the Letter Grades scale.
- Use the **TODAY** function to insert the current date.
- Review common **Error Messages** using Smart Lookup to get definitions of some of the terms in your spreadsheet.
- Apply **Data Bars** to the Total Points values.
- Apply **Conditional Formatting** to the Percentage, Pass/Fail, and Letter Grade columns.
- **Printing Review** – Change to Landscape, Scale to Fit Columns on One Page and Set Print Area.

Figure 3-1 shows the completed workbook that will be demonstrated in this chapter. Notice the techniques used in columns O and R that highlight the results of your calculations. Notice, also that there are more numbers on this version of the file than you will see in your original data file. These are all completed using Excel calculations.

Figure 3.1 Completed Gradebook Worksheet



CAS 170 Grades

Thursday, August 25, 2016

Student Name	CH1	CH2	CH3	Test 1	CH4	CH5	CH6	Test 2	CH7	CH8	CH9	Test 3	Final Exam	Total Points	Percentage	Pass/Fail	Letter Grade
Andrews, DeShea	10	10	10	48	10	9	7	50	10	10	10	47		231	59%	Fail	F
Coffey, Amber	8	7	8	38	8	7	7	36	8	8	8	39	113	295	76%	Pass	C
Denson, Edward	9	8	8	35	6	5	0	30	0	0	0	0	0	101	26%	Fail	F
Di, Nina	9	8	9	50	9	8	9	48	8	8	9	45	140	360	92%	Pass	A
Gashi, Moesha	10	9	8	49	10	9	10	43	9	9	8	40	129	343	88%	Pass	B
Gray, Emanuel	10	5	10	39	10	5	9	41	5	5	9	38	115	301	77%	Pass	C
Klein, Tamar	9	10	8	42	9	8	6	33	10	5	0	31	99	270	69%	Fail	D
Mansur, Yusuf	10	7	7	48	10	7	8	50	7	7	0	47	150	358	92%	Pass	A
Naharro, Claudia	10	10	6	44	10	8	7	46	10	10	9	43	130	343	88%	Pass	B
Persson, Thea	6	6	4	26	6	5	4	28	6	6	5	26	78	206	53%	Fail	F
Popov, Olga	9	10	8	47	9	7	0	49	10	10	8	46	139	352	90%	Pass	A
Prinosil, Jonas	7	5	5	30	7	5	7	31	5	5	6	29	88	230	59%	Fail	F
Riley, Jordyn	10	9	9	46	10	6	9	48	9	9	9	45	136	355	91%	Pass	A
Smirnov, Yuri	9	9	9	42	9	5	9	44	9	9	8	41	124	327	84%	Pass	B
Sokolov, Yegor	10	8	10	48	10	7	10	44	8	8	10	41	130	344	88%	Pass	B
Tan, Duong	10	9	8	41	9	9	8	43	9	9	10	40	121	326	84%	Pass	B
Taylor, Jaquoya	9	10	9	50	5	10	9	50	10	10	7	50	148	377	97%	Pass	A
Trong, Nguyen	9	10	7	44	10	0	7	32	0	0	6	29	101	255	65%	Fail	D
Vesely, Katerina	8	6	6	38	8	6	6	40	6	6	0	38	114	282	72%	Pass	C
Weller, Elijah	9	10	10	49	8	10	10	49	10	10	8	46	141	370	95%	Pass	A
Points Possible	**10**	**10**	**10**	**50**	**10**	**10**	**10**	**50**	**10**	**10**	**10**	**50**	**150**	390			

Letter Grades	
0%	F
60%	D
70%	C
80%	B
90%	A

Grades

Attribution

5.XLSX.1 MORE ON FORMULAS AND FUNCTIONS

Noreen Brown; Barbara Lave; Hallie Puncochar; Julie Romey; Mary Schatz; Art Schneider; Diane Shingledecker; and Jennifer Evans

Learning Objectives

1. Review the use of the =MAX function.
2. Examine the Quick Analysis Tool to create standard calculations, formatting, and charts very quickly.
3. Create Percentage calculation.
   – Use the Smart Lookup tool to acquire additional information about percentage calculations.
   – Review the use of Absolute cell reference in a division formula.

Another use for =MAX

Before we move on to the more interesting calculations we will be discussing in this chapter, we need to determine how many points it is possible for each student to earn for each of the assignments. This information will go into Row 25. The **=MAX** function is our tool of choice.

Download Data File: CH3 Data

1. Open the data file **CH3 Data** and save the file to your computer as **CH3 Gradebook and Parks**.
2. Make B25 your active cell.
3. Start typing **=MAX** (See **Figure 3.2**) Note the explanation you see on the offered list of functions. You can either keep typing **(** or double click MAX from the list.

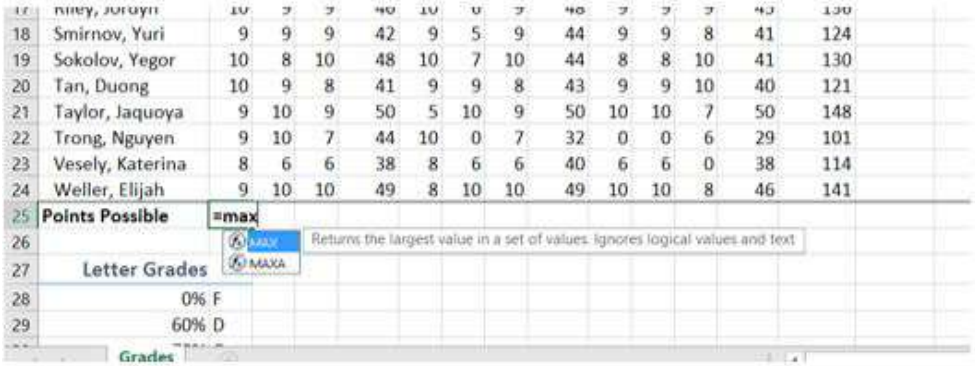


Figure 3.2 Entering a function

4. Select the range of numbers above row 25. Your calculation will be: **=MAX(B5:B24).**
5. Press Enter after selecting the range.
6. Now, use the Fill Handle to copy the calculation from Column B through Column N.
   Note that as you copy the calculation from one column to the next, the cell references change. The calculation in column B reads: **=MAX(B5:B24)**. The one in column N reads: **=MAX(N5:N24)**. These cell references are relative references.

By default, the calculations that Excel copies change their cell references **relative** to the row or column you copy them to. That makes sense. You wouldn't want column N to display an answer that uses the values in column L.

Want to see all the calculations you have just created? Press **Ctrl ~** (See **Figure 3.3**.) **Ctrl ~** displays your calculations (formulas). Pressing **Ctrl ~** a second time will display your calculations in the default view – as values.

Figure 3.3 Relative References – Displayed as calculations.

Quick Analysis Tool

The Quick Analysis Tool allows you to create standard calculations, formatting, and charts very quickly. In this exercise we will use it to insert the Total Points for each student in Column O.

Mac Users: the Quick Analysis Tool is not available with Excel for Mac. We have alternate steps for Mac Users below. Skip down below Figure 3.5 to continue.)

Be sure to press **Ctrl ~** to return your spreadsheet to the normal view (the formula results should display, not the formulas themselves).

1. Select the range of cells **B5:N25**
2. In the lower right corner of your selection, you will see the Quick Analysis tool (see **Figure 3.4**).

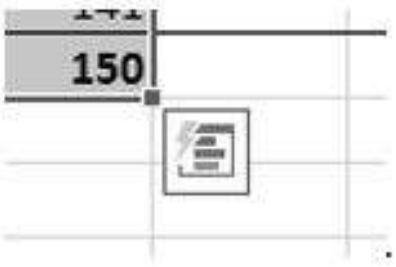


Figure 3.4 Quick Analysis Tool

3. When you click on it, you will see that there are a number of different options. This time we will be using the **Totals** option. In future exercises, we will use other options.
4. Select **Totals**, and then the **SUM** option that highlights the right column (see **Figure 3.5**). Selecting that SUM option places =SUM() calculations in column O.

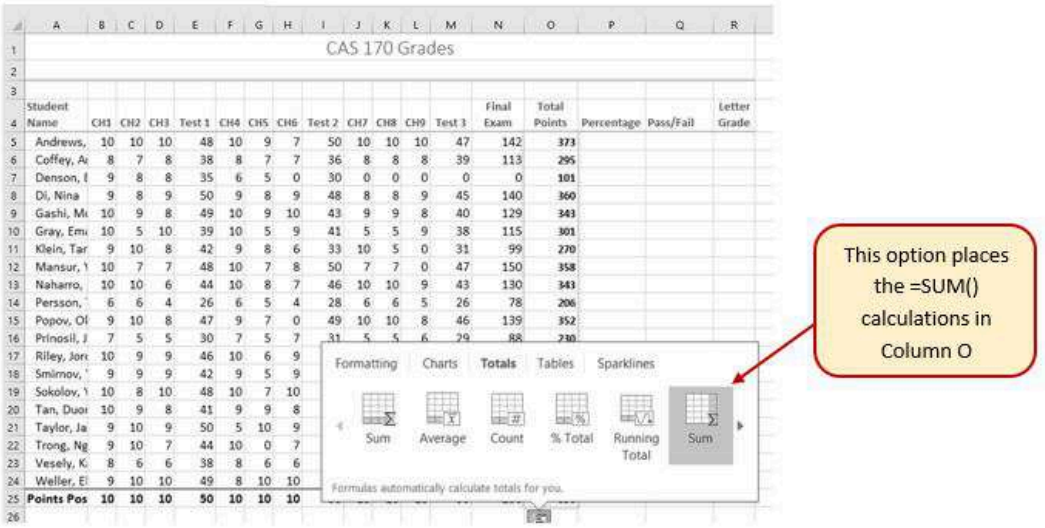


Figure 3.5 Quick Analysis Tool – Totals, Sum Column

Alternate steps for Mac Users:

1. Select the range B5:O25 then click the AutoSum button on the Ribbon (Home tab or Formulas tab)
2. Select the range O5:O25 and click the Bold button.

Percentage calculation

Column P requires a Percentage calculation. Before we launch into creating a calculation for this, it might be handy to know precisely what it is we are looking for. If you are connected to the internet and are using Excel 365, you can use the **Smart Lookup** tool to get some more information about calculating percentages.

In general, the Smart Lookup tool allows you to get more information and definitions about unfamiliar terms or features. This tool is available in all of the Microsoft Office applications.

1. Select cell P4.
2. Find the **Smart Lookup** tool on the **Review** tab (see **Figure 3.6**) and click it. You can also "Right-click" the specific cell and choose **Smart Lookup.**

   Mac Users: The **Smart Lookup** tool is only on the **Review** tab in the latest versions of Excel for Mac. If you can't find the **Smart Lookup** tool on the **Review** tab, you will find it by clicking on the "**Tools**" menu bar option.

   Note for all users: there is a keyboard shortcut for using the Smart Lookup tool. You can hold down the Control key and click in the specific cell (in this case, P4)
3. If this is the first time you have used the Smart Lookup tool, you may need to respond to a statement about your privacy. Press the **Got it** button. We think the Wikipedia article does a pretty good job explaining the calculation, don't you?
4. Close the Smart Lookup pane after reading through the definitions.

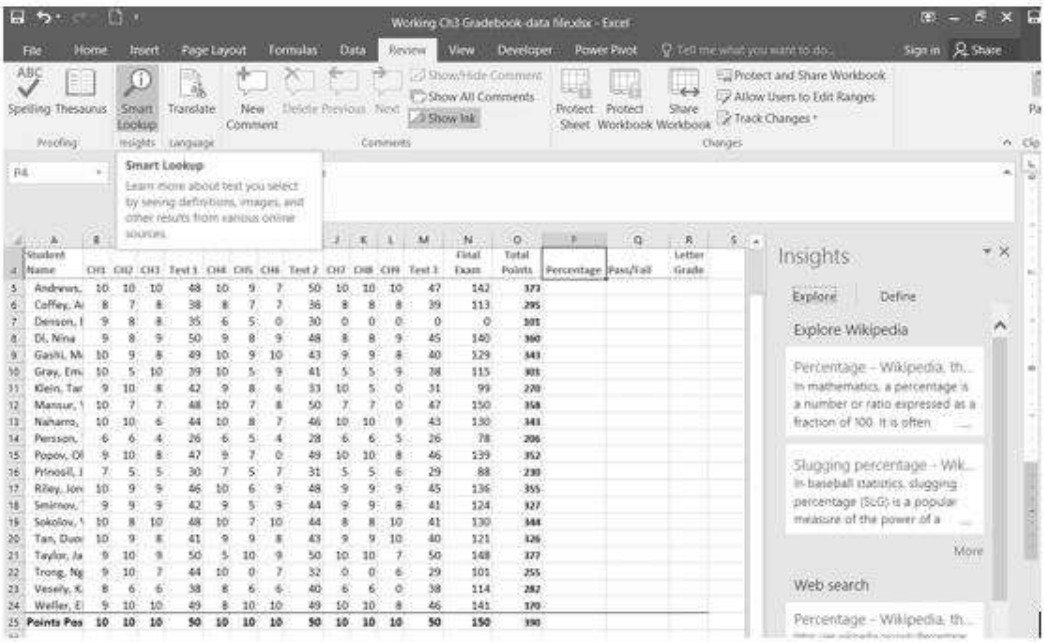


Figure 3.6 Smart Lookup tool

Now that we know what is needed for the Percentage calculation, we can have Excel do the calculation for us.

We need to divide the **Total Points** for each student by the **Total Points** of all the **Points Possible**. Notice that there is a different number on each row – for each student. But, there is only one **Total Points Possible** – the value that is in cell **O25**.

1. Make sure that P5 is your active cell.
2. Press = then select cell **O5**. Press /, then cell **O25**. Your calculation should look like this: **=O5/O25**. The result of the formula should be **0.95641026**. (So far, so good. DeShea Andrews is doing well in this class – with a percentage grade of almost 96%. Definitely an "A"!)
3. Next use the Fill handle to copy the calculation down through row 24 to calculate the other students' grades. You should get the error message **#DIV/0!**. This error message reminds us that you can't divide a number by 0 (zero). And that is just what is happening. If you look at the calculation in P9, the calculation reads: =O9/O29. The first cell reference is correct – it points to Moesha Gashi's total points for the class. But the second reference is wrong. It points to an empty cell – O29.

Before copying the calculation, we have to make the second reference (O25) an **absolute cell reference**. That way, when we copy the formula down, the cell reference for O25 will be locked and will not change.

1. Make P5 your active cell. In the Formula Bar click on O25 (see **Figure 3.7**).
2. Press F4 (on the function keys at the top of your keyboard). That will make the O25 reference absolute. It will not change when you copy the calculation (see **Figure 3.8**). (If you are working on a laptop and do not have an F4 function key, you can type in a $ before the O and another one before the 25.)
3. The calculation now looks like this: **=O5/$O$25.**
4. Use the Fill Handle to copy the formula down through P24 again. Now, when you copy the formula, you will get correct values for all of the students.

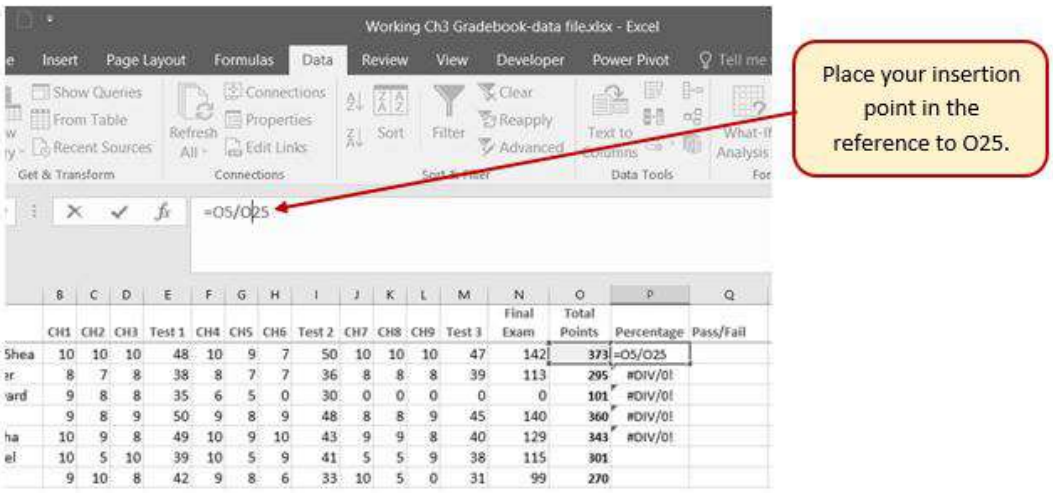


Figure 3.7 Editing a formula

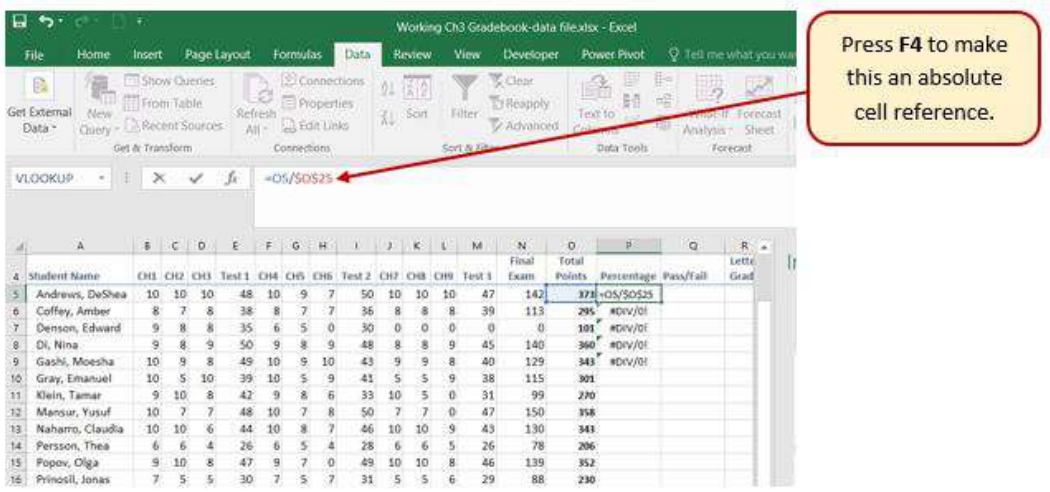


Figure 3.8 Absolute Cell reference – press F4

Those long decimals are a bit nonstandard. Let's change them to % by applying cell formatting.

1. Select the range P5:P24.
2. On the Home tab, in the Number Group, select the % (Percent Style) button.

Skill Refresher

Absolute References

1. Click in front of the column letter of a cell reference in a formula or function that you do not want altered when the formula or function is pasted into a new cell location.
2. Press the F4 key or type a dollar sign ($) in front of the column letter and row number of the cell reference.

Keyboard Shortcuts

Smart Lookup Tool

- Hold down the CTRL key and click the specific cell that you are working with. Then choose "**Smart Lookup**"

- **Mac Users**: Same as above

Key Takeaways

- Functions can be created using cell ranges or selected cell locations separated by commas. Make sure you use a cell range (two cell locations separated by a colon) when applying a statistical function to a contiguous range of cells.
- To prevent Excel from changing the cell references in a formula or function when they are pasted to a new cell location, you must use an absolute reference. You can do this by placing a dollar sign ($) in front of the column letter and row number of a cell reference or by using the F4 function key.
- The #DIV/0 error appears if you create a formula that attempts to divide a constant or the value in a cell reference by zero.

More functions:

Create the Sample Worksheet

This section uses a sample worksheet to illustrate Excel built-in functions. Consider the example of referencing a name from column A and returning the age of that person from column C. To create this worksheet, enter the following data into a blank Excel worksheet.

You will type the value that you want to find into cell E2. You can type the formula in any blank cell in the same worksheet.

	A	B	C	D	E
1	Name	Dept	Age		Find Value
2	Henry	501	28		Mary
3	Stan	201	19		
4	Mary	101	22		
5	Larry	301	29		

Term	Definition	Example
Table Array	The whole lookup table	A2:C5
Lookup_Value	The value to be found in the first column of Table_Array.	E2
Lookup_Array -or- Lookup_Vector	The range of cells that contains possible lookup values.	A2:A5
Col_Index_Num	The column number in Table_Array the matching value should be returned for.	3 (third column in Table_Array)
Result_Array -or- Result_Vector	A range that contains only one row or column. It must be the same size as Lookup_Array or Lookup_Vector.	C2:C5
Range_Lookup	A logical value (TRUE or FALSE). If TRUE or omitted, an approximate match is returned. If FALSE, it will look for an exact match.	FALSE
Top_cell	This is the reference from which you want to base the offset. Top_Cell must refer to a cell or range of adjacent cells. Otherwise, OFFSET returns the #VALUE! error value.	
Offset_Col	This is the number of columns, to the left or right, that you want the upper-left cell of the result to refer to. For example, "5" as the Offset_Col argument specifies that the upper-left cell in the reference is five columns to the right of reference. Offset_Col can be positive (which means to the right of the starting reference) or negative (which means to the left of the starting reference).	
CONCAT	This is used for text that needs to be merged into one cell. You can type data into cells, then by using the CONCAT function and the range or cells you want to use, the data will be merged into the cell reference. For example, if you have the words "Red" in cell C2 and "Cat" in cell C3, by using CONCAT in cell C4, you can have the words Red Cat appear in that cell.	

5.XLSX.2 LOGICAL AND LOOKUP FUNCTIONS

Noreen Brown; Barbara Lave; Hallie Puncochar; Julie Romey; Mary Schatz; Art Schneider; Diane Shingledecker; and Jennifer Evans

Learning Objectives

- Use an IF Function to make logical comparisons between a value and what you expect.
- Create a VLOOKUP calculation to look up information in a table.
- Understand error messages.
- Understand how to enter and format Date/Time Functions.

In addition to doing arithmetic, Excel can do other kinds of functions based on the data in your spreadsheet. In this section, we will use an **=IF** function to determine whether a student is passing or failing the class. Then, we will use a **=VLOOKUP** function to determine what grade each student has earned.

IF Function

The IF function is one of the most popular functions in Excel. It allows you to make logical comparisons between a value and what you expect. In its simplest form, the IF function says something like:
If the value in a cell is what you expect (true) – do this. If not – do that.

The IF function has three arguments:

- **Logical test** – Here, we can test to see if the value in a selected cell is what we expect. You could use something like "B7=14" or "B7>12" or "B7<6"
- **Value_if_true** – If the requirements in the logical test are met – if B7 is equal to 14 – then it is said to be true. For this argument, you can type text – "True", or "On budget!" Or you could insert a calculation, like B7*2 (If B7 does equal 14, multiply it by 2). Or, if you want Excel to put nothing at all in the cell, type "" (two quotes).
- **Value_if_false** – If the requirements in the logical test are not met – if B7 does not equal 14 – then it is

said to be false. You can enter the same instructions here as you did above. Let's say that you type the double quotes here. Then, if B7 does not equal 14, nothing will be displayed in this cell.

In column Q we would like Excel to tell us whether a student is passing – or failing the class. If the student scores **70%** or better, he/she will pass the class. But, if he/she scores less than 70%, he/she is failing.

1. Make sure that Q5 is your active cell.
2. On the Formulas tab, in the Function Library group, find the IF function on the Logical pulldown menu (see **Figure 3.9**).

   Mac Users: There is no "Function Library" group for Excel for Mac. Mac Users should click on the Formulas tab, then click the "Logical" tool 

list arrow, and choose IF (see **Figure 3.9).**

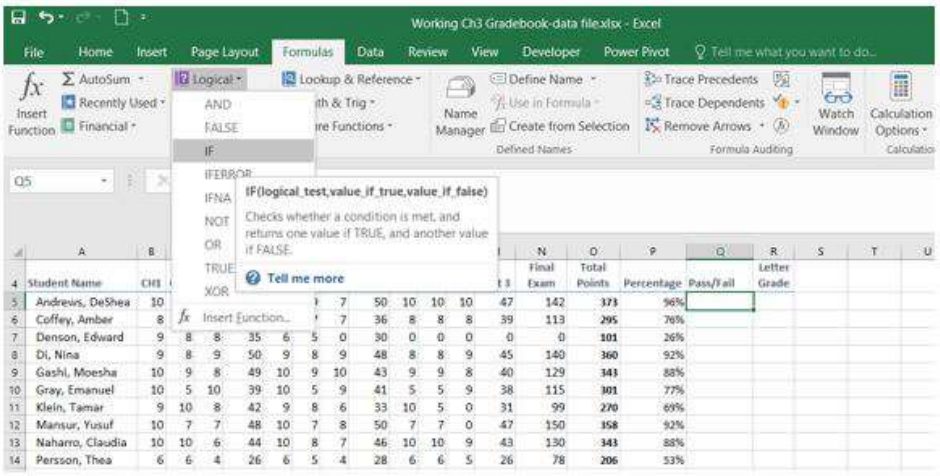


Figure 3.9 IF Function

Now you will see the IF Function dialog box, with a place to enter each of the three arguments.

Mac Users: There is no "dialog box". The "Formula Builder" pane will display at the right side of the Excel window. It has the same layout as Figure 3.10 below.

1. Click in the box for **Logical Test**. We need to test whether a student's score is less than **.7**. So, in this box, type **P5<.7**
2. Click in the box for **Value_if_true**. If the student's score is less than .7, then they are failing the class. In this box, type **Fail**.
3. Click in the box for **Value_if_false**. If the student's score is NOT less than .7, then they are passing the class. In this box, type **Pass**.
4. Make sure that your dialog box matches **Figure 3.10**.

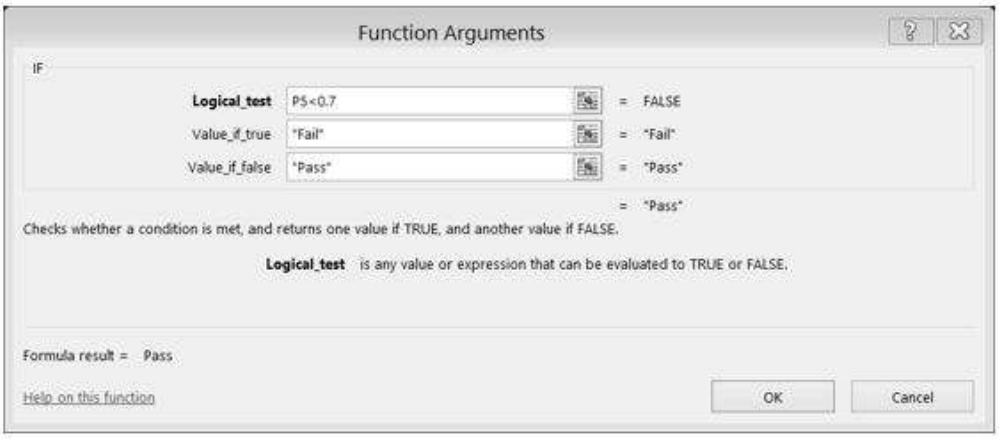


Figure 3.10 IF Function Dialog Box

While we are here, let's take a look at the dialog box. Notice that as you click in each box, Excel gives you a brief explanation of the contents (in the middle below the boxes.) In the lower left-hand corner, you can see the results of the calculation. In this case, DeShae is passing the class. Below that is a link to **Help on this function**. Selecting this link will take you to the Excel help for this function – with detailed information on how it works.

5. Once you have typed in the required arguments and reviewed to make sure they are correct, press OK.
   Mac Users should click the "Done" button, then close the Formula Builder pane.
   (The text Pass should be displayed in Q5 because DeShae is passing the class.)
6. Use the Fill handle to copy the IF function down through row 24.
7. Click on Q5. When you look in the formula bar, you will see the IF calculation: **=IF(P5<0.7,"Fail","Pass")** (see **Figure 3.11**).

<img class="wp-image-179 size-full" src="https://spscc.pressbooks.pub/app/uploads/sites/50/2021/05/Figure-3-11.jpg" alt="Formula bar shows IF function (=IF(PS Figure 3.11 IF Function Results

VLOOKUP Function

You need to use a VLOOKUP function to look up information in a table. Sometimes that table is on a different sheet in your workbook. Sometimes it is in another file entirely. In this case, we need to know what grade each student is getting based on their percentage score. You will find the table that defines the scores and the grades in **A28:B32**.

There are four pieces of information that you will need in order to build the VLOOKUP syntax. These are the four arguments of a VLOOKUP function:

- The value you want to lookup, also called the **Lookup_value**. In our example, the lookup value will be the student's percentage score in column P.

- The **Table_array** is the range (table) where the lookup values and the values you want returned by the function are located. In our example, this is the table of percentages and corresponding letter grades in the range A28:B32. The lookup value should always be in the first column in the table array for VLOOKUP to work correctly. For example, in our table_array the lookup value is in cell A28, so the range should start with A.
- The **Col_index_num** is the column number in the range that contains the value to return. In our example, when you specify A28:B32 as the Table_array range, you should count A as the first column (1), B as the second column (2), and so on. You will enter the appropriate column number in this box as 1, 2, or 3 and so on.
- In the **Range_lookup**, you can optionally specify TRUE if you want an approximate match or FALSE if you want an exact match of the return value. If you leave this blank, the default value will always be TRUE, or approximate match.

Let's create the VLOOKUP to display the correct Letter Grade in column R.

1. Make sure that R5 is your active cell.
2. On the Formulas tab, in the Function Library, find the **VLOOKUP** function on the Lookup & Reference pull-down menu (see **Figure 3.12**).

   Mac Users should click the Lookup and Reference tool list arrow 

   to find the **VLOOKUP** function.

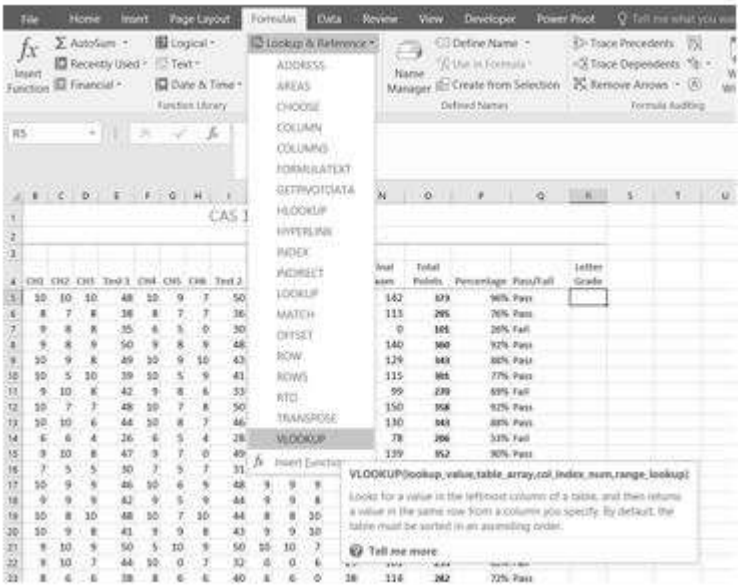


Figure 3.12 VLOOKUP Function

3. Fill in the dialog box so that it looks like the image in **Figure 3.13**.

Mac Users will use the "Formula Builder" pane at the right side of the Excel Window.

- **Lookup_value** – In this case, we will use the Percentage score. So, **P5** for the first lookup value.
- **Table_array** – This is the range that contains the value you want returned by the function. In this case, that range is A28:B32. Note that this range does NOT include the label in row 27; just the actual data. The cell references for the Table_array need to be absolute – $A$28:$B$32. When we copy this function to the other cells, we do not want these cell references to change. It should always be $A$28:$B$32. **This is very important! They must have the absolute reference symbols or the calculations will not work.**
- **Col_index_number** – This is the column in the table array range that includes the information that we are looking up. In our case, the actual grades are in the 2nd column of the range. So, the column index will be 2.
- **Range_lookup** – In some cases, you will need something in the Range_lookup box. Since we are looking for an approximate match for the percentages, we want the default value of TRUE, so we do not need to enter anything for this argument.

2. While you are in the dialog box, be sure to look at all the helpful definitions that Excel offers.
3. When you have filled in the dialog box, press OK.

   Mac Users should click the "Done" button, then close the Formula Builder pane.
4. The calculation you will see in the formula bar is: =VLOOKUP(P5,$A$28:$B$32,2)
5. Use the fill handle to copy the function down through row 24. The results displayed should match **Figure 3.14**.

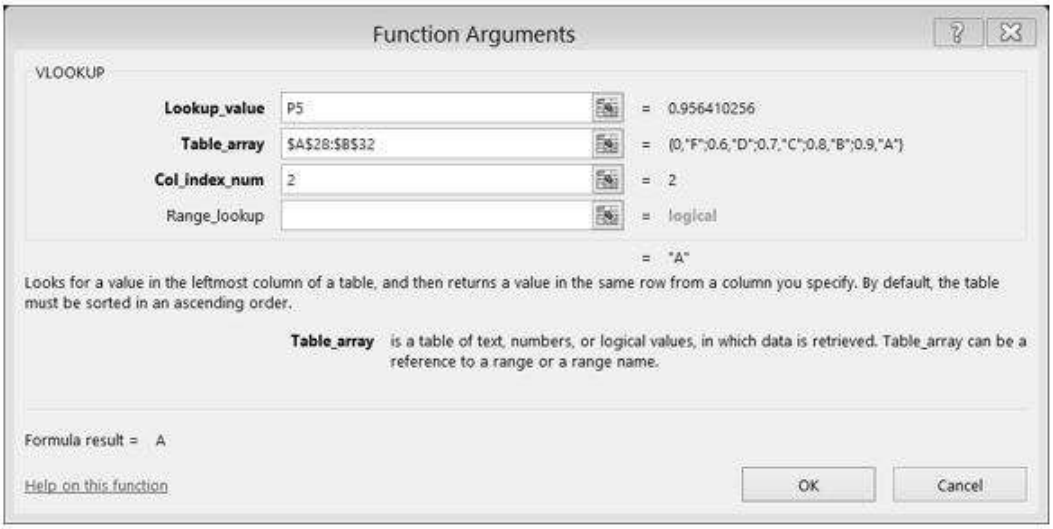
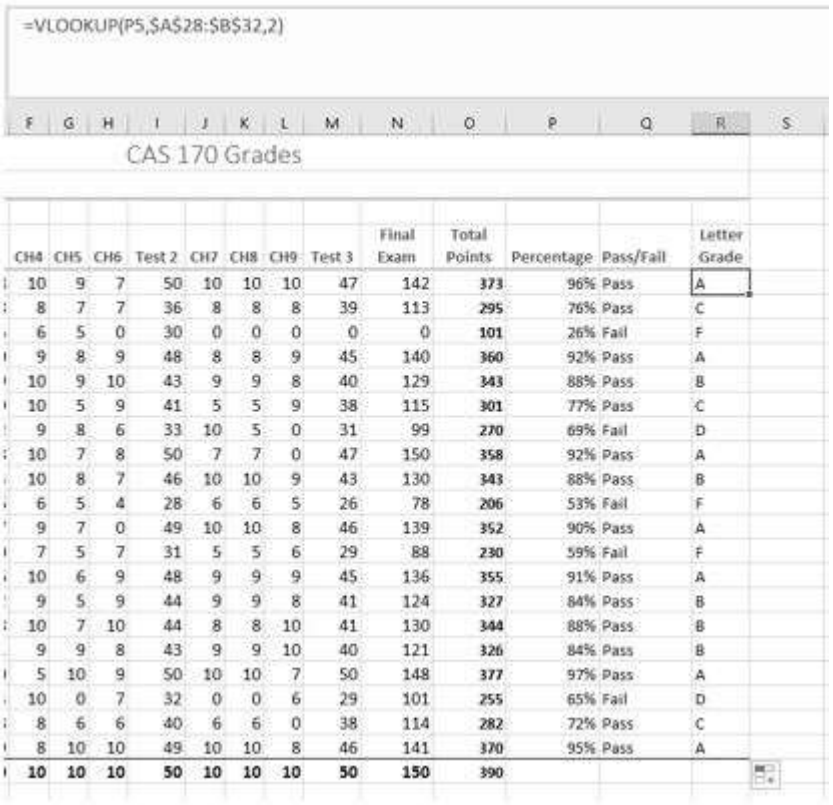


Figure 3.13 VLOOKUP completed dialog box



CH4	CH5	CH6	Test 2	CH7	CH8	CH9	Test 3	Final Exam	Total Points	Percentage	Pass/Fail	Letter Grade
10	9	7	50	10	10	10	47	142	373	96%	Pass	A
8	7	7	36	8	8	8	39	113	295	76%	Pass	C
6	5	0	30	0	0	0	0	0	101	26%	Fail	F
9	8	9	48	8	8	9	45	140	360	92%	Pass	A
10	9	10	43	9	9	8	40	129	343	88%	Pass	B
10	5	9	41	5	5	9	38	115	301	77%	Pass	C
9	8	6	33	10	5	0	31	99	270	69%	Fail	D
10	7	8	50	7	7	0	47	150	358	92%	Pass	A
10	8	7	46	10	10	9	43	130	343	88%	Pass	B
6	5	4	28	6	6	5	26	78	206	53%	Fail	F
9	7	0	49	10	10	8	46	139	352	90%	Pass	A
7	5	7	31	5	5	6	29	88	230	59%	Fail	F
10	6	9	48	9	9	9	45	136	355	91%	Pass	A
9	5	9	44	9	9	8	41	124	327	84%	Pass	B
10	7	10	44	8	8	10	41	130	344	88%	Pass	B
9	9	8	43	9	9	10	40	121	326	84%	Pass	B
5	10	9	50	10	10	7	50	148	377	97%	Pass	A
10	0	7	32	0	0	6	29	101	255	65%	Fail	D
8	6	6	40	6	6	0	38	114	282	72%	Pass	C
8	10	10	49	10	10	8	46	141	370	95%	Pass	A
10	**10**	**10**	**50**	**10**	**10**	**10**	**50**	**150**	**390**			

Figure 3.14 VLOOKUP Complete

Note: What if it didn't work? What if you get a result different from the one predicted? In this case, either you have made a previous error, resulting in different % scores than this exercise anticipated, or you made a mistake entering your VLOOKUP function.

To make repairs in the function, make sure that R5 is your active cell. On the Formula bar, press the Insert Function button (see **Figure 3.15**). That will reopen the dialog box so you can make your repairs. Did you forget to make the cell references for the Table_array absolute? Did you use the wrong cell for the Lookup_value? Press OK when you are done and recopy the corrected function.

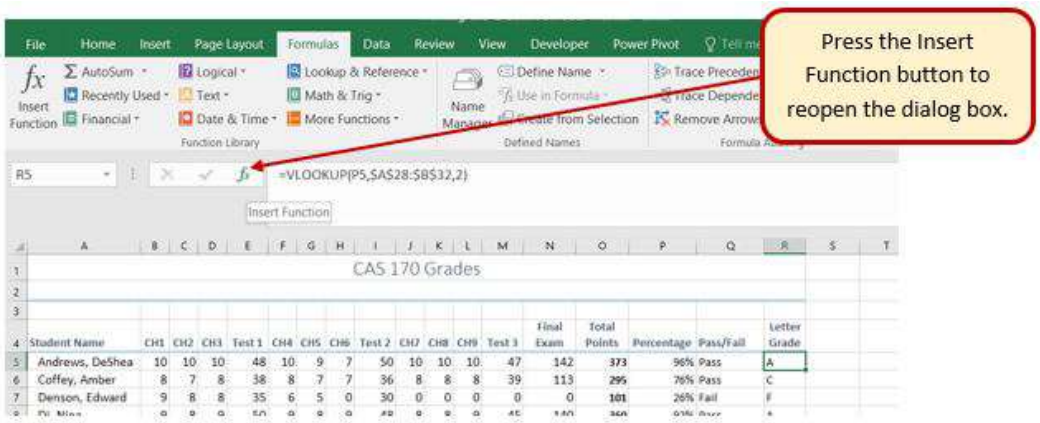


Figure 3.15 Insert Function

Error Messages

Sometimes Excel notices that you have made errors in your calculations before you do. In those cases Excel alerts you with some slightly mysterious error messages. A list of common error messages can be found in **Table 3.1** below.

Table 3.1 – Common Error Messages

Message	What Went Wrong
#DIV/0!	You tried to divide a number by a zero (0) or an empty cell.
#NAME	You used a cell range name in the formula, but the name isn't defined. Sometimes this error occurs because you type the name incorrectly.
#N/A	The formula refers to an empty cell, so no data is available for computing the formula. Sometimes people enter N/A in a cell as a placeholder to signal the fact that data isn't entered yet. Revise the formula or enter a number or formula in the empty cells.
#NULL	The formula refers to a cell range that Excel can't understand. Make sure that the range is entered correctly.
#NUM	An argument you use in your formula is invalid.
#REF	The cell or range of cells that the formula refers to aren't there.
#VALUE	The formula includes a function that was used incorrectly, takes an invalid argument, or is misspelled. Make sure that the function uses the right argument and is spelled correctly.

This table was copied from the internet. Look here for additional information.
http://www.dummies.com/software/microsoft-office/excel/how-to-detect-and-correct-formula-errors-in-excel-2016/

Date Functions

Very often dates and times are an important part of Excel data. Numbers that are correct today may not be accurate tomorrow. So, it is frequently useful to include dates and times on your spreadsheets.

These dates and times fall into two categories – ones that:

- **Remain the same**. For instance, if a spreadsheet includes data for May 15th, you don't want the date to change each time you re-open the spreadsheet.

- **Change to reflect the current date/time**. When it is important to have the current date or time on a spreadsheet, you want Excel to update the information regularly.

Take a look at the list of Date and Time functions offered in the Function Library on the Formulas tab (see **Figure 3.16**).

Figure 3.16 Date & Time Functions

For our gradebook, we want the date and time to be displayed in A2, and it needs to update whenever the workbook file is opened.

1. Make A2 your active cell. Notice that A2 extends all the way from column A to Column R. Previously, someone has used the Merge & Center tool on this cell to make it match the title above.
2. On the Formulas tab, in the Function Library, select **NOW** from the Date & Time drop-down list and then click OK.

   Mac Users click the "Done" button in the "Formula Builder" pane at the right side of the Excel window; then close the pane.
3. The result you will see in the formula bar is: =NOW(). The result you will see in A2 depends on the current date and time. The NOW function is a very handy function. It takes no arguments and is Volatile! That is not as alarming as it may seem. This just means that you don't need to give it any more information to do its job and that your results will change frequently. You can update the date and time whenever you want – you don't have to wait until you open the workbook again.
4. Make sure that A1 is your active cell and press the **F9** function key (along the top of your keyboard.) The time will update.

Excel will update this field independently whenever you save and re-open the file, or print it. It may happen more frequently than that – depending on how you have set this up in your installation of Excel.

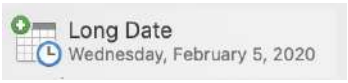
Another variation of the current date is the **TODAY** function. Let's try that one next.

1. Make sure A2 is your active cell. Press Delete to remove the NOW function.
2. From the Date & Time drop-down list in the Function Library on the Formulas tab (see **Figure 3.16**), select TODAY and then click OK.

   Mac Users click the "Done" button in the "Formula Builder" pane; then close the pane.
3. The result you will see in the formula bar is: =TODAY(). The result you will see in A2 depends on the current date. Since we haven't asked for the time, the time you are seeing is likely 12:00. That is not very helpful so we need to change the format of the date.
4. On the Home tab, in the Number group, press the Number Format Launcher button (see **Figure 3.17**).
5. In the **Format Cells** dialog box, click the Number tab. Choose the **Date** category and select **Wednesday, March 14, 2012** (this format is called **Long Date**).

   Mac Users: there is no Number Format Launcher button or "Format Cells" dialog box.

   Click the list arrow next to "Date" and choose **"Long Date"**



6. The current day and date will display in A2.

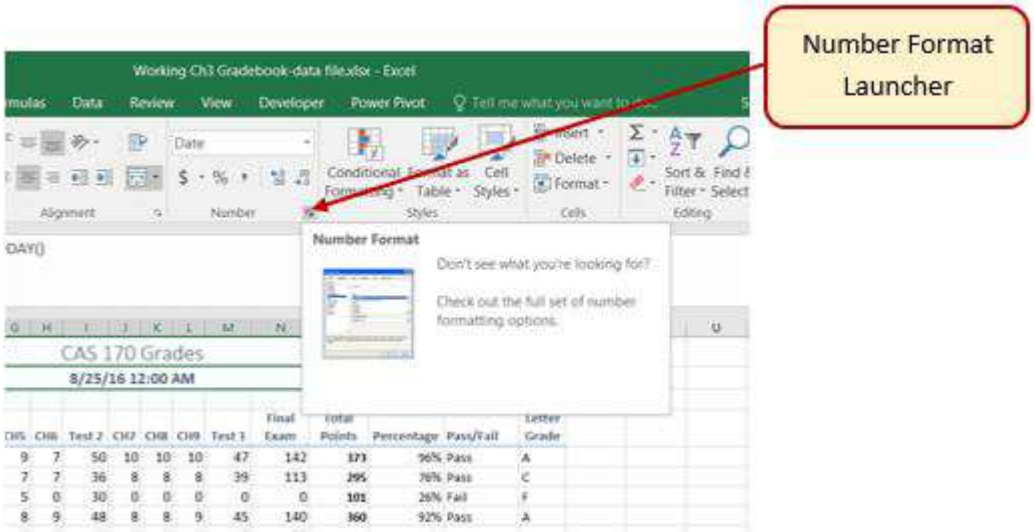


Figure 3.17 Number Format Launcher

Keyboard Shortcuts

Sometimes you want the date or the time to show up in your spreadsheet, but you don't want it to change. You can simply type in the date or time. Or, you can use shortcut keys.

- CTRL ; (semi colon) will bring you the current date
  Mac Users: same as above
- CTRL : (colon or CTRL SHIFT ;) will bring you the current time.
  Mac Users: SHIFT COMMAND :

Key Takeaways

- Functions don't always have to be about arithmetic. Excel provides functions that will help you perform logical evaluations, look things up, and work with dates and times.
- Excel displays error messages when your formulas and functions are not constructed properly.

Attribution

5.XLSX.3 CONDITIONAL FORMATTING

Noreen Brown; Barbara Lave; Hallie Puncochar; Julie Romey; Mary Schatz; Art Schneider; Diane Shingledecker; and Jennifer Evans

Learning Objectives

- Use Conditional Formatting techniques to provide flexible highlighting, applying specified formatting only when certain conditions are met. Techniques include:
  - Data bars — to make it easy to visualize values in a range of cells.
  - Cells Rules — to highlight values that match the requirements you specify.

You now have all the calculations you need in your CAS 170 Grades spreadsheet. There is a lot of data here. To make it easier to pick out the most important pieces of data, Excel provides **Conditional Formatting**. The best thing about Conditional Formatting is that it is flexible, applying specified formatting only when certain conditions are met.

1. Select the values in the Total Points column (**O5:O24**).
2. At the bottom of your selection, click on the Quick Analysis Tool. On the Formatting tab, select Data Bars (see **Figure 3.18**).

   Mac Users: as stated previously, there is no Quick Analysis Tool for Excel for Mac. Use the alternate steps as shown below.

Excel places blue bars on top of your values; long blue bars for larger numbers, shorter ones for smaller numbers. This makes it easier to see how well each student did in the class – without having to look at the specific numbers.

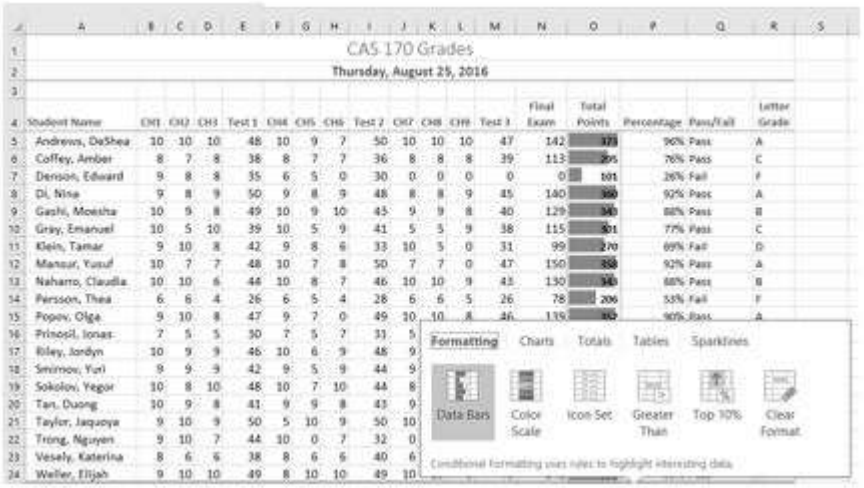


Figure 3.18 Data Bars on the Quick Analysis tool

Another way to apply Data Bars is to:

- Select the range that needs data bars
- On the Home tab, in the Styles group, select Data Bars from the Conditional Formatting tool.
- From there you can select data bars of different colors and opacities (see **Figure 3.19**).

Mac Users: Alternate Steps:

- On the Home tab select Data Bars from the Conditional Formatting tool 

.
- From there you can select data bars of different colors and opacities (see **Figure 3.19**).

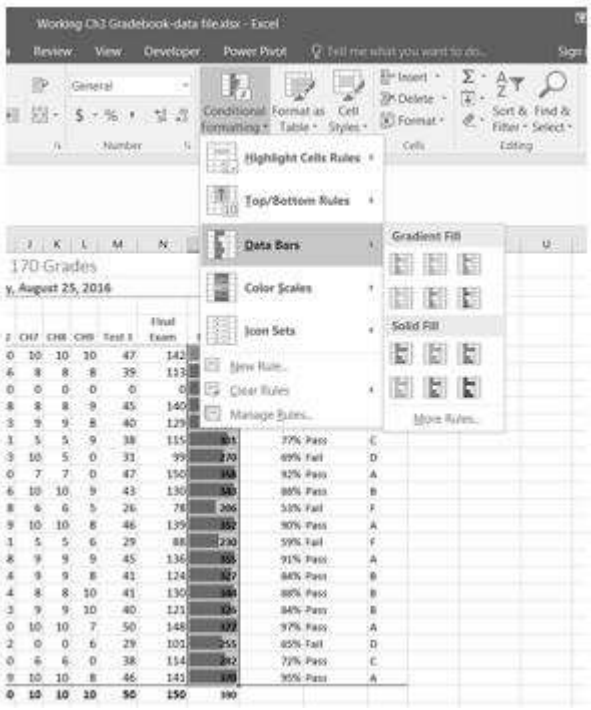


Figure 3.19 Data Bars on the Conditional Formatting tool

It is even more important to highlight the students who are failing in the class. To practice further with Conditional Formatting we will do that in two places, in the Percentages column and on the Letter Grade column. To start with, we want any **F** letter grades to be formatted with a light red fill color and dark red text.

1. Select the Letter Grades (**R5:R24**).
2. On the Home tab, in the Styles group, select **Highlight Cell Rules** from the Conditional Formatting tool (see **Figure 3.20**).
3. Select **Equal To**
4. Fill out the Equal to dialog box so that cells that are equal to: **F** have **Light Red Fill with Dark Red Text** (see **Figure 3.21**).

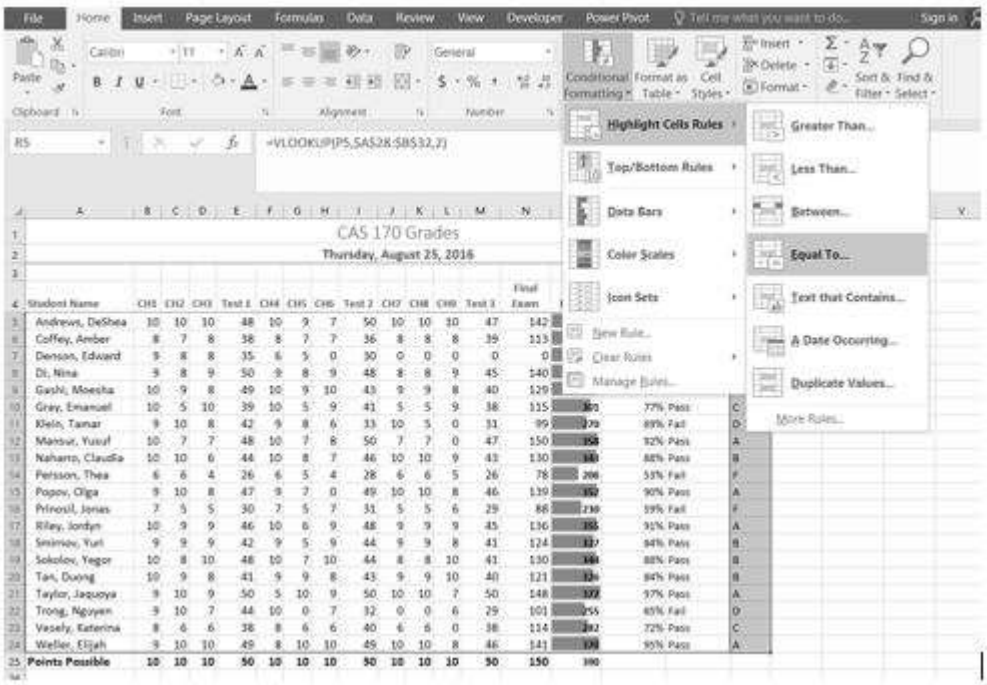


Figure 3.20 Conditional Formatting Equal To

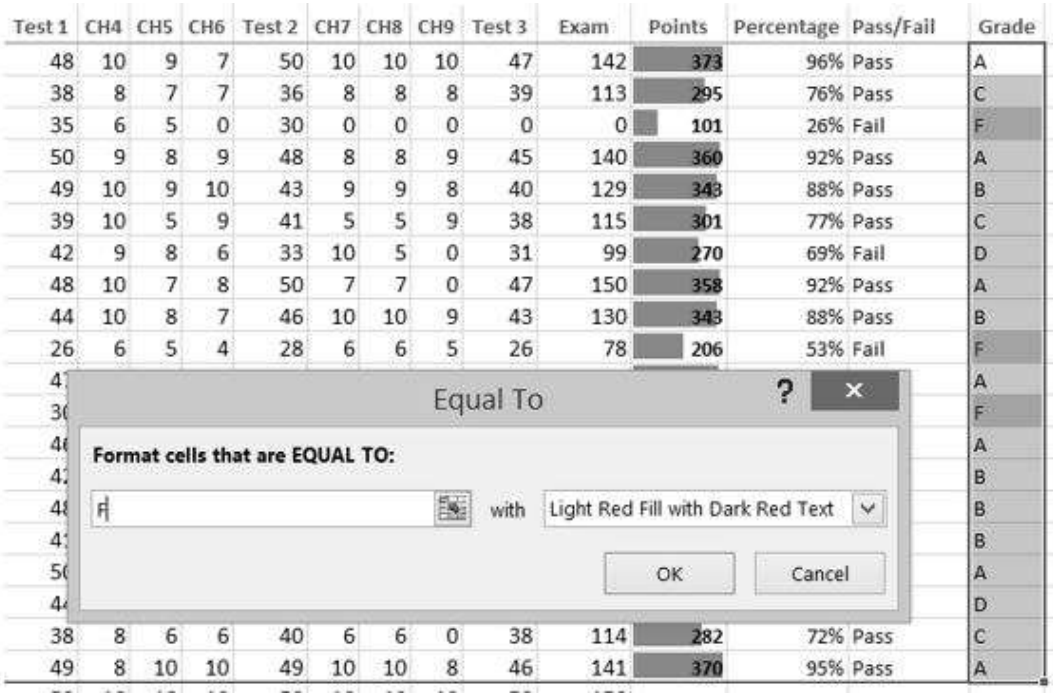


Figure 3.21 Conditional Formatting Equal To Dialog Box

Let's try that one more time – to highlight those students who are passing the class. This time we will use the Pass/Fail text in the Pass/Fail column. If the text for a student is Pass we want the cell to be formatted with a yellow fill with dark yellow text.

1. Select the Pass/Fail grades (**Q5:Q24**).
2. On the Home tab, in the Styles group, select **Highlight Cell Rules** from the Conditional Formatting

tool (see **Figure 3.20**).
3. Select **Equal To**
4. Fill out the Equal to dialog box so that cells that are equal to: **Pass** have **Yellow Fill with Dark Yellow Text**. (To find the Yellow Fill with Dark Yellow text option, click the the down arrow at the end of the last (with) box).

You do not have to use the default styles to make your data stand out. You can set any formatting you want. When you do, it is probably a good idea to include other styling in addition to color. Your spreadsheet might be printed in black and white. You would hate to lose your Conditional formatting. Now we are going to use conditional formatting to display any Percentages that are less than 60% with red text formatted in bold and italic.

1. Select the Percentage grades (**P5:P24**).
2. On the Home tab, in the Styles group, select Highlight Cell Rules from the Conditional Formatting tool (see **Figure 3.20**).
3. Select **Less Than**
4. Fill out the Less Than dialog box so that cells that are less than **.6** will be have conditional formatting. But, instead of using the default red text on a light red fill, press the down arrow at the end of that box and select **Custom Format**.
5. On the **Font** tab of the Format Cells dialog box, in the Font style box, select **Bold Italic**. In the Color box, select **Red** (see **Figure 3.22**).
6. Press OK. Then press OK again.

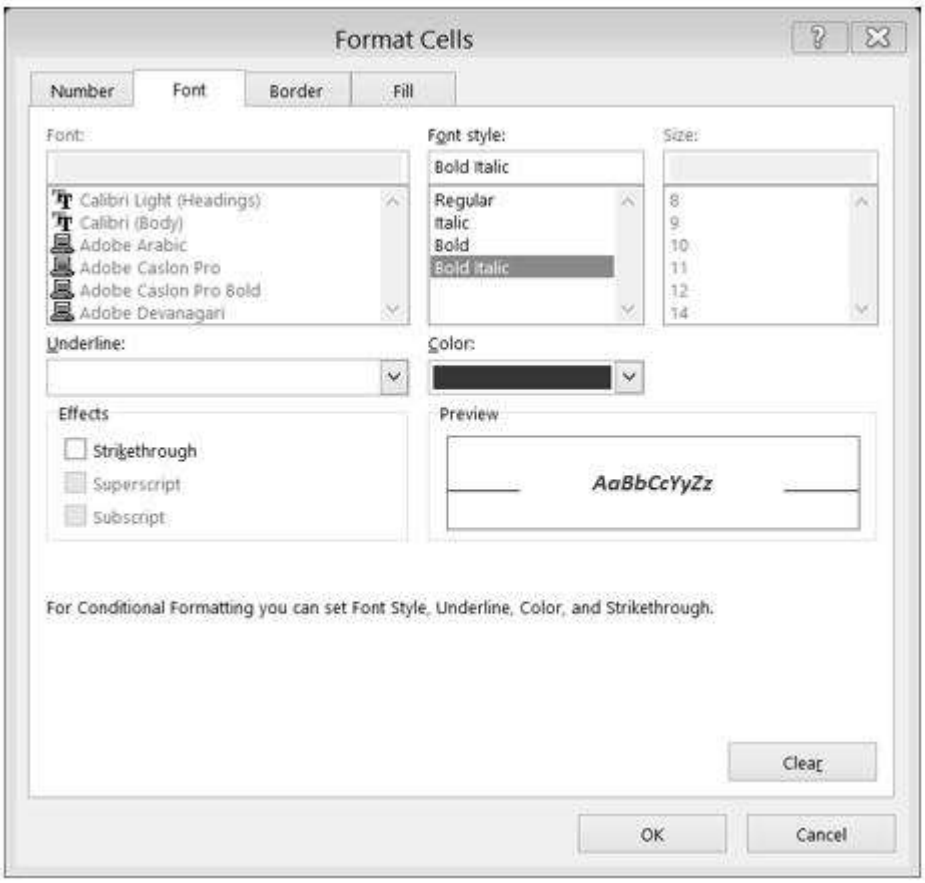


Figure 3.22 Conditional Formatting Custom Format Cells Dialog box

Conditional Formatting is valuable in that it reflects the current data. It changes to reflect changes in the data. To test this, delete DeShea's final exam score. (Select N5. Press Delete on your keyboard.) Suddenly, DeShae is failing the course and the Conditional Formatting reflects that. This is a little unfair to DeShae – who has worked so hard this quarter. Let's give him back his grade. Press CTRL Z (Undo). His test score reappears and the Conditional formatting reflects that as well.

Making Changes

What if you have made a mistake with your Conditional Formatting? Or, you want to delete it altogether? You can use the **Conditional Formatting Manage Rules** tool. In our example, we want to remove the conditional formatting rule that formats the **Pass** text with yellow. We are also going to modify the minimum passing percentage for the conditional formatting rule that is applied to the percentages.

1. On the Home Tab, in the Styles Group, select **Manage Rules** at the very bottom of the Conditional Formatting drop-down list.
2. Show formatting rules for: **This Worksheet** (see **Figure 3.23**).
3. We don't really need to highlight the students who are passing the class, so select that rule in the Rules

Manager and press the **Delete Rule** button. Mac Users should click the minus symbol – at bottom left corner to delete the rule.

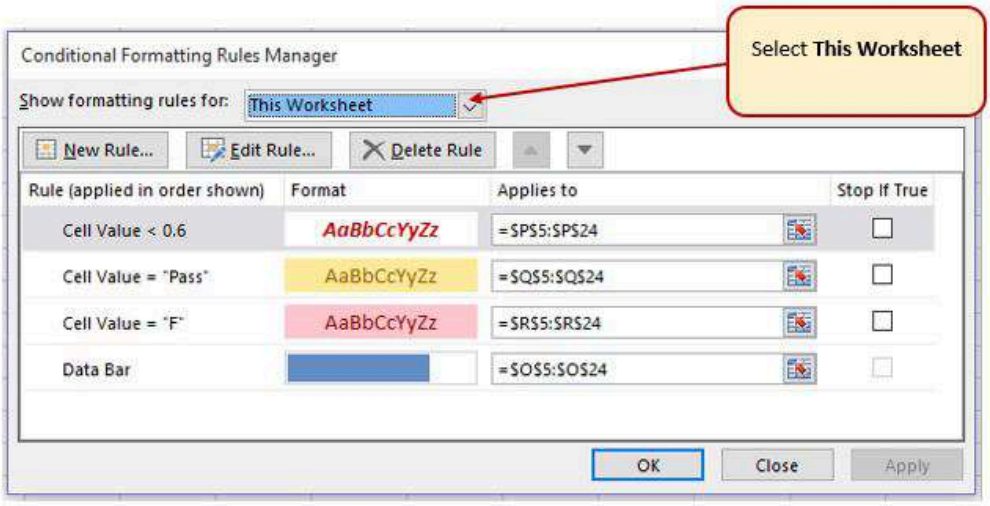


Figure 3.23 Conditional Formatting Manage Rules

In a previous exercise (the IF function), we decided that students were failing if they got a percentage score of less than 70%, so the Conditional Formatting rule in the Percentage column needs repair.

4. Select the rule that reads Cell Value <0.6.
5. Select the Edit Rule button, and change the .6 to .7 (see **Figure 3.24**).
6. Click OK (or Apply) twice. Double check that your completed workbook matches **Figure 3.25**.

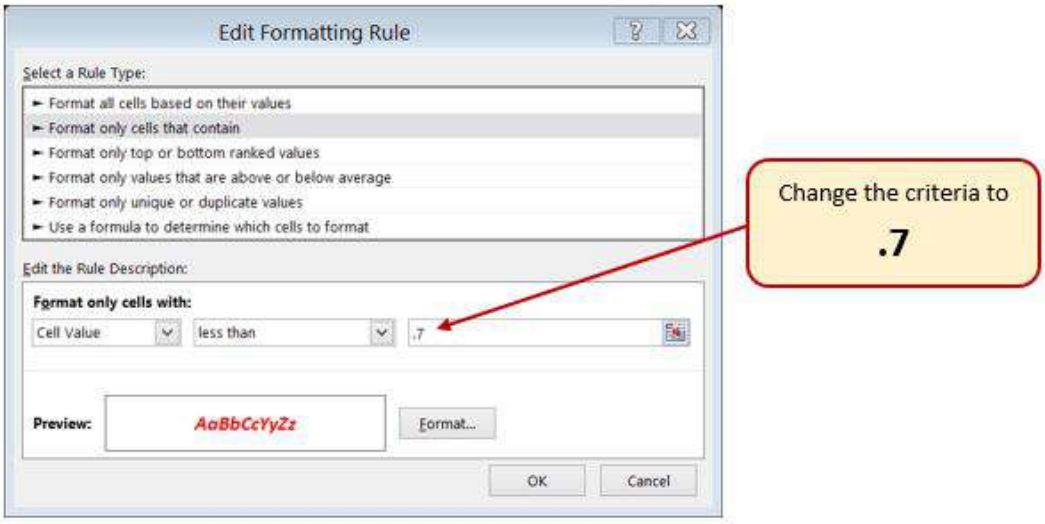
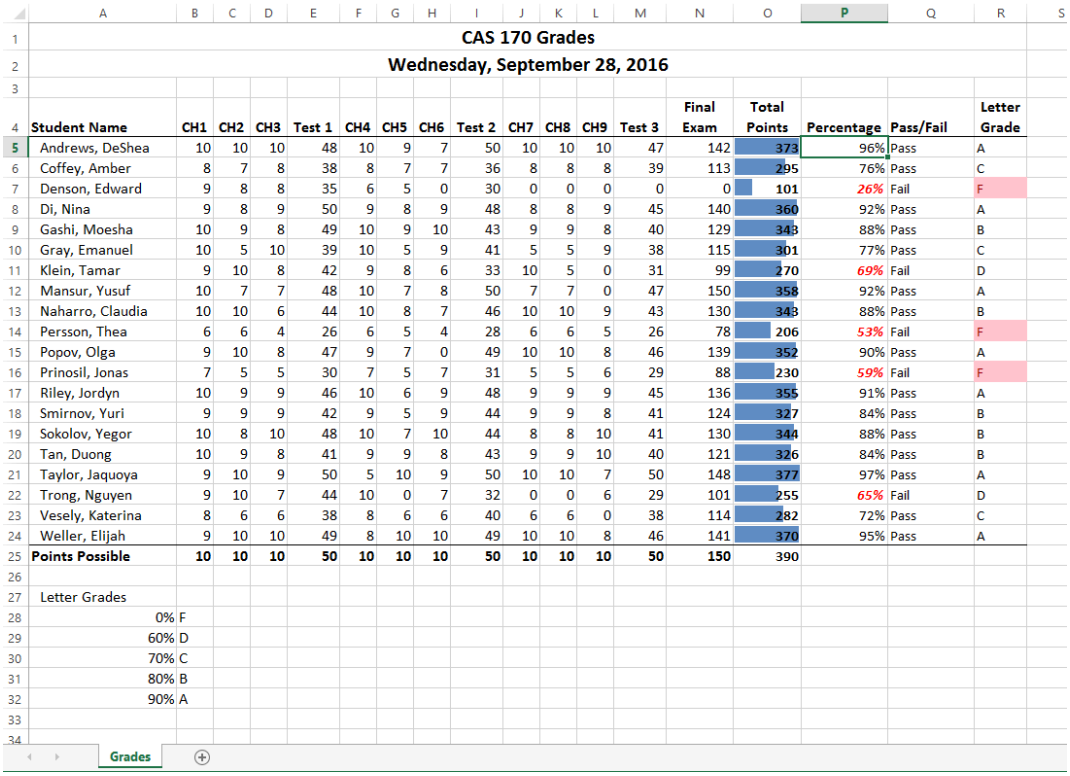


Figure 3.24 Conditional Formatting Edit Formatting Rule Dialog box



CAS 170 Grades

Wednesday, September 28, 2016

Student Name	CH1	CH2	CH3	Test 1	CH4	CH5	CH6	Test 2	CH7	CH8	CH9	Test 3	Final Exam	Total Points	Percentage	Pass/Fail	Letter Grade
Andrews, DeShea	10	10	10	48	10	9	7	50	10	10	10	47	142	373	96%	Pass	A
Coffey, Amber	8	7	8	38	8	7	7	36	8	8	8	39	113	295	76%	Pass	C
Denson, Edward	9	8	8	35	6	5	0	30	0	0	0	0	0	101	26%	Fail	F
Di, Nina	9	8	9	50	9	8	9	48	8	8	9	45	140	360	92%	Pass	A
Gashi, Moesha	10	9	8	49	10	9	10	43	9	9	8	40	129	343	88%	Pass	B
Gray, Emanuel	10	5	10	39	10	5	9	41	5	5	9	38	115	301	77%	Pass	C
Klein, Tamar	9	10	8	42	9	8	6	33	10	5	0	31	99	270	69%	Fail	D
Mansur, Yusuf	10	7	7	48	10	7	8	50	7	7	0	47	150	358	92%	Pass	A
Naharro, Claudia	10	10	6	44	10	8	7	46	10	10	9	43	130	343	88%	Pass	B
Persson, Thea	6	6	4	26	6	5	4	28	6	6	5	26	78	206	53%	Fail	F
Popov, Olga	9	10	8	47	9	7	0	49	10	10	8	46	139	352	90%	Pass	A
Prinosil, Jonas	7	5	5	30	7	5	7	31	5	5	6	29	88	230	59%	Fail	F
Riley, Jordyn	10	9	9	46	10	6	9	48	9	9	9	45	136	355	91%	Pass	A
Smirnov, Yuri	9	9	9	42	9	5	9	44	9	9	8	41	124	327	84%	Pass	B
Sokolov, Yegor	10	8	10	48	10	7	10	44	8	8	10	41	130	344	88%	Pass	B
Tan, Duong	10	9	8	41	9	9	8	43	9	9	10	40	121	326	84%	Pass	B
Taylor, Jaquoya	9	10	9	50	5	10	9	50	10	10	7	50	148	377	97%	Pass	A
Trong, Nguyen	9	10	7	44	10	0	7	32	0	0	6	29	101	255	65%	Fail	D
Vesely, Katerina	8	6	6	38	8	6	6	40	6	6	0	38	114	282	72%	Pass	C
Weller, Elijah	9	10	10	49	8	10	10	49	10	10	8	46	141	370	95%	Pass	A
Points Possible	**10**	**10**	**10**	**50**	**10**	**10**	**10**	**50**	**10**	**10**	**10**	**50**	**150**	390			

Letter Grades	
0%	F
60%	D
70%	C
80%	B
90%	A

Grades

Figure 3.25 Completed Ch3 Gradebook

Setting the Print Area

Before you consider this workbook finished, you need to prepare it for printing. The first thing you will do is set the Print Area so that the table of Letter Grades in A27:B32 does not print.

1. Select A1:R25. This is the only part of the worksheet that you want to have print.
2. On the **Page Layout** ribbon, click the **Print Area** button. Choose **Set Print Area** from the menu.

Next you will preview the worksheet in Print Preview to check that the print area setting worked, as well as make sure it is printing on one page.

1. View the workbook in Print Preview.

   Mac Users should choose **Print** from the **File** menu to view Print Preview.
2. Set the page orientation to **Landscape**.
3. Change the page scaling if needed so that the entire worksheet prints on one page.
4. Save the **CH3 Gradebook and Parks** workbook.

Attribution

5.XLSX.4 PREPARING TO PRINT

Noreen Brown; Barbara Lave; Hallie Puncochar; Julie Romey; Mary Schatz; Art Schneider; Diane Shingledecker; and Jennifer Evans

Learning Objectives

- Locate and fix formatting consistency errors.
- Apply new formatting techniques.
- Use Print Titles to repeat rows and columns on each page of a multiple page worksheet.
- Control where page breaks occur in a multiple page worksheet.

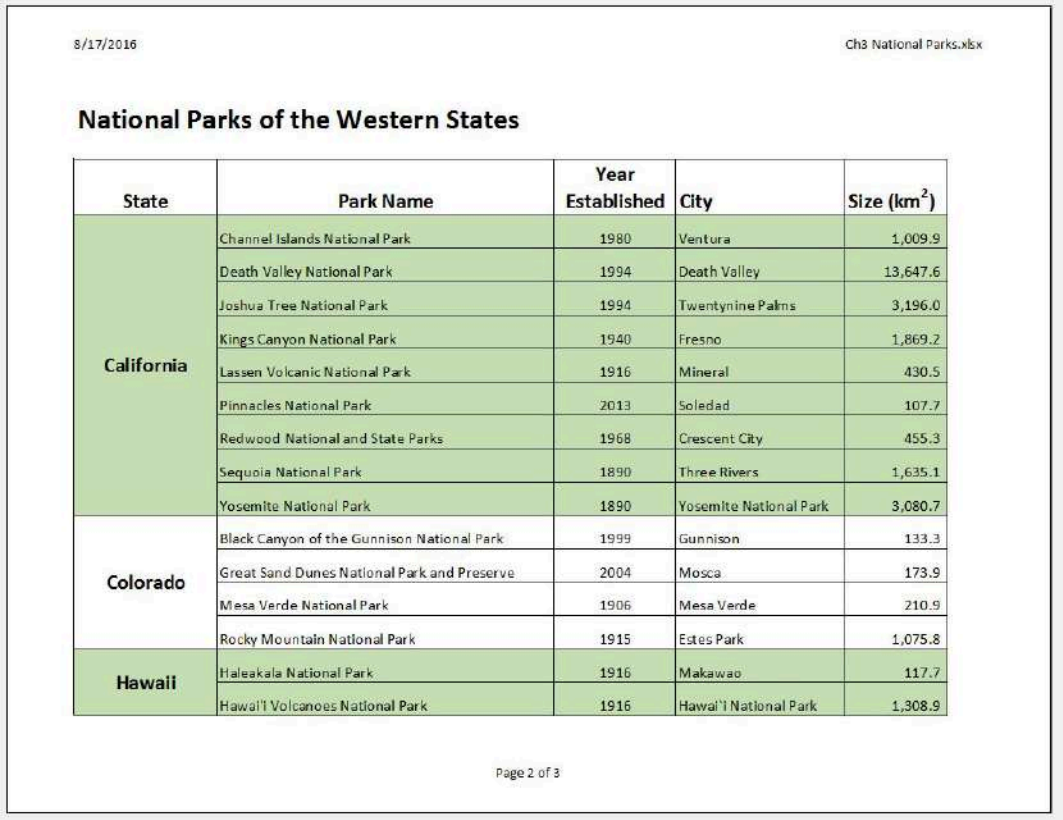
In this section, we will review a worksheet for formatting consistency, as well as learn two new formatting techniques. This worksheet currently prints on four pages, so we will learn new page setup options to control how these pages print. A new data file will be used for this section.

Reviewing Formatting for Consistency

Open the **"CH3-Gradebook and Parks"** workbook if it isn't already open.

*Click on the **"Park Size"** sheet tab within your **"CH3-Gradebook and Parks"** workbook*

You have been given a spreadsheet with data about the national parks in the western United States. Your coworker formatted the workbook and has asked you to review it for consistency. You also need to prepare it for printing. **Figure 3.26** shows how the second page of the finished worksheet will appear in Print Preview.



8/17/2016 Ch3 National Parks.xlsx

National Parks of the Western States

State	Park Name	Year Established	City	Size (km^2)
California	Channel Islands National Park	1980	Ventura	1,009.9
	Death Valley National Park	1994	Death Valley	13,647.6
	Joshua Tree National Park	1994	Twentynine Palms	3,196.0
	Kings Canyon National Park	1940	Fresno	1,869.2
	Lassen Volcanic National Park	1916	Mineral	430.5
	Pinnacles National Park	2013	Soledad	107.7
	Redwood National and State Parks	1968	Crescent City	455.3
	Sequoia National Park	1890	Three Rivers	1,635.1
	Yosemite National Park	1890	Yosemite National Park	3,080.7
Colorado	Black Canyon of the Gunnison National Park	1999	Gunnison	133.3
	Great Sand Dunes National Park and Preserve	2004	Mosca	173.9
	Mesa Verde National Park	1906	Mesa Verde	210.9
	Rocky Mountain National Park	1915	Estes Park	1,075.8
Hawaii	Haleakala National Park	1916	Makawao	117.7
	Hawai'i Volcanoes National Park	1916	Hawai`i National Park	1,308.9

Page 2 of 3

Figure 3.26 Completed National Parks worksheet

Reviewing Formatting for Inconsistencies

The first thing you are going to do is review the worksheet for formatting inconsistencies.

1. Scroll through the worksheet and locate the following formatting errors:

    - The formatting of the Utah label does not match the other states.
    - The Year Established values for Hawaii are not center aligned like the other years.
    - The cells for the Nevada data should have the same green fill color as the other alternating states.
    - The number of digits after the decimal place for the Size values is inconsistent. Also, these values should be formatted with Comma style to make them easier to read.

2. To fix these errors, complete the following steps:

    - Merge & Center A34:A38. Change the font size to 16 and apply Bold format.
    - Center align C28:C29.
    - Apply the green fill color to A31:E31 (be sure to match the green fill color of the other states).
    - Select E4:E43 and apply Comma Style. Use Increase Decimal and/or Decrease Decimal until one

digit appears after the decimal place for all values.

3. While you're fixing errors, proofread the sheet and correct any typos.
4. Finally, let's add color to the two sheet tabs. The use of colored tabs assists in navigating between sheet tabs.

    - Right-click the **"Park Size"** sheet tab (**Mac users** hold down **Ctrl key and click the sheet tab)**
    - Point to **Tab Color** and choose a "**blue**" color.
    - Now right-click the "**Grades"** sheet tab, point to **Tab Color** and choose an "**orange**" color. That's it!

Fine-tuning Formatting

Now that you have fixed the inconsistencies in the formatting, you decide to apply some formatting techniques to make the worksheet look even better. You are going to start by **vertically aligning** the names of the states within the cells.

1. Select A4:A43 (the cells with the state labels).
2. Click the Home tab on the ribbon.
3. In the Alignment group, click the **Middle Align** button (see **Figure 3.26**). Notice that the names of the states are now centered between the top and bottom borders of the cells.

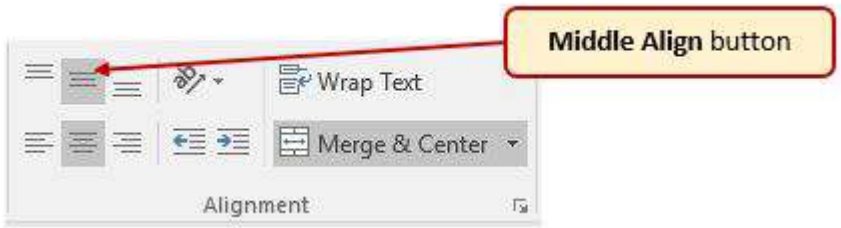
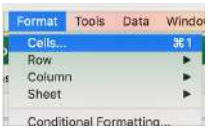


Figure 3.26
Alignment Group

The next new formatting skill is to change the label in E3 from Size (km2) to Size (km^2) with the 2 after km formatted with **superscript**.

1. Double-click on cell E3 to enter Edit mode
2. Select just the **2** (be careful not to select anything else).
3. On the ribbon (Home tab) click the dialog box launcher arrow in the Font group.

Mac Users: there is no dialog box launcher for Excel for Mac. Instead, choose **Format** from the Menu Bar, click **Cells**: 

then continue with Steps 4 and 5

4. In the Effects section of the Format Cells dialog box, check the box for **Superscript** (see **Figure 3-27**). Click OK.
5. Save the **CH3 Gradebook and Parks** file.

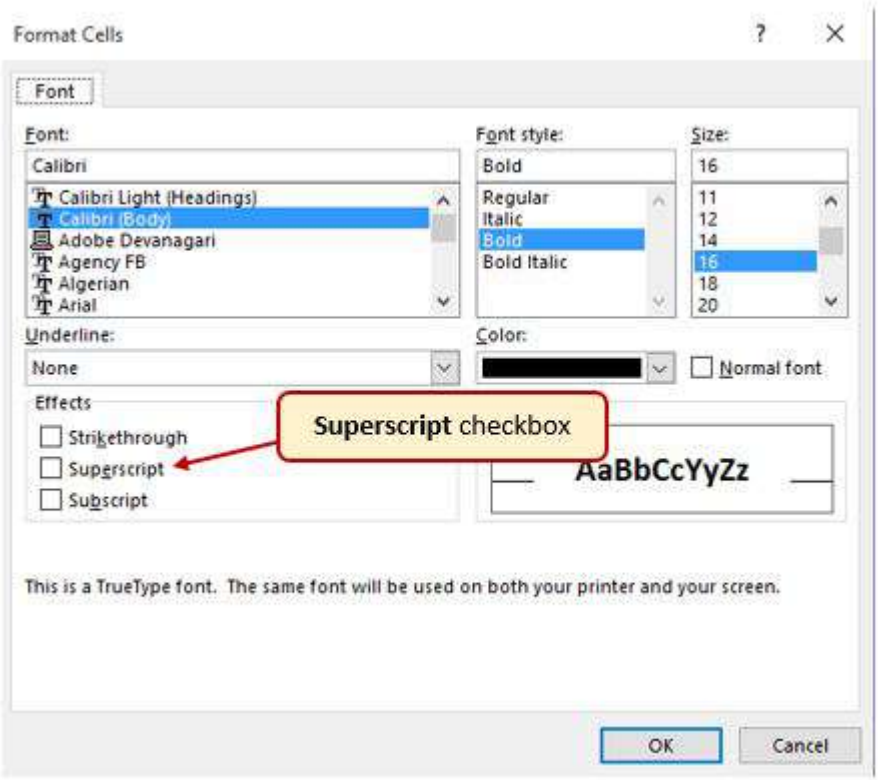


Figure 3.27 Font Tab in Format Cells Dialog Box

Repeating Column (and Row) Labels

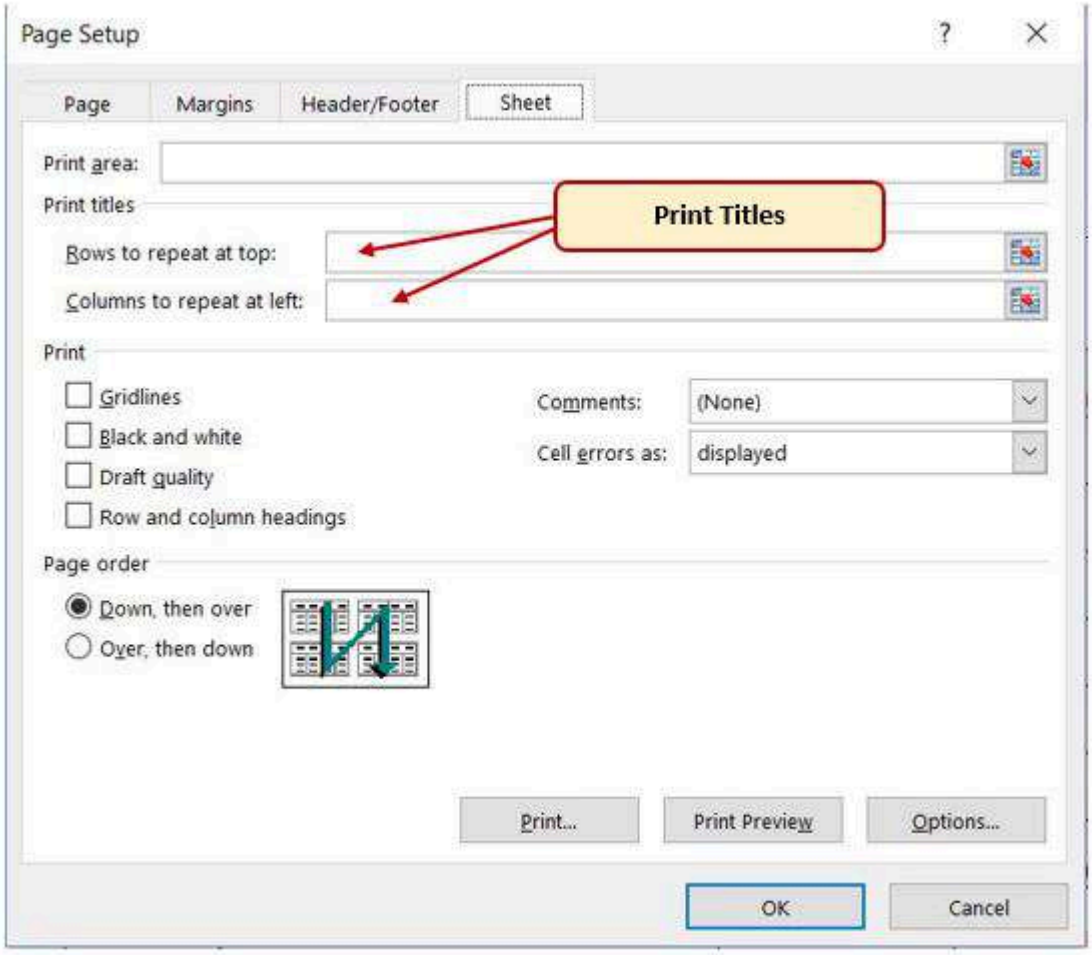
Now that you have fixed the cell and text formatting, you are ready to review the worksheet in Print Preview. You will notice that the worksheet is printing on multiple pages, and you cannot tell what each column of data represents on some of the pages.

1. With the ***CH3-Gradebook and Parks*** file still open, and the Parks tab selected, go to Backstage View by clicking the File tab on the ribbon. Select Print from the menu.

   Mac Users: choose **File** from the Menu Bar, and then choose **Print**
2. Click through each of the pages. The worksheet is currently printing on four pages (Mac users may only see three pages but that is ok), with the City and Sizes columns printing on separate pages from the

rest of the data.

3. Change the Orientation from Portrait to Landscape. This fits all of the columns on one page. All of the columns are now on the same page, but the second and third pages have no column labels to identify what information is in each column. You are going to use **Print Titles** to repeat the first three rows of the worksheet on each of the printed pages. To set **Print Titles** you need to exit Print Preview.
4. Exit Backstage View then click the Page Layout tab on the ribbon.
5. Click the Print Titles button in the Page Setup group on the ribbon. The dialog box shown in **Figure 3.28** should appear.
6. Click the Sheet tab if necessary.



7. Click in the **Rows to repeat at top:** box. *Be sure your insertion point is blinking in that box before moving on to the next step.*
8. In the worksheet, select Rows 1 through 3. The text **$1:$3** should now appear in the **Rows to repeat at top:** box.
9. Click OK.

You will not see a change to the worksheet in Normal view, so you will need to return to Print Preview. While looking in Print Preview, you will notice that the pages are breaking in inconvenient places.

1. Go to Print Preview and look at each of the pages. Notice that the first three rows are now repeated at the top of each page.
2. Exit Backstage View.

Skill Refresher

Creating Print Titles

1. Open the Page Setup dialog box and click the Sheet tab.
2. Click in the Rows to repeat at top: box or the Columns to repeat at left: box.
3. Click in the worksheet and select the row(s) or column(s) that you want to repeat on each page.

Inserting Page Breaks

Notice that the data for California is split between the first and second pages. You want all of the data for each state to be together on the same page, so you need to control the page breaks. You are going to start by inserting a page break before the California data to force it to start on the second page, then you will move the page break for the third page if needed. To make these changes you are going to work in **Page Break Preview**.

1. Click the View tab on the ribbon then click **Page Break Preview** in the Workbook Views Group. Your screen should look similar to **Figure 3.29**.

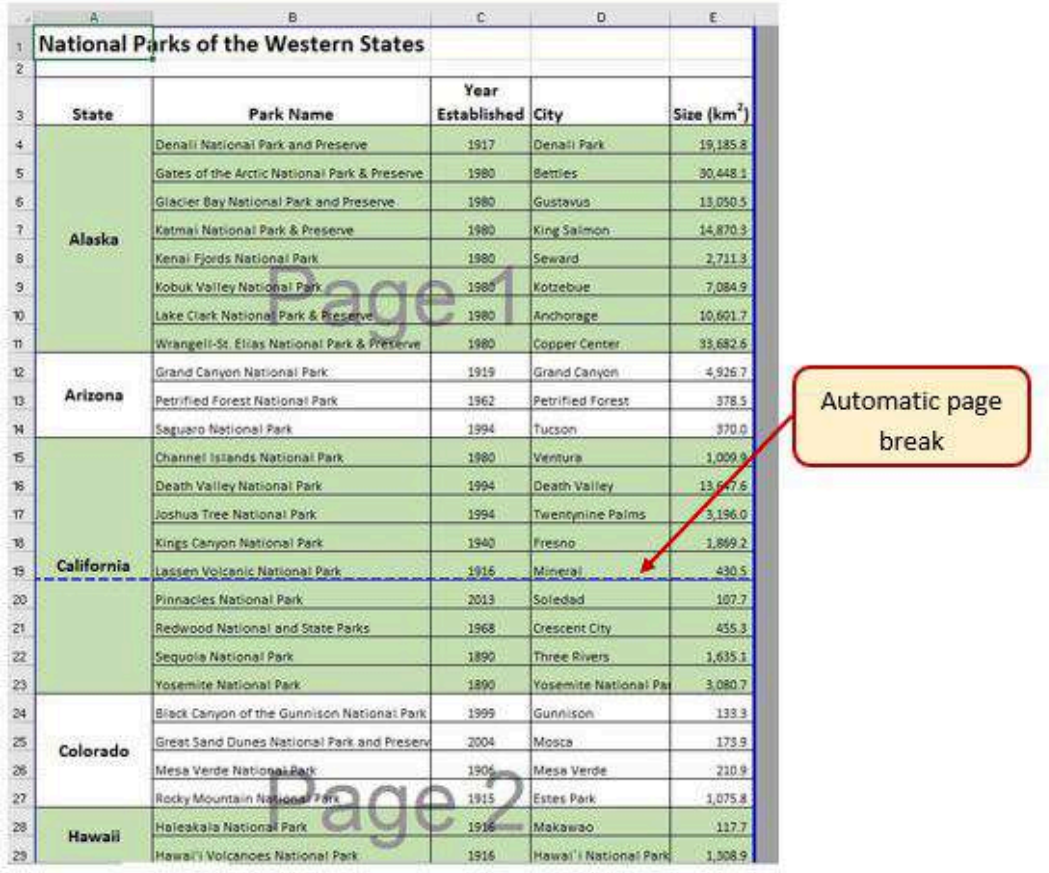


Figure 3.29 Page Break Preview

Mac Users: in the next paragraph below, the location of the automatic page breaks may be in different locations. That's ok.

In Page Break Preview, automatic page breaks are displayed as dotted blue lines. Notice the dotted blue lines after rows 13 and 28. These lines indicate where Excel will start a new page. For this worksheet, you want the first page to break before the California data, so you are going to insert a **manual page break**.

1. Select cell A15. When inserting a page break, you select the cell below where you want the page break to appear.
2. Click the Page Layout tab on the ribbon.
3. Click the Breaks button in the Page Setup group (see **Figure 3.30**).
4. Select **Insert Page Break** from the menu. There is now a solid blue line after row 14, which indicates a manual page break that was inserted.
5. Go to Print Preview. Notice that the California data now starts on the second page.

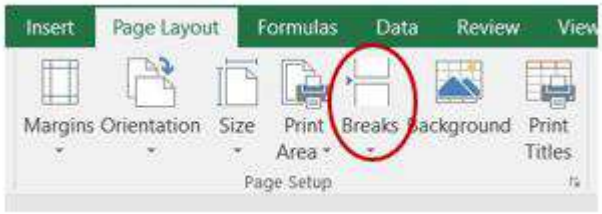


Figure 3.30 Breaks Button on Page Layout tab

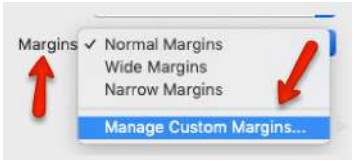
While looking at each page in Print Preview you decide that the third page should start with Montana. To make this change you are going to move the automatic page break that appears after Nevada.

1. Exit Backstage View. Switch back to Page Break Preview if needed.
2. Locate the next dotted blue line (automatic page break).
3. Put your pointer over the dotted blue line and it will switch to a vertical double-headed arrow. Click on the dotted blue line and drag it **above** Montana.
4. Release the mouse button when the line is above row 30 (above Montana). The line will now be a solid blue line, indicating a manual page break.
5. Go to Print Preview. The Montana data now appears at the top of the third page.

While evaluating the pages in Print Preview you decide that there is too much white space at the bottom of the pages. To fix this, you are going to center the contents vertically on the pages.

1. Click the Page Setup link at the bottom of the Settings section of Backstage View to open the Page Setup dialog box.

   Mac Users: there is no "Page Setup link" in Print Preview for Excel for Mac. Click the Margins list arrow instead, and choose "Manage Custom Margins" then continue with the steps below.



2. Click on the Margins tab.
3. In the Center on page section, check the box for Vertically then click OK.
4. Review each page in Print Preview to see the changes. Exit Backstage View.

Creating a Header and Footer using Page Layout View

Now that the worksheet is printing on three pages, with page breaks in appropriate places, you are ready to

add a header with the current date and filename. You will also add a footer with the page number and the total number of pages that will appear as **Page 1 of 3**. You are going to edit the header and footer in Page Layout View.

1. Click the View tab on the ribbon and click the Page Layout button in the Workbook Views group.
2. The white space at the top of the worksheet should say Add header. Place the mouse pointer over the left section of the Header and click to activate that section.

   Mac Users should make sure the mouse pointer turns into a small page icon  then click in the left section of the Header
3. Click the Header & Footer Tools Design tab on the ribbon.
4. Click the **Current Date** button in the Header & Footer Elements group (see **Figure 3.31**). Inserting the date this way will insert a field that will update every time the workbook is opened.
5. Click in the right section of the Header. Click the **Filename** button in the Header & Footer Elements group (see **Figure 3.31**). Inserting the filename this way will insert a field that will update if the filename is changed.
6. Click the **Go to Footer** button in the Navigation group of commands.
7. In the center section of the footer, type the word *Page* with a space after it.
8. Click the **Page Number** button in the Header & Footer Elements group (see **Figure 3.31**), then type a space after the **&[Page]** code that appears.
9. Type the word *of* with a space after it, then click the **Number of Pages** button in the Header & Footer Elements group (see **Figure 3.31**). The footer should match **Figure 3.32**.
10. Click anywhere on the worksheet to close the Footer editing.
11. Review the worksheet again in Print Preview. Pay careful attention to the page numbers in the footer to ensure they will print correctly, then exit Backstage View.
12. View the correct print preview screenshot below in Figure 3.33
13. Check the spelling on all of the worksheets and make any necessary changes. Save and submit the ***CH3-Gradebook and Parks*** workbook.

5.XLSX.5 CHAPTER PRACTICE

Noreen Brown; Barbara Lave; Hallie Puncochar; Julie Romey; Mary Schatz; Art Schneider; Diane Shingledecker; and Jennifer Evans

Household Budget

Download Data File: PR3 Data

Etta and Lucian Redding are a recently married couple living in Portland, Oregon. Lucian works part time and attends the local community college. Etta works as a marketing manager at a clothing company in North Portland. They are trying to decide if they can afford to move to a better apartment, one that is closer to work and school. They want to use Excel to examine their household budget. They have started their budget spreadsheet, but they need your help with it.

1. Open the file named **PR3 Data** and then save it as **PR3 Redding**.
2. Insert two new rows at the top of the worksheet.
3. Enter the following text:

A2	Category
B2	Item
C2	January
O2	Yearly Total (adjust column width as needed to fit this text)

1. Using the text in cell C2, use Autofill to fill in the months February through December in cells D2:N2. Adjust column widths as needed to fit the names of the months in these columns.
2. Bold and center align all of the headings in Row 2.
3. Type "**Redding Family Budget"** in A1. Merge & Center A1:O1. Make this text 22 point **bold**.
4. Next you need to complete the monthly values for some of the income and expense items. In the rows for Income #1, Income #2, Mortgage/Rent, Homeowners/Rent Insurance, Car Insurance, Car Payment, and Gym Fees/Memberships, copy the values for January to the cells for February through December.
5. Use the Totals tab in the Quick Analysis tool to add the SUM to Column O. Delete the formulas from O7, O17, O24, O32, and O38.

   Mac Users should use the AutoSum tool to calculate the totals in Column O. Since you are using the AutoSum tool, you may not have to delete any formulas in the cells listed in Step 5 above. Also, the

Quick Analysis tool will automatically bold the values in Column O. Mac Users should bold cells O3:O45.

6. In C6: N6, use the SUM function to calculate the Total Income for each month.
7. Similar to step 6, use the SUM function to calculate the Total Home Expenses, Total Daily Living Expenses, Total Transportation Expenses, Total Entertainment Expenses, and Total Personal Expenses for each month.
8. Use the SUM function to calculate the Yearly Total Personal Expenses in cell O45.
9. Format the numerical data in Row 3 as Currency with no decimal places, and with a top border. Format all the total rows as Currency with no decimal places and with a top border (Rows 6, 16, 23, 31, 37, and 45).
10. Apply the Comma format with no decimal places in all the other rows.
11. In A47, type "**Total Expenses**".
12. In C47, enter a formula that adds together all of the **expense category totals** for January. Copy the formula in C47 to D47:O47.
13. In A49, type "**Net Income**". Bold and indent this text. Also bold C49:N49
14. In C49, enter a formula that calculates the difference between Total Income and Total Expenses (=Total Income-Total Expenses) for January. Copy this formula to D49:O49.
15. Format the data in Rows 47 and 49 as Currency with no decimal places. Bold O47 and O49. Add a Top and Double Bottom Border to the data in Row 49.
16. Select C49:N49. Use the Quick Analysis tool to add data bars to this data.  Mac Users should use the Conditional Formatting tool on the Ribbon. The negative values should automatically have a red data bar and the positive values will have a blue data bar.
17. In B50, type "**New Home?**". Enter an IF statement in C50 that displays the word "No" if the amount in C49 is less than or equal to zero and "Maybe" if the amount is greater than zero. Copy C50 to D50:N50.
18. Check to see if your IF statement worked correctly in row 50. If the cells say "No" when the data bar in the cell above it is red and "Maybe" when the data bar in the cell above it is blue, your IF statement is correct.
19. Review the worksheet in Print Preview. Make any changes needed to make the worksheet print on one page with landscape orientation.
20. Rename the **"Sheet 1"** sheet tab:

    - Double-click the "Sheet 1" tab
    - Type: **Budget** and press Enter

21. Change the color of the sheet tab:

- Right-click the "Budget" sheet tab
- Mac Users should hold down the **CTRL key** and click the Budget sheet tab
- Point to "Tab Color", choose a green color

22. Check the spelling on all of the worksheets and make any necessary changes. Save the **PR3 Redding** workbook.
23. Compare your work with the self-check answer key **below** and then submit the **PR3 Redding** workbook as directed by your instructor.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1							Redding Family Budget									
2	Category	Item	January	February	March	April	May	June	July	August	September	October	November	December	Yearly Total	
3	Income	Income #1	$1,645	$1,645	$1,645	$1,645	$1,645	$1,645	$1,645	$1,645	$1,645	$1,645	$1,645	$1,645	$19,740	
4		Income #2	2,010	2,010	2,010	2,010	2,010	2,010	2,010	2,010	2,010	2,010	2,010	2,010	24,120	
5		Other Income													-	
6	Total Income		$3,655	$3,655	$3,655	$3,655	$3,655	$3,655	$3,655	$3,655	$3,655	$3,655	$3,655	$3,655	$43,860	
7																
8	Home	Mortgage/Rent	920	920	920	920	920	920	920	920	920	920	920	920	11,040	
9		Homeowners/Rent Insurance	8	8	8	8	8	8	8	8	8	8	8	8	96	
10		Utilities	255	230	200	195	150	165	175	165	160	160	200	235	2,290	
11		Cable/Internet													-	
12		Trash													-	
13		Home Maintenance													-	
14		Misc. House Stuff													-	
15		Lawn/Garden													-	
16	Total Home Expenses		$1,183	$1,158	$1,128	$1,123	$1,078	$1,093	$1,103	$1,093	$1,088	$1,088	$1,128	$1,163	$13,426	
17																
18	Daily Living	Groceries	875	730	795	825	855	815	885	920	942	875	975	1,050	10,542	
19		Clothing													-	
20		Books & Supplies	2,115					2,419		3,275					7,809	
21		Child Care													-	
22		Pets													-	
23	Total Daily Living Expenses		$2,990	$730	$795	$825	$855	$3,234	$885	$4,195	$942	$875	$975	$1,050	$18,351	
24																
25	Transportation	Car Insurance	265	265	265	265	265	265	265	265	265	265	265	265	3,180	
26		Car Maintenance													-	
27		Car Payments	210	210	210	210	210	210	210	210	210	210	210	210	2,520	
28		Bus/Train Fair													-	
29		Gas (Car)													-	
30		Parking													-	
31	Total Transportation Expenses		$475	$475	$475	$475	$475	$475	$475	$475	$475	$475	$475	$475	$5,700	
32																
33	Entertainment	Travel													-	
34		Movies													-	
35		Special Occasions													-	
36		Gifts													-	
37	Total Entertainment Expenses		$0	$0	$0	$0	$0	$0	$0	$0	$0	$0	$0	$0	$0	
38																
39	Personal	Subscriptions													-	
40		Phone	120	131	125	138	120	145	140	135	145	135	165	175	1,674	
41		Gym Fees/Memberships	60	60	60	60	60	60	60	60	60	60	60	60	720	
42		Salon/Barber													-	
43		Eating Out													-	
44		Miscellaneous	200	250	120	180	250	120	150	135	200	250	270	350	2,475	
45	Total Personal Expenses		$380	$441	$305	$378	$430	$325	$350	$330	$405	$445	$495	$585	$4,869	
46																
47	Total Expenses		$5,028	$2,804	$2,703	$2,801	$2,838	$5,127	$2,813	$6,093	$2,910	$2,883	$3,073	$3,273	$42,346	
48																
49	Net Income		-$1,373	$851	$952	$854	$817	-$1,472	$842	-$2,438	$745	$772	$582	$382	$1,514	
50		New Home?	No	Maybe	Maybe	Maybe	Maybe	No	Maybe	No	Maybe	Maybe	Maybe	Maybe		
51																

Budget

Attribution

5.XLSX.6 CHAPTER SCORED

Noreen Brown; Barbara Lave; Hallie Puncochar; Julie Romey; Mary Schatz; Art Schneider; Diane Shingledecker; and Jennifer Evans

MidasCoffee Company

Download Data File: SC3 data

MidasCoffee: Ruth Kobran owns a coffee supply company named MidasCoffee. She needs some help writing the formulas for the order form she uses to invoice customers. You will need to write the formulas for all of the calculations on the form. Some of the more complex parts are determining if the customer will get a discount (based on the customer status) as well as the shipping charge (orders over $199 get free shipping). You will use IF functions for both of those calculations.

1. Open the **SC3 Data** workbook and save the workbook as **SC3 MidasCoffee**.
2. Enter the following order information:
   Order #: **56894**
   Order Date: use a function that displays the current date
3. Enter the following Billing Information:
   Samantha Raitt
   4270 SW Cooper Ln, Portland, OR 97225
   503-674-1632
   samantha.raitt@zmail.com
4. For the Shipping Information, create formulas using cell references to display the corresponding information from the Billing Information section. For example, the Customer cell will display the name of the customer in cell C11.
5. In the range B19:E22, enter the following item orders:

Item #	Description	Qty	Unit Price
K56	Dark Mocha K-Cups (12 pack)	1	11.99
G03	Decaf Dark Roast – Ground (1 lb.)	3	12.99
B07	Organic Dark Roast – Whole Bean (1 lb.)	2	14.99
K52	Chai Latte K-Cups (12 pack)	3	10.99

6. In cell F19, enter an IF function that tests whether the order quantity in cell D19 is greater than 0 (zero). lf it is, return the value of the Qty (in D19) multiplied by the Unit Price (in E19); otherwise, return no text by entering "". *Hint: You will need to use a formula for the Value if True argument.*
7. Copy/fill this formula into the other cells in the range F19:F25. *Hint: be sure to copy the formula to all of the Item Total cells, even if it is a blank row. You want the worksheet to be prepared for orders with more items in the future.*
8. In cell F26, calculate the sum of all of the Item Total cells.
9. In cell F27, use an IF function to calculate the discount amount for this order based on the customer's status (which is found in F16). If the customer's status is Preferred, the discount amount will be the Order Subtotal times the discount percentage found in cell B29; otherwise the discount amount will be 0 (zero). *Hint: You will need to use a formula for the Value if True argument.*
10. Test your IF function to make sure that it still works if the customer is NOT preferred by deleting the word Preferred in F16. Make sure you do not end up with an error message! If you get an error message, check the IF function and make the changes needed.
11. Calculate the Discounted Total for this order in cell F28. *Hint: Use a simple subtraction formula.*
12. In cell F29, use an IF function to display the correct Shipping Charge, based on the amount of the Discounted Total. If the Discounted Total is greater than or equal to the Free Shipping Minimum found in cell B28, the Shipping Charge is 0 (zero); otherwise, the Shipping Charge is 5% of the Discounted Total. *Hint: You will need to use a formula for the Value if False to calculate what 5% of the Discounted Total will be.*
13. Calculate the Invoice Total in cell F31. *Hint: This will be the total of the Discounted Total and the Shipping Charge.*
14. Take a critical look at your worksheet to ensure that all of the number and cell formatting is professional.
15. Review the worksheet in Print Preview. Make any changes needed to make the worksheet print on one page.
16. Check the spelling on all of the worksheets and make any necessary changes. Save the **SC3 MidasCoffee** workbook.
17. Submit the **SC3 Midas Coffee** workbook as directed by your instructor.

Attribution

SECTION X

6. MEASURES OF VARIATION

6.1 Describing Variability

6.1 DESCRIBING VARIABILITY

6.1: Describing Variability

6.1.1: Range

The range is a measure of the total spread of values in a quantitative dataset.

Learning Objective

Interpret the range as the overall dispersion of values in a dataset

Key Takeaways

Key Points

- Unlike other more popular measures of dispersion, the range actually measures total dispersion (between the smallest and largest values) rather than relative dispersion around a measure of central tendency.
- The range is measured in the same units as the variable of reference and, thus, has a direct interpretation as such.
- Because the information the range provides is rather limited, it is seldom used in statistical analyses.

- The mid-range of a set of statistical data values is the arithmetic mean of the maximum and minimum values in a data set.

Key Terms

dispersion
: the degree of scatter of data

range
: the length of the smallest interval which contains all the data in a sample; the difference between the largest and smallest observations in the sample

In statistics, the range is a measure of the total spread of values in a quantitative dataset. Unlike other more popular measures of dispersion, the range actually measures total dispersion (between the smallest and largest values) rather than relative dispersion around a measure of central tendency.

Interpreting the Range

The range is interpreted as t*he overall dispersion* of values in a dataset or, more literally, as the *difference between the largest and the smallest value* in a dataset. The range is measured in the same units as the variable of reference and, thus, has a direct interpretation as such. This can be useful when comparing similar variables but of little use when comparing variables measured in different units. However, because the information the range provides is rather limited, it is seldom used in statistical analyses.

For example, if you read that the age range of two groups of students is 3 in one group and 7 in another, then you know that the second group is more spread out (there is a difference of seven years between the youngest and the oldest student) than the first (which only sports a difference of three years between the youngest and the oldest student).

Mid-Range

The mid-range of a set of statistical data values is the arithmetic mean of the maximum and minimum values in a data set, defined as:

M=xmax+xmin2">$M = \frac{X_{max}+X_{min}}{2}$

The mid-range is the midpoint of the range; as such, it is a measure of central tendency. The mid-range is rarely used in practical statistical analysis, as it lacks efficiency as an estimator for most distributions of

interest because it ignores all intermediate points. The mid-range also lacks robustness, as outliers change it significantly. Indeed, it is one of the least efficient and least robust statistics.

However, it finds some use in special cases:

- It is the maximally efficient estimator for the center of a uniform distribution
- Trimmed mid-ranges address robustness
- As an L-estimator, it is simple to understand and compute.

6.1.2: Variance

Variance is the sum of the probabilities that various outcomes will occur multiplied by the squared deviations from the average of the random variable.

Learning Objective

Calculate variance to describe a population

Key Takeaways

Key Points

- When determining the "spread" of the population, we want to know a measure of the possible distances between the data and the population mean.
- When trying to determine the risk associated with a given set of options, the variance is a very useful tool.
- When dealing with the complete population the (population) variance is a constant, a

parameter which helps to describe the population.
- When dealing with a sample from the population the (sample) variance is actually a random variable, whose value differs from sample to sample.

Key Terms

deviation
: For interval variables and ratio variables, a measure of difference between the observed value and the mean.

spread
: A numerical difference.

When describing data, it is helpful (and in some cases necessary) to determine the spread of a distribution. In describing a complete population, the data represents all the elements of the population. When determining the spread of the population, we want to know a measure of the possible distances between the data and the population mean. These distances are known as deviations.

The variance of a data set measures the average square of these deviations. More specifically, the variance is the sum of the probabilities that various outcomes will occur multiplied by the squared deviations from the average of the random variable. When trying to determine the risk associated with a given set of options, the variance is a very useful tool.

Calculating the Variance

Calculating the variance begins with finding the mean. Once the mean is known, the variance is calculated by finding the average squared deviation of each number in the sample from the mean. For the numbers 1, 2, 3, 4, and 5, the mean is 3. The calculation for finding the mean is as follows:

1+2+3+4+55=155=3″>$\frac{1+2+3+4+5}{5} = \frac{15}{3} = 3$

Once the mean is known, the variance can be calculated. The variance for the above set of numbers is:

σ2=(1−3)2+(2−3)2+(3−3)2+(4−3)2+(5−3)25″>$\sigma^2 = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5}$

σ2=(−2)2+(−1)2+(0)2+(1)2+(2)25″>$\sigma^2 = \frac{(-2)^2+(-1)^2+(0)^2+(1)^2+(2)^2}{5}$

σ2=4+1+0+1+45″>$\sigma^2 = \frac{4+1+0+1+4}{5}$

σ2=105=2″>$\sigma^2 = \frac{10}{5} = 2$

A clear distinction should be made between dealing with the population or with a sample from it. When dealing with the complete population the (population) variance is a constant, a parameter which helps to

describe the population. When dealing with a sample from the population the (sample) variance is actually a random variable, whose value differs from sample to sample.

Population of Cheetahs

The population variance can be very helpful in analyzing data of various wildlife populations.

6.1.3: Standard Deviation: Definition and Calculation

Standard deviation is a measure of the average distance between the values of the data in the set and the mean.

Learning Objective

Contrast the usefulness of variance and standard deviation

Key Takeaways

Key Points

- A low standard deviation indicates that the data points tend to be very close to the mean; a high standard deviation indicates that the data points are spread out over a large range of values.
- In addition to expressing the variability of a population, standard deviation is commonly used to measure confidence in statistical conclusions.
- To calculate the population standard deviation, first compute the difference of each data point from the mean, and square the result of each. Next, compute the average of these values, and take the square root.
- The standard deviation is a "natural" measure of statistical dispersion if the center of the data is measured about the mean because the standard deviation from the mean is smaller than from any other point.

Key Terms

normal distribution
: A family of continuous probability distributions such that the probability density function is the normal (or Gaussian) function.

coefficient of variation
: The ratio of the standard deviation to the mean.

mean squared error

A measure of the average of the squares of the "errors"; the amount by which the value implied by the estimator differs from the quantity to be estimated.

standard deviation

a measure of how spread out data values are around the mean, defined as the square root of the varianc

Example

The average height for adult men in the United States is about 70 inches, with a standard deviation of around 3 inches. This means that most men (about 68%, assuming a normal distribution) have a height within 3 inches of the mean (67–73 inches) – one standard deviation – and almost all men (about 95%) have a height within 6 inches of the mean (64–76 inches) – two standard deviations. If the standard deviation were zero, then all men would be exactly 70 inches tall. If the standard deviation were 20 inches, then men would have much more variable heights, with a typical range of about 50–90 inches. Three standard deviations account for 99.7% of the sample population being studied, assuming the distribution is normal (bell-shaped).

Since the variance is a squared quantity, it cannot be directly compared to the data values or the mean value of a data set. It is therefore more useful to have a quantity that is the square root of the variance. The standard error is an estimate of how close to the population mean your sample mean is likely to be, whereas the standard deviation is the degree to which individuals within the sample differ from the sample mean. This quantity is known as the standard deviation.

Standard deviation (represented by the symbol sigma, σ">σ) shows how much variation or dispersion exists from the average (mean), or expected value. More precisely, it is a measure of the average distance between the values of the data in the set and the mean. A low standard deviation indicates that the data points tend to be very close to the mean; a high standard deviation indicates that the data points are spread out over a large range of values. A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data.

In statistics, the standard deviation is the most common measure of statistical dispersion. However, in addition to expressing the variability of a population, standard deviation is commonly used to measure

confidence in statistical conclusions. For example, the margin of error in polling data is determined by calculating the expected standard deviation in the results if the same poll were to be conducted multiple times.

Basic Calculation

Consider a population consisting of the following eight values:

2, 4, 4, 4, 5, 5, 7, 9

These eight data points have a mean (average) of 5:

2+4+4+4+5+5+7+98=5">$\frac{2+4+4+4+5+5+7+9}{8} = 5$

To calculate the population standard deviation, first compute the difference of each data point from the mean, and square the result of each:

(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">$(2-5)^2=9$

(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">$(4-5)^2=1$

(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">$(4-5)^2=1$

(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">$(4-5)^2=1$

(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">$(5-5)^2=0$

(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">$(5-5)^2=0$

(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">$(7-5)^2=4$

(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">(2−5)2=9(4−5)2=1(4−5)2=1(4−5)2=1(5−5)2=0(5−5)2=0(7−5)2=4(9−5)2=16">$(9-5)^2=16$

Next, compute the average of these values, and take the square root:

9+1+1+1+0+0+4+168=2">$\sqrt{\frac{9+1+1+1+0+0+4+16}{8}} = 2$

This quantity is the population standard deviation, and is equal to the square root of the variance. The formula is valid only if the eight values we began with form the complete population. If the values instead were a random sample drawn from some larger parent population, then we would have divided by 7 (which is n−1">n−1) instead of 8 (which is n">n) in the denominator of the last formula, and then the quantity thus obtained would be called the sample standard deviation.

Estimation

The sample standard deviation, s">s, is a statistic known as an estimator. In cases where the standard deviation of an entire population cannot be found, it is estimated by examining a random sample taken from the population and computing a statistic of the sample. Unlike the estimation of the population mean, for which the sample mean is a simple estimator with many desirable properties (unbiased, efficient, maximum likelihood), there is no single estimator for the standard deviation with all these properties. Therefore, unbiased estimation of standard deviation is a very technically involved problem.

As mentioned above, most often the standard deviation is estimated using the corrected sample standard deviation (using N−1">N−1). However, other estimators are better in other respects:

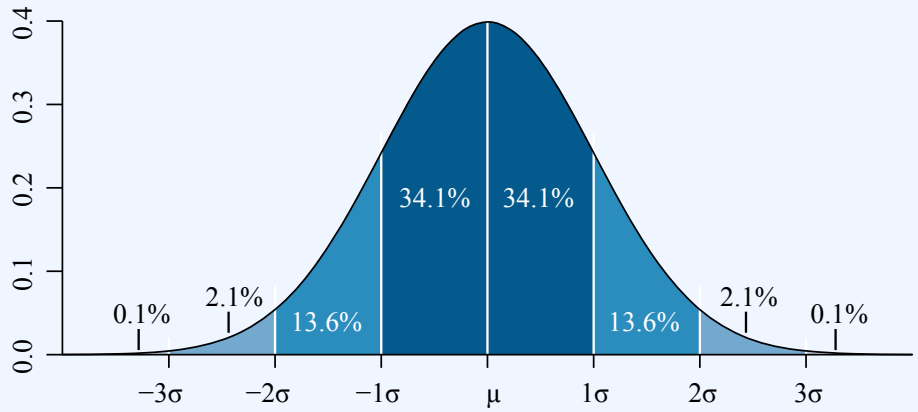
- Using the uncorrected estimator (using N">N) yields lower mean squared error.
- Using N−1.5">N−1.5 (for the normal distribution) almost completely eliminates bias.

Relationship with the Mean

The mean and the standard deviation of a set of data are usually reported together. In a certain sense, the standard deviation is a “natural” measure of statistical dispersion if the center of the data is measured about the mean. This is because the standard deviation from the mean is smaller than from any other point. Variability can also be measured by the coefficient of variation, which is the ratio of the standard deviation to the mean.

Often, we want some information about the precision of the mean we obtained. We can obtain this by determining the standard deviation of the sample mean, which is the standard deviation divided by the square root of the total amount of numbers in a data set:

σmean=σN">$\sigma_{mean} = \frac{\sigma}{\sqrt{N}}$



Standard Deviation Diagram

Dark blue is one standard deviation on either side of the mean. For the normal distribution, this accounts for 68.27 percent of the set; while two standard deviations from the mean (medium and dark blue) account for 95.45 percent; three standard deviations (light, medium, and dark blue) account for 99.73 percent; and four standard deviations account for 99.994 percent.

6.1.4: Interpreting the Standard Deviation

The practical value of understanding the standard deviation of a set of values is in appreciating how much variation there is from the mean.

Learning Objective

Derive standard deviation to measure the uncertainty in daily life examples

Key Takeaways

Key Points

- A large standard deviation indicates that the data points are far from the mean, and a small standard deviation indicates that they are clustered closely around the mean.
- When deciding whether measurements agree with a theoretical prediction, the standard deviation of those measurements is of crucial importance.
- In finance, standard deviation is often used as a measure of the risk associated with price-fluctuations of a given asset (stocks, bonds, property, etc.), or the risk of a portfolio of assets.

Key Terms

disparity
: the state of being unequal; difference

standard deviation
: a measure of how spread out data values are around the mean, defined as the square root of the variance

Example

In finance, standard deviation is often used as a measure of the risk associated with price-fluctuations of a given asset (stocks, bonds, property, etc.), or the risk of a portfolio of assets. Risk is an important factor in determining how to efficiently manage a portfolio of investments because it determines the variation in returns on the asset and/or portfolio and gives investors a mathematical basis for investment decisions. When evaluating investments, investors should estimate both the expected return and the uncertainty of future returns. Standard deviation provides a quantified estimate of the uncertainty of future returns.

A large standard deviation, which is the square root of the variance, indicates that the data points are far from the mean, and a small standard deviation indicates that they are clustered closely around the mean. For example, each of the three populations {0,0,14,14}">{0,0,14,14}, {0,6,8,14}">{0,6,8,14}, and {6,6,8,8}">{6,6,8,8} has a mean of 7. Their standard deviations are 7, 5, and 1, respectively. The third population has a much smaller standard deviation than the other two because its values are all close to 7.

Standard deviation may serve as a measure of uncertainty. In physical science, for example, the reported standard deviation of a group of repeated measurements gives the precision of those measurements. When deciding whether measurements agree with a theoretical prediction, the standard deviation of those measurements is of crucial importance. If the mean of the measurements is too far away from the prediction (with the distance measured in standard deviations), then the theory being tested probably needs to be revised. This makes sense since they fall outside the range of values that could reasonably be expected to occur, if the prediction were correct and the standard deviation appropriately quantified.

Application of the Standard Deviation

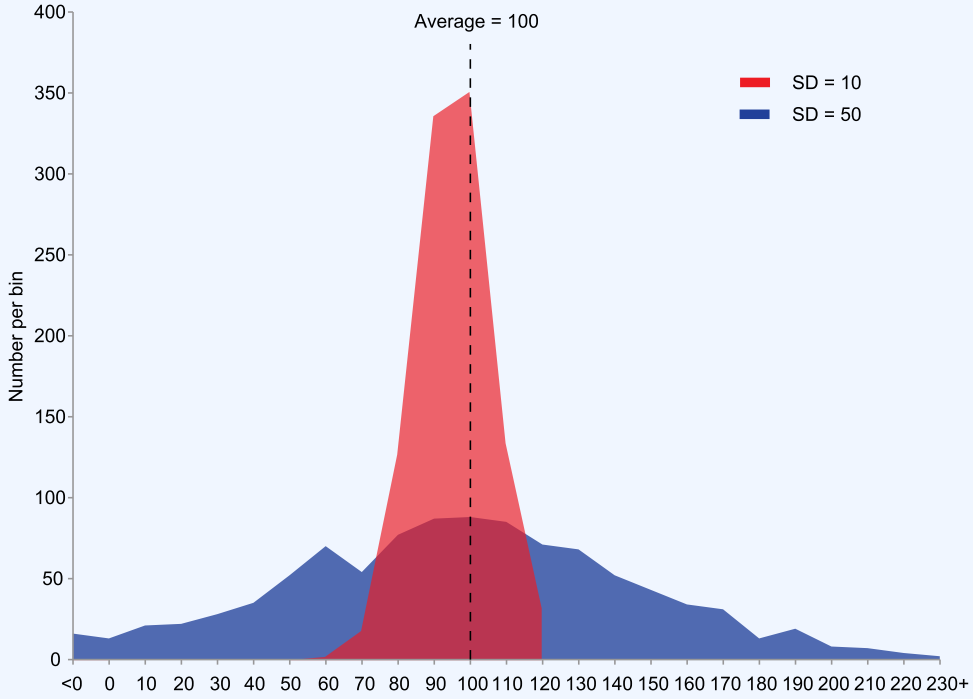
The practical value of understanding the standard deviation of a set of values is in appreciating how much variation there is from the average (mean).

Climate

As a simple example, consider the average daily maximum temperatures for two cities, one inland and one on the coast. It is helpful to understand that the range of daily maximum temperatures for cities near the coast is smaller than for cities inland. Thus, while these two cities may each have the same average maximum temperature, the standard deviation of the daily maximum temperature for the coastal city will be less than that of the inland city as, on any particular day, the actual maximum temperature is more likely to be farther from the average maximum temperature for the inland city than for the coastal one.

Sports

Another way of seeing it is to consider sports teams. In any set of categories, there will be teams that rate highly at some things and poorly at others. Chances are, the teams that lead in the standings will not show such disparity but will perform well in most categories. The lower the standard deviation of their ratings in each category, the more balanced and consistent they will tend to be. Teams with a higher standard deviation, however, will be more unpredictable.



Comparison of Standard Deviations

Example of two samples with the same mean and different standard deviations. The red sample has a mean of 100 and a SD of 10; the blue sample has a mean of 100 and a SD of 50. Each sample has 1,000 values drawn at random from a Gaussian distribution with the specified parameters.

6.1.5: Using a Statistical Calculator

For advanced calculating and graphing, it is often very helpful for students and statisticians to have access to statistical calculators.

Learning Objective

Analyze the use of R statistical software and TI-83 graphing calculators

Key Takeaways

Key Points

- Two of the most common calculators in use are the TI-83 series and the R statistical software environment.
- The TI-83 includes many features, including function graphing, polar/parametric/sequence graphing modes, statistics, trigonometric, and algebraic functions, along with many useful applications.
- The R language is widely used among statisticians and data miners for developing statistical software and data analysis.
- R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, and clustering.
- Another strength of R is static graphics, which can produce publication-quality graphs, including mathematical symbols.

Key Terms

TI-83
: A calculator manufactured by Texas Instruments that is one of the most popular graphing calculators for statistical purposes.

R

A free software programming language and a software environment for statistical computing and graphics.

For many advanced calculations and/or graphical representations, statistical calculators are often quite helpful for statisticians and students of statistics. Two of the most common calculators in use are the TI-83 series and the R statistical software environment.

TI-83

The TI-83 series of graphing calculators, shown in , is manufactured by Texas Instruments. Released in 1996, it was one of the most popular graphing calculators for students. In addition to the functions present on normal scientific calculators, the TI-83 includes many andvanced features, including function graphing, polar/parametric/sequence graphing modes, statistics, trigonometric, and algebraic functions, along with many useful applications.

The TI-83 has a handy statistics mode (accessed via the "STAT" button) that will perform such functions as manipulation of one-variable statistics, drawing of histograms and box plots, linear regression, and even distribution tests.

R

R is a free software programming language and a software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls and surveys of data miners are showing R's popularity has increased substantially in recent years.

R is an implementation of the S programming language, which was created by John Chambers while he was at Bell Labs. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R

TI-83

The TI-83 series of graphing calculators is one of the most popular calculators for statistics students.

Development Core Team, of which Chambers is a member. R is a GNU project, which means it's source code is freely available under the GNU General Public License.

R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, and clustering. Another strength of R is static graphics, which can produce publication-quality graphs, including mathematical symbols. Dynamic and interactive graphics are available through additional packages.

R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. These packagers allow specialized statistical techniques, graphical devices, import/export capabilities, reporting tools, et cetera. Due to its S heritage, R has stronger object-oriented programming facilities than most statistical computing languages.

6.1.6: Degrees of Freedom

The number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary.

Learning Objective

Outline an example of "degrees of freedom"

Key Takeaways

Key Points

- The degree of freedom can be defined as the minimum number of independent coordinates which can specify the position of the system completely.
- A parameter is a characteristic of the variable under examination as a whole; it is part of describing the overall distribution of values.
- As more degrees of freedom are lost, fewer and fewer different situations are accounted for by a model since fewer and fewer pieces of information could, in principle, be different from what is actually observed.
- Degrees of freedom can be seen as linking sample size to explanatory power.

Key Terms

residual
: The difference between the observed value and the estimated function value.

vector
: in statistics, a set of real-valued random variables that may be correlated

The number of independent ways by which a dynamical system can move without violating any constraint imposed on it is known as "degree of freedom. " The degree of freedom can be defined as the minimum number of independent coordinates that completely specify the position of the system.

Consider this example: To compute the variance, first sum the square deviations from the mean. The mean is a parameter, a characteristic of the variable under examination as a whole, and a part of describing the overall distribution of values. Knowing all the parameters, you can accurately describe the data. The more known (fixed) parameters you know, the fewer samples fit this model of the data. If you know only the mean, there will be many possible sets of data that are consistent with this model. However, if you know the mean and the standard deviation, fewer possible sets of data fit this model.

In computing the variance, first calculate the mean, then you can vary any of the scores in the data except one. This one score left unexamined can always be calculated accurately from the rest of the data and the mean itself.

As an example, take the ages of a class of students and find the mean. With a fixed mean, how many of the other scores (there are N of them remember) could still vary? The answer is N-1 independent pieces of information (degrees of freedom) that could vary while the mean is known. One piece of information cannot vary because its value is fully determined by the parameter (in this case the mean) and the other scores. Each parameter that is fixed during our computations constitutes the loss of a degree of freedom.

Imagine starting with a small number of data points and then fixing a relatively large number of parameters as we compute some statistic. We see that as more degrees of freedom are lost, fewer and fewer different situations are accounted for by our model since fewer and fewer pieces of information could, in principle, be different from what is actually observed.

Put informally, the "interest" in our data is determined by the degrees of freedom. If there is nothing that can vary once our parameter is fixed (because we have so very few data points, maybe just one) then there is nothing to investigate. Degrees of freedom can be seen as linking sample size to explanatory power.

The degrees of freedom are also commonly associated with the squared lengths (or "sum of squares" of the coordinates) of random vectors and the parameters of chi-squared and other distributions that arise in associated statistical testing problems.

Notation and Residuals

In equations, the typical symbol for degrees of freedom is ν">ν (lowercase Greek letter nu). In text and tables, the abbreviation "*d.f.* " is commonly used.

In fitting statistical models to data, the random vectors of residuals are constrained to lie in a space of smaller dimension than the number of components in the vector. That smaller dimension is the number of degrees of freedom for error. In statistical terms, a random vector is a list of mathematical variables each of whose value is unknown, either because the value has not yet occurred or because there is imperfect knowledge of its value. The individual variables in a random vector are grouped together because there may be correlations among them. Often they represent different properties of an individual statistical unit (e.g., a particular person, event, etc.).

A residual is an observable estimate of the unobservable statistical error. Consider an example with men's heights and suppose we have a random sample of n people. The sample mean could serve as a good estimator of the population mean. The difference between the height of each man in the sample and the observable sample mean is a residual. Note that the sum of the residuals within a random sample is necessarily zero, and thus the residuals are necessarily not independent.

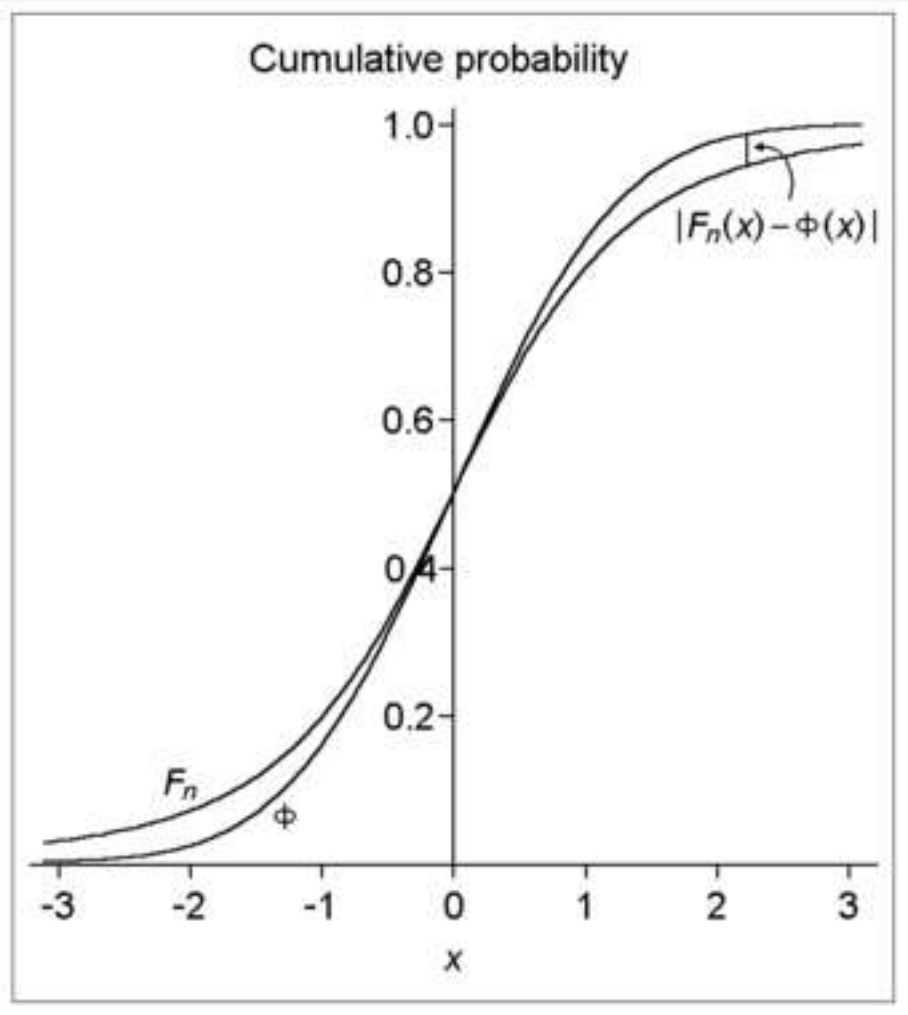
Perhaps the simplest example is this. Suppose $X_1,...,X_n$ are random variables each with expected value μ, and let

$\bar{X_n} = \frac{X_1+...+X_n}{n}$

be the "sample mean. " Then the quantities

Xi−Xn¯">$X_i - \bar{X_n}$

are residuals that may be considered estimates of the errors $X_i - \mu$. The sum of the residuals is necessarily 0. If one knows the values of any $n - 1$ of the residuals, one can thus find the last one. That means they are constrained to lie in a space of dimension $n - 1$, and we say that "there are $n - 1$ degrees of freedom for error. "



Degrees of Freedom

This image illustrates the difference (or distance) between the cumulative distribution functions of the standard normal distribution (Φ) and a hypothetical distribution of a standardized sample mean (Fn). Specifically, the plotted hypothetical distribution is a t distribution with 3 degrees of freedom.

6.1.7: Interquartile Range

The interquartile range (IQR) is a measure of statistical dispersion, or variability, based on dividing a data set into quartiles.

Learning Objectives

Calculate interquartile range based on a given data set

Key Takeaways

Key Points

- The interquartile range is equal to the difference between the upper and lower quartiles: IQR = Q3 − Q1.
- It is a trimmed estimator, defined as the 25% trimmed mid-range, and is the most significant basic robust measure of scale.
- The IQR is used to build box plots, which are simple graphical representations of a probability distribution.

Key Terms

quartile
: any of the three points that divide an ordered distribution into four parts, each containing a quarter of the population

outlier
: a value in a statistical sample which does not fit a pattern that describes most other data

points; specifically, a value that lies 1.5 IQR beyond the upper or lower quartile

The interquartile range (IQR) is a measure of statistical dispersion, or variability, based on dividing a data set into quartiles. Quartiles divide an ordered data set into four equal parts. The values that divide these parts are known as the first quartile, second quartile and third quartile (Q1, Q2, Q3). The interquartile range is equal to the difference between the upper and lower quartiles:

IQR = Q3 – Q1

It is a trimmed estimator, defined as the 25% trimmed mid-range, and is the most significant basic robust measure of scale. As an example, consider the following numbers:

1, 13, 6, 21, 19, 2, 137

Put the data in numerical order: 1, 2, 6, 13, 19, 21, 137

Find the median of the data: 13

Divide the data into four quartiles by finding the median of all the numbers below the median of the full set, and then find the median of all the numbers above the median of the full set.

To find the lower quartile, take all of the numbers below the median: 1, 2, 6

Find the median of these numbers: take the first and last number in the subset and add their positions (not values) and divide by two. This will give you the position of your median:

1+3 = 4/2 = 2

The median of the subset is the second position, which is two. Repeat with numbers above the median of the full set: 19, 21, 137. Median is 1+3 = 4/2 = 2nd position, which is 21. This median separates the third and fourth quartiles.

Subtract the lower quartile from the upper quartile: 21-2=19. This is the Interquartile range, or IQR.

If there is an even number of values, then the position of the median will be in between two numbers. In that case, take the average of the two numbers that the median is between. Example: 1, 3, 7, 12. Median is 1+4=5/2=2.5th position, so it is the average of the second and third positions, which is 3+7=10/2=5. This median separates the first and second quartiles.

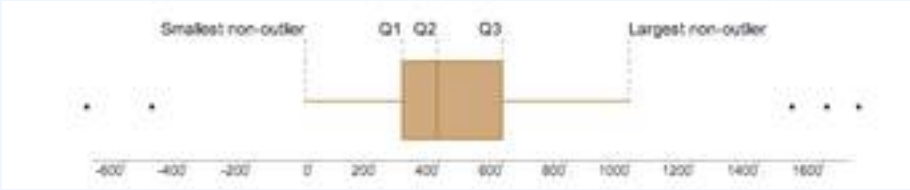
Uses

Unlike (total) range, the interquartile range has a breakdown point of 25%. Thus, it is often preferred to the total range. In other words, since this process excludes outliers, the interquartile range is a more accurate representation of the "spread" of the data than range.

The IQR is used to build box plots, which are simple graphical representations of a probability distribution. A box plot separates the quartiles of the data. All outliers are displayed as regular points on the graph. The

vertical line in the box indicates the location of the median of the data. The box starts at the lower quartile and ends at the upper quartile, so the difference, or length of the boxplot, is the IQR.

On this boxplot in , the IQR is about 300, because Q1 starts at about 300 and Q3 ends at 600, and 600 – 300 = 300.



Interquartile Range

The IQR is used to build box plots, which are simple graphical representations of a probability distribution.

In a boxplot, if the median (Q2 vertical line) is in the center of the box, the distribution is symmetrical. If the median is to the left of the data (such as in the graph above), then the distribution is considered to be skewed right because there is more data on the right side of the median. Similarly, if the median is on the right side of the box, the distribution is skewed left because there is more data on the left side.

The range of this data is 1,700 (biggest outlier) – 500 (smallest outlier) = 2,200. If you wanted to leave out the outliers for a more accurate reading, you would subtract the values at the ends of both "whiskers:"

1,000 – 0 = 1,000

To calculate whether something is truly an outlier or not you use the formula 1.5 x IQR. Once you get that number, the range that includes numbers that are not outliers is [Q1 – 1.5(IQR), Q3 + 1.5(IQR)]. Anything lying outside those numbers are true outliers.

6.1.8: Measures of Variability of Qualitative and Ranked Data

Variability for qualitative data is measured in terms of how often observations differ from one another.

Learning Objective

Assess the use of IQV in measuring statistical dispersion in nominal distributions

Key Takeaways

Key Points

- The notion of "how far apart" does not make sense when evaluating qualitative data. Instead, we should focus on the unlikeability, or how often observations differ.
- An index of qualitative variation (IQV) is a measure of statistical dispersion in nominal distributions–or those dealing with qualitative data.
- The variation ratio is the simplest measure of qualitative variation. It is defined as the proportion of cases which are not the mode.

Key Terms

qualitative data
: data centered around descriptions or distinctions based on some quality or characteristic rather than on some quantity or measured value

variation ratio
: the proportion of cases not in the mode

The study of statistics generally places considerable focus upon the distribution and measure of variability of quantitative variables. A discussion of the variability of qualitative–or categorical– data can sometimes be absent. In such a discussion, we would consider the variability of qualitative data in terms of unlikeability. Unlikeability can be defined as the frequency with which observations differ from one another. Consider this

in contrast to the variability of quantitative data, which ican be defined as the extent to which the values differ from the mean. In other words, the notion of "how far apart" does not make sense when evaluating qualitative data. Instead, we should focus on the unlikeability.

In qualitative research, two responses differ if they are in different categories and are the same if they are in the same category. Consider two polls with the simple parameters of "agree" or "disagree. " These polls question 100 respondents. The first poll results in 75 "agrees" while the second poll only results in 50 "agrees. " The first poll has less variability since more respondents answered similarly.

Index of Qualitative Variation

An index of qualitative variation (IQV) is a measure of statistical dispersion in nominal distributions–or those dealing with qualitative data. The following standardization properties are required to be satisfied:

- Variation varies between 0 and 1.
- Variation is 0 if and only if all cases belong to a single category.
- Variation is 1 if and only if cases are evenly divided across all categories.

In particular, the value of these standardized indices does not depend on the number of categories or number of samples. For any index, the closer to uniform the distribution, the larger the variance, and the larger the differences in frequencies across categories, the smaller the variance.

Variation Ratio

The variation ratio is a simple measure of statistical dispersion in nominal distributions. It is the simplest measure of qualitative variation. It is defined as the proportion of cases which are not the mode:

v=1−fmN">$v = 1 - \frac{f_m}{N}$

Just as with the range or standard deviation, the larger the variation ratio, the more differentiated or dispersed the data are; and the smaller the variation ratio, the more concentrated and similar the data are.

For example, a group which is 55% female and 45% male has a proportion of 0.55 females and, therefore, a variation ratio of:

1.0−0.55=0.45">1.0−0.55=0.45

This group is more dispersed in terms of gender than a group which is 95% female and has a variation ratio of only 0.05. Similarly, a group which is 25% Catholic (where Catholic is the modal religious preference) has a variation ratio of 0.75. This group is much more dispersed, religiously, than a group which is 85% Catholic and has a variation ratio of only 0.15.

6.1.9: Distorting the Truth with Descriptive Statistics

Descriptive statistics can be manipulated in many ways that can be misleading, including the changing of scale and statistical bias.

Learning Objectives

Assess the significance of descriptive statistics given its limitations

Key Takeaways

Key Points

- Descriptive statistics is a powerful form of research because it collects and summarizes vast amounts of data and information in a manageable and organized manner.
- Descriptive statistics, however, lacks the ability to identify the cause behind the phenomenon, correlate (associate) data, account for randomness, or provide statistical calculations that can lead to hypothesis or theories of populations studied.
- A statistic is biased if it is calculated in such a way that is systematically different from the population parameter of interest.
- Every time you try to describe a large set of observations with a single descriptive statistics indicator, you run the risk of distorting the original data or losing important detail.

Key Terms

null hypothesis
: A hypothesis set up to be refuted in order to support an alternative hypothesis; presumed true until statistical evidence in the form of a hypothesis test indicates otherwise.

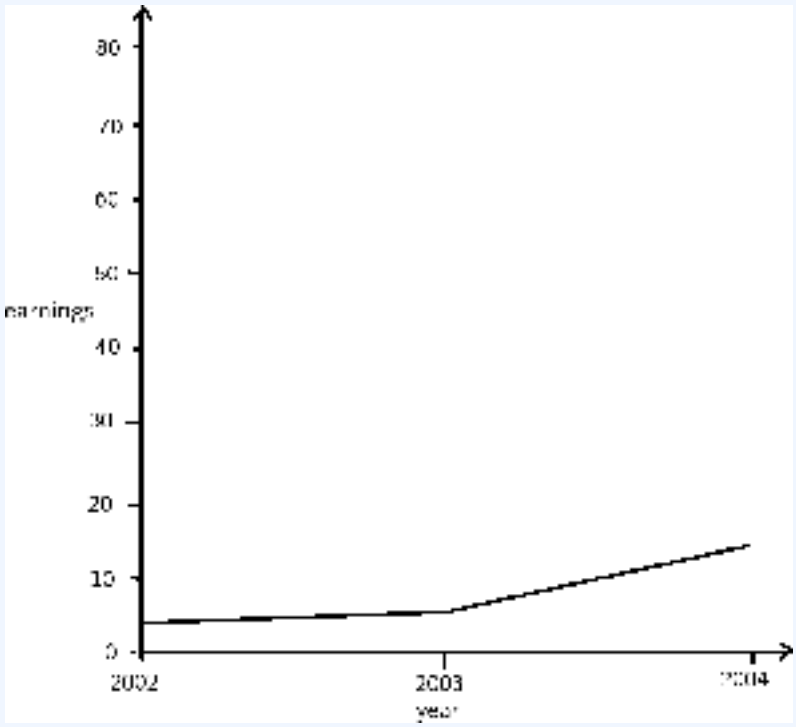
descriptive statistics
: A branch of mathematics dealing with summarization and description of collections of data sets, including the concepts of arithmetic mean, median, and mode.

bias
: (Uncountable) Inclination towards something; predisposition, partiality, prejudice, preference, predilection.

Descriptive statistics can be manipulated in many ways that can be misleading. Graphs need to be carefully analyzed, and questions must always be asked about "the story behind the figures. " Potential manipulations include:

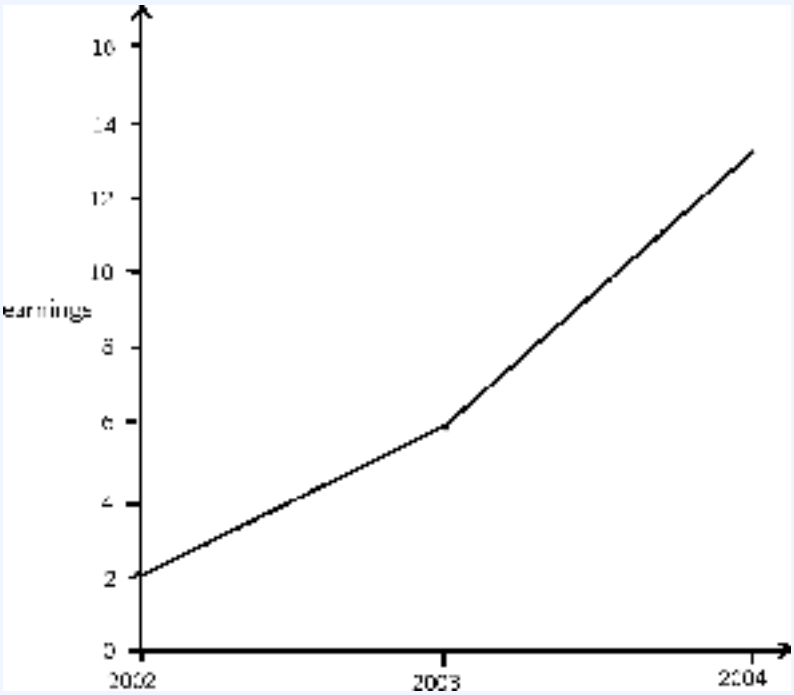
- changing the scale to change the appearence of a graph
- omissions and biased selection of data
- focus on particular research questions
- selection of groups

As an example of changing the scale of a graph, consider the following two figures, and .



Effects of Changing Scale

In this graph, the earnings scale is greater.



Effects of Changing Scale

This is a graph plotting yearly earnings.

Both graphs plot the years 2002, 2003, and 2004 along the x-axis. However, the y-axis of the first graph presents earnings from "0 to 10," while the y-axis of the second graph presents earnings from "0 to 30. " Therefore, there is a distortion between the two of the rate of increased earnings.

Statistical Bias

Bias is another common distortion in the field of descriptive statistics. A statistic is biased if it is calculated in such a way that is systematically different from the population parameter of interest. The following are examples of statistical bias.

- *Selection bias* occurs when individuals or groups are more likely to take part in a research project than others, resulting in biased samples.
- *Spectrum bias* arises from evaluating diagnostic tests on biased patient samples, leading to an overestimate of the sensitivity and specificity of the test.
- The *bias of an estimator* is the difference between an estimator's expectations and the true value of the parameter being estimated.

- *Omitted*-variable *bias* appears in estimates of parameters in a regression analysis when the assumed specification is incorrect, in that it omits an independent variable that should be in the model.
- In *statistical hypothesis testing*, a test is said to be unbiased when the probability of rejecting the null hypothesis is less than or equal to the significance level when the null hypothesis is true, and the probability of rejecting the null hypothesis is greater than or equal to the significance level when the alternative hypothesis is true.
- *Detection bias* occurs when a phenomenon is more likely to be observed and/or reported for a particular set of study subjects.
- *Funding bias* may lead to selection of outcomes, test samples, or test procedures that favor a study's financial sponsor.
- *Reporting bias* involves a skew in the availability of data, such that observations of a certain kind may be more likely to be reported and consequently used in research.
- *Data-snooping bias* comes from the misuse of data mining techniques.
- *Analytical bias* arises due to the way that the results are evaluated.
- *Exclusion bias* arises due to the systematic exclusion of certain individuals from the study

Limitations of Descriptive Statistics

Descriptive statistics is a powerful form of research because it collects and summarizes vast amounts of data and information in a manageable and organized manner. Moreover, it establishes the standard deviation and can lay the groundwork for more complex statistical analysis.

However, what descriptive statistics lacks is the ability to:

1. identify the cause behind the phenomenon because it only describes and reports observations;
2. correlate (associate) data or create any type of statistical relationship modeling relationship among variables;
3. account for randomness; and
4. provide statistical calculations that can lead to hypothesis or theories of populations studied.

To illustrate you can use descriptive statistics to calculate a raw GPA score, but a raw GPA does not reflect:

1. how difficult the courses were, or
2. the identity of major fields and disciplines in which courses were taken.

In other words, every time you try to describe a large set of observations with a single descriptive statistics indicator, you run the risk of distorting the original data or losing important detail.

6.1.10: Exploratory Data Analysis (EDA)

Exploratory data analysis is an approach to analyzing data sets in order to summarize their main characteristics, often with visual methods.

Learning Objectives

Explain how the techniques of EDA achieve its objectives

Key Takeaways

Key Points

- EDA is concerned with uncovering underlying structure, extracting important variables, detecting outliers and anomalies, testing underlying assumptions, and developing models.
- Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data and possibly formulate hypotheses that could lead to new data collection and experiments.
- Robust statistics and nonparametric statistics both try to reduce the sensitivity of statistical inferences to errors in formulating statistical models.
- Many EDA techniques have been adopted into data mining and are being taught to young students as a way to introduce them to statistical thinking.

Key Terms

skewed
: Biased or distorted (pertaining to statistics or information).

data mining
: a technique for searching large-scale databases for patterns; used mainly to find previously unknown correlations between variables that may be commercially useful

exploratory data analysis
: an approach to analyzing data sets that is concerned with uncovering underlying structure, extracting important variables, detecting outliers and anomalies, testing underlying assumptions, and developing models

Exploratory data analysis (EDA) is an approach to analyzing data sets in order to summarize their main characteristics, often with visual methods. It is a statistical practice concerned with (among other things):

- uncovering underlying structure,
- extracting important variables,
- detecting outliers and anomalies,
- testing underlying assumptions, and
- developing models.

Primarily, EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, handling missing values, and making transformations of variables as needed. EDA encompasses IDA.

Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data and possibly formulate hypotheses that could lead to new data collection and experiments. Tukey's EDA was related to two other developments in statistical theory: robust statistics and nonparametric statistics. Both of these try to reduce the sensitivity of statistical inferences to errors in formulating statistical models. Tukey promoted the use of the five number summary of numerical data:

- the two extremes (maximum and minimum),
- the median, and
- the quartiles.

His reasoning was that the median and quartiles, being functions of the empirical distribution, are defined for all distributions, unlike the mean and standard deviation. Moreover, the quartiles and median are more robust to skewed or heavy-tailed distributions than traditional summaries (the mean and standard deviation).

Exploratory data analysis, robust statistics, and nonparametric statistics facilitated statisticians' work on

scientific and engineering problems. Such problems included the fabrication of semiconductors and the understanding of communications networks. These statistical developments, all championed by Tukey, were designed to complement the analytic theory of testing statistical hypotheses.

Objectives of EDA

Tukey held that too much emphasis in statistics was placed on statistical hypothesis testing (confirmatory data analysis) and more emphasis needed to be placed on using data to suggest hypotheses to test. In particular, he held that confusing the two types of analyses and employing them on the same set of data can lead to systematic bias owing to the issues inherent in testing hypotheses suggested by the data.

Subsequently, the objectives of EDA are to:

1. suggest hypotheses about the causes of observed phenomena,
2. assess assumptions on which statistical inference will be based,
3. support the selection of appropriate statistical tools and techniques, and
4. provide a basis for further data collection through surveys or experiments.

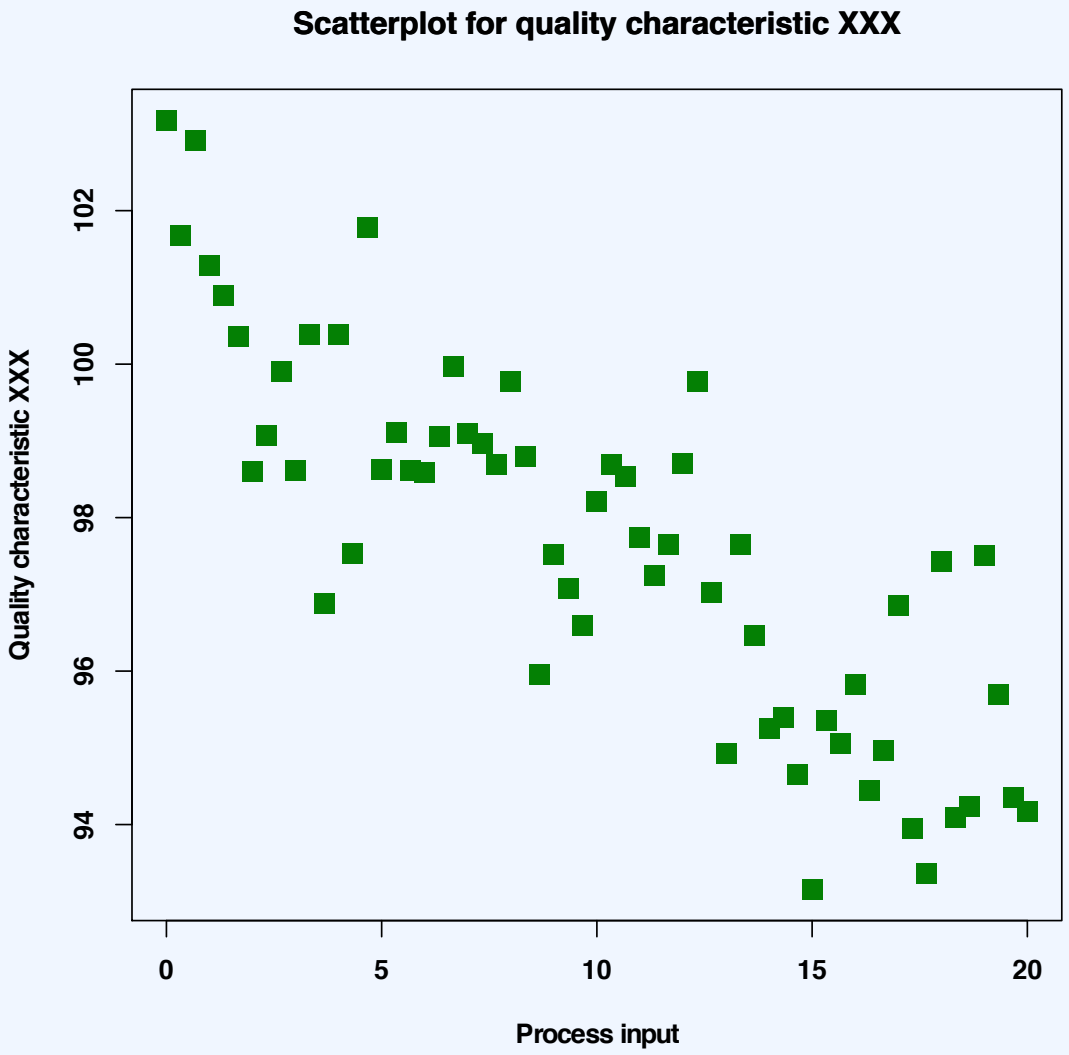
Techniques of EDA

Although EDA is characterized more by the attitude taken than by particular techniques, there are a number of tools that are useful. Many EDA techniques have been adopted into data mining and are being taught to young students as a way to introduce them to statistical thinking. Typical graphical techniques used in EDA are:

- Box plots
- Histograms
- Multi-vari charts
- Run charts
- Pareto charts
- Scatter plots
- Stem-and-leaf plots
- Parallel coordinates
- Odds ratios
- Multidimensional scaling
- Targeted projection pursuits
- Principal component analyses
- Parallel coordinate plots

- Interactive versions of these plots
- Projection methods such as grand tour, guided tour and manual tour

These EDA techniques aim to position these plots so as to maximize our natural pattern-recognition abilities. *A clear picture is worth a thousand words*!



Scatter Plots

A scatter plot is one visual statistical technique developed from EDA.

Attributions

- Range
  - “Boundless.” http://www.boundless.com/. Boundless Learning CC BY-SA 3.0.
  - “Mid-range.” http://en.wikipedia.org/wiki/Mid-range. Wikipedia CC BY-SA 3.0.
  - “Range (statistics).” http://en.wikipedia.org/wiki/Range_(statistics). Wikipedia CC BY-SA 3.0.
  - “Descriptive statistics.” http://en.wikipedia.org/wiki/Descriptive_statistics. Wikipedia CC BY-SA 3.0.
  - “range.” http://en.wiktionary.org/wiki/range. Wiktionary CC BY-SA 3.0.
  - “dispersion.” http://en.wiktionary.org/wiki/dispersion. Wiktionary CC BY-SA 3.0.
  - “20130216 – Range – WikiofScience.” http://wikiofscience.wikidot.com/print:20130216-range. Wikidot CC BY-SA.
- Variance
  - “Boundless.” http://www.boundless.com/. Boundless Learning CC BY-SA 3.0.
  - “Error 404.”

http://www.abs.gov.au/websitedbs/a3121120.nsf/89a5f3d8684682b6ca256de4002c809b/8a79c9ade4ea90ccca25794900128238!OpenDocument.
Austrailian Bureau of Statistics
CC BY.

- "spread."
  http://en.wiktionary.org/wiki/spread.
  Wiktionary
  CC BY-SA 3.0.
- "deviation."
  http://en.wiktionary.org/wiki/deviation.
  Wiktionary
  CC BY-SA 3.0.
- "variance."
  http://mbaecon.wikispaces.com/variance.
  mbaecon Wikispace
  CC BY-SA 3.0.
- "Statistics/Summary/Variance."
  http://en.wikibooks.org/wiki/Statistics/Summary/Variance.
  Wikibooks
  CC BY-SA 3.0.
- "CheetahsSerengetiNationalParkApr2011."
  http://en.wikipedia.org/wiki/File:CheetahsSerengetiNationalParkApr2011.jpg.
  Wikipedia
  CC BY-SA.
- "Statistics/Summary/Variance."
  http://en.wikibooks.org/wiki/Statistics/Summary/Variance.
  Wikibooks
  CC BY-SA 3.0.

- Standard Deviation: Definition and Calculation
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Error 404."
    http://www.abs.gov.au/websitedbs/a3121120.nsf/89a5f3d8684682b6ca256de4002c809b/8a79c9ade4ea90ccca25794900128238!OpenDocument.
    Austrailian Bureau of Statistics

CC BY.
- "coefficient of variation."
  http://en.wikipedia.org/wiki/coefficient%20of%20variation.
  Wikipedia
  CC BY-SA 3.0.
- "normal distribution."
  http://en.wiktionary.org/wiki/normal_distribution.
  Wiktionary
  CC BY-SA 3.0.
- "mean squared error."
  http://en.wikipedia.org/wiki/mean%20squared%20error.
  Wikipedia
  CC BY-SA 3.0.
- "Standard deviation."
  http://en.wikipedia.org/wiki/Standard_deviation.
  Wikipedia
  CC BY-SA 3.0.
- "Standard deviation."
  http://en.wikipedia.org/wiki/Standard_deviation.
  Wikipedia
  CC BY-SA 3.0.
- "Standard deviation."
  http://en.wikipedia.org/wiki/Standard_deviation.
  Wikipedia
  CC BY-SA 3.0.
- "Standard deviation."
  http://en.wikipedia.org/wiki/Standard_deviation.
  Wikipedia
  CC BY-SA 3.0.
- "standard deviation."
  http://en.wiktionary.org/wiki/standard_deviation.
  Wiktionary
  CC BY-SA 3.0.
- "Free High School Science Texts Project, Statistics: Standard Deviation and Variance. September 17, 2013."
  http://cnx.org/content/m38858/latest/.
  OpenStax CNX

CC BY 3.0.

- "Standard deviation diagram."
  http://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg.
  Wikimedia
  CC BY.

- Interpreting the Standard Deviation
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "standard deviation."
    http://en.wiktionary.org/wiki/standard_deviation.
    Wiktionary
    CC BY-SA 3.0.
  - "Standard deviation."
    http://en.wikipedia.org/wiki/Standard_deviation.
    Wikipedia
    CC BY-SA 3.0.
  - "disparity."
    http://en.wiktionary.org/wiki/disparity.
    Wiktionary
    CC BY-SA 3.0.
  - "Comparison standard deviations."
    http://commons.wikimedia.org/wiki/File:Comparison_standard_deviations.svg.
    Wikimedia
    Public domain.
- Using a Statistical Calculator
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "TI-83 series."
    http://en.wikipedia.org/wiki/TI-83_series.
    Wikipedia
    CC BY-SA 3.0.
  - "R."
    http://en.wikipedia.org/wiki/R.

Wikipedia
CC BY-SA 3.0.

- "R Statistics."
  http://en.wikipedia.org/wiki/R_Statistics.
  Wikipedia
  CC BY-SA 3.0.
- "TI-83."
  http://commons.wikimedia.org/wiki/File:TI-83.png.
  Wikimedia
  CC BY-SA.

- Degrees of Freedom
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Errors and residuals in statistics."
    http://en.wikipedia.org/wiki/Errors_and_residuals_in_statistics.
    Wikipedia
    CC BY-SA 3.0.
  - "Random vector."
    http://en.wikipedia.org/wiki/Random_vector.
    Wikipedia
    CC BY-SA 3.0.
  - "Random variable."
    http://en.wikipedia.org/wiki/Random_variable.
    Wikipedia
    CC BY-SA 3.0.
  - "residual."
    http://en.wikipedia.org/wiki/residual.
    Wikipedia
    CC BY-SA 3.0.
  - "vector."
    http://en.wiktionary.org/wiki/vector.
    Wiktionary
    CC BY-SA 3.0.
  - "Degrees of freedom (statistics)."
    http://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics).

Wikipedia
CC BY-SA 3.0.

- "Degrees of freedom (statistics)."
  http://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics).
  Wikipedia
  CC BY-SA 3.0.
- "Statistics Ground Zero/Degrees of freedom."
  http://en.wikibooks.org/wiki/Statistics_Ground_Zero/Degrees_of_freedom.
  Wikibooks
  CC BY-SA 3.0.
- "BerryEsseenTheoremCDFGraphExample."
  http://commons.wikimedia.org/wiki/File:BerryEsseenTheoremCDFGraphExample.png.
  Wikimedia
  Public domain.

- Interquartile Range
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "2. Range and Interquartile Range."
    http://killianhma0910.wikispaces.com/2.+Range+and+Interquartile+Range.
    killianhma0910 Wikispace
    CC BY-SA 3.0.
  - "outlier."
    http://en.wiktionary.org/wiki/outlier.
    Wiktionary
    CC BY-SA 3.0.
  - "quartile."
    http://en.wiktionary.org/wiki/quartile.
    Wiktionary
    CC BY-SA 3.0.
  - "Interquartile range."
    http://en.wikipedia.org/wiki/Interquartile_range.
    Wikipedia
    CC BY-SA 3.0.
  - "killianHMA0910 – 2.
    Range and Interquartile Range."

http://killianhma0910.wikispaces.com/2.+Range+and+Interquartile+Range.
Wikispaces
CC BY-SA.

- Measures of Variability of Qualitative and Ranked Data
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Qualitative variation."
    http://en.wikipedia.org/wiki/Qualitative_variation.
    Wikipedia
    CC BY-SA 3.0.
  - "qualitative data."
    http://en.wikipedia.org/wiki/qualitative%20data.
    Wikipedia
    CC BY-SA 3.0.
  - "variation ratio."
    http://en.wiktionary.org/wiki/variation_ratio.
    Wiktionary
    CC BY-SA 3.0.
  - "Variation ratio."
    http://en.wikipedia.org/wiki/Variation_ratio.
    Wikipedia
    CC BY-SA 3.0.
- Distorting the Truth with Descriptive Statistics
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "descriptive statistics."
    http://en.wiktionary.org/wiki/descriptive_statistics.
    Wiktionary
    CC BY-SA 3.0.
  - "Bias (statistics)."
    http://en.wikipedia.org/wiki/Bias_(statistics).
    Wikipedia
    CC BY-SA 3.0.

- "Descriptive Statistics."
  http://medanth.wikispaces.com/Descriptive+Statistics.
  medanth Wikispace
  CC BY-SA 3.0.
- "null hypothesis."
  http://en.wiktionary.org/wiki/null_hypothesis.
  Wiktionary
  CC BY-SA 3.0.
- "Free High School Science Texts Project, Statistics: Misuse of Statistics. September 17, 2013."
  http://cnx.org/content/m38864/latest/.
  OpenStax CNX
  CC BY 3.0.
- "bias."
  http://en.wiktionary.org/wiki/bias.
  Wiktionary
  CC BY-SA 3.0.
- "Free High School Science Texts Project, Statistics: Misuse of Statistics. April 29, 2013."
  http://cnx.org/content/m38864/latest/.
  OpenStax CNX
  CC BY 3.0.
- "Free High School Science Texts Project, Statistics: Misuse of Statistics. April 29, 2013."
  http://cnx.org/content/m38864/latest/.
  OpenStax CNX
  CC BY 3.0.

- Exploratory Data Analysis (EDA)
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - Bioinformatics.

    http://bioinformatics.ca//files/CBW%20-%20presentations/Stats_Toronto2011_Module%202/Stats_Toronto2011_Module%202.pdf.
    CC BY-SA.
  - "exploratory data analysis."
    http://en.wikipedia.org/wiki/exploratory%20data%20analysis.

Wikipedia
CC BY-SA 3.0.

- "Exploratory data analysis."
  http://en.wikipedia.org/wiki/Exploratory_data_analysis.
  Wikipedia
  CC BY-SA 3.0.
- "data mining."
  http://en.wiktionary.org/wiki/data_mining.
  Wiktionary
  CC BY-SA 3.0.
- "skewed."
  http://en.wiktionary.org/wiki/skewed.
  Wiktionary
  CC BY-SA 3.0.
- "Scatter diagram for quality characteristic XXX."
  http://en.wikipedia.org/wiki/File:Scatter_diagram_for_quality_characteristic_XXX.svg.
  Wikipedia
  CC BY-SA.

SECTION XI

7. SAMPLING

7.1 Populations and Samples
7.2 Sample Surveys
7.3 Sampling Distributions
7.4 Errors in Sampling
7.5 Sampling Examples

7.1 POPULATIONS AND SAMPLES

7.1: Populations and Samples

7.1.1: Populations

In statistics, a population includes all members of a defined group that we are studying for data driven decisions.

Learning Objectives

Give examples of a statistical populations and sub-populations

Key Takeaways

Key Points

- It is often impractical to study an entire population, so we often study a sample from that population to infer information about the larger population as a whole.
- Sometimes a government wishes to try to gain information about all the people living within an area with regard to gender, race, income, and religion. This type of information gathering over a whole population is called a census.
- A subset of a population is called a sub-population.

Key Terms

heterogeneous
: diverse in kind or nature; composed of diverse parts

sample
: a subset of a population selected for measurement, observation, or questioning to provide statistical information about the population

Populations

When we hear the word *population*, we typically think of all the people living in a town, state, or country. This is one type of population. In statistics, the word takes on a slightly different meaning.

Census

This is the logo for the Bureau of the Census in the United States.

A statistical population is a set of entities from which statistical inferences are to be drawn, often based on a random sample taken from the population. For example, if we are interested in making generalizations about all crows, then the statistical population is the set of all crows that exist now, ever existed, or will exist in the future. Since in this case and many others it is impossible to observe the entire statistical population, due to time constraints, constraints of geographical accessibility, and constraints on the researcher's resources, a researcher would instead observe a statistical sample from the population in order to attempt to learn something about the population as a whole.

Sometimes a government wishes to try to gain information about all the people living within an area with regard to gender, race, income, and religion. This type of information gathering over a whole population is called a census.

Sub-Populations

A subset of a population is called a sub-population. If different sub-populations have different properties, so that the overall population is heterogeneous, the properties and responses of the overall population can often be better understood if the population is first separated into distinct sub-populations. For instance, a

particular medicine may have different effects on different sub-populations, and these effects may be obscured or dismissed if such special sub-populations are not identified and examined in isolation.

Similarly, one can often estimate parameters more accurately if one separates out sub-populations. For example, the distribution of heights among people is better modeled by considering men and women as separate sub-populations.

7.1.2: Samples

A sample is a set of data collected and/or selected from a population by a defined procedure.

Learning Objective

Differentiate between a sample and a population

Key Takeaways

Key Points

- A complete sample is a set of objects from a parent population that includes all such objects that satisfy a set of well-defined selection criteria.
- An unbiased (representative) sample is a set of objects chosen from a complete sample using a selection process that does not depend on the properties of the objects.
- A random sample is defined as a sample where each individual member of the population has a known, non-zero chance of being selected as part of the sample.

Key Terms

census
: an official count of members of a population (not necessarily human), usually residents or citizens in a particular region, often done at regular intervals

population
: a group of units (persons, objects, or other items) enumerated in a census or from which a sample is drawn

unbiased
: impartial or without prejudice

What is a Sample?

In statistics and quantitative research methodology, a data sample is a set of data collected and/or selected from a population by a defined procedure.

Typically, the population is very large, making a census or a complete enumeration of all the values in the population impractical or impossible. The sample represents a subset of manageable size. Samples are collected and statistics are calculated from the samples so that one can make inferences or extrapolations from the sample to the population. This process of collecting information from a sample is referred to as sampling.

Types of Samples

A complete sample is a set of objects from a parent population that includes all such objects that satisfy a set of well-defined selection criteria. For example, a complete sample of Australian men taller than 2 meters would consist of a list of every Australian male taller than 2 meters. It wouldn't include German males, or tall Australian females, or people shorter than 2 meters. To compile such a complete sample requires a complete list of the parent population, including data on height, gender, and nationality for each member of that parent population. In the case of human populations, such a complete list is unlikely to exist, but such complete samples are often available in other disciplines, such as complete magnitude-limited samples of astronomical objects.

An unbiased (representative) sample is a set of objects chosen from a complete sample using a selection process that does not depend on the properties of the objects. For example, an unbiased sample of Australian men taller than 2 meters might consist of a randomly sampled subset of 1% of Australian males taller than 2 meters. However, one chosen from the electoral register might not be unbiased since, for example, males aged under 18 will not be on the electoral register. In an astronomical context, an unbiased sample might consist of that fraction of a complete sample for which data are available, provided the data availability is not biased by individual source properties.

The best way to avoid a biased or unrepresentative sample is to select a random sample, also known as a probability sample. A random sample is defined as a sample wherein each individual member of the population has a known, non-zero chance of being selected as part of the sample. Several types of random samples are simple random samples, systematic samples, stratified random samples, and cluster random samples.

A sample that is not random is called a non-random sample, or a non-probability sampling. Some examples of nonrandom samples are convenience samples, judgment samples, and quota samples.

Samples

Online and phone-in polls produce biased samples because the respondents are self-selected. In self-selection bias, those individuals who are highly motivated to respond– typically individuals who have strong opinions– are over-represented, and individuals who are indifferent or apathetic are less likely to respond.

7.1.3: Random Sampling

A random sample, also called a probability sample, is taken when each individual has an equal probability of being chosen for the sample.

Learning Objectives

Categorize a random sample as a simple random sample, a stratified random sample, a cluster sample, or a systematic sample

Key Takeaways

Key Points

- A simple random sample (SRS) of size n consists of n individuals from the population chosen in such a way that every set on n individuals has an equal chance of being in the selected sample.
- Stratified sampling occurs when a population embraces a number of distinct categories and is divided into sub-populations, or strata. At this stage, a simple random sample would be chosen from each stratum and combined to form the full sample.
- Cluster sampling divides the population into groups, or clusters. Some of these clusters are randomly selected. Then, all the individuals in the chosen cluster are selected to be in the sample.
- Systematic sampling relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list.

Key Terms

stratum
: a category composed of people with certain similarities, such as gender, race, religion, or even grade level

population
: a group of units (persons, objects, or other items) enumerated in a census or from which a sample is drawn

cluster
: a significant subset within a population

Simple Random Sample (SRS)

There is a variety of ways in which one could choose a sample from a population. A simple random sample (SRS) is one of the most typical ways. Also commonly referred to as a probability sample, a simple random sample of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance of being in the selected sample. An example of an SRS would be drawing names from a hat. An online poll in which a person is asked to given their opinion about something is not random because only those people with strong opinions, either positive or negative, are likely to respond. This type of poll doesn't reflect the opinions of the apathetic .

Simple random samples are not perfect and should not always be used. They can be vulnerable to sampling error because the randomness of the selection may result in a sample that doesn't reflect the makeup of the population. For instance, a simple random sample of ten people from a given country will on average produce five men and five women, but any given trial is likely to over-represent one sex and under-represent the other. Systematic and stratified techniques, discussed below, attempt to overcome this problem by using information about the population to choose a more representative sample.

In addition, SRS may also be cumbersome and tedious when sampling from an unusually large target population. In some cases, investigators are interested in research questions specific to subgroups of the population. For example, researchers might be interested in examining whether cognitive ability as a predictor of job performance is equally applicable across racial groups. SRS cannot accommodate the needs of researchers in this situation because it does not provide sub-samples of the population. Stratified sampling, which is discussed below, addresses this weakness of SRS.

Stratified Random Sample

When a population embraces a number of distinct categories, it can be beneficial to divide the population in sub-populations called strata. These strata must be in some way important to the response the researcher is studying. At this stage, a simple random sample would be chosen from each stratum and combined to form the full sample.

For example, let's say we want to sample the students of a high school to see what type of music they like to listen to, and we want the sample to be representative of all grade levels. It would make sense to divide the students into their distinct grade levels and then choose an SRS from each grade level. Each sample would be combined to form the full sample.

Cluster Sample

Cluster sampling divides the population into groups, or clusters. Some of these clusters are randomly selected. Then, all the individuals in the chosen cluster are selected to be in the sample. This process is often used because it can be cheaper and more time-efficient.

For example, while surveying households within a city, we might choose to select 100 city blocks and then interview every household within the selected blocks, rather than interview random households spread out over the entire city.

Systematic Sample

Systematic sampling relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every th element from then onward. In this case,
. It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the th element in the list. A simple example would be to select every 10^{th} name from the telephone directory (an 'every 10^{th}' sample, also referred to as 'sampling with a skip of 10').

7.1.4: Random Assignment of Subjects

Random assignment helps eliminate the differences between the experimental group and the control group.

Learning Objective

Discover the importance of random assignment of subjects in experiments

Key Takeaways

Key Points

- Researchers randomly assign participants in a study to either the experimental group or the control group. Dividing the participants randomly reduces group differences, thereby reducing the possibility that confounding factors will influence the results.
- By randomly assigning subjects to groups, researchers are able to feel confident that the groups are the same in terms of all variables except the one which they are manipulating.
- A randomly assigned group may statistically differ from the mean of the overall population, but this is rare.
- Random assignment became commonplace in experiments in the late 1800s due to the influence of researcher Charles S. Peirce.

Key Terms

null hypothesis
: A hypothesis set up to be refuted in order to support an alternative hypothesis; presumed true until statistical evidence in the form of a hypothesis test indicates otherwise.

control
: a separate group or subject in an experiment against which the results are compared where the primary variable is low or nonexistence

Importance of Random Assignment

When designing controlled experiments, such as testing the effects of a new drug, statisticians often employ an experimental design, which by definition involves random assignment. Random assignment, or random placement, assigns subjects to treatment and control (no treatment) group(s) on the basis of chance rather than any selection criteria. The aim is to produce experimental groups with no statistically significant characteristics prior to the experiment so that any changes between groups observed after experimental activities have been completed can be attributed to the treatment effect rather than to other, pre-existing differences among individuals between the groups.

Control Group

Take identical growing plants, randomly assign them to two groups, and give fertilizer to one of the groups. If there are differences between the fertilized plant group and the unfertilized "control" group, these differences may be due to the fertilizer.

In experimental design, random assignment of participants in experiments or treatment and control groups help to ensure that any differences between or within the groups are not systematic at the outset of the experiment. Random assignment does not guarantee that the groups are "matched" or equivalent; only that any differences are due to chance.

Random assignment is the desired assignment method because it provides control for all attributes of the members of the samples—in contrast to matching on only one or more variables—and provides the mathematical basis for estimating the likelihood of group equivalence for characteristics one is interested in, both for pre-treatment checks on equivalence and the evaluation of post treatment results using inferential statistics.

Random Assignment Example

Consider an experiment with one treatment group and one control group. Suppose the experimenter has recruited a population of 50 people for the experiment—25 with blue eyes and 25 with brown eyes. If the experimenter were to assign all of the blue-eyed people to the treatment group and the brown-eyed people to the control group, the results may turn out to be biased. When analyzing the results, one might question whether an observed effect was due to the application of the experimental condition or was in fact due to eye color.

With random assignment, one would randomly assign individuals to either the treatment or control group, and therefore have a better chance at detecting if an observed change were due to chance or due to the experimental treatment itself.

If a randomly assigned group is compared to the mean, it may be discovered that they differ statistically, even though they were assigned from the same group. To express this same idea statistically–if a test of statistical significance is applied to randomly assigned groups to test the difference between sample means against the null

hypothesis that they are equal to the same population mean (i.e., population mean of differences = 0), given the probability distribution, the null hypothesis will sometimes be "rejected"–that is, deemed implausible. In other words, the groups would be sufficiently different on the variable tested to conclude statistically that they did not come from the same population, even though they were assigned from the same total group. In the example above, using random assignment may create groups that result in 20 blue-eyed people and 5 brown-eyed people in the same group. This is a rare event under random assignment, but it could happen, and when it does, it might add some doubt to the causal agent in the experimental hypothesis.

History of Random Assignment

Randomization was emphasized in the theory of statistical inference of Charles S. Peirce in "Illustrations of the Logic of Science" (1877–1878) and "A Theory of Probable Inference" (1883). Peirce applied randomization in the Peirce-Jastrow experiment on weight perception. Peirce randomly assigned volunteers to a blinded, repeated-measures design to evaluate their ability to discriminate weights. His experiment inspired other researchers in psychology and education, and led to a research tradition of randomized experiments in laboratories and specialized textbooks in the nineteenth century.

7.1.5: Surveys or Experiments?

Surveys and experiments are both statistical techniques used to gather data, but they are used in different types of studies.

Learning Objective

Distinguish between when to use surveys and when to use experiments

Key Takeaways

Key Points

- A survey is a technique that involves questionnaires and interviews of a sample population with the intention of gaining information, such as opinions or facts, about the general population.
- An experiment is an orderly procedure carried out with the goal of verifying, falsifying, or establishing the validity of a hypothesis.
- A survey would be useful if trying to determine whether or not people would be interested in trying out a new drug for headaches on the market. An experiment would test the effectiveness of this new drug.

Key Term

placebo
: an inactive substance or preparation used as a control in an experiment or test to determine the effectiveness of a medicinal drug

What is a Survey?

Survey methodology involves the study of the sampling of individual units from a population and the associated survey data collection techniques, such as questionnaire construction and methods for improving the number and accuracy of responses to surveys.

Statistical surveys are undertaken with a view towards making statistical inferences about the population being studied, and this depends strongly on the survey questions used. Polls about public opinion, public health surveys, market research surveys, government surveys, and censuses are all examples of quantitative research that use contemporary survey methodology to answers questions about a population. Although censuses do not include a "sample," they do include other aspects of survey methodology, like questionnaires, interviewers, and nonresponse follow-up techniques. Surveys provide important information for all kinds of public information and research fields, like marketing research, psychology, health, and sociology.

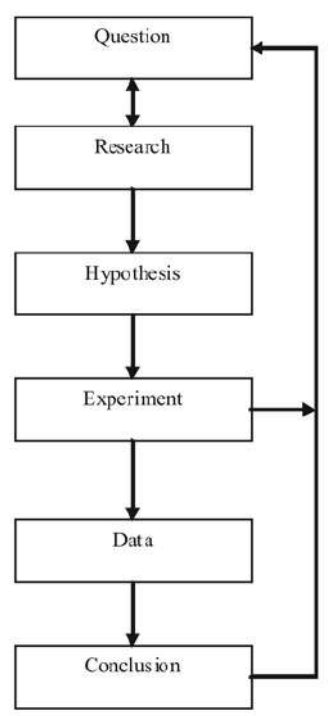
Since survey research is almost always based on a sample of the population, the success of the research

is dependent on the representativeness of the sample with respect to a target population of interest to the researcher.

What is an Experiment?

An experiment is an orderly procedure carried out with the goal of verifying, falsifying, or establishing the validity of a hypothesis. Experiments provide insight into cause-and-effect by demonstrating what outcome occurs when a particular factor is manipulated. Experiments vary greatly in their goal and scale, but always rely on repeatable procedure and logical analysis of the results in a method called the scientific method . A child may carry out basic experiments to understand the nature of gravity, while teams of scientists may take years of systematic investigation to advance the understanding of a phenomenon. Experiments can vary from personal and informal (e.g. tasting a range of chocolates to find a favorite), to highly controlled (e.g. tests requiring a complex apparatus overseen by many scientists that hope to discover information about subatomic particles). Uses of experiments vary considerably between the natural and social sciences.

In statistics, controlled experiments are often used. A controlled experiment generally compares the results obtained from an experimental sample against a control sample, which is practically identical to the experimental sample except for the one aspect whose effect is being tested (the independent variable). A good example of this would be a drug trial, where the effects of the actual drug are tested against a placebo.



Scientific Method

This flow chart shows the steps of the scientific method.

When is One Technique Better Than the Other?

Surveys and experiments are both techniques used in statistics. They have similarities, but an in depth look into these two techniques will reveal how different they are. When a businessman wants to market his products, it's a survey he will need and not an experiment. On the other hand, a scientist who has discovered a new element or drug will need an experiment, and not a survey, to prove its usefulness. A survey involves asking different people about their opinion on a particular product or about a particular issue, whereas an experiment is a

comprehensive study about something with the aim of proving it scientifically. They both have their place in different types of studies.

Attributions

- Populations
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Statistical population."
    http://en.wikipedia.org/wiki/Statistical_population.
    Wikipedia
    CC BY-SA 3.0.
  - "heterogeneous."
    http://en.wiktionary.org/wiki/heterogeneous.
    Wiktionary
    CC BY-SA 3.0.
  - "sample."
    http://en.wiktionary.org/wiki/sample.
    Wiktionary
    CC BY-SA 3.0.
  - "Census Bureau seal."
    http://commons.wikimedia.org/wiki/File:Census_Bureau_seal.jpg.
    Wikimedia
    CC BY-SA.
- Samples
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Sample (statistics)."
    http://en.wikipedia.org/wiki/Sample_(statistics).
    Wikipedia
    CC BY-SA 3.0.
  - "census."
    http://en.wiktionary.org/wiki/census.

Wiktionary
CC BY-SA 3.0.

- "unbiased."
  http://en.wiktionary.org/wiki/unbiased.
  Wiktionary
  CC BY-SA 3.0.
- "population."
  http://en.wiktionary.org/wiki/population.
  Wiktionary
  CC BY-SA 3.0.
- "Boundless."
  https://www.boundless.com/psychology/psychology-as-science/descriptive-techniques/explanation-random-sampling/.
  Boundless Learning
  CC BY.

- Random Sampling
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Sampling (statistics)."
    http://en.wikipedia.org/wiki/Sampling_(statistics)%23Sampling_methods.
    Wikipedia
    CC BY-SA 3.0.
  - "cluster."
    http://en.wiktionary.org/wiki/cluster.
    Wiktionary
    CC BY-SA 3.0.
  - "population."
    http://en.wiktionary.org/wiki/population.
    Wiktionary
    CC BY-SA 3.0.
  - "Boundless."
    https://www.boundless.com/psychology/psychology-as-science/descriptive-techniques/explanation-random-sampling/.
    Boundless Learning
    CC BY.

- Random Assignment of Subjects
  - “Boundless.”
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - “Placebo.”
    http://en.wikipedia.org/wiki/Placebo.
    Wikipedia
    CC BY-SA 3.0.
  - “Blind experiment.”
    http://en.wikipedia.org/wiki/Blind_experiment.
    Wikipedia
    CC BY-SA 3.0.
  - “Random assignment.”
    http://en.wikipedia.org/wiki/Random_assignment.
    Wikipedia
    CC BY-SA 3.0.
  - “control.”
    http://en.wiktionary.org/wiki/control.
    Wiktionary
    CC BY-SA 3.0.
  - “null hypothesis.”
    http://en.wiktionary.org/wiki/null_hypothesis.
    Wiktionary
    CC BY-SA 3.0.
  - “Starr 011107-0010 Argyroxiphium sandwicense subsp. macrocephalum.”
    http://en.wikipedia.org/wiki/File:Starr_011107-0010_Argyroxiphium_sandwicense_subsp._macrocephalum.jpg.
    Wikipedia
    CC BY-SA.
- Surveys or Experiments?
  - “Boundless.”
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - “Survey methodology.”

http://en.wikipedia.org/wiki/Survey_methodology.
Wikipedia
CC BY-SA 3.0.

- "Experiment."
  http://en.wikipedia.org/wiki/Experiment.
  Wikipedia
  CC BY-SA 3.0.
- "placebo."
  http://en.wikipedia.org/wiki/placebo.
  Wikipedia
  CC BY-SA 3.0.
- "The Scientific Method."
  http://commons.wikimedia.org/wiki/File:The_Scientific_Method.png.
  Wikimedia
  CC BY-SA.

7.2 SAMPLE SURVEYS

7.2: Sample Surveys

7.2.1: The Literary Digest Poll

Incorrect polling techniques used during the 1936 presidential election led to the demise of the popular magazine, *The Literary Digest*.

Learning Objective

Critique the problems with the techniques used by the Literary Digest Poll

Key Takeaways

Key Points

- As it had done in 1920, 1924, 1928 and 1932, *The Literary Digest* conducted a straw poll regarding the likely outcome of the 1936 presidential election. Before 1936, it had always correctly predicted the winner. It predicted Landon would beat Roosevelt.
- In November, Landon carried only Vermont and Maine; President F. D. Roosevelt carried the 46 other states. Landon's electoral vote total of eight is a tie for the record low for a major-

party nominee since the American political paradigm of the Democratic and Republican parties began in the 1850s.

- The polling techniques used were to blame, even though they polled 10 million people and got a response from 2.4 million.They polled mostly their readers, who had more money than the typical American during the Great Depression. Higher income people were more likely to vote Republican.
- Subsequent statistical analysis and studies have shown it is not necessary to poll ten million people when conducting a scientific survey. A much lower number, such as 1,500 persons, is adequate in most cases so long as they are appropriately chosen.
- This debacle led to a considerable refinement of public opinion polling techniques and later came to be regarded as ushering in the era of modern scientific public opinion research.

Key Terms

bellwether
: anything that indicates future trends

straw poll
: a survey of opinion which is unofficial, casual, or ad hoc

The Literary Digest

The Literary Digest was an influential general interest weekly magazine published by Funk & Wagnalls. Founded by Isaac Kaufmann Funk in 1890, it eventually merged with two similar weekly magazines, Public Opinion and Current Opinion.

The Literary Digest

Cover of the February 19, 1921 edition of The Literary Digest.

History

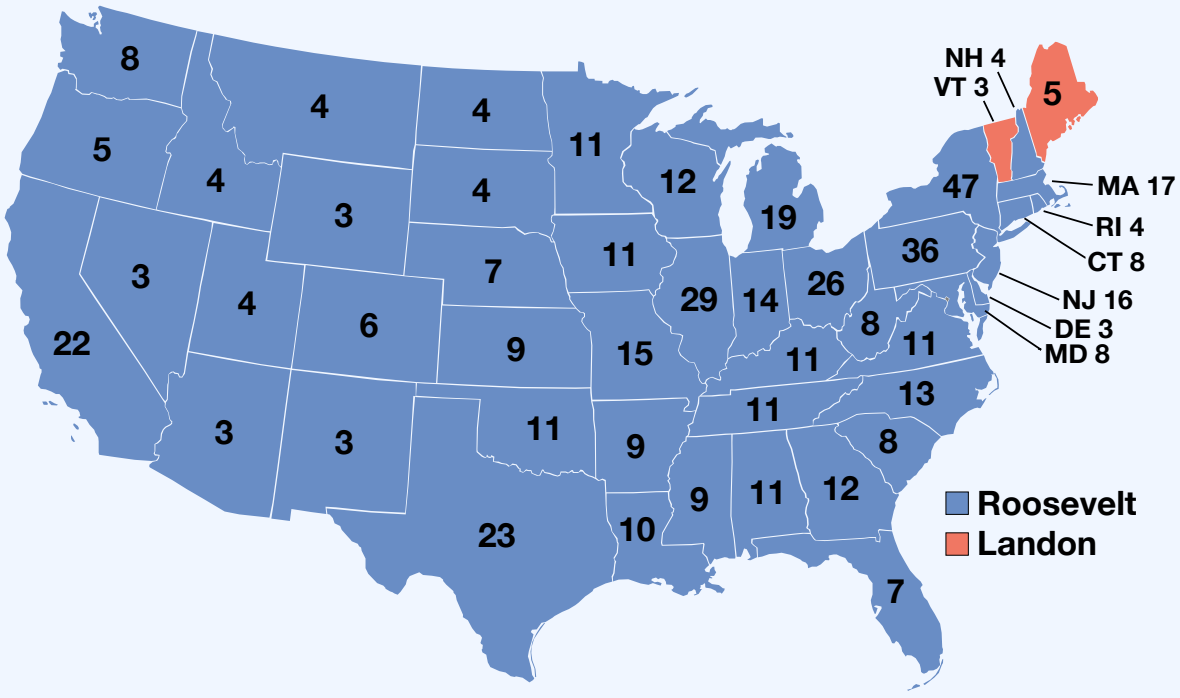
Beginning with early issues, the emphasis of *The Literary Digest* was on opinion articles and an analysis of news events. Established as a weekly news magazine, it offered condensations of articles from American, Canadian, and European publications. Type-only covers gave way to illustrated covers during the early 1900s. After Isaac Funk's death in 1912, Robert Joseph Cuddihy became the editor. In the 1920s, the covers carried full-color reproductions of famous paintings . By 1927, *The Literary Digest* climbed to a circulation of over one million. Covers of the final issues displayed various photographic and photo-montage techniques. In 1938, it merged with the *Review of Reviews*, only to fail soon after. Its subscriber list was bought by *Time*.

Presidential Poll

The Literary Digest is best-remembered today for the circumstances surrounding its demise. As it had done in 1920, 1924, 1928 and 1932, it conducted a straw poll regarding the likely outcome of the 1936 presidential election. Before 1936, it had always correctly predicted the winner.

The 1936 poll showed that the Republican candidate, Governor Alfred Landon of Kansas, was likely to be the overwhelming winner. This seemed possible to some, as the Republicans had fared well in Maine, where the congressional and gubernatorial elections were then held in September, as opposed to the rest of the nation, where these elections were held in November along with the presidential election, as they are today. This outcome seemed especially likely in light of the conventional wisdom, "As Maine goes, so goes the nation," a saying coined because Maine was regarded as a "bellwether" state which usually supported the winning candidate's party.

In November, Landon carried only Vermont and Maine; President Franklin Delano Roosevelt carried the 46 other states . Landon's electoral vote total of eight is a tie for the record low for a major-party nominee since the American political paradigm of the Democratic and Republican parties began in the 1850s. The Democrats joked, "As goes Maine, so goes Vermont," and the magazine was completely discredited because of the poll, folding soon thereafter.



1936 Presidential Election

This map shows the results of the 1936 presidential election. Red denotes states won by Landon/Knox, blue denotes those won by Roosevelt/Garner. Numbers indicate the number of electoral votes allotted to each state.

In retrospect, the polling techniques employed by the magazine were to blame. Although it had polled ten million individuals (of whom about 2.4 million responded, an astronomical total for any opinion poll), it had surveyed firstly its own readers, a group with disposable incomes well above the national average of the time, shown in part by their ability still to afford a magazine subscription during the depths of the Great Depression, and then two other readily available lists: that of registered automobile owners and that of telephone users. While such lists might come close to providing a statistically accurate cross-section of Americans today, this assumption was manifestly incorrect in the 1930s. Both groups had incomes well above the national average of the day, which resulted in lists of voters far more likely to support Republicans than a truly typical voter of the time. In addition, although 2.4 million responses is an astronomical number, it is only 24% of those surveyed, and the low response rate to the poll is probably a factor in the debacle. It is erroneous to assume that the responders and the non-responders had the same views and merely to extrapolate the former on to the latter. Further, as subsequent statistical analysis and study have shown, it is not necessary to poll ten million people when conducting a scientific survey . A much lower number, such as 1,500 persons, is adequate in most cases so long as they are appropriately chosen.

George Gallup's American Institute of Public Opinion achieved national recognition by correctly predicting the result of the 1936 election and by also correctly predicting the quite different results of the Literary Digest poll to within about 1%, using a smaller sample size of 50,000. This debacle led to a considerable refinement of public opinion polling techniques and later came to be regarded as ushering in the era of modern scientific public opinion research.

7.2.2: The Year the Polls Elected Dewey

In the 1948 presidential election, the use of quota sampling led the polls to inaccurately predict that Dewey would defeat Truman.

Learning Objective

Criticize the polling methods used in 1948 that incorrectly predicted that Dewey would win the presidency

Key Takeaways

Key Points

- Many polls, including Gallup, Roper, and Crossley, wrongfully predicted the outcome of the election due to their use of quota sampling.
- Quota sampling is when each interviewer polls a certain number of people in various categories that are representative of the whole population, such as age, race, sex, and income.
- One major problem with quota sampling includes the possibility of missing an important representative category that is key to how people vote. Another is the human element involved.
- Truman, as it turned out, won the electoral vote by a 303-189 majority over Dewey, although a swing of just a few thousand votes in Ohio, Illinois, and California would have produced a Dewey victory.
- One of the most famous blunders came when the Chicago Tribune wrongfully printed the inaccurate headline, “Dewey Defeats Truman” on November 3, 1948, the day after Truman defeated Dewey.

Key Terms

quota sampling
: a sampling method that chooses a representative cross-section of the population by taking into consideration each important characteristic of the population proportionally, such as income, sex, race, age, etc.

margin of error
: An expression of the lack of precision in the results obtained from a sample.

quadrennial
: happening every four years

1948 Presidential Election

The United States presidential election of 1948 was the 41st quadrennial presidential election, held on Tuesday, November 2, 1948. Incumbent President Harry S. Truman, the Democratic nominee, successfully ran for election against Thomas E. Dewey, the Republican nominee.

This election is considered to be the greatest election upset in American history. Virtually every prediction (with or without public opinion polls) indicated that Truman would be defeated by Dewey. Both parties had severe ideological splits, with the far left and far right of the Democratic Party running third-party campaigns. Truman's surprise victory was the fifth consecutive presidential win for the Democratic Party, a record never surpassed since contests against the Republican Party began in the 1850s. Truman's feisty campaign style energized his base of traditional Democrats, most of the white South, Catholic and Jewish voters, and—in a surprise—Midwestern farmers. Thus, Truman's election confirmed the Democratic Party's status as the nation's majority party, a status it would retain until the conservative realignment in 1968.

Incorrect Polls

As the campaign drew to a close, the polls showed Truman was gaining. Though Truman lost all nine of the Gallup Poll's post-convention surveys, Dewey's Gallup lead dropped from 17 points in late September to 9% in mid-October to just 5 points by the end of the month, just above the poll's margin of error. Although Truman was gaining momentum, most political analysts were reluctant to break with the conventional wisdom and say that a Truman victory was a serious possibility. The Roper Poll had suspended its presidential polling at the end of September, barring "some development of outstanding importance," which, in their subsequent view,

never occurred. Dewey was not unaware of his slippage, but he had been convinced by his advisers and family not to counterattack the Truman campaign.

Let's take a closer look at the polls. The Gallup, Roper, and Crossley polls all predicted a Dewey win. The actual results are shown in the following table: . How did this happen?

Candidate	Crossley Poll	Gallup Poll	Roper Poll	Election Results
Truman	45	44	38	50
Dewey	50	50	53	45
Others	5	6	9	5

1948 Election

The table shows the results of three polls against the actual results in the 1948 presidential election. Notice that Dewey was ahead in all three polls, but ended up losing the election.

The Crossley, Gallup, and Roper organizations all used quota sampling. Each interviewer was assigned a specified number of subjects to interview. Moreover, the interviewer was required to interview specified numbers of subjects in various categories, based on residential area, sex, age, race, economic status, and other variables. The intent of quota sampling is to ensure that the sample represents the population in all essential respects.

This seems like a good method on the surface, but where does one stop? What if a significant criterion was left out–something that deeply affected the way in which people vote? This would cause significant error in the results of the poll. In addition, quota sampling involves a human element. Pollsters, in reality, were left to poll whomever they chose. Research shows that the polls tended to overestimate the Republican vote. In earlier years, the margin of error was large enough that most polls still accurately predicted the winner, but in 1948, their luck ran out. Quota sampling had to go.

Mistake in the Newspapers

One of the most famous blunders came when the Chicago Tribune wrongfully printed the inaccurate headline, "Dewey Defeats Truman" on November 3, 1948, the day after incumbent United States President Harry S. Truman beat Republican challenger and Governor of New York Thomas E. Dewey.

The paper's erroneous headline became notorious after a jubilant Truman was photographed holding a copy

of the paper during a stop at St. Louis Union Station while returning by train from his home in Independence, Missouri to Washington, D.C .

Dewey Defeats Truman

President Truman holds up the newspaper that wrongfully reported his defeat.

Truman, as it turned out, won the electoral vote by a 303-189 majority over Dewey, although a swing of just a few thousand votes in Ohio, Illinois, and California would have produced a Dewey victory.

7.2.3: Using Chance in Survey Work

When conducting a survey, a sample can be chosen by chance or by more methodical methods.

Learning Objective

Distinguish between probability samples and non-probability samples for surveys

Key Takeaways

Key Points

- A probability sampling is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined.
- Probability sampling includes simple random sampling, systematic sampling, stratified sampling, and cluster sampling. These various ways of probability sampling have two things in common: every element has a known nonzero probability of being sampled, and random selection is involved at some point.
- Non-probability sampling is any sampling method wherein some elements of the population have no chance of selection (these are sometimes referred to as 'out of coverage'/'undercovered'), or where the probability of selection can't be accurately determined.

Key Terms

purposive sampling
: occurs when the researchers choose the sample based on who they think would be appropriate for the study; used primarily when there is a limited number of people that have expertise in the area being researched

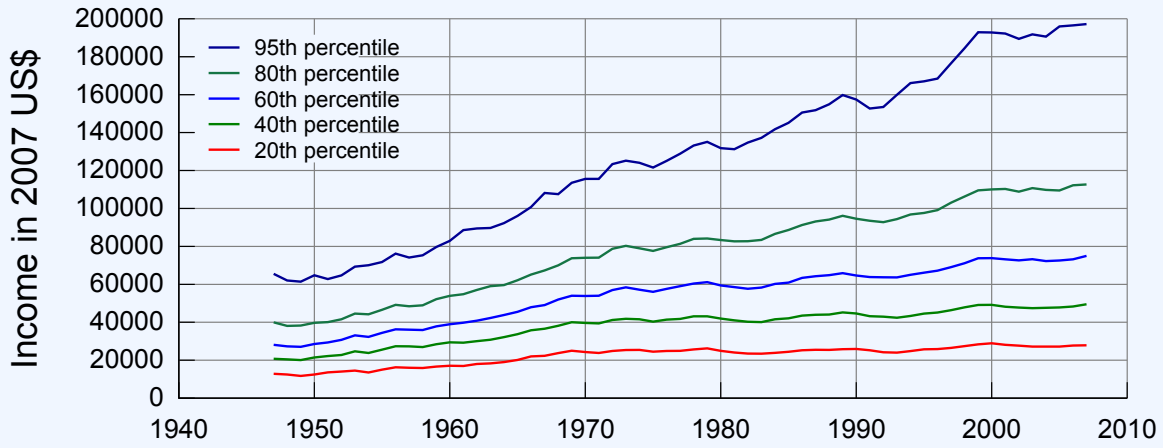
nonresponse
: the absence of a response

In order to conduct a survey, a sample from the population must be chosen. This sample can be chosen using chance, or it can be chosen more systematically.

Probability Sampling for Surveys

A probability sampling is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

Let's say we want to estimate the total income of adults living in a given street by using a survey with questions. We visit each household in that street, identify all adults living there, and randomly select one adult from each household. (For example, we can allocate each person a random number, generated from a uniform distribution between 0 and 1, and select the person with the highest number in each household). We then interview the selected person and find their income. People living on their own are certain to be selected, so we simply add their income to our estimate of the total. But a person living in a household of two adults has only a one-in-two chance of selection. To reflect this, when we come to such a household, we would count the selected person's income twice towards the total. (The person who is selected from that household can be loosely viewed as also representing the person who isn't selected.)



Income in the United States

Graph of United States income distribution from 1947 through 2007 inclusive, normalized to 2007 dollars. The data is from the US Census, which is a survey over the entire population, not just a sample.

In the above example, not everybody has the same probability of selection; what makes it a probability sample is the fact that each person's probability is known. When every element in the population does have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

Probability sampling includes: Simple Random Sampling, Systematic Sampling, Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling. These various ways of probability sampling have two things in common: every element has a known nonzero probability of being sampled, and random selection is involved at some point.

Non-Probability Sampling for Surveys

Non-probability sampling is any sampling method wherein some elements of the population have no chance of selection (these are sometimes referred to as 'out of coverage'/'undercovered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, non-probability sampling does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population.

Let's say we visit every household in a given street and interview the first person to answer the door. In any household with more than one occupant, this is a non-probability sample, because some people are more likely to answer the door (e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities.

Non-probability sampling methods include accidental sampling, quota sampling, and purposive sampling. In addition, nonresponse effects may turn any probability design into a non-probability design if the characteristics of nonresponse are not well understood, since nonresponse effectively modifies each element's probability of being sampled.

7.2.4: How Well Do Probability Methods Work?

Even when using probability sampling methods, bias can still occur.

Learning Objective

Analyze the problems associated with probability sampling

Key Takeaways

Key Points

- Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.
- Nonresponse occurs when an individual chosen for the sample can't be contacted or does not cooperate.
- Response bias occurs when a respondent lies about his or her true beliefs.
- The wording of questions–especially if they are leading questions– can affect the outcome of a survey.
- The larger the sample size, the more accurate the survey.

Key Terms

undercoverage
: Occurs when a survey fails to reach a certain portion of the population.

nonresponse
: the absence of a response

response bias
: Occurs when the answers given by respondents do not reflect their true beliefs.

Probability vs. Non-probability Sampling

In earlier sections, we discussed how samples can be chosen. Failure to use probability sampling may result in bias or systematic errors in the way the sample represents the population. This is especially true of voluntary response samples–in which the respondents choose themselves if they want to be part of a survey– and convenience samples–in which individuals easiest to reach are chosen.

However, even probability sampling methods that use chance to select a sample are prone to some problems. Recall some of the methods used in probability sampling: simple random samples, stratified samples, cluster samples, and systematic samples. In these methods, each member of the population has a chance of being chosen for the sample, and that chance is a known probability.

Problems With Probability Sampling

Random sampling eliminates some of the bias that presents itself in sampling, but when a sample is chosen by human beings, there are always going to be some unavoidable problems. When a sample is chosen, we first need an accurate and complete list of the population. This type of list is often not available, causing most samples to suffer from undercoverage. For example, if we chose a sample from a list of households, we will miss those who are homeless, in prison, or living in a college dorm. In another example, a telephone survey calling landline phones will potentially miss those who are unlisted, those who only use a cell phone, and those who do not have a phone at all. Both of these examples will cause a biased sample in which poor people, whose opinions may very well differ from those of the rest of the population, are underrepresented.

Another source of bias is nonresponse, which occurs when a selected individual cannot be contacted or refuses to participate in the survey. Many people do not pick up the phone when they do not know the person who is calling . Nonresponse is often higher in urban areas, so most researchers conducting surveys will substitute other people in the same area to avoid favoring rural areas. However, if the people eventually contacted differ from those who are rarely at home or refuse to answer questions for one reason or another, some bias will still be present.

A third example of bias is called response bias. Respondents may not answer questions truthfully, especially if the survey asks about illegal or unpopular behavior. The race and sex of the interviewer may influence people to respond in a way that is more extreme than their true beliefs. Careful training of pollsters can greatly reduce response bias.

Finally, another source of bias can come in the wording of questions. Confusing or leading questions can strongly influence the way a respondent answers questions.

Ringing Phone

Conclusion

When reading the results of a survey, it is important to know the exact questions asked, the rate of non-response, and the method of survey before you trust a poll. In addition, remember that a larger sample size will provide more accurate results.

7.2.5: The Gallup Poll

The Gallup Poll is a public opinion poll that conducts surveys in 140 countries around the world.

Learning Objective

Examine the pros and cons of the way in which the Gallup Poll is conducted

Key Takeaways

Key Points

- The Gallup Poll measures and tracks the public's attitudes concerning virtually every political, social, and economic issues of the day in 140 countries around the world.
- The Gallup Polls have been traditionally known for their accuracy in predicting presidential elections in the United States from 1936 to 2008. They were only incorrect in 1948 and 1976.
- Today, Gallup samples people using both landline telephones and cell phones. They have gained much criticism for not adapting quickly enough for a society that is growing more and more towards using only their cell phones over landlines.

Key Terms

Objective
: not influenced by the emotions or prejudices

public opinion polls
: surveys designed to represent the beliefs of a population by conducting a series of questions and then extrapolating generalities in ratio or within confidence intervals

Overview of the Gallup Organization

Gallup, Inc. is a research-based performance-management consulting company. Originally founded by George Gallup in 1935, the company became famous for its public opinion polls, which were conducted in the United States and other countries. Today, Gallup has more than 40 offices in 27 countries. The world headquarters are located in Washington, D.C. , while the operational headquarters are in Omaha, Nebraska. Its current Chairman and CEO is Jim Clifton.

The Gallup Organization

History of Gallup

George Gallup founded the American Institute of Public Opinion, the precursor to the Gallup Organization, in Princeton, New Jersey in 1935. He wished to objectively determine the opinions held by the people. To ensure his independence and objectivity, Dr. Gallup resolved that he would undertake no polling that was paid for or sponsored in any way by special interest groups such as the Republican and Democratic parties, a commitment that Gallup upholds to this day.

In 1936, Gallup successfully predicted that Franklin Roosevelt would defeat Alfred Landon for the U.S. presidency; this event quickly popularized the company. In 1938, Dr. Gallup and Gallup Vice President David Ogilvy began conducting market research for advertising companies and the film industry. In 1958, the modern Gallup Organization was formed when George Gallup grouped all of his polling operations into one organization. Since then, Gallup has seen huge expansion into several other areas.

The Gallup Poll

The Gallup Poll is the division of Gallup that regularly conducts public opinion polls in more than 140 countries around the world. Gallup Polls are often referenced in the mass media as a reliable and objective audience measurement of public opinion. Gallup Poll results, analyses, and videos are published daily on Gallup.com in the form of data-driven news. The poll loses about $10 million a year but gives the company the visibility of a very well-known brand.

Historically, the Gallup Poll has measured and tracked the public's attitudes concerning virtually every political, social, and economic issue of the day, including highly sensitive and controversial subjects. In 2005, Gallup began its World Poll, which continually surveys citizens in more than 140 countries, representing 95% of the world's adult population. General and regional-specific questions, developed in collaboration with the world's leading behavioral economists, are organized into powerful indexes and topic areas that correlate with real-world outcomes.

Reception of the Poll

The Gallup Polls have been recognized in the past for their accuracy in predicting the outcome of United States presidential elections, though they have come under criticism more recently. From 1936 to 2008, Gallup correctly predicted the winner of each election–with the notable exceptions of the 1948 Thomas Dewey-Harry S. Truman election, when nearly all pollsters predicted a Dewey victory, and the 1976 election, when they inaccurately projected a slim victory by Gerald Ford over Jimmy Carter. For the 2008 U.S. presidential election, Gallup correctly predicted the winner, but was rated 17th out of 23 polling organizations in terms of the precision of its pre-election polls relative to the final results. In 2012, Gallup's final election survey had Mitt Romney 49% and Barack Obama 48%, compared to the election results showing Obama with 51.1% to Romney's 47.2%. Poll analyst Nate Silver found that Gallup's results were the least accurate of the 23 major polling firms Silver analyzed, having the highest incorrect average of being 7.2 points away from the final result. Frank Newport, the Editor-in-Chief of Gallup, responded to the criticism by stating that Gallup simply makes an estimate of the national popular vote rather than predicting the winner, and that their final poll was within the statistical margin of error.

In addition to the poor results of the poll in 2012, many people are criticizing Gallup on their sampling techniques. Gallup conducts 1,000 interviews per day, 350 days out of the year, among both landline and cell phones across the U.S., for its health and well-being survey. Though Gallup surveys both landline and cell phones, they conduct only 150 cell phone samples out of 1000, making up 15%. The population of the U.S. that relies only on cell phones (owning no landline connections) makes up more than double that number, at 34%. This fact has been a major criticism in recent times of the reliability Gallup polling, compared to other polls, in its failure to compensate accurately for the quick adoption of "cell phone only" Americans.

7.2.6: Telephone Surveys

Telephone surveys can reach a wide range of people very quickly and very inexpensively.

Learning Objective

Identify the advantages and disadvantages of telephone surveys

Key Takeaways

Key Points

- About 95% of people in the United States have a telephone (see, so conducting a poll by calling people is a good way to reach nearly every part of the population.
- Calling people by telephone is a quick process, allowing researches to gain a lot of data in a short amount of time.
- In certain polls, the interviewer or interviewee (or both) may wish to remain anonymous, which can be achieved if the poll is conducted via telephone by a third party.
- Non-response bias is one of the major problems with telephone surveys as many people do not answer calls from people they do not know.
- Due to certain uncontrollable factors (e.g., unlisted phone numbers, people who only use cell phones, or instances when no one is home/available to take pollster calls), undercoverage can negatively affect the outcome of telephone surveys.

Key Terms

undercoverage
: Occurs when a survey fails to reach a certain portion of the population.

response bias
: Occurs when the answers given by respondents do not reflect their true beliefs.

non-response bias
: Occurs when the sample becomes biased because some of those initially selected refuse to respond.

A telephone survey is a type of opinion poll used by researchers. As with other methods of polling, their are advantages and disadvantages to utilizing telephone surveys.

Advantages

- *Large scale accessibility*. About 95% of people in the United States have a telephone (see), so conducting a poll by via telephone is a good way to reach most parts of the population.
- *Efficient data collection*. Conducting calls via telephone produces a quick process, allowing researches to gain a large amount of data in a short amount of time. Previously, pollsters physically had to go to each interviewee's home (which, obviously, was more time consuming).
- *Inexpensive*. Phone interviews are not costly (e.g., telephone researchers do not pay for travel).
- *Anonymity*. In certain polls, the interviewer or interviewee (or both) may wish to remain anonymous, which can be achieved if the poll is conducted over the phone by a third party.

Disadvantages

- *Lack of visual materials*. Depending on what the researchers are asking, sometimes it may be helpful for the respondent to see a product in person, which of course, cannot be done over the phone.
- *Call screening*. As some people do not answer calls from strangers, or may refuse to answer the poll, poll samples are not always representative samples from a population due to what is known as non-response bias. In this type of bias, the characteristics of those who agree to be interviewed may be markedly different from those who decline. That is, the actual sample is a biased version of the population the pollster wants to analyze. If those who refuse to answer, or are never reached, have the same characteristics as those who do answer, then the final results should be unbiased. However, if those who

do not answer have different opinions, then the results have bias. In terms of election polls, studies suggest that bias effects are small, but each polling firm has its own techniques for adjusting weights to minimize selection bias.

- *Undercoverage*. Undercoverage is a highly prevalent source of bias. If the pollsters only choose telephone numbers from a telephone directory, they miss those who have unlisted landlines or only have cell phones (which is is becoming more the norm). In addition, if the pollsters only conduct calls between 9:00 a.m and 5:00 p.m, Monday through Friday, they are likely to miss a huge portion of the working population—those who may have very different opinions than the non-working population.

7.2.7: Chance Error and Bias

Chance error and bias are two different forms of error associated with sampling.

Learning Objective

Differentiate between random, or chance, error and bias

Key Takeaways

Key Points

- The error that is associated with the unpredictable variation in the sample is called a random, or chance, error. It is only an "error" in the sense that it would automatically be corrected if we could survey the entire population.
- Random error cannot be eliminated completely, but it can be reduced by increasing the sample size.

- A sampling bias is a bias in which a sample is collected in such a way that some members of the intended population are less likely to be included than others.
- There are various types of bias, including selection from a specific area, self-selection, pre-screening, and exclusion.

Key Terms

bias
: (Uncountable) Inclination towards something; predisposition, partiality, prejudice, preference, predilection.

random sampling
: a method of selecting a sample from a statistical population so that every subject has an equal chance of being selected

standard error
: A measure of how spread out data values are around the mean, defined as the square root of the variance.

Sampling Error

In statistics, a sampling error is the error caused by observing a sample instead of the whole population. The sampling error can be found by subtracting the value of a parameter from the value of a statistic. The variations in the possible sample values of a statistic can theoretically be expressed as sampling errors, although in practice the exact sampling error is typically unknown.

In sampling, there are two main types of error: systematic errors (or biases) and random errors (or chance errors).

Random/Chance Error

Random sampling is used to ensure that a sample is truly representative of the entire population. If we were to select a perfect sample (which does not exist), we would reach the same exact conclusions that we would have reached if we had surveyed the entire population. Of course, this is not possible, and the error that is associated with the unpredictable variation in the sample is called random, or chance, error. This is only an "error" in the sense that it would automatically be corrected if we could survey the entire population rather than just a sample taken from it. It is not a mistake made by the researcher.

Random error always exists. The size of the random error, however, can generally be controlled by taking a large enough random sample from the population. Unfortunately, the high cost of doing so can be prohibitive. If the observations are collected from a random sample, statistical theory provides probabilistic estimates of the likely size of the error for a particular statistic or estimator. These are often expressed in terms of its standard error:

SEx¯=sn">$SE_x = \frac{s}{\sqrt{n}}$

Bias

In statistics, sampling bias is a bias in which a sample is collected in such a way that some members of the intended population are less likely to be included than others. It results in a biased sample, a non-random sample of a population in which all individuals, or instances, were not equally likely to have been selected. If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling.

There are various types of sampling bias:

- *Selection from a specific real area*. For example, a survey of high school students to measure teenage use of illegal drugs will be a biased sample because it does not include home-schooled students or dropouts.
- *Self-selection bias, which is possible whenever the group of people being studied has any form of* control *over whether to participate.* Participants' decision to participate may be correlated with traits that affect the study, making the participants a non-representative sample. For example, people who have strong opinions or substantial knowledge may be more willing to spend time answering a survey than those who do not.
- *Pre-screening of trial participants, or advertising for volunteers within particular groups.* For example, a study to "prove" that smoking does not affect fitness might recruit at the local fitness center, but advertise for smokers during the advanced aerobics class and for non-smokers during the weight loss sessions.
- *Exclusion bias, or exclusion of particular groups from the sample.* For example, subjects may be left out if they either migrated into the study area or have moved out of the area.

Attributions

- The Literary Digest Poll
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.

- "The Literary Digest."
  http://en.wikipedia.org/wiki/The_Literary_Digest.
  Wikipedia
  CC BY-SA 3.0.
- "straw poll."
  http://en.wiktionary.org/wiki/straw_poll.
  Wiktionary
  CC BY-SA 3.0.
- "bellwether."
  http://en.wiktionary.org/wiki/bellwether.
  Wiktionary
  CC BY-SA 3.0.
- "LiteraryDigest-19210219."
  http://en.wikipedia.org/wiki/File:LiteraryDigest-19210219.jpg.
  Wikipedia
  CC BY-SA.
- "ElectoralCollege1936."
  http://en.wikipedia.org/wiki/File:ElectoralCollege1936.svg.
  Wikipedia
  CC BY-SA.

- The Year the Polls Elected Dewey
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Dewey Defeats Truman."
    http://en.wikipedia.org/wiki/Dewey_Defeats_Truman.
    Wikipedia
    CC BY-SA 3.0.
  - "United States presidential election, 1948."
    http://en.wikipedia.org/wiki/United_States_presidential_election,_1948.
    Wikipedia
    CC BY-SA 3.0.
  - "quadrennial."
    http://en.wiktionary.org/wiki/quadrennial.
    Wiktionary
    CC BY-SA 3.0.

- "margin of error."
  http://en.wiktionary.org/wiki/margin_of_error.
  Wiktionary
  CC BY-SA 3.0.
- "The 1948 Presidential Election Polls."
  http://www.math.uah.edu/stat/data/1948Election.html.
  The University of Alabama in Huntsville
  CC BY.
- "Deweytruman12."
  http://en.wikipedia.org/wiki/File:Deweytruman12.jpg.
  Wikipedia
  CC BY-SA.
- "The 1948 Presidential Election Polls."
  http://www.math.uah.edu/stat/data/1948Election.html.
  The University of Alabama in Huntsville
  CC BY.

- Using Chance in Survey Work
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Sampling (statistics)."
    http://en.wikipedia.org/wiki/Sampling_(statistics)%23Probability_and_nonprobability_sampling.
    Wikipedia
    CC BY-SA 3.0.
  - "nonresponse."
    http://en.wiktionary.org/wiki/nonresponse.
    Wiktionary
    CC BY-SA 3.0.
  - "United States Income Distribution 1947-2007."
    http://commons.wikimedia.org/wiki/File:United_States_Income_Distribution_1947-2007.svg.
    Wikimedia
    CC BY-SA.
- How Well Do Probability Methods Work?
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning

CC BY-SA 3.0.

- "Sampling (statistics)."
  http://en.wikipedia.org/wiki/Sampling_(statistics).
  Wikipedia
  CC BY-SA 3.0.
- "nonresponse."
  http://en.wiktionary.org/wiki/nonresponse.
  Wiktionary
  CC BY-SA 3.0.
- "Tower, Phone, Mail, Icon, Rings – Free image – 25477."
  http://pixabay.com/en/tower-phone-mail-icon-rings-25477/.
  Pixabay
  CC BY.

- The Gallup Poll
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Gallup (company)."
    http://en.wikipedia.org/wiki/Gallup_(company).
    Wikipedia
    CC BY-SA 3.0.
  - "Objective."
    http://en.wiktionary.org/wiki/Objective.
    Wiktionary
    CC BY-SA 3.0.
  - "Gallup Portrait."
    http://en.wikipedia.org/wiki/File:Gallup_Portrait.jpg.
    Wikipedia
    CC BY-SA.
- Telephone Surveys
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Opinion poll."
    http://en.wikipedia.org/wiki/Opinion_poll.

Wikipedia
CC BY-SA 3.0.

- Chance Error and Bias
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "standard error."
    http://en.wiktionary.org/wiki/standard_error.
    Wiktionary
    CC BY-SA 3.0.
  - "Sampling error."
    http://en.wikipedia.org/wiki/Sampling_error.
    Wikipedia
    CC BY-SA 3.0.
  - "Sampling bias."
    http://en.wikipedia.org/wiki/Sampling_bias.
    Wikipedia
    CC BY-SA 3.0.
  - "bias."
    http://en.wiktionary.org/wiki/bias.
    Wiktionary
    CC BY-SA 3.0.

7.3 SAMPLING DISTRIBUTIONS

7.3: Sampling Distributions

7.3.1: What Is a Sampling Distribution?

The sampling distribution of a statistic is the distribution of the statistic for all possible samples from the same population of a given size.

Learning Objective

Recognize the characteristics of a sampling distribution

Key Takeaways

Key Points

- A critical part of inferential statistics involves determining how far sample statistics are likely to vary from each other and from the population parameter.
- The sampling distribution of a statistic is the distribution of that statistic, considered as a random variable, when derived from a random sample of size n.
- Sampling distributions allow analytical considerations to be based on the sampling

distribution of a statistic rather than on the joint probability distribution of all the individual sample values.

- The sampling distribution depends on: the underlying distribution of the population, the statistic being considered, the sampling procedure employed, and the sample size used.

Key Terms

inferential statistics
: A branch of mathematics that involves drawing conclusions about a population based on sample data drawn from it.

sampling distribution
: The probability distribution of a given statistic based on a random sample.

Suppose you randomly sampled 10 women between the ages of 21 and 35 years from the population of women in Houston, Texas, and then computed the mean height of your sample. You would not expect your sample mean to be equal to the mean of all women in Houston. It might be somewhat lower or higher, but it would not equal the population mean exactly. Similarly, if you took a second sample of 10 women from the same population, you would not expect the mean of this second sample to equal the mean of the first sample.

Houston Skyline

Suppose you randomly sampled 10 people from the population of women in Houston, Texas between the ages of 21 and 35 years and computed the mean height of your sample. You would not expect your sample mean to be equal to the mean of all women in Houston.

Inferential statistics involves generalizing from a sample to a population. A critical part of inferential statistics involves determining how far sample statistics are likely to vary from each other and from the population parameter. These determinations are based on sampling distributions. The sampling distribution of a statistic is the distribution of that statistic, considered as a random variable, when derived from a random sample of size n">n. It may be considered as the distribution of the statistic for all possible samples from the same population of a given size. Sampling distributions allow analytical considerations to be based on the sampling distribution of a statistic rather than on the joint probability distribution of all the individual sample values.

The sampling distribution depends on: the underlying distribution of the population, the statistic being considered, the sampling procedure employed, and the sample size used. For example, consider a normal population with mean μ">μ and variance σ">σ. Assume we repeatedly take samples of a given size from this population and calculate the arithmetic mean for each sample. This statistic is then called the sample mean. Each sample has its own average value, and the distribution of these averages is called the "sampling distribution of the sample mean. " This distribution is normal since the underlying population

is normal, although sampling distributions may also often be close to normal even when the population distribution is not.

An alternative to the sample mean is the sample median. When calculated from the same population, it has a different sampling distribution to that of the mean and is generally not normal (but it may be close for large sample sizes).

7.3.2: Properties of Sampling Distributions

Knowledge of the sampling distribution can be very useful in making inferences about the overall population.

Learning Objective

Describe the general properties of sampling distributions and the use of standard error in analyzing them

Key Takeaways

Key Points

- In practice, one will collect sample data and, from these data, estimate parameters of the population distribution.
- Knowing the degree to which means from different samples would differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean.
- The standard deviation of the sampling distribution of a statistic is referred to as the standard error of that quantity.

- If all the sample means were very close to the population mean, then the standard error of the mean would be small.
- On the other hand, if the sample means varied considerably, then the standard error of the mean would be large.

Key Terms

inferential statistics
: A branch of mathematics that involves drawing conclusions about a population based on sample data drawn from it.

sampling distribution
: The probability distribution of a given statistic based on a random sample.

Sampling Distributions and Inferential Statistics

Sampling distributions are important for inferential statistics. In practice, one will collect sample data and, from these data, estimate parameters of the population distribution. Thus, knowledge of the sampling distribution can be very useful in making inferences about the overall population.

For example, knowing the degree to which means from different samples differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean. Fortunately, this information is directly available from a sampling distribution. The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean. This standard deviation is called the standard error of the mean.

Standard Error

The standard deviation of the sampling distribution of a statistic is referred to as the standard error of that quantity. For the case where the statistic is the sample mean, and samples are uncorrelated, the standard error is:

SEx¯=sn">$SE_x = \frac{s}{\sqrt{n}}$

Where S is the sample standard deviation and n is the size (number of items) in the sample. An important implication of this formula is that the sample size must be quadrupled (multiplied by 4) to achieve half the measurement error. When designing statistical studies where cost is a factor, this may have a role in understanding cost-benefit tradeoffs.

If all the sample means were very close to the population mean, then the standard error of the mean would be small. On the other hand, if the sample means varied considerably, then the standard error of the mean would be large. To be specific, assume your sample mean is 125 and you estimated that the standard error of the mean is 5. If you had a normal distribution, then it would be likely that your sample mean would be within 10 units of the population mean since most of a normal distribution is within two standard deviations of the mean.

More Properties of Sampling Distributions

1. The overall shape of the distribution is symmetric and approximately normal.
2. There are no outliers or other important deviations from the overall pattern.
3. The center of the distribution is very close to the true population mean.

A statistical study can be said to be biased when one outcome is systematically favored over another. However, the study can be said to be unbiased if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

Finally, the variability of a statistic is described by the spread of its sampling distribution. This spread is determined by the sampling design and the size of the sample. Larger samples give smaller spread. As long as the population is much larger than the sample (at least 10 times as large), the spread of the sampling distribution is approximately the same for any population size

7.3.3: Creating a Sampling Distribution

Learn to create a sampling distribution from a discrete set of data.

Learning Objective

Differentiate between a frequency distribution and a sampling distribution

Key Takeaways

Key Points

- Consider three pool balls, each with a number on it.
- Two of the balls are selected randomly (with replacement), and the average of their numbers is computed.
- The relative frequencies are equal to the frequencies divided by nine because there are nine possible outcomes.
- The distribution created from these relative frequencies is called the sampling distribution of the mean.
- As the number of samples approaches infinity, the frequency distribution will approach the sampling distribution.

Key Terms

sampling distribution
: The probability distribution of a given statistic based on a random sample.

frequency distribution
: a representation, either in a graphical or tabular format, which displays the number of observations within a given interval

We will illustrate the concept of sampling distributions with a simple example. Consider three pool balls, each with a number on it. Two of the balls are selected randomly (with replacement), and the average of their numbers is computed. All possible outcomes are shown below.

Outcome	Ball 1	Ball 2	Mean
1	1	1	1.0
2	1	2	1.5
3	1	3	2.0
4	2	1	1.5
5	2	2	2.0
6	2	3	2.5
7	3	1	2.0
8	3	2	2.5
9	3	3	3.0

Pool Ball Example 1

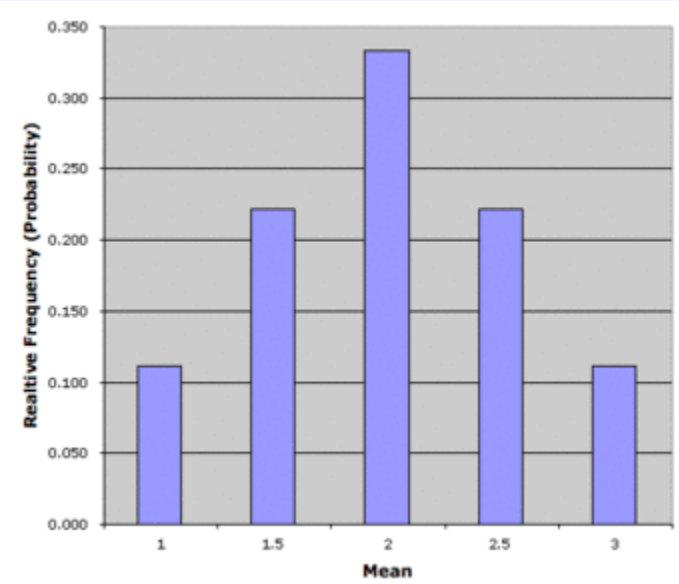
This table shows all the possible outcome of selecting two pool balls randomly from a population of three.

Notice that all the means are either 1.0, 1.5, 2.0, 2.5, or 3.0. The frequencies of these means are shown below. The relative frequencies are equal to the frequencies divided by nine because there are nine possible outcomes.

Pool Ball Example 2

This table shows the frequency of means for N=2.

The figure below shows a relative frequency distribution of the means. This distribution is also a probability distribution since the y-axis is the probability of obtaining a given mean from a sample of two balls in addition to being the relative frequency.



Relative Frequency Distribution

Relative frequency distribution of our pool ball example.

The distribution shown in the above figure is called the sampling distribution of the mean. Specifically, it is the sampling distribution of the mean for a sample size of 2 (N=2″>N=2). For this simple example, the distribution of pool balls and the sampling distribution are both discrete distributions. The pool balls have only the numbers 1, 2, and 3, and a sample mean can have one of only five possible values.

There is an alternative way of conceptualizing a sampling distribution that will be useful for more complex distributions. Imagine that two balls are sampled (with replacement), and the mean of the two balls is computed and recorded. This process is repeated for a second sample, a third sample, and eventually thousands of samples. After thousands of samples are taken and the mean is computed for each, a relative frequency distribution is drawn. The more samples, the closer the relative frequency distribution will come to the sampling distribution shown in the above figure. As the number of samples approaches infinity , the frequency distribution will approach the sampling distribution. This means that you can conceive of a sampling distribution as being a frequency distribution based on a very large number of samples. To be strictly correct, the sampling distribution only equals the frequency distribution exactly when there is an infinite number of samples.

7.3.4: Continuous Sampling Distributions

When we have a truly continuous distribution, it is not only impractical but actually impossible to enumerate all possible outcomes.

Learning Objective

Differentiate between discrete and continuous sampling distributions

Key Takeaways

Key Points

- In continuous distributions, the probability of obtaining any single value is zero.
- Therefore, these values are called probability densities rather than probabilities.
- A probability density function, or density of a continuous random variable, is a function that describes the relative likelihood for this random variable to take on a given value.

Key Term

probability density function
: any function whose integral over a set gives the probability that a random variable has a value in that set

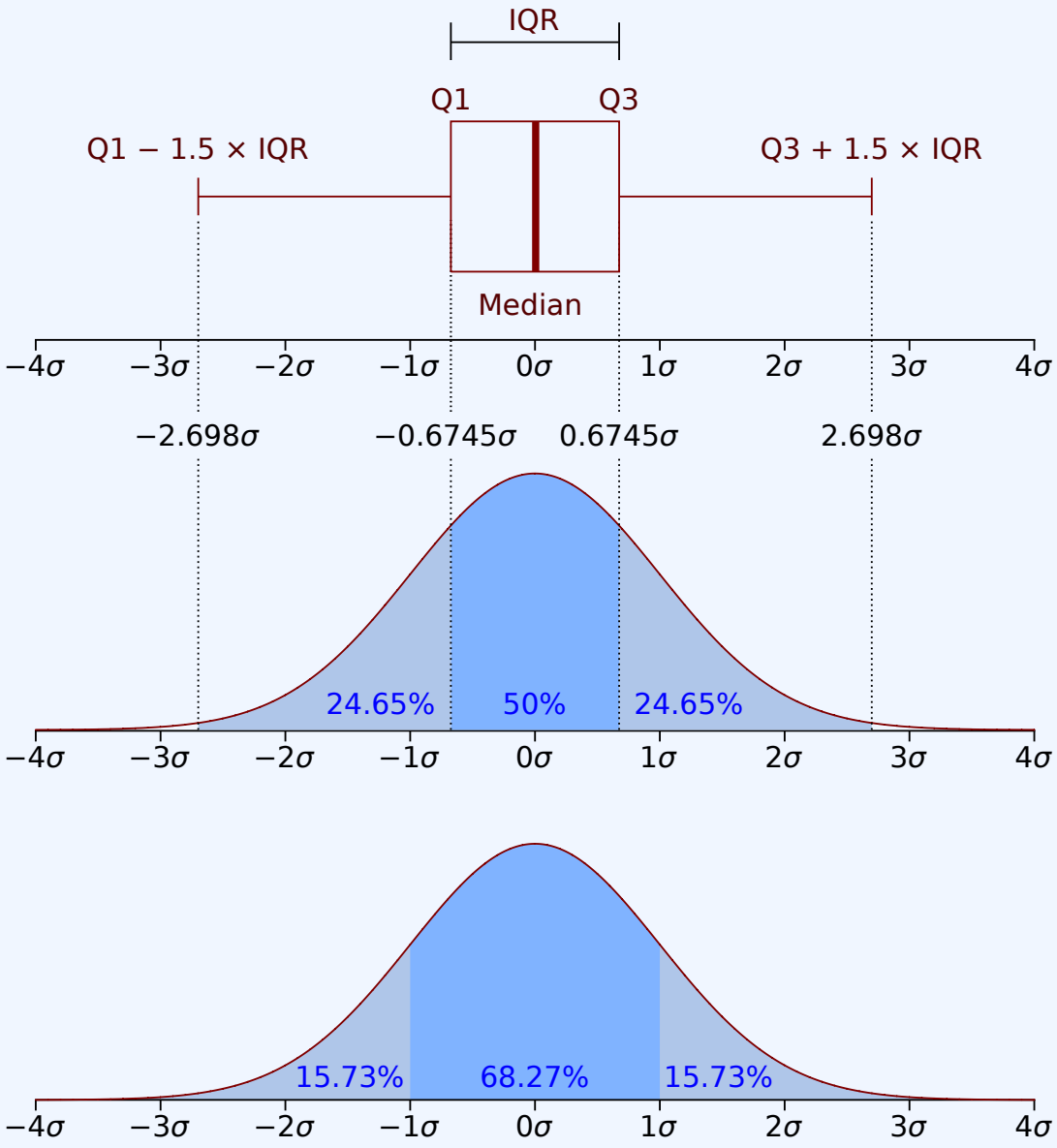
In the previous section, we created a sampling distribution out of a population consisting of three pool balls. This distribution was discrete, since there were a finite number of possible observations. Now we will consider sampling distributions when the population distribution is continuous.

What if we had a thousand pool balls with numbers ranging from 0.001 to 1.000 in equal steps? Note that although this distribution is not really continuous, it is close enough to be considered continuous for practical purposes. As before, we are interested in the distribution of the means we would get if we sampled two balls and computed the mean of these two. In the previous example, we started by computing the mean for each of the nine possible outcomes. This would get a bit tedious for our current example since there are 1,000,000 possible outcomes (1,000 for the first ball multiplied by 1,000 for the second.) Therefore, it is more convenient to use our second conceptualization of sampling distributions, which conceives of sampling distributions in terms of relative frequency distributions– specifically, the relative frequency distribution that would occur if samples of two balls were repeatedly taken and the mean of each sample computed.

Probability Density Function

When we have a truly continuous distribution, it is not only impractical but actually impossible to enumerate all possible outcomes. Moreover, in continuous distributions, the probability of obtaining any single value is zero. Therefore, these values are called probability densities rather than probabilities.

A probability density function, or density of a continuous random variable, is a function that describes the relative likelihood for this random variable to take on a given value. The probability for the random variable to fall within a particular region is given by the integral of this variable's density over the region .



Probability Density Function

Boxplot and probability density function of a normal distribution N(0, 2).

7.3.5: Mean of All Sample Means (μ x)

The mean of the distribution of differences between sample means is equal to the difference between population means.

Learning Objectives

Discover that the mean of the distribution of differences between sample means is equal to the difference between population means

Key Takeaways

Key Points

- Statistical analysis are very often concerned with the difference between means.
- The mean of the sampling distribution of the mean is $\mu_{M1-M2} = \mu_{1-2}$.
- The variance sum law states that the variance of the sampling distribution of the difference between means is equal to the variance of the sampling distribution of the mean for Population 1 plus the variance of the sampling distribution of the mean for Population 2.

Key Term

sampling distribution
: The probability distribution of a given statistic based on a random sample.

Statistical analyses are, very often, concerned with the difference between means. A typical example is an experiment designed to compare the mean of a control group with the mean of an experimental group.

Inferential statistics used in the analysis of this type of experiment depend on the sampling distribution of the difference between means.

The sampling distribution of the difference between means can be thought of as the distribution that would result if we repeated the following three steps over and over again:

1. Sample n_1 scores from Population 1 and n_2 scores from Population 2;
2. Compute the means of the two samples (M_1 and M_2);
3. Compute the difference between means $M_1 – M_2$. The distribution of the differences between means is the sampling distribution of the difference between means.

The mean of the sampling distribution of the mean is:

$\mu_{M1-M2} = \mu_{1-2}$,

which says that the mean of the distribution of differences between sample means is equal to the difference between population means. For example, say that mean test score of all 12-year olds in a population is 34 and the mean of 10-year olds is 25. If numerous samples were taken from each age group and the mean difference computed each time, the mean of these numerous differences between sample means would be 34 – 25 = 9.

The variance sum law states that the variance of the sampling distribution of the difference between means is equal to the variance of the sampling distribution of the mean for Population 1 plus the variance of the sampling distribution of the mean for Population 2. The formula for the variance of the sampling distribution of the difference between means is as follows:

σM1−M22=σM12n1+σM22n2">$\sigma^2_{M_1-M_2} = \frac{\sigma^2_{M1}}{n_1} + \frac{\sigma^2_{M2}}{n_2}$

Recall that the standard error of a sampling distribution is the standard deviation of the sampling distribution, which is the square root of the above variance.

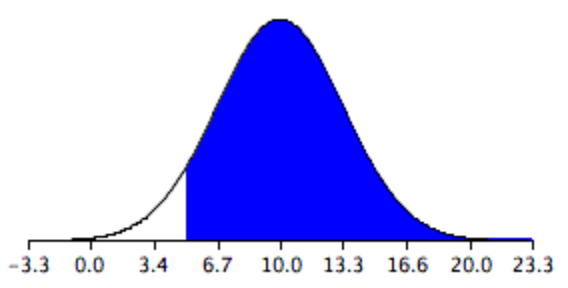
Let's look at an application of this formula to build a sampling distribution of the difference between means. Assume there are two species of green beings on Mars. The mean height of Species 1 is 32, while the mean height of Species 2 is 22. The variances of the two species are 60 and 70, respectively, and the heights of both species are normally distributed. You randomly sample 10 members of Species 1 and 14 members of Species 2.

The difference between means comes out to be 10, and the standard error comes out to be 3.317.

$\mu_{M1-M2} = 32 – 22 = 10$

Standard error equals the square root of (60 / 10) + (70 / 14) = 3.317.

The resulting sampling distribution as diagramed in , is normally distributed with a mean of 10 and a standard deviation of 3.317.



Sampling Distribution of the Difference Between Means

The distribution is normally distributed with a mean of 10 and a standard deviation of 3.317.

7.3.6: Shapes of Sampling Distributions

The overall shape of a sampling distribution is expected to be symmetric and approximately normal.

Learning Objective

Give examples of the various shapes a sampling distribution can take on

Key Takeaways

Key Points

- The concept of the shape of a distribution refers to the shape of a probability distribution.
- It most often arises in questions of finding an appropriate distribution to use to model the statistical properties of a population, given a sample from that population.
- A sampling distribution is assumed to have no outliers or other important deviations from the overall pattern.
- When calculated from the same population, the sample median has a different sampling distribution to that of the mean and is generally not normal; although, it may be close for large sample sizes.

Key Terms

normal distribution
: A family of continuous probability distributions such that the probability density function is the normal (or Gaussian) function.

skewed
: Biased or distorted (pertaining to statistics or information).

Pareto Distribution
: The Pareto distribution, named after the Italian economist Vilfredo Pareto, is a power law probability distribution that is used in description of social, scientific, geophysical, actuarial, and many other types of observable phenomena.

probability distribution
: A function of a discrete random variable yielding the probability that the variable will have a given value.

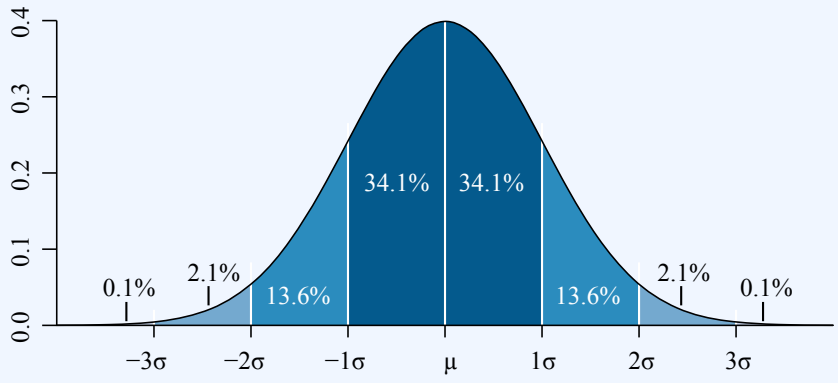
The "shape of a distribution" refers to the shape of a probability distribution. It most often arises in questions of finding an appropriate distribution to use in order to model the statistical properties of a population, given a sample from that population. The shape of a distribution will fall somewhere in a continuum where a flat distribution might be considered central; and where types of departure from this include:

- mounded (or unimodal)
- u-shaped
- j-shaped
- reverse-j-shaped
- multi-modal

The shape of a distribution is sometimes characterized by the behaviors of the tails (as in a long or short tail). For example, a flat distribution can be said either to have no tails or to have short tails. A normal distribution is usually regarded as having short tails, while a Pareto distribution has long tails. Even in the relatively simple case of a mounded distribution, the distribution may be skewed to the left or skewed to the right (with symmetric corresponding to no skew).

As previously mentioned, the overall shape of a sampling distribution is expected to be symmetric and approximately normal. This is due to the fact, or assumption, that there are no outliers or other important deviations from the overall pattern. This fact holds true when we repeatedly take samples of a given size from a population and calculate the arithmetic mean for each sample.

An alternative to the sample mean is the sample median. When calculated from the same population, it has a different sampling distribution to that of the mean and is generally not normal; although, it may be close for large sample sizes.



The Normal Distribution

Sample distributions, when the sampling statistic is the mean, are generally expected to display a normal distribution.

7.3.7: Sampling Distributions and the Central Limit Theorem

The central limit theorem for sample means states that as larger samples are drawn, the sample means form their own normal distribution.

Learning Objective

Illustrate that as the sample size gets larger, the sampling distribution approaches normality

Key Takeaways

Key Points

- The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by n, the sample size.
- is the number of values that are averaged together not the number of times the experiment is done.
- The usefulness of the theorem is that the sampling distribution approaches normality regardless of the shape of the population distribution.

Key Terms

central limit theorem
: The theorem that states: If the sum of independent identically distributed random variables has a finite variance, then it will be (approximately) normally distributed.

sampling distribution
: The probability distribution of a given statistic based on a random sample.

Example

Imagine rolling a large number of identical, unbiased dice. The distribution of the sum (or average) of the rolled numbers will be well approximated by a normal distribution. Since real-world quantities are often the balanced sum of many unobserved random events, the central limit theorem also provides a partial explanation for the prevalence of the normal probability distribution. It also justifies the approximation of large-sample statistics to the normal distribution in controlled experiments.

The central limit theorem states that, given certain conditions, the mean of a sufficiently large number of independent random variables, each with a well-defined mean and well-defined variance, will be (approximately) normally distributed. The central limit theorem has a number of variants. In its common form, the random variables must be identically distributed. In variants, convergence of the mean to the normal distribution also occurs for non-identical distributions, given that they comply with certain conditions.

The central limit theorem for sample means specifically says that if you keep drawing larger and larger samples (like rolling 1, 2, 5, and, finally, 10 dice) and calculating their means the sample means form their own normal distribution (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by n″>n, the sample size. n″>n is the number of values that are averaged together not the number of times the experiment is done.

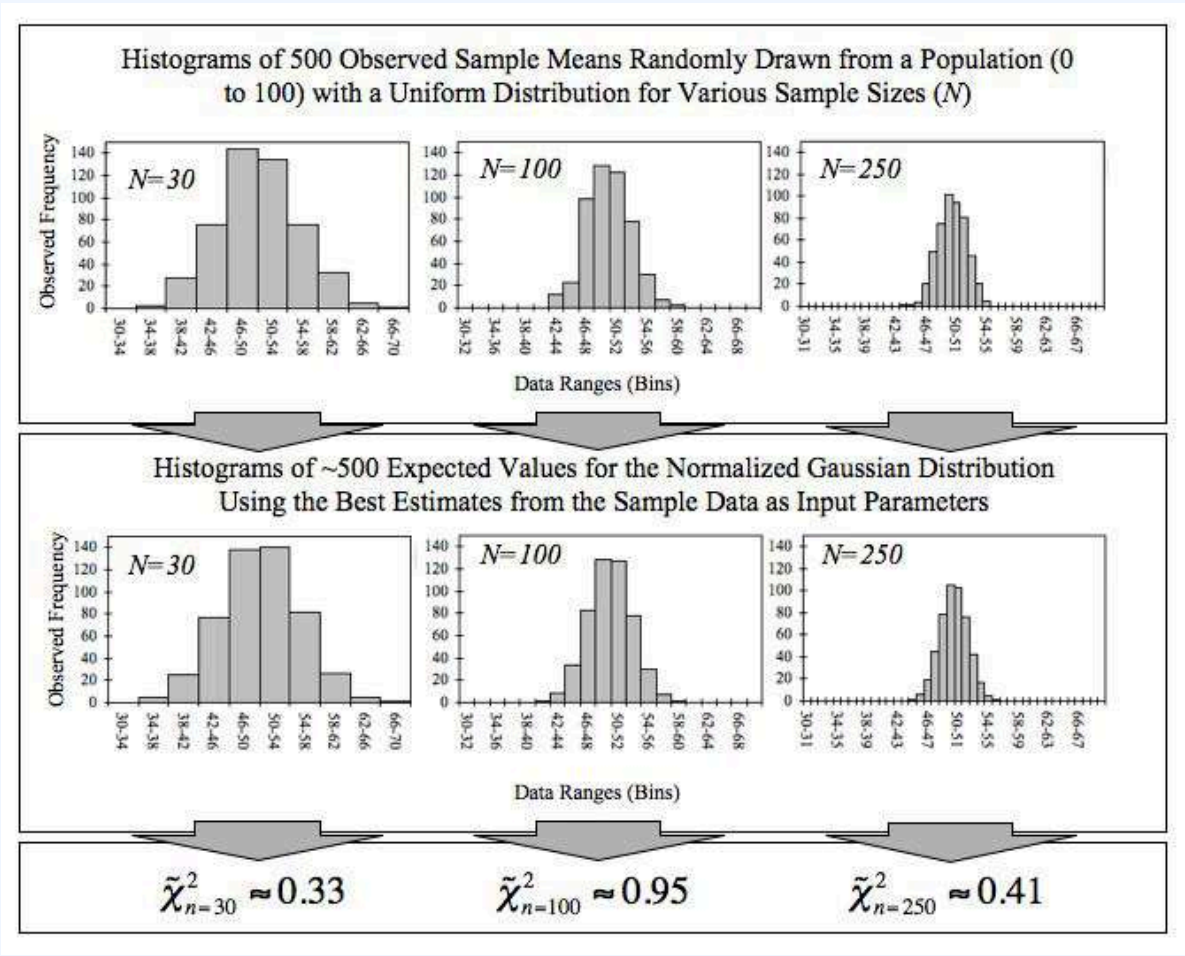
Classical Central Limit Theorem

Consider a sequence of independent and identically distributed random variables drawn from distributions of expected values given by μ″>μ and finite variances given by σ2″>σ2. Suppose we are interested in the sample average of these random variables. By the law of large numbers, the sample averages converge in probability and almost surely to the expected value μ″>μ as n→∞″>n→∞. The classical central limit theorem describes the size and the distributional form of the stochastic fluctuations around the deterministic number μ″>μ during this convergence. More precisely, it states that as n″>n gets larger, the distribution of the difference between the sample average Sn″>Sn and its limit μ″>μ approximates the normal distribution with mean 0 and variance σ2″>σ2. For large enough n″>n, the distribution of Sn″>Sn is close to the normal distribution with mean μ″>μ and variance

σ2n″>$\frac{\sigma^2}{n}$

The upshot is that the sampling distribution of the mean approaches a normal distribution as n″>n, the

sample size, increases. The usefulness of the theorem is that the sampling distribution approaches normality regardless of the shape of the population distribution.



Empirical Central Limit Theorem

This figure demonstrates the central limit theorem. The sample means are generated using a random number generator, which draws numbers between 1 and 100 from a uniform probability distribution. It illustrates that increasing sample sizes result in the 500 measured sample means being more closely distributed about the population mean (50 in this case).

Attributions

- What Is a Sampling Distribution?

- "Boundless."
  http://www.boundless.com/.
  Boundless Learning
  CC BY-SA 3.0.
- "David Lane, Introduction to Sampling Distributions. September 17, 2013."
  http://cnx.org/content/m11130/latest/.
  OpenStax CNX
  CC BY 3.0.
- "Sampling distribution."
  http://en.wikipedia.org/wiki/Sampling_distribution.
  Wikipedia
  CC BY-SA 3.0.
- "inferential statistics."
  http://en.wiktionary.org/wiki/inferential_statistics.
  Wiktionary
  CC BY-SA 3.0.
- "sampling distribution."
  http://en.wikipedia.org/wiki/sampling%20distribution.
  Wikipedia
  CC BY-SA 3.0.
- "All sizes | Houston Skyline | Flickr – Photo Sharing!."
  http://www.flickr.com/photos/mrbill/4803461/sizes/o/in/photostream/.
  Flickr
  CC BY.

- Properties of Sampling Distributions
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "inferential statistics."
    http://en.wiktionary.org/wiki/inferential_statistics.
    Wiktionary
    CC BY-SA 3.0.
  - "Standard error."
    http://en.wikipedia.org/wiki/Standard_error.
    Wikipedia
    CC BY-SA 3.0.

- "sampling distribution."
  http://en.wikipedia.org/wiki/sampling%20distribution.
  Wikipedia
  CC BY-SA 3.0.
- "Chapter 9-Sampling Distributions."
  http://mrschasesstatspage.wikispaces.com/Chapter+9-Sampling+Distributions.
  mrschasesstatspage Wikispace
  CC BY-SA 3.0.
- "David Lane, Introduction to Sampling Distributions. September 17, 2013."
  http://cnx.org/content/m11130/latest/.
  OpenStax CNX
  CC BY 3.0.

- Creating a Sampling Distribution
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "sampling distribution."
    http://en.wikipedia.org/wiki/sampling%20distribution.
    Wikipedia
    CC BY-SA 3.0.
  - "David Lane, Introduction to Sampling Distributions. September 17, 2013."
    http://cnx.org/content/m11130/latest/.
    OpenStax CNX
    CC BY 3.0.
  - "David Lane, Introduction to Sampling Distributions. May 8, 2013."
    http://cnx.org/content/m11130/latest/.
    OpenStax CNX
    CC BY 3.0.
  - "David Lane, Introduction to Sampling Distributions. May 8, 2013."
    http://cnx.org/content/m11130/latest/.
    OpenStax CNX
    CC BY 3.0.
  - "David Lane, Introduction to Sampling Distributions. May 8, 2013."
    http://cnx.org/content/m11130/latest/.
    OpenStax CNX
    CC BY 3.0.

- Continuous Sampling Distributions
  - “Boundless.”
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - “Probability density function.”
    http://en.wikipedia.org/wiki/Probability_density_function.
    Wikipedia
    CC BY-SA 3.0.
  - “probability density function.”
    http://en.wiktionary.org/wiki/probability_density_function.
    Wiktionary
    CC BY-SA 3.0.
  - “David Lane, Introduction to Sampling Distributions. September 17, 2013.”
    http://cnx.org/content/m11130/latest/.
    OpenStax CNX
    CC BY 3.0.
  - “Boxplot vs PDF.”
    http://commons.wikimedia.org/wiki/File:Boxplot_vs_PDF.svg.
    Wikimedia
    CC BY-SA.
- Mean of All Sample Means ($\mu_{\bar{x}}$)
  - “Boundless.”
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - “sampling distribution.”
    http://en.wikipedia.org/wiki/sampling%20distribution.
    Wikipedia
    CC BY-SA 3.0.
  - “David Lane, Sampling Distribution of Difference Between Means September 17, 2013.”
    http://cnx.org/content/m11132/latest/.
    OpenStax CNX
    CC BY 3.0.
  - “David Lane, Sampling Distribution of Difference Between Means May 10, 2013.”
    http://cnx.org/content/m11132/latest/.
    OpenStax CNX

CC BY 3.0.

- Shapes of Sampling Distributions
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Sampling distribution."
    http://en.wikipedia.org/wiki/Sampling_distribution.
    Wikipedia
    CC BY-SA 3.0.
  - "Shape of the distribution."
    http://en.wikipedia.org/wiki/Shape_of_the_distribution.
    Wikipedia
    CC BY-SA 3.0.
  - "normal distribution."
    http://en.wiktionary.org/wiki/normal_distribution.
    Wiktionary
    CC BY-SA 3.0.
  - "probability distribution."
    http://en.wiktionary.org/wiki/probability_distribution.
    Wiktionary
    CC BY-SA 3.0.
  - "skewed."
    http://en.wiktionary.org/wiki/skewed.
    Wiktionary
    CC BY-SA 3.0.
  - "Standard deviation diagram."
    http://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg.
    Wikimedia
    CC BY.
- Sampling Distributions and the Central Limit Theorem
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Sampling distribution."
    http://en.wikipedia.org/wiki/Sampling_distribution.

Wikipedia
CC BY-SA 3.0.

- "Central limit theorem."
  http://en.wikipedia.org/wiki/Central_limit_theorem.
  Wikipedia
  CC BY-SA 3.0.
- "central limit theorem."
  http://en.wiktionary.org/wiki/central_limit_theorem.
  Wiktionary
  CC BY-SA 3.0.
- "sampling distribution."
  http://en.wikipedia.org/wiki/sampling%20distribution.
  Wikipedia
  CC BY-SA 3.0.
- "Susan Dean and Barbara Illowsky, Central Limit Theorem: Central Limit Theorem for Sample Means. September 17, 2013."
  http://cnx.org/content/m16947/latest/.
  OpenStax CNX
  CC BY 3.0.
- "Empirical CLT – Figure – 040711."
  http://commons.wikimedia.org/wiki/File:Empirical_CLT_-_Figure_-_040711.jpg.
  Wikimedia
  CC BY-SA.

7.4 ERRORS IN SAMPLING

7.4: Errors in Sampling

7.4.1: Expected Value and Standard Error

Expected value and standard error can provide useful information about the data recorded in an experiment.

Learning Objective

Solve for the standard error of a sum and the expected value of a random variable

Key Takeaways

Key Points

- The expected value (or expectation, mathematical expectation, EV, mean, or first moment) of a random variable is the weighted average of all possible values that this random variable can take on.
- The expected value may be intuitively understood by the law of large numbers: the expected value, when it exists, is almost surely the limit of the sample mean as sample size grows to infinity.
- The standard error is the standard deviation of the sampling distribution of a statistic.

- The standard error of the sum represents how much one can expect the actual value of a repeated experiment to vary from the expected value of that experiment.

Key Terms

standard deviation
: a measure of how spread out data values are around the mean, defined as the square root of the variance

continuous random variable
: obtained from data that can take infinitely many values

discrete random variable
: obtained by counting values for which there are no in-between values, such as the integers 0, 1, 2, ...

Expected Value

In probability theory, the expected value (or expectation, mathematical expectation, EV, mean, or first moment) of a random variable is the weighted average of all possible values that this random variable can take on. The weights used in computing this average are probabilities in the case of a discrete random variable, or values of a probability density function in the case of a continuous random variable.

The expected value may be intuitively understood by the law of large numbers: the expected value, when it exists, is almost surely the limit of the sample mean as sample size grows to infinity. More informally, it can be interpreted as the long-run average of the results of many independent repetitions of an experiment (e.g. a dice roll). The value may not be expected in the ordinary sense—the "expected value" itself may be unlikely or even impossible (such as having 2.5 children), as is also the case with the sample mean.

The expected value of a random variable can be calculated by summing together all the possible values with their weights (probabilities):

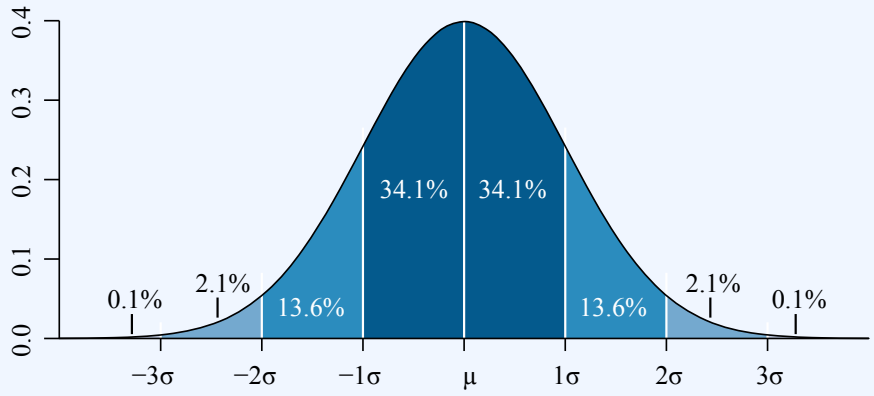
E[X]=x1p1+x2p2+…+xkpk">$E[X] = X_1P_1 + X_2P_2 + ... + X_kP_k$

where x">x represents a possible value and p">p represents the probability of that possible value.

Standard Error

The standard error is the standard deviation of the sampling distribution of a statistic. For example, the sample mean s the usual estimator of a population mean. However, different samples drawn from that same

population would in general have different values of the sample mean. The standard error of the mean (i.e., of using the sample mean as a method of estimating the population mean) is the standard deviation of those sample means over all possible samples of a given size drawn from the population.



Standard Deviation

This is a normal distribution curve that illustrates standard deviations. The likelihood of being further away from the mean diminishes quickly on both ends.

Expected Value and Standard Error of a Sum

Suppose there are five numbers in a box: 1, 1, 2, 3, and 4. If we were to selected one number from the box, the expected value would be:

E[X]=1⋅15+1⋅15+2⋅15+3⋅15+4⋅15=2.2">
$E[X] = 1 \cdot \frac{1}{5} + 1 \cdot \frac{1}{5} + 2 \cdot \frac{1}{5} + 3 \cdot \frac{1}{5} + 4 \cdot \frac{1}{5} = 2.2$

Now, let's say we draw a number from the box 25 times (with replacement). The new expected value of the sum of the numbers can be calculated by the number of draws multiplied by the expected value of the box: 25⋅2.2=55">25·2.2=55. The standard error of the sum can be calculated by the square root of number of draws multiplied by the standard deviation of the box: 25⋅SD of box=5⋅1.17=5.8">$\sqrt{25} \cdot SD$of box=5·1.17=5.8. This means that if this experiment were to be repeated many times, we could expect the sum of 25 numbers chosen to be within 5.8 of the expected value of 55, either higher or lower.

7.5.2: Using the Normal Curve

The normal curve is used to find the probability that a value falls within a certain standard deviation away from the mean.

Learning Objective

Calculate the probability that a variable is within a certain range by finding its z-value and using the Normal curve

Key Takeaways

Key Points

- In order to use the normal curve to find probabilities, the observed value must first be standardized using the following formula: z=x−μσ">$z = \frac{x-\mu}{\sigma}$.
- To calculate the probability that a variable is within a range, we have to find the area under the curve. Luckily, we have tables to make this process fairly easy.
- When reading the table, we must note that the leftmost column tells you how many sigmas above the the mean the value is to one decimal place (the tenths place), the top row gives the second decimal place (the hundredths), and the intersection of a row and column gives the probability.
- It is important to remember that the table only gives the probabilities to the left of the z-value and that the normal curve is symmetrical.
- In a normal distribution, approximately 68% of values fall within one standard deviation of the mean, approximately 95% of values fall with two standard deviations of the mean, and

approximately 99.7% of values fall within three standard of the mean.

Key Terms

standard deviation
: a measure of how spread out data values are around the mean, defined as the square root of the variance

z-value
: the standardized value of an observation found by subtracting the mean from the observed value, and then dividing that value by the standard deviation; also called z-score

Z-Value

The functional form for a normal distribution is a bit complicated. It can also be difficult to compare two variables if their mean and or standard deviations are different. For example, heights in centimeters and weights in kilograms, even if both variables can be described by a normal distribution. To get around both of these conflicts, we can define a new variable:

z=x−μσ">$z = \frac{x-\mu}{\sigma}$

This variable gives a measure of how far the variable is from the mean (x−μ">x−μ), then "normalizes" it by dividing by the standard deviation (σ">σ). This new variable gives us a way of comparing different variables. The z">z-value tells us how many standard deviations, or "how many sigmas", the variable is from its respective mean.

Areas Under the Curve

To calculate the probability that a variable is within a range, we have to find the area under the curve. Normally, this would mean we'd need to use calculus. However, statisticians have figured out an easier method, using tables, that can typically be found in your textbook or even on your calculator.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.5636	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999
3.1	0.99903	0.99906	0.9991	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.9994	0.99942	0.99944	0.99946	0.99948	0.9995
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.9996	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.9997	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.9998	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.9999	0.9999	0.9999	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998	0.99998	0.99998	0.99998

Standard Normal Table

This table can be used to find the cumulative probability up to the standardized normal value z. You can use common search engines to find Z-score tables as needed.

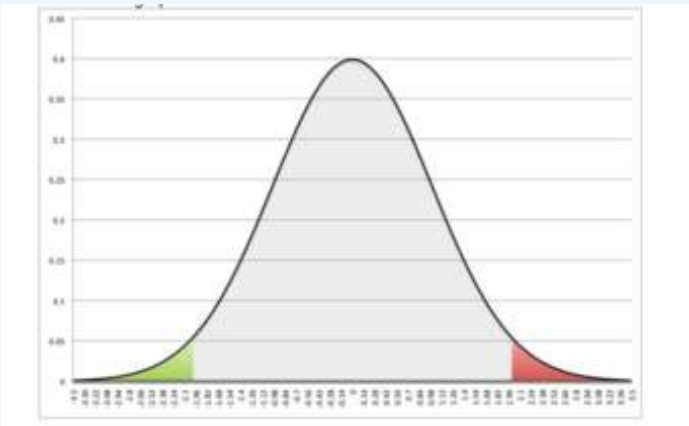
These tables can be a bit intimidating, but you simply need to know how to read them. The leftmost column tells you how many sigmas above the the mean to one decimal place (the tenths place).The top row gives the second decimal place (the hundredths).The intersection of a row and column gives the probability.

For example, if we want to know the probability that a variable is no more than 0.51 sigmas above the mean, P(z<0.51)">P(z<0.51), we look at the 6^{th} row down (corresponding to 0.5) and the 2^{nd} column (corresponding to 0.01). The intersection of the 6^{th} row and 2^{nd} column is 0.6950, which tells us that there is a 69.50% percent chance that a variable is less than 0.51 sigmas (or standard deviations) above the mean.

A common mistake is to look up a z">z-value in the table and simply report the corresponding entry, regardless of whether the problem asks for the area to the left or to the right of the z">z-value. The table only gives the probabilities to the *left* of the z">z-value. Since the total area under the curve is 1, all we need to do is subtract the value found in the table from 1. For example, if we wanted to find out the probability that a variable is more than 0.51 sigmas above the mean, P(z>0.51)">P(z>0.51), we just need to calculate 1−P(z<0.51)=1−0.6950=0.3050">1−P(z<0.51)=1−0.6950=0.3050, or 30.5%.

There is another note of caution to take into consideration when using the table: The table provided only

gives values for positive z">z-values, which correspond to values above the mean. What if we wished instead to find out the probability that a value falls *below* a z">z-value of −0.51">−0.51, or 0.51 standard deviations below the mean? We must remember that the standard normal curve is symmetrical, meaning that P(z<−0.51)=P(z>0.51)">P(z<−0.51)=P(z>0.51), which we calculated above to be 30.5%.



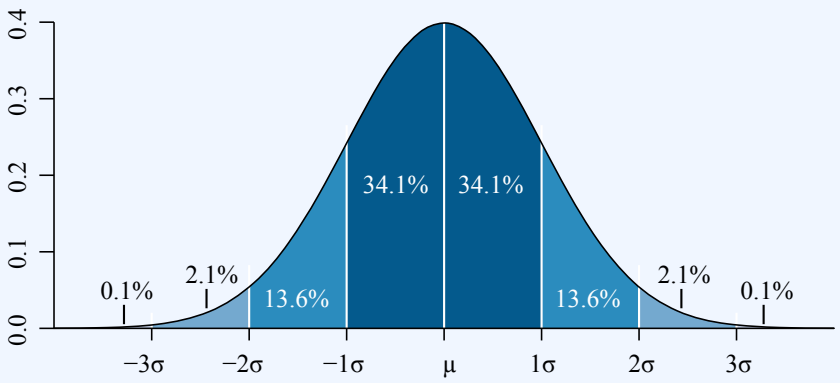
Symmetrical Normal Curve

This images shows the symmetry of the normal curve. In this case, P(z2.01).

We may even wish to find the probability that a variable is between two z-values, such as between 0.50 and 1.50, or P(0.50).

68-95-99.7 Rule

Although we can always use the z">z-score table to find probabilities, the 68-95-99.7 rule helps for quick calculations. In a normal distribution, approximately 68% of values fall within one standard deviation of the mean, approximately 95% of values fall with two standard deviations of the mean, and approximately 99.7% of values fall within three standard deviations of the mean.



68-95-99.7 Rule

Dark blue is less than one standard deviation away from the mean. For the normal distribution, this accounts for about 68% of the set, while two standard deviations from the mean (medium and dark blue) account for about 95%, and three standard deviations (light, medium, and dark blue) account for about 99.7%.

7.4.3: The Correction Factor

The expected value is a weighted average of all possible values in a data set.

Learning Objective

Recognize when the correction factor should be utilized when sampling

Key Takeaways

Key Points

- The expected value refers, intuitively, to the value of a random variable one would "expect" to find if one could repeat the random variable process an infinite number of times and take the average of the values obtained.
- The intuitive explanation of the expected value above is a consequence of the law of large numbers: the expected value, when it exists, is almost surely the limit of the sample mean as the sample size grows to infinity.
- From a rigorous theoretical standpoint, the expected value of a continuous variable is the integral of the random variable with respect to its probability measure.
- A positive value for r indicates a positive association between the variables, and a negative value indicates a negative association.
- Correlation does not necessarily imply causation.

Key Terms

integral
: the limit of the sums computed in a process in which the domain of a function is divided into small subsets and a possibly nominal value of the function on each subset is multiplied by the measure of that subset, all these products then being summed

random variable
: a quantity whose value is random and to which a probability distribution is assigned, such as the possible outcome of a roll of a die

weighted average
: an arithmetic mean of values biased according to agreed weightings

In probability theory, the expected value refers, intuitively, to the value of a random variable one would "expect" to find if one could repeat the random variable process an infinite number of times and take the average of the values obtained. More formally, the expected value is a weighted average of all possible values. In

other words, each possible value the random variable can assume is multiplied by its assigned weight, and the resulting products are then added together to find the expected value.

The weights used in computing this average are the probabilities in the case of a discrete random variable (that is, a random variable that can only take on a finite number of values, such as a roll of a pair of dice), or the values of a probability density function in the case of a continuous random variable (that is, a random variable that can assume a theoretically infinite number of values, such as the height of a person).

From a rigorous theoretical standpoint, the expected value of a continuous variable is the integral of the random variable with respect to its probability measure. Since probability can never be negative (although it can be zero), one can intuitively understand this as the area under the curve of the graph of the values of a random variable multiplied by the probability of that value. Thus, for a continuous random variable the expected value is the limit of the weighted sum, i.e. the integral.

Simple Example

Suppose we have a random variable X, which represents the number of girls in a family of three children. Without too much effort, you can compute the following probabilities:

P[X=0]=0.125P[X=1]=0.375P[X=2]=0.375P[X=3]=0.125">P[X=0]=0.125

P[X=0]=0.125P[X=1]=0.375P[X=2]=0.375P[X=3]=0.125">P[X=1]=0.375

P[X=0]=0.125P[X=1]=0.375P[X=2]=0.375P[X=3]=0.125">P[X=2]=0.375

P[X=0]=0.125P[X=1]=0.375P[X=2]=0.375P[X=3]=0.125">P[X=3]=0.125

The expected value of X, $E[X]$, is computed as:

E[X]=∑x=03xP[X=x]">$E[X] = \sum_{x=0}^{3} xP[X = x]$

=0⋅0.125+1⋅0.375+2⋅0.375+3⋅0.125">=0⋅0.125+1⋅0.375+2⋅0.375 +3⋅0.125

=1.5">=1.5

This calculation can be easily generalized to more complicated situations. Suppose that a rich uncle plans to give you $2,000 for each child in your family, with a bonus of $500 for each girl. The formula for the bonus is:

Y=1,000+500X">Y=1,000+500X

What is your expected bonus?

E[1000+500X]=∑x=03(1000+500x)P[X=x]">

$E[1000 + 500X] = \sum_{x=0}^{3}(1000 + 500x)P[X = x]$

=1000⋅0.125+1500⋅0.375+2000⋅0.375+2500⋅0.125">=1000⋅0.125 +1500⋅0.375+2000⋅0.375+2500⋅0.125

=1750">=1750

We could have calculated the same value by taking the expected number of children and plugging it into the equation:

E[1,000+500X]=1,000+500E[X]">E[1,000+500X]=1,000+500E[X]

Expected Value and the Law of Large Numbers

The intuitive explanation of the expected value above is a consequence of the law of large numbers: the expected value, when it exists, is almost surely the limit of the sample mean as the sample size grows to infinity. More informally, it can be interpreted as the long-run average of the results of many independent repetitions of an experiment (e.g. a dice roll). The value may not be expected in the ordinary sense—the "expected value" itself may be unlikely or even impossible (such as having 2.5 children), as is also the case with the sample mean.

Uses and Applications

To empirically estimate the expected value of a random variable, one repeatedly measures observations of the variable and computes the arithmetic mean of the results. If the expected value exists, this procedure estimates the true expected value in an unbiased manner and has the property of minimizing the sum of the squares of the residuals (the sum of the squared differences between the observations and the estimate). The law of large numbers demonstrates (under fairly mild conditions) that, as the size of the sample gets larger, the variance of this estimate gets smaller.

This property is often exploited in a wide variety of applications, including general problems of statistical estimation and machine learning, to estimate (probabilistic) quantities of interest via Monte Carlo methods.

The expected value plays important roles in a variety of contexts. In regression analysis, one desires a formula in terms of observed data that will give a "good" estimate of the parameter giving the effect of some explanatory variable upon a dependent variable. The formula will give different estimates using different samples of data, so the estimate it gives is itself a random variable. A formula is typically considered good in this context if it is an unbiased estimator—that is, if the expected value of the estimate (the average value it would give over an arbitrarily large number of separate samples) can be shown to equal the true value of the desired parameter.

In decision theory, and in particular in choice under uncertainty, an agent is described as making an optimal choice in the context of incomplete information. For risk neutral agents, the choice involves using the expected values of uncertain quantities, while for risk averse agents it involves maximizing the expected value of some objective function such as a von Neumann-Morgenstern utility function.

7.4.4: A Closer Look at the Gallup Poll

The Gallup Poll is an opinion poll that uses probability samples to try to accurately represent the attitudes and beliefs of a population.

Learning Objective

Examine the errors that can still arise in the probability samples chosen by Gallup

Key Takeaways

Key Points

- The Gallup Poll has transitioned over the years from polling people in their residences to using phone calls. Today, both landlines and cell phones are called, and are selected randomly using a technique called random digit dialing.
- Opinion polls like Gallup face problems such as nonresponse bias, response bias, undercoverage, and poor wording of questions.
- Contrary to popular belief, sample sizes as small as 1,000 can accurately represent the views of the general population within 4 percentage points, if chosen properly.
- To make sure that the sample is representative of the whole population, each respondent is assigned a weight so that demographic characteristics of the weighted sample match those of the entire population. Gallup weighs for gender, race, age, education, and region.

Key Terms

probability sample
: a sample in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined

nonresponse
: the absence of a response

undercoverage
: Occurs when a survey fails to reach a certain portion of the population.

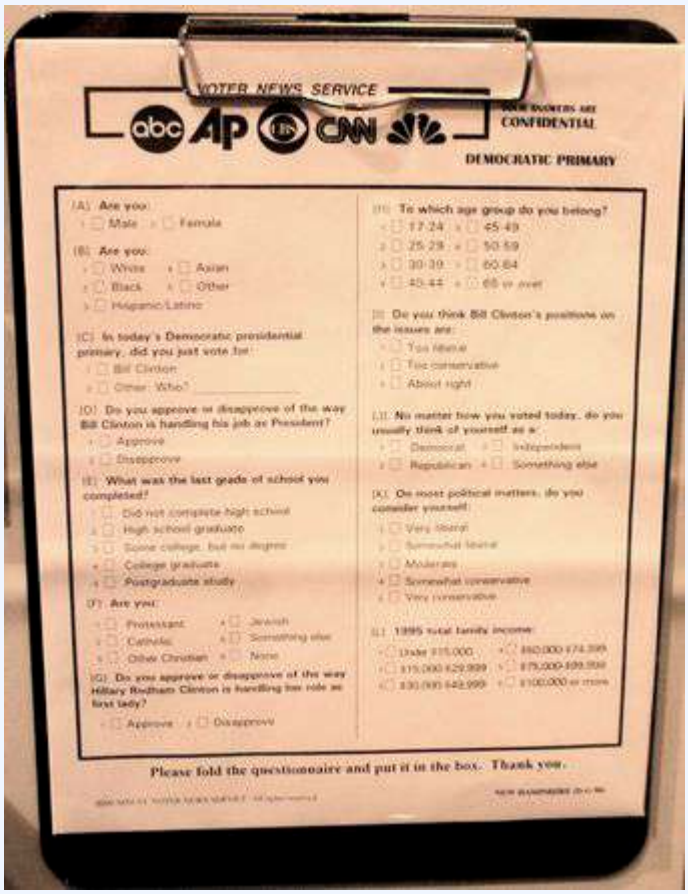
Overview of the Gallup Poll

The Gallup Poll is the division of Gallup, Inc. that regularly conducts public opinion polls in more than 140 countries around the world. Historically, the Gallup Poll has measured and tracked the public's attitudes concerning virtually every political, social, and economic issue of the day, including highly sensitive or controversial subjects. It is very well known when it comes to presidential election polls and is often referenced in the mass media as a reliable and objective audience measurement of public opinion. Its results, analyses, and videos are published daily on Gallup.com in the form of data-driven news. The poll has been around since 1935.

How Does Gallup Choose its Samples?

The Gallup Poll is an opinion poll that uses probability sampling. In a probability sample, each individual has an equal opportunity of being selected. This helps generate a sample that can represent the attitudes, opinions, and behaviors of the entire population.

In the United States, from 1935 to the mid-1980s, Gallup typically selected its sample by selecting residences from all geographic locations. Interviewers would go to the selected houses and ask whatever questions were included in that poll, such as who the interviewee was planning to vote for in an upcoming election .



Voter Polling Questionnaire

This questionnaire asks voters about their gender, income, religion, age, and political beliefs.

There were a number of problems associated with this method. First of all, it was expensive and inefficient. Over time, Gallup realized that it needed to come up with a more effective way to collect data rapidly. In addition, there was the problem of non-response. Certain people did not wish to answer the door to a stranger, or simply declined to answer the questions the interviewer asked.

In 1986, Gallup shifted most of its polling to the telephone. This provided a much quicker way to poll many people. In addition, it was less expensive because interviewers no longer had to travel all over the nation to go to someone's house. They simply had to make phone calls. To make sure that every person had an equal opportunity of being selected, Gallup used a technique called random digit dialing. A computer would randomly generate phone numbers found from telephone exchanges for the sample. This method prevented problems such as under-coverage, which could occur if Gallup had chosen to select numbers from a phone

book (since not all numbers are listed). When a house was called, the person over eighteen with the most recent birthday would be the one to respond to the questions.

A major problem with this method arose in the mid-late 2000s, when the use of cell phones spiked. More and more people in the United States were switching to using only their cell phones over landline telephones. Now, Gallup polls people using a mix of landlines and cell phones. Some people claim that the ratio they use is incorrect, which could result in a higher percentage of error.

Sample Size and Error

A lot of people incorrectly assume that in order for a poll to be accurate, the sample size must be huge. In actuality, small sample sizes that are chosen well can accurately represent the entire population, with, of course, a margin of error. Gallup typically uses a sample size of 1,000 people for its polls. This results in a margin of error of about 4%. To make sure that the sample is representative of the whole population, each respondent is assigned a weight so that demographic characteristics of the weighted sample match those of the entire population (based on information from the US Census Bureau). Gallup weighs for gender, race, age, education, and region.

Potential for Inaccuracy

Despite all the work done to make sure a poll is accurate, there is room for error. Gallup still has to deal with the effects of nonresponse bias, because people may not answer their cell phones. Because of this selection bias, the characteristics of those who agree to be interviewed may be markedly different from those who decline. Response bias may also be a problem, which occurs when the answers given by respondents do not reflect their true beliefs. In addition, it is well established that the wording of the questions, the order in which they are asked, and the number and form of alternative answers offered can influence results of polls. Finally, there is still the problem of coverage bias. Although most people in the United States either own a home phone or a cell phone, some people do not (such as the homeless). These people can still vote, but their opinions would not be taken into account in the polls.

Attributions

- Expected Value and Standard Error
  - “Boundless.” http://www.boundless.com/. Boundless Learning CC BY-SA 3.0.
  - “Expected value.”

http://en.wikipedia.org/wiki/Expected_value.
Wikipedia
CC BY-SA 3.0.

- "Standard error."
  http://en.wikipedia.org/wiki/Standard_error.
  Wikipedia
  CC BY-SA 3.0.
- "standard deviation."
  http://en.wiktionary.org/wiki/standard_deviation.
  Wiktionary
  CC BY-SA 3.0.
- "Standard deviation diagram."
  http://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg.
  Wikimedia
  CC BY-SA.

- Using the Normal Curve
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Using Normal Distributions – IB Math Stuff."
    http://ibmathstuff.wikidot.com/usingnormaldistributions.
    Wikidot
    CC BY-SA.
  - "Chapter 7: Normal distribution – Statistics."
    http://statistics.wikidot.com/ch7.
    Wikidot
    CC BY-SA.
  - "standard deviation."
    http://en.wiktionary.org/wiki/standard_deviation.
    Wiktionary
    CC BY-SA 3.0.
  - "Standard deviation diagram."
    https://en.wikipedia.org/wiki/File:Standard_deviation_diagram.svg.
    Wikipedia
    CC BY-SA.
  - "Using Normal Distributions – IB Math Stuff."

http://ibmathstuff.wikidot.com/usingnormaldistributions.
Wikidot
CC BY-SA.

- "Chapter 7: Normal distribution – Statistics."
http://statistics.wikidot.com/ch7.
Wikidot
CC BY-SA.

- The Correction Factor
  - "Boundless."
http://www.boundless.com/.
Boundless Learning
CC BY-SA 3.0.
  - "Expected value."
http://en.wikipedia.org/wiki/Expected_value.
Wikipedia
CC BY-SA 3.0.
  - "weighted average."
http://en.wiktionary.org/wiki/weighted_average.
Wiktionary
CC BY-SA 3.0.
  - "random variable."
http://en.wiktionary.org/wiki/random_variable.
Wiktionary
CC BY-SA 3.0.
  - "integral."
http://en.wiktionary.org/wiki/integral.
Wiktionary
CC BY-SA 3.0.
  - "Stats: Expected value and moments (July 29, 2005)."
http://www.pmean.com/05/Moments.asp.
P.Mean Website
CC BY.
- A Closer Look at the Gallup Poll
  - "Boundless."
http://www.boundless.com/.
Boundless Learning
CC BY-SA 3.0.

- "Opinion poll."
  http://en.wikipedia.org/wiki/Opinion_poll.
  Wikipedia
  CC BY-SA 3.0.
- "Gallup (company)."
  http://en.wikipedia.org/wiki/Gallup_(company).
  Wikipedia
  CC BY-SA 3.0.
- "Nonprobability sampling."
  http://en.wikipedia.org/wiki/Nonprobability_sampling.
  Wikipedia
  CC BY-SA 3.0.
- "nonresponse."
  http://en.wiktionary.org/wiki/nonresponse.
  Wiktionary
  CC BY-SA 3.0.
- "Voter poll."
  http://en.wikipedia.org/wiki/File:Voter_poll.jpg.
  Wikipedia
  CC BY-SA.

7.5 SAMPLING EXAMPLES

7.5: Sampling Examples

7.5.1: Measuring Unemployment

Labor force surveys are the most preferred method of measuring unemployment due to their comprehensive results and categories such as race and gender.

Learning Objective

Analyze how the United States measures unemployment

Key Takeaways

Key Points

- As defined by the International Labour Organization (ILO), "unemployed workers" are those who are currently not working but are willing and able to work for pay, currently available to work, and have actively searched for work.
- The unemployment rate is calculated as a percentage by dividing the number of unemployed individuals by all individuals currently in the labor force.

- Though many people care about the number of unemployed individuals, economists typically focus on the unemployment rate.
- In the U.S., the Current Population Survey (CPS) conducts a survey based on a sample of 60,000 households.
- The Current Employment Statistics survey (CES) conducts a survey based on a sample of 160,000 businesses and government agencies that represent 400,000 individual employers.
- The Bureau of Labor Statistics also calculates six alternate measures of unemployment, U1 through U6, that measure different aspects of unemployment.

Key Terms

unemployment
: The level of joblessness in an economy, often measured as a percentage of the workforce.

labor force
: The collective group of people who are available for employment, whether currently employed or unemployed (though sometimes only those unemployed people who are seeking work are included).

Unemployment, for the purposes of this atom, occurs when people are without work and actively seeking work. The unemployment rate is a measure of the prevalence of unemployment. It is calculated as a percentage by dividing the number of unemployed individuals by all individuals currently in the labor force.

Though many people care about the number of unemployed individuals, economists typically focus on the unemployment rate. This corrects for the normal increase in the number of people employed due to increases in population and increases in the labor force relative to the population.

As defined by the International Labour Organization (ILO), "unemployed workers" are those who are currently not working but willing and able to work for pay, those who are currently available to work, and those who have actively searched for work. Individuals who are actively seeking job placement must make the following efforts:

- be in contact with an employer
- have job interviews
- contact job placement agencies
- send out resumes
- submit applications
- respond to advertisements (or some other means of active job searching) within the prior four weeks

There are different ways national statistical agencies measure unemployment. These differences may limit the validity of international comparisons of unemployment data. To some degree, these differences remain despite national statistical agencies increasingly adopting the definition of unemployment by the International Labor Organization. To facilitate international comparisons, some organizations, such as the OECD, Eurostat, and International Labor Comparisons Program, adjust data on unemployment for comparability across countries.

The ILO describes 4 different methods to calculate the unemployment rate:

1. Labor Force Sample Surveys are the most preferred method of unemployment rate calculation since they give the most comprehensive results and enable calculation of unemployment by different group categories such as race and gender. This method is the most internationally comparable.
2. Official Estimates are determined by a combination of information from one or more of the other three methods. The use of this method has been declining in favor of labor surveys.
3. Social Insurance Statistics, such as unemployment benefits, are computed base on the number of persons insured representing the total labor force and the number of persons who are insured that are collecting benefits. This method has been heavily criticized due to the expiration of benefits before the person finds work.
4. Employment Office Statistics are the least effective, being that they only include a monthly tally of unemployed persons who enter employment offices. This method also includes unemployed who are not unemployed per the ILO definition.

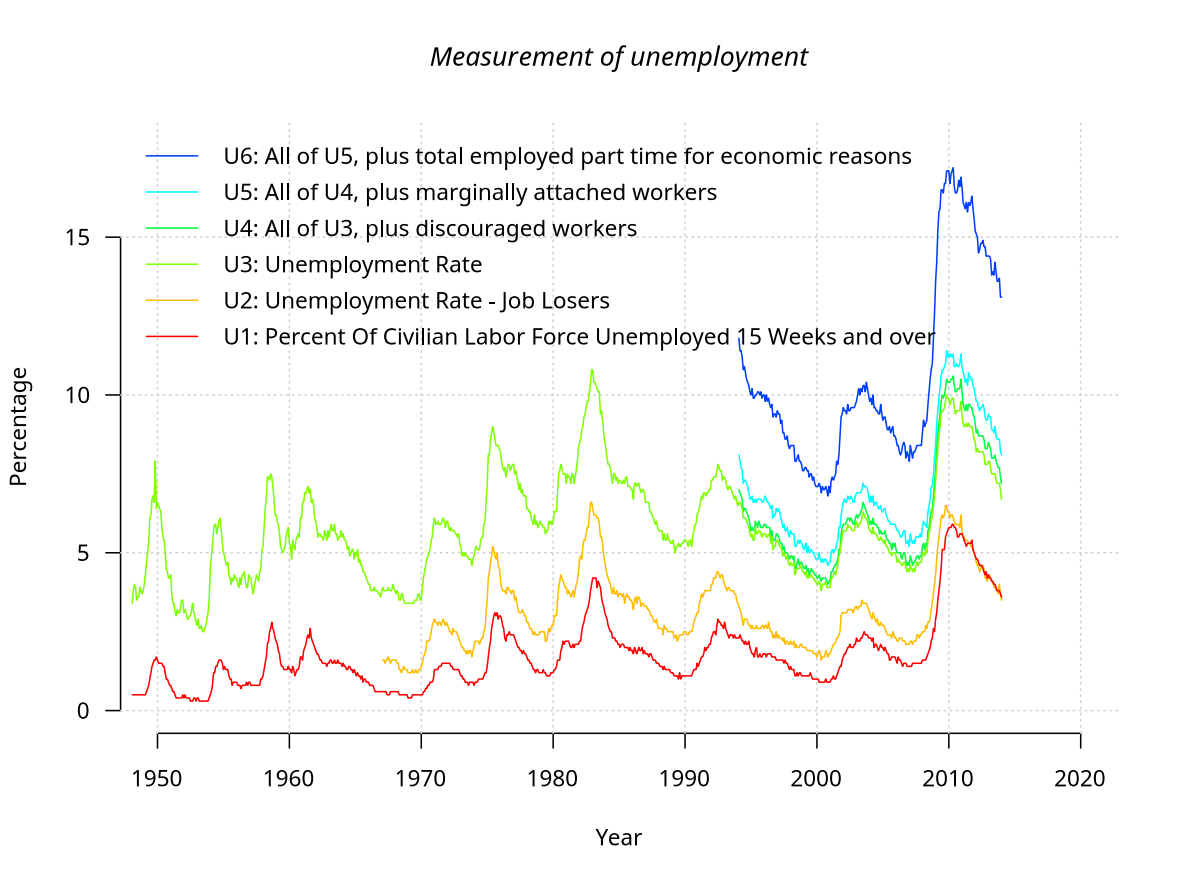
Unemployment in the United States

The Bureau of Labor Statistics measures employment and unemployment (of those over 15 years of age) using two different labor force surveys conducted by the United States Census Bureau (within the United States Department of Commerce) and/or the Bureau of Labor Statistics (within the United States Department of Labor). These surveys gather employment statistics monthly. The Current Population Survey (CPS), or "Household Survey," conducts a survey based on a sample of 60,000 households. This survey measures the unemployment rate based on the ILO definition.

The Current Employment Statistics survey (CES), or "Payroll Survey", conducts a survey based on a sample of 160,000 businesses and government agencies that represent 400,000 individual employers. This survey measures only civilian nonagricultural employment; thus, it does not calculate an unemployment rate, and it differs from the ILO unemployment rate definition.

These two sources have different classification criteria and usually produce differing results. Additional data are also available from the government, such as the unemployment insurance weekly claims report available from the Office of Workforce Security, within the U.S. Department of Labor Employment & Training Administration.

The Bureau of Labor Statistics also calculates six alternate measures of unemployment, U1 through U6 (as diagramed in the following images), that measure different aspects of unemployment:



U.S. Unemployment Measures

U1–U6 from 1950–2010, as reported by the Bureau of Labor Statistics.

- U1: Percentage of labor force unemployed 15 weeks or longer.
- U2: Percentage of labor force who lost jobs or completed temporary work.
- U3: Official unemployment rate per the ILO definition occurs when people are without jobs and they have actively looked for work within the past four weeks.
- U4: U3 + "discouraged workers", or those who have stopped looking for work because current economic conditions make them believe that no work is available for them.
- U5: U4 + other "marginally attached workers," "loosely attached workers," or those who "would like" and are able to work, but have not looked for work recently.

- U6: U5 + Part-time workers who want to work full-time, but cannot due to economic reasons (underemployment).

7.5.2: Chance Models in Genetics

Gregor Mendel's work on genetics acted as a proof that application of statistics to inheritance could be highly useful.

Learning Objective

Examine the presence of chance models in genetics

Key Takeaways

Key Points

- In breeding experiments between 1856 and 1865, Gregor Mendel first traced inheritance patterns of certain traits in pea plants and showed that they obeyed simple statistical rules.
- Mendel conceived the idea of heredity units, which he called "factors," one of which is a recessive characteristic, and the other of which is dominant.
- Mendel found that recessive traits not visible in first generation hybrid seeds reappeared in the second, but the dominant traits outnumbered the recessive by a ratio of 3:1.
- Genetical theory has developed largely due to the use of chance models featuring randomized draws, such as pairs of chromosomes.

Key Terms

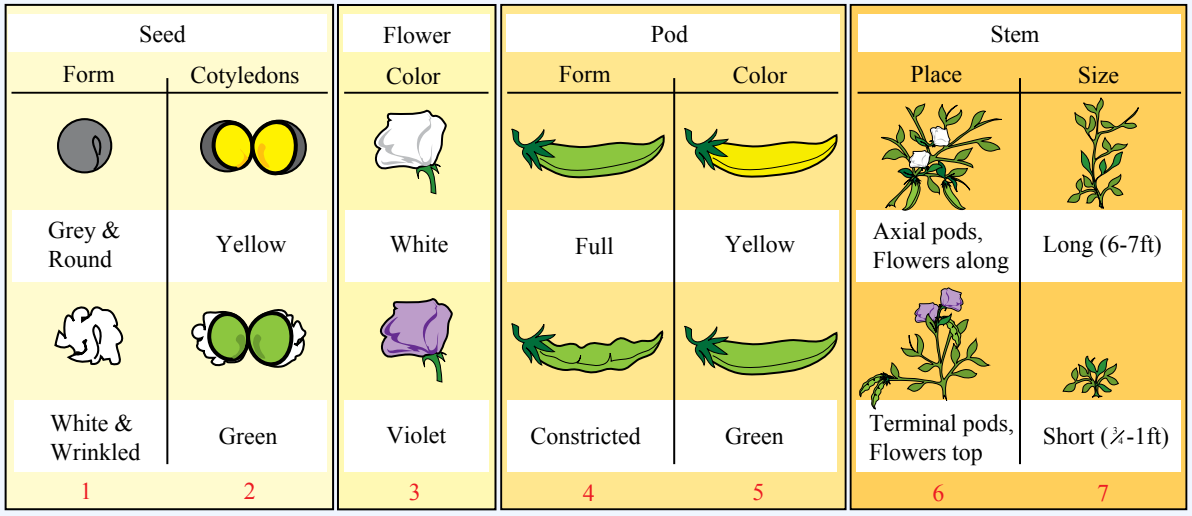
chi-squared test
: In probability theory and statistics, refers to a test in which the chi-squared distribution (also chi-square or χ-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.

gene
: a unit of heredity; a segment of DNA or RNA that is transmitted from one generation to the next, and that carries genetic information such as the sequence of amino acids for a protein

chromosome
: A structure in the cell nucleus that contains DNA, histone protein, and other structural proteins.

Gregor Mendel is known as the "father of modern genetics. " In breeding experiments between 1856 and 1865, Gregor Mendel first traced inheritance patterns of certain traits in pea plants and showed that they obeyed simple statistical rules. Although not all features show these patterns of "Mendelian Inheritance," his work served as a proof that application of statistics to inheritance could be highly useful. Since that time, many more complex forms of inheritance have been demonstrated.

In 1865, Mendel wrote the paper *Experiments on Plant Hybridization.* Mendel read his paper to the Natural History Society of Brünn on February 8 and March 8, 1865. It was published in the *Proceedings of the Natural History Society of Brünn* the following year. In his paper, Mendel compared seven discrete characters (as diagramed in):



Mendel's Seven Characters

This diagram shows the seven genetic "characters" observed by Mendel.

1. color and smoothness of the seeds (yellow and round or green and wrinkled)
2. color of the cotyledons (yellow or green)
3. color of the flowers (white or violet)
4. shape of the pods (full or constricted)
5. color of unripe pods (yellow or green)
6. position of flowers and pods on the stems
7. height of the plants (short or tall)

Mendel's work received little attention from the scientific community and was largely forgotten. It was not until the early 20th century that Mendel's work was rediscovered, and his ideas used to help form the modern synthesis.

The Experiment

Mendel discovered that when crossing purebred white flower and purple flower plants, the result is not a blend. Rather than being a mixture of the two plants, the offspring was purple-flowered. He then conceived the idea of heredity units, which he called "factors", one of which is a recessive characteristic and the other of which is dominant. Mendel said that factors, later called genes, normally occur in pairs in ordinary body cells, yet

segregate during the formation of sex cells. Each member of the pair becomes part of the separate sex cell. The dominant gene, such as the purple flower in Mendel's plants, will hide the recessive gene, the white flower.

When Mendel grew his first generation hybrid seeds into first generation hybrid plants, he proceeded to cross these hybrid plants with themselves, creating second generation hybrid seeds. He found that recessive traits not visible in the first generation reappeared in the second, but the dominant traits outnumbered the recessive by a ratio of 3:1.

After Mendel self-fertilized the F1 generation and obtained the 3:1 ratio, he correctly theorized that genes can be paired in three different ways for each trait: AA, aa, and Aa. The capital "A" represents the dominant factor and lowercase "a" represents the recessive. Mendel stated that each individual has two factors for each trait, one from each parent. The two factors may or may not contain the same information. If the two factors are identical, the individual is called homozygous for the trait. If the two factors have different information, the individual is called heterozygous. The alternative forms of a factor are called alleles. The genotype of an individual is made up of the many alleles it possesses.

An individual possesses two alleles for each trait; one allele is given by the female parent and the other by the male parent. They are passed on when an individual matures and produces gametes: egg and sperm. When gametes form, the paired alleles separate randomly so that each gamete receives a copy of one of the two alleles. The presence of an allele does not mean that the trait will be expressed in the individual that possesses it. In heterozygous individuals, the allele that is expressed is the dominant. The recessive allele is present but its expression is hidden

Relation to Statistics

The upshot is that Mendel observed the presence of chance in relation to which gene-pairs a seed would get. Because the number of pollen grains is large in comparison to the number of seeds, the selection of gene-pairs is essentially independent. Therefore, the second generation hybrid seeds are determined in a way similar to a series of draws from a data set, with replacement. Mendel's interpretation of the hereditary chain was based on this sort of statistical evidence.

In 1936, the statistician R.A. Fisher used a chi-squared test to analyze Mendel's data, and concluded that Mendel's results with the predicted ratios were far too perfect; this indicated that adjustments (intentional or unconscious) had been made to the data to make the observations fit the hypothesis. However, later authors have claimed Fisher's analysis was flawed, proposing various statistical and botanical explanations for Mendel's numbers. It is also possible that Mendel's results were "too good" merely because he reported the best subset of his data — Mendel mentioned in his paper that the data was from a subset of his experiments.

In summary, the field of genetics has become one of the most fulfilling arenas in which to apply statistical methods. Genetical theory has developed largely due to the use of chance models featuring randomized draws, such as pairs of chromosomes.

Attributions

- Measuring Unemployment
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Unemployment."
    http://en.wikipedia.org/wiki/Unemployment.
    Wikipedia
    CC BY-SA 3.0.
  - "labor force."
    http://en.wiktionary.org/wiki/labor_force.
    Wiktionary
    CC BY-SA 3.0.
  - "US Unemployment measures."
    http://commons.wikimedia.org/wiki/File:US_Unemployment_measures.svg.
    Wikimedia
    CC BY-SA.
- Chance Models in Genetics
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Experiments on Plant Hybridization."
    http://en.wikipedia.org/wiki/Experiments_on_Plant_Hybridization.
    Wikipedia
    CC BY-SA 3.0.
  - "History of genetics."
    http://en.wikipedia.org/wiki/History_of_genetics.
    Wikipedia
    CC BY-SA 3.0.
  - "chi-squared test."
    http://en.wikipedia.org/wiki/chi-squared%20test.
    Wikipedia
    CC BY-SA 3.0.
  - "Mendelian inheritance."

http://en.wikipedia.org/wiki/Mendelian_inheritance.
Wikipedia
CC BY-SA 3.0.

- "Gregor Mendel."
  http://en.wikipedia.org/wiki/Gregor_Mendel.
  Wikipedia
  CC BY-SA 3.0.
- "gene."
  http://en.wiktionary.org/wiki/gene.
  Wiktionary
  CC BY-SA 3.0.
- "chromosome."
  http://en.wiktionary.org/wiki/chromosome.
  Wiktionary
  CC BY-SA 3.0.
- "Mendel seven characters."
  http://commons.wikimedia.org/wiki/File:Mendel_seven_characters.svg.
  Wikimedia
  Public domain.

SECTION XII

7.XLSX – EXCEL CHALLENGE - TABLES

Excel is the leading application for storing, managing and analyzing data. In Chapter 5, you will explore how to import, organize, and analyze data effectively. To manage and analyze a group of related data, users can turn a range of cells into an Excel table.

A table, also called a database, is an organized structure of rows and columns of related data in a worksheet; for example, a list of employee information. In a table of employees, each employee would have a separate record; as shown below, each record might include several fields, such as the Employee ID Number, their Last Name, and First Name, etc. Each row of a table stores records, and each column stores one field for the record. A record also can include fields that contain references, formulas, and functions. Additionally, a row of column headings at the top of the table stores field names that identify the data being collected and stored.

Excel has a vast collection of database and tabling tools that allow users to import, clean, sort, filter, total, subtotal, analyze, visualize, and report. This chapter explores how to import, insert, edit, and examine data with Excel table and PivotTable tools. Demonstrate skills by studying the provided 2017-2018 employee database. Examine employee relations, payroll, benefits, and training options.

Attribution

7.XLSX.1 TABLE BASICS

Noreen Brown; Barbara Lave; Hallie Puncochar; Julie Romey; Mary Schatz; Art Schneider; Diane Shingledecker; and Jennifer Evans

Learning Objectives

1. Understand table properties and structure.
2. Format data as a table.
3. Use Freeze Panes.
4. Work with the Table Tools Design tab.

Organizing, maintaining, analyzing, and reporting human resources data is essentials across industries. In this chapter, we will import data, and demonstrate tabling skills by examining employee relations, payroll, benefits, and training options.

WCM Analytics

Employee Database

Employee ID	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary
1102	Vanesa	Allen	7/10/2012	3.48	4/11/1961	55	Portland	FT	$ 106,010
1106	Elizabeth	Allen	11/6/2015	2.15	11/23/1991	25	Seattle	FT	$ 42,182
1110	James	Anderson	12/4/2015	2.08	10/15/1966	50	Portland	FT	$ 92,254
1114	Katherine	Baker	3/24/2003	14.78	12/8/1964	52	Seattle	FT	$ 69,250
1118	Ina	Baker	5/23/2009	6.61	2/15/1962	54	Portland	FT	$ 102,567
1122	Brandon	Barnes	8/12/2002	15.40	10/15/1968	48	San Francisco	FT	$ 94,517
1126	Paul	Benham	11/6/2015	2.15	3/20/1973	43	San Diego	FT	$ 51,791
1130	Santos	Bennett	6/10/2010	7.56	4/20/1966	50	San Diego	FT	$ 32,530
1134	James	Bennett	1/20/2016	1.95	3/21/1957	59	Seattle	FT	$ 94,502
1138	June	Bennett	5/4/2012	5.66	6/28/1967	49	Portland	PT	$ 45,671
1142	Gregory	Blackshear	7/16/2009	6.47	2/8/1986	30	Seattle	FT	$ 70,346
1146	Thomas	Bradley	4/12/2008	9.73	7/13/1986	30	San Diego	FT	$ 34,685
1150	Linda	Brown	3/13/2012	5.81	7/8/1949	67	Portland	FT	$ 96,944
1154	Santina	Bryant	8/8/2015	2.40	12/1/1956	60	Portland	FT	$ 92,091
1158	Charlotte	Burgess	7/17/2015	2.46	7/12/1959	57	San Diego	FT	$ 30,150
1162	Patricia	Butler	1/8/2015	2.98	3/6/1970	46	Portland	FT	$ 81,536
1166	Ramon	Cannon	10/4/2013	4.24	10/25/1959	57	San Francisco	FT	$ 96,021
1170	Antolin	Casas	9/7/2012	5.32	3/11/1961	55	San Diego	FT	$ 58,720

Figure 5.1 Table Example

TABLE PROPERTIES & STRUCTURE

Turning a range of cells into an Excel table makes related data easier to analyze, visualize, and report.

Structuring and planning table layouts are vital for data integrity. Below are guidelines to consider when designing and building a table from scratch:

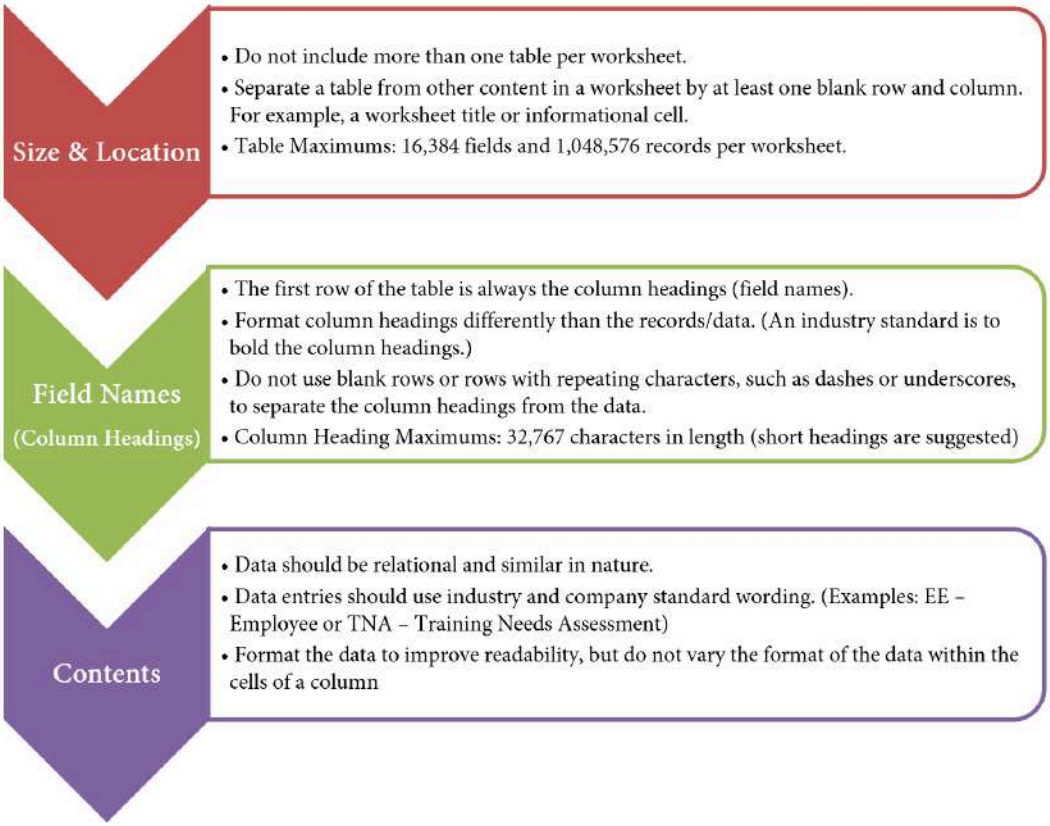


Figure 5.2 Table Layout Guidelines

OVERVIEW

Excel tables behave independently from the rest of the information on the worksheet. Excel treats the table area as a database locking the record entries together. There are several advantages of Excel treating the data independently. For example, using integrated filters and sort functions you can effortlessly drill down data

based on questions and in return get results. Excel will also automatically expand the table to accommodate new data entries and allows for automatic formatting, such as recoloring of banded rows or columns.

You will also notice Excel treats formulas and calculations differently in a table, showing structured column names, along with automatically filling a calculated field to the entire table or offering quick and easy table totaling tools.

When graphing and charting table data you will also see Excel automatically adjusts of associated charts and ranges based on what the user is sorting or filtering at the time.

In industry, data is commonly stored in databases or multiple Excel files. Databases vary drastically, therefore in some cases, it is necessary to import data types into Excel. In our example, we will work with an Excel file that has imported data from a human resources database. The data downloaded from the database is stored in an Excel workbook, however, it's in a Comma Separated Values (CSV) format. We will import the Excel file into our CH 5 Data file, turn the data into a table for further analysis.

IMPORT AND FORMAT DATA AS A TABLE

Download Data file: *CH5 Data*

CH-5-HR_

Keeping the above table guidelines in mind, import human resource data into Excel, as a table. Demonstrate tabling skills by examining employee relations, payroll, and benefits. Note you will need to save the **CH 5 HR** file on your computer as you will import this file into the **CH 5 Data** file in the below steps.

1. Open data file **CH 5 Data** and save the file as **CH5 HR Report**.
2. In the **EmployeeData** sheet, click on cell **A5**.

   Mac Users: Excel for Mac does not have the tool for "Getting Data" from an Excel Workbook. You will set up this data using alternate steps. **Please skip steps 3-11.** The alternate steps can be found below after Step 11.
3. From the Data tab, choose **Get Data**.
4. From the Get Data menu, choose **From File,** then **From Workbook**.

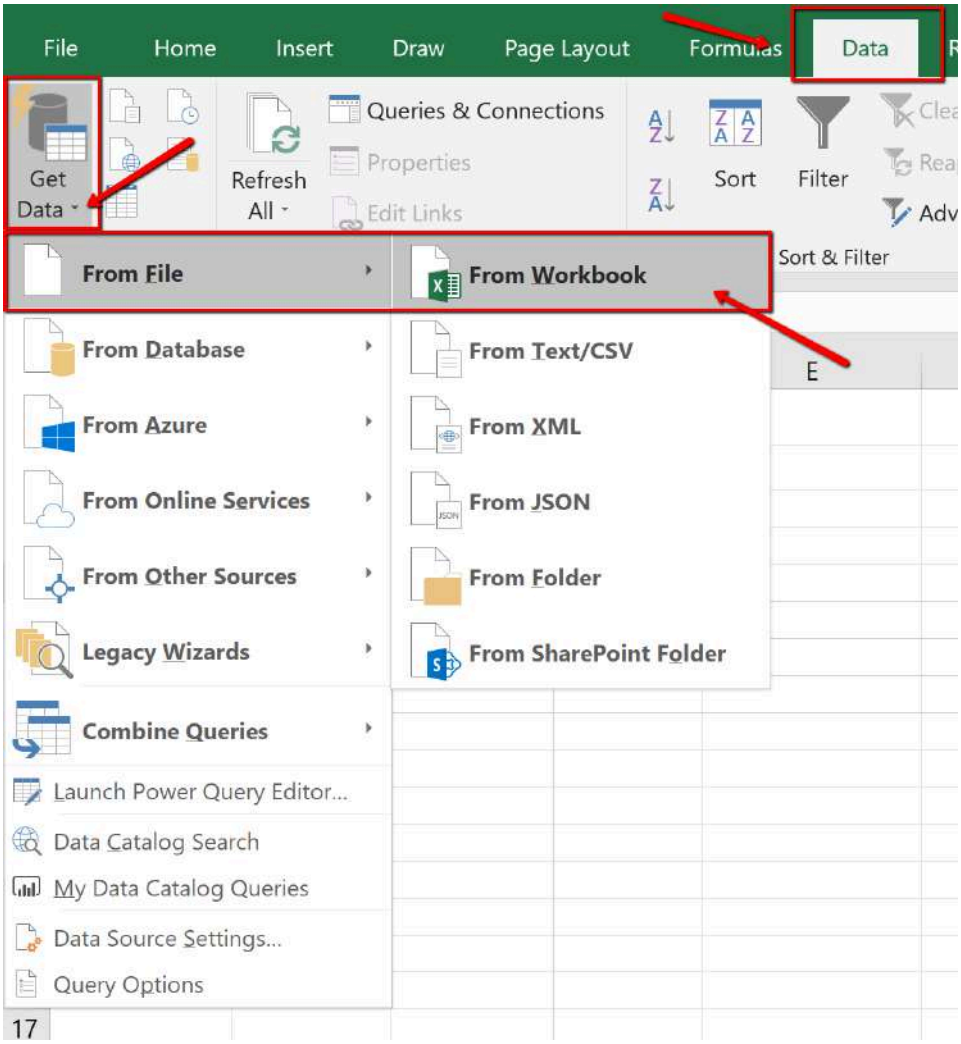


Figure 5.3 Get Data From File, From Workbook

5. Navigate to the course data files. Find, and select the **CH 5 HR** file.
6. Click **Import.**

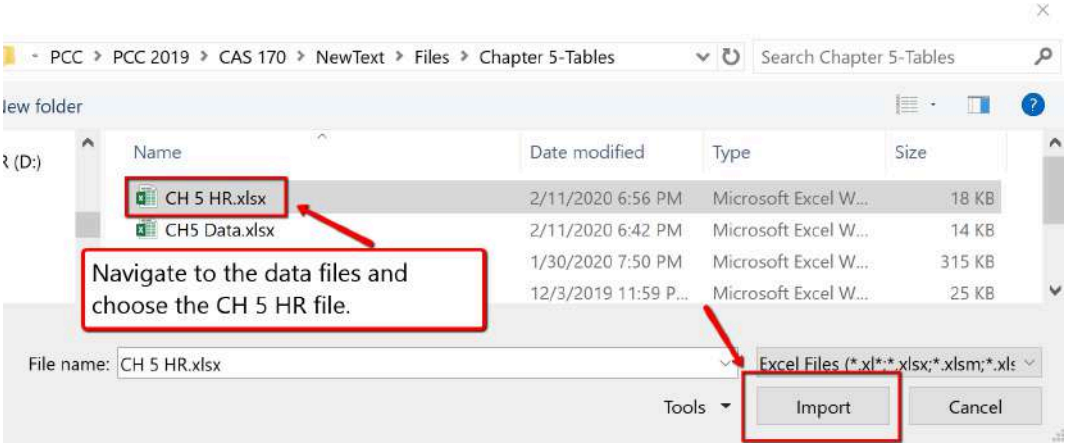


Figure 5.4 Import File Data CH 5 HR

7. The Navigator dialogue box will open. Select the **CH5 CSV File** listed in the Display Options pane.

8. At the bottom of the Navigator dialogue box, select **Load** to expand the menu and choose **Load To**...

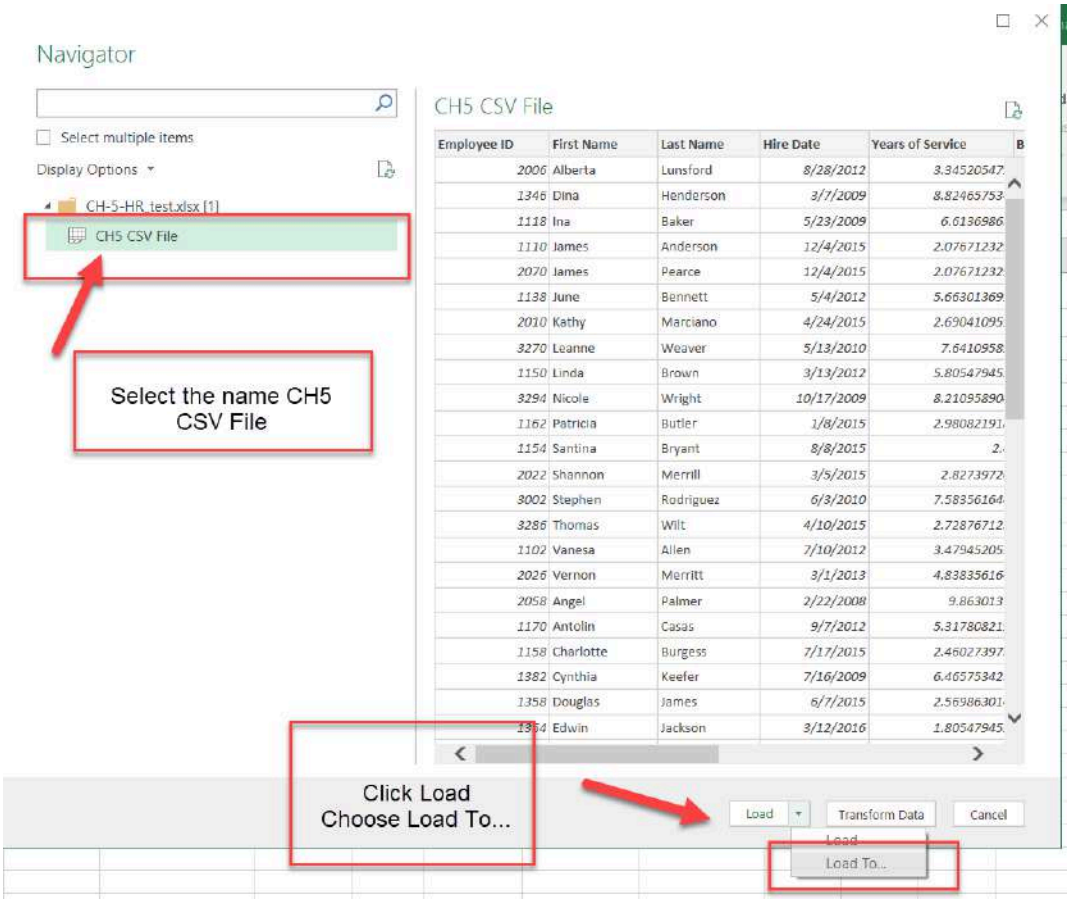


Figure 5.5 Navigator Window

9. The Import dialogue box will open. In the "**Where do you want to put the data?**" section choose **Existing worksheet:**

10. In the above steps A5 was already selected when we started the import, so Excel will indicate we want the information to import and display starting at cell =$A$5. If you did not click cell A5, then select the cell now. Click **OK.**

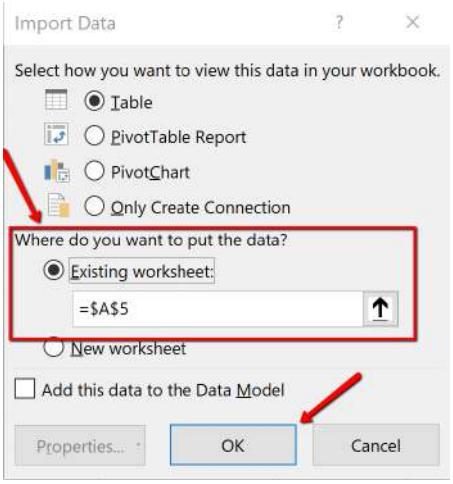


Figure 5.6 Import Data Dialogue Box

11. The data imports as a table. Close the Queries & Connections dialogue box.

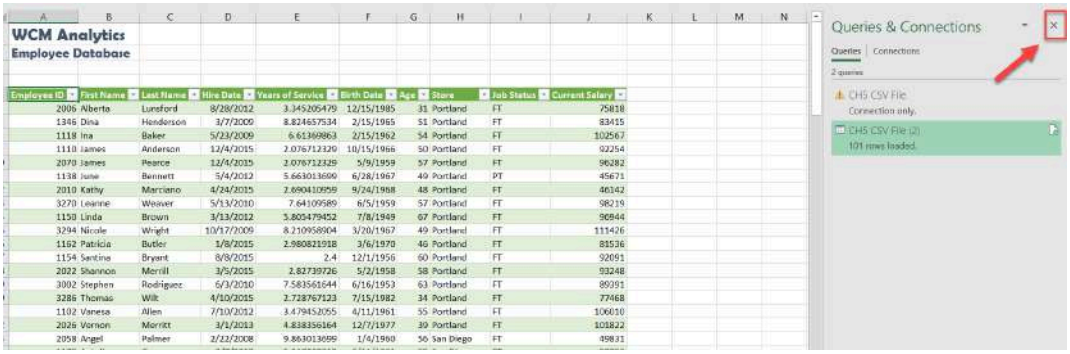
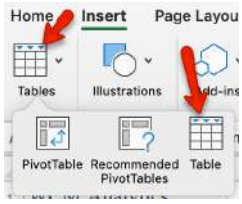


Figure 5.7 Close Queries & Connections

These are the alternate steps for Mac Users Only. If you are using Excel for Windows, please continue with the "Table Tools Design Tab" section below these alternate steps.

1. Only complete the following steps if you are using a Mac. If you are using a PC, you have already inserted the table. You should already have the **CH 5 HR Report** open and you should have clicked into cell A5 in the **EmploymentData sheet.** If you did not do this, please do it now.
2. Open the **CH 5 HR** workbook that you downloaded
3. Use the keyboard shortcut of **Ctrl key + letter A** to select all of the data in the worksheet. That should be cells A1:J102
4. Copy this data
5. Switch back to the **CH 5 Report** workbook and make sure cell A5 is the active cell
6. Paste the data into the **Employment Data sheet at cell A5**
7. Make sure Cell A5 is still the active cell and click the **Insert tab** from the Ribbon



8. Click the **Tables** button from the Ribbon and then click the **Tables** icon
9. The **Create Table** dialog box should appear with the cell range of **A5:J106 as shown here in Figure 5.8.** Click **"OK"** to accept this range for your table**.**

Figure 5.8 Excel for Mac Create Table Dialog Box

10. **Congrats!** You just converted the data to an Excel table. Continue following the steps in the section below.

TABLE TOOLS DESIGN TAB

Excel tables require specific tools. The Table Tools Design tab houses these specific tools used for formatting and editing tables. The Table Tools tab is considered a contextual tab; meaning the tabs appear when you are clicked in a table area. When you click out of a table, the Table Tools disappear.

Explore the table tools now. Notice the specific checkboxes to turn on table options, for example, you can choose to display banded rows or banded columns, or a total row etc. We will explore table tools in the following steps.

When importing data as a table, Excel automatically applied table formatting. Follow the below steps to format and edit the table.

1. Click the **Table Tools/Design** tab on the ribbon.

   Mac Users: you don't have a Table Tools/Design tab. Just make sure the Table tab is selected.
2. From the provided **Table Styles**, choose the **Blue, Table Style Medium 2** option.

   Mac Users: the table you just created may already have the "Blue, Table Style Medium 2 option.

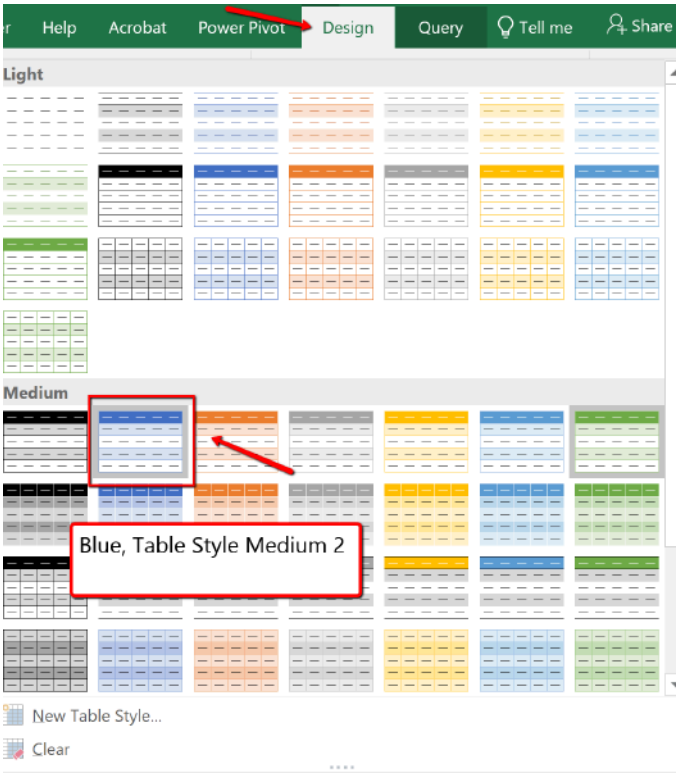


Figure 5.9 Blue Table Style Medium 2 Option

Another option for inserting a table is using the Insert button. The Insert Table button, located on the Insert tab will turn a range of information into an unformatted table. We will use the insert table option later on in the chapter.

Skill Refresher

Format Data as a Table

1. Click on the top-left cell in your data.
2. Click the Format As Table menu from the Home tab on the Ribbon. Choose a style.
3. Make sure "My table has headers" is checked. Click OK.
4. Click on the top-left cell again.
5. Adjust the widths of the columns so that you can see the complete headings with the filter arrows showing.

VIEWING table data

USING PANES

Data sets can bridge thousands of records with dozens of fields and extend beyond a workbook window. It can be difficult to compare fields and records in widely separated columns and rows. One way of dealing with this problem is by dividing the workbook window into viewing panes by using the **Split** view option. Excel can split the workbook window into four sections called panes with each pane offering a separate view into the worksheet. By scrolling through the contents of individual panes, you can compare cells from different sections of the worksheet side-by-side within the workbook window.

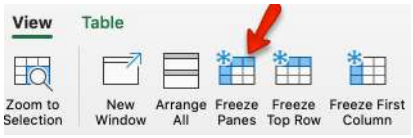
To split the workbook window into four panes, select any cell or range in the worksheet, and then on the View tab, in the Window group, click the Split button. Split bars divide the workbook window along the top and left border of the selected cell or range. To split the window into two vertical panes displayed side-by-side, select any cell in the first row of the worksheet and then click the Split button. To split the window into two stacked horizontal panes, select any cell in the first column and then click the Split button. To turn off the Spilt window option, simply click Split again on the View tab.

In our specific example the data set is manageable, however freezing the first column, and the top heading could be useful when scrolling through data.

FREEZE PANES

To keep an area of a worksheet visible while you scroll to another area of the worksheet use Freeze Panes. Follow the steps below to freeze, based on selection, the first column, and heading row.

1. If needed, adjust column widths so all heading names in row 5 are visible.
2. Click cell **B6** in the table. (By selecting this specific cell, when we apply the freeze pane option, Excel will freeze the table where the first column ends and the heading row is viewable.)
3. Click the View tab.
4. Select Freeze Panes, and for the listed options choose Freeze Panes (See **Figure 5.10 below)**. The column and rows will remain visible based on the cell that was selected above.



Mac Users should just click the **Freeze Panes** button under the View tab.

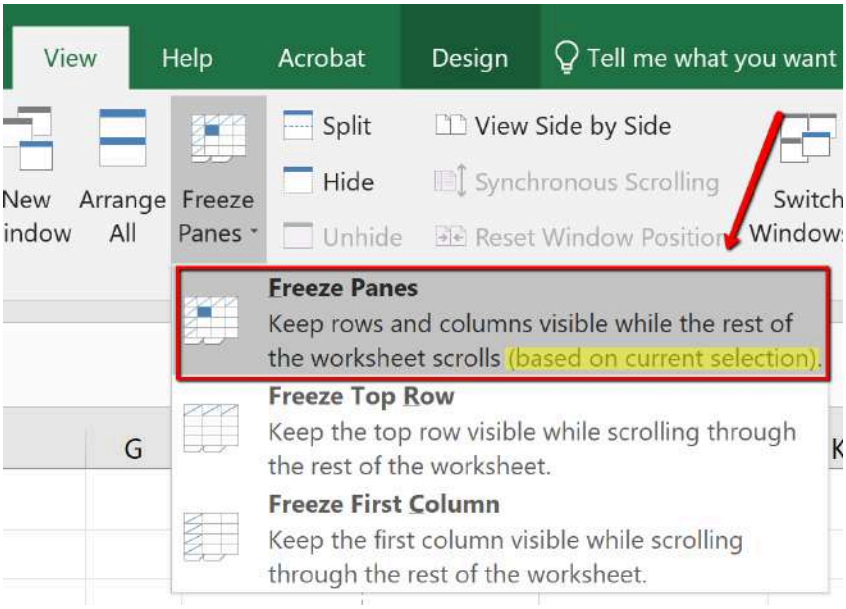


Figure 5.10 Freeze Panes

FORMATTING TABLE DATA

After reviewing the table, two columns have data that need to be formatted accordingly. In large data sets, it is useful to know data selection short cuts. In this example, we are going to use keyboard short cuts to select a column of information in the table and apply number formatting.

Format data by following the below steps:

1. In the EmployeeData sheet, click cell **E6**.
2. On the keyboard press and hold the **CTRL** and **SHIFT** and **DOWN** keys.
3. With the "Years of Service" data selected, click the Home tab. In the Numbers category, format the data as a Number. The number should automatically decrease the decimal to two decimal places.

   Mac Users: click the **"list arrow"** next to **"General,** 

   and then choose **"Number"** from the list.
4. Click in cell J6. *(Be sure you have clicked J6 so that you are in the first cell in the Current Salary column)*. Using the same selection process, select the Current Salary column, and format the data as Currency, zero decimal place.
5. Using the non-adjacent selection method, select column headings E, G, and I, and center the data.

WCM Analytics
Employee Database

Employee ID	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary
2006	Alberta	Lunsford	8/28/2012	3.35	12/15/1985	31	Portland	FT	$75,818
1346	Dina	Henderson	3/7/2009	8.82	2/15/1965	51	Portland	FT	$83,415
1118	Ina	Baker	5/23/2009	6.61	2/15/1962	54	Portland	FT	$102,567
1110	James	Anderson	12/4/2015	2.08	10/15/1966	50	Portland	FT	$92,254
2070	James	Pearce	12/4/2015	2.08	5/9/1959	57	Portland	FT	$96,282
1138	June	Bennett	5/4/2012	5.66	6/28/1967	49	Portland	PT	$45,671
2010	Kathy	Marciano	4/24/2015	2.69	9/24/1968	48	Portland	FT	$46,142
3270	Leanne	Weaver	5/13/2010	7.64	6/5/1959	57	Portland	FT	$98,219
1150	Linda	Brown	3/13/2012	5.81	7/8/1949	67	Portland	FT	$96,944
3294	Nicole	Wright	10/17/2009	8.21	3/20/1967	49	Portland	FT	$111,426
1162	Patricia	Butler	1/8/2015	2.98	3/6/1970	46	Portland	FT	$81,536
1154	Santina	Bryant	8/8/2015	2.40	12/1/1956	60	Portland	FT	$92,091
2022	Shannon	Merrill	3/5/2015	2.83	5/2/1958	58	Portland	FT	$93,248
3002	Stephen	Rodriguez	6/3/2010	7.58	6/16/1953	63	Portland	FT	$89,391
3286	Thomas	Wilt	4/10/2015	2.73	7/15/1982	34	Portland	FT	$77,468
1102	Vanesa	Allen	7/10/2012	3.48	4/11/1961	55	Portland	FT	$106,010
2026	Vernon	Merritt	3/1/2013	4.84	12/7/1977	39	Portland	FT	$101,822

Figure 5.11 Number Formatting

NAMING A TABLE

Each time a table is created, Excel assigns a default name. The default naming convention is similar to the way new workbooks are named (Book1, Book2, etc.), however in this case Excel recognizes the area as a table and will assign the name table instead of book: Table1, Table2, Table3, and so on.

Why name a table range? Referring to the table by name rather than by range will make it easier to refer to a table in the future, for example, in a workbook that contains many tables. Seeing tables named Jan or Feb is more informational then seeing Table1 or Table 2. You can custom name each table and in the future connect named tables for reporting purposes.

There are two rules to consider when naming tables. One, Excel does not allow spaces in table names, and two, Excel also requires that table names begin with a letter or underscore.

Follow the next step to assign a custom name to the table.

1. Click anywhere in the table and then display the Table Tools Design tab.

   Mac Users: there is no "Table Tools Design" tab in Excel for Mac. Simply click the **Table** tab and follow steps 2 and 3 below to give the table a new name.

2. Click the Table Name text box, in the Properties group.
3. Type **Employee_DB** and then press enter to name the table.

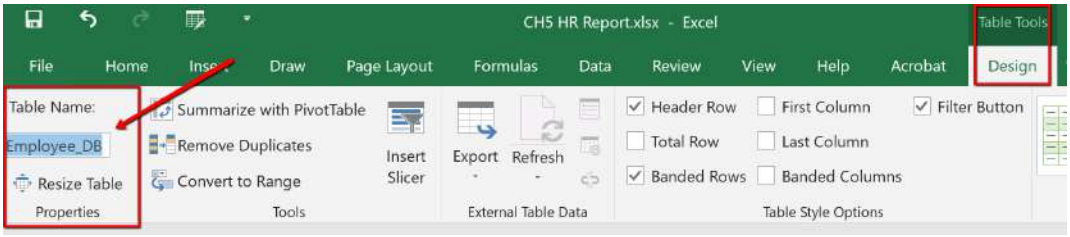


Figure 5.12 Name a Table Range

ENTERING & DELETING RECORDS

Tables require constant updating and may need calculations. When your table needs updating you can add/delete data, by adding/deleting rows, or columns. Excel adjusts the table automatically to the new content. The format applied to the banded rows updates to accommodate the new data set size.

When calculations are needed you can create a calculated column or use the built-in Total Row tool. Excel tables are a fantastic tool for entering formulas efficiently in a calculated column. Excel allows you to enter a single formula in one cell, and then that formula will automatically expand to the rest of the column by itself. There's no need to use the Fill or Copy commands. This feature can be incredibly time-saving, especially if you have a lot of rows. And the same thing happens when you change a formula; the change will also expand to the rest of the calculated column. The Total Row tool, available on the Table Tools Design tab automatically adds a total row to the bottom of the table. To add a new row, uncheck the Total Row checkbox, add the row, and then recheck the Total Row checkbox. From the total row drop-down, you can select a function, like Average, Count, Count Numbers, Max, Min, Sum, StdDev, Var, and more.

Follow the steps below to update the employee table. You will insert new information just below the table. Data entered in rows or columns adjacent to the table becomes part of the table. Excel will format the new table data automatically.

1. Press and hold the **Ctrl** and **End** button to move to the last record in the table.

   Mac Users: there is no "End" key on most Mac keyboards. Press and hold the **"Command" key a**nd tap the **right arrow key.** Then press and hold the **Command key,** again, and tap the **down arrow key.** That should move to the last record in the table.
2. Click tab to start a new record.
3. Type the new entries below. Click tab to move to the next column.

3297	Alfred	Yelnats	5/29/2015	2.59	2/19/1953	63	Seattle	FT	$95,552
3299	Jackson	Brown	7/15/2013	4	3/16/1953	63	Portland	FT	$98,655

As you enter the data, notice that Excel tries to complete your fields based on previous common entries.

REMOVE DUPLICATES

Duplicate entries may appear in tables. Why? Duplicates sometimes happen when data is entered incorrectly, by more than one person, or from more than one source. The following steps remove duplicate records in the table. In this particular table, Robert Griffin was entered twice by mistake. Delete the duplicate record by following the below steps:

1. Click anywhere in the table.
2. From the Table Tools Design tab click the **Remove Duplicates** button.

   Mac Users: Click the Table tab and click the Remove Duplicates button
3. The Remove Duplicates dialog box will open.
4. If necessary, click the Select All button to deselect all columns.
5. Click **OK** to remove duplicate records from the table.

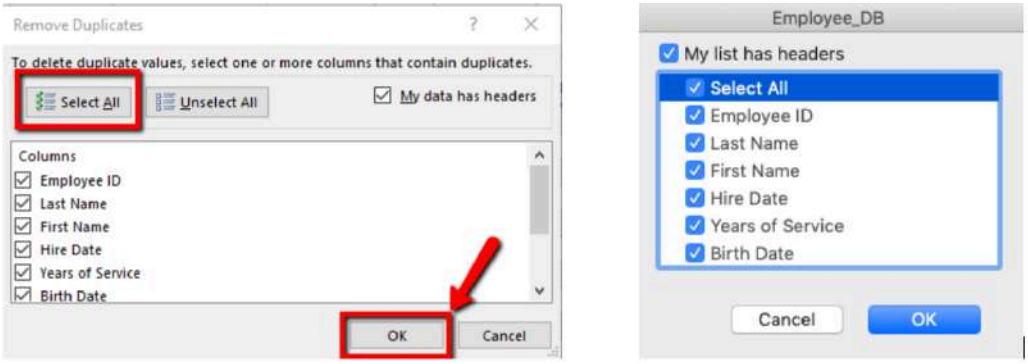


Figure 5.13 Excel for Windows and Excel for Mac Remove Duplicates Dialog Box

6. Excel notifies you that 1 duplicate record was removed.

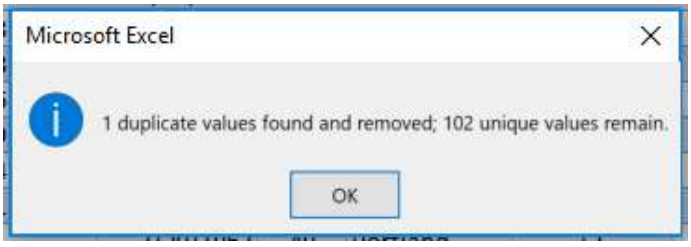


Figure 5.14 Results of Remove Duplicates

CREATE NEW COLUMNS

In this next exercise, we will explore how to add two new columns in the table. Take note, Excel automatically adds the column to the table's range and copies the format of the existing table heading to the new column heading. The first new column will use the VLOOKUP function to determine what cost of living

adjustment (COLA) the employee qualifies for based on the region the employee lives in. The second column added will calculate the projected salary increase based on the COLA. When you use a formula in a table it is considered a calculated column.

A calculated column uses a single formula that adjusts for each row and automatically expands to include additional rows in that column. The formula is immediately extended to those rows. You only need to enter a formula to have it automatically filled down to create a calculated column—there's no need to use the Fill or Copy commands.

As mentioned in the previous section, Excel assigns a name to the table, and to each column header in the table. When you add formulas to an Excel table, those names can appear automatically as you enter the formula and select the cell references in the table instead of manually entering them.

As a visual reference compare the differences to a formula entered in a cell, compared to in a table:

Formula – Cell References	**Formula – Table: Excel shows field names**
=SUM(J6:K6)	=SUM([Current Salary]:[COLA])

Excel displaying table and or field names in a formula is called a structured reference. The names in structured references adjust whenever you add or remove data from the table headings. Structured references also appear when you create a formula outside of an Excel table that references table data. The references can make it easier to locate tables in a large workbook. To include structured references in your formula, use point mode method to click the cells you want to reference instead of typing their cell reference in the formula.

Complete the following steps to enter two new columns to determine each employee's COLA and their projected salaries.

1. Click cell K5, and type **COLA**. Autofit the column width.
2. Click cell L5, and type **Projected Salary Increase**. Autofit the column width.
3. Click cell K6. From the Formulas tab, choose the VLOOKUP function (it is located within the "Lookup and Reference" tool) to look up each employee's **Store** location. Matching their store location to the COLA table, located on the COLA sheet, bring over their percentage of increase listed in the second (2) column of the col_index. Note this is an EXACT match, so eliminate all **FALSE** possibilities in the Range_lookup area:

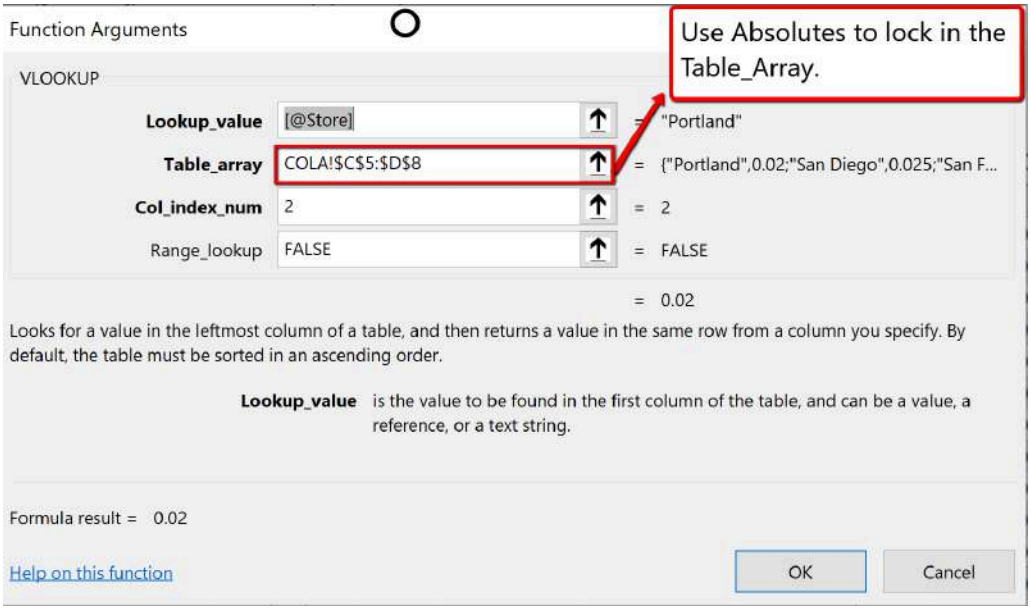


Figure 5.15 COLA VLOOKUP

4. The Excel table will request you to overwrite all cells in the column with the formula. Click the icon, and choose the Overwrite command as shown below:

Mac Users: Excel for Mac will automatically fill in the rest of the cells in the column. You do not have to click the icon. Close the Formula Builder pane.

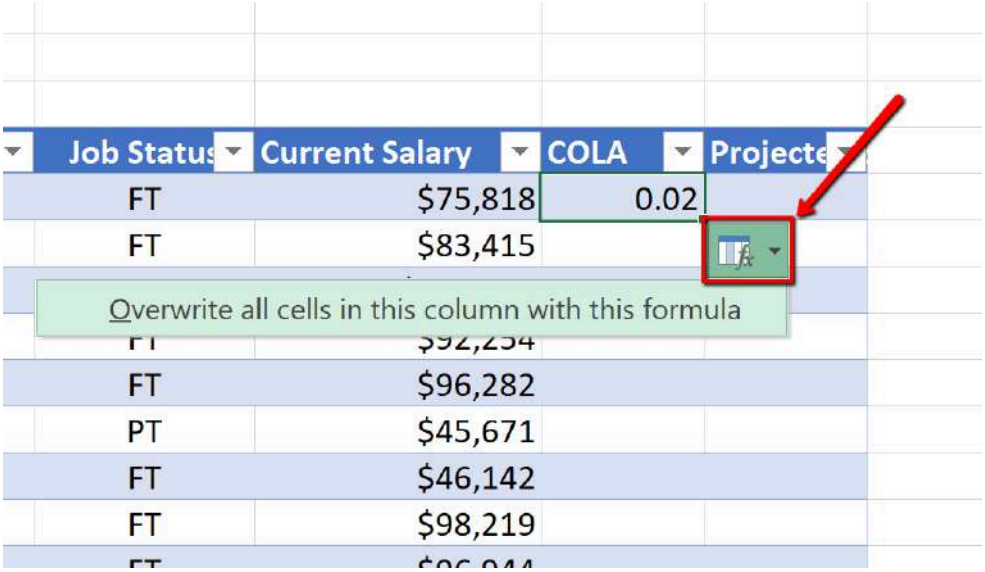


Figure 5.16 Table AutoCorrect Option

5. Using point mode method click the table cells to calculate the employees Projected Salary Increase by multiplying the Current Salary by the COLA increase:

=[@[Current Salary]]*[@COLA]

6. The Excel table will again request you to overwrite all cells in the column with the formula. Click the icon, and choose the Overwrite command.

Mac Users: You do not have to click the icon. Excel for Mac will auto-fill the rest of the cells in the column.

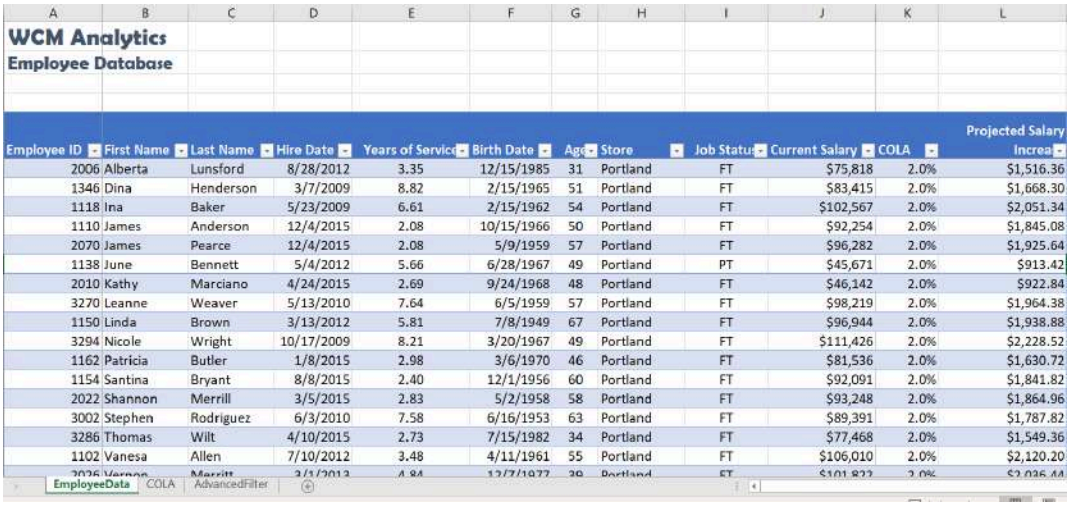
7. Format the COLA and Projected Salary Increase columns by selecting **K6:K107**, and applying the percentage number format, and increase the decimal to one place. Autofit the column widths.

(Suggestion: Use the short cut selection method; click in K6, press and hold the **CTRL** and **SHIFT** and **DOWN** arrow keys to select the column data.)

8. Select **L6:L107**, and apply the Currency number format.

(Suggestion: Use the short cut selection method; click in L6, press and hold the **CTR**L and **SHIFT** and **DOWN** arrow keys to select the column data.)

9. Select **L5.** Wrap, and right-align the text, then decrease the column width, and increase the row height to show the contents of the heading row wrapped on two lines.



WCM Analytics
Employee Database

Employee ID	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary	COLA	Projected Salary Increase
2006	Alberta	Lunsford	8/28/2012	3.35	12/15/1985	31	Portland	FT	$75,818	2.0%	$1,516.36
1346	Dina	Henderson	3/7/2009	8.82	2/15/1965	51	Portland	FT	$83,415	2.0%	$1,668.30
1118	Ina	Baker	5/23/2009	6.61	2/15/1962	54	Portland	FT	$102,567	2.0%	$2,051.34
1110	James	Anderson	12/4/2015	2.08	10/15/1966	50	Portland	FT	$92,254	2.0%	$1,845.08
2070	James	Pearce	12/4/2015	2.08	5/9/1959	57	Portland	FT	$96,282	2.0%	$1,925.64
1138	June	Bennett	5/4/2012	5.66	6/28/1967	49	Portland	PT	$45,671	2.0%	$913.42
2010	Kathy	Marciano	4/24/2015	2.69	9/24/1968	48	Portland	FT	$46,142	2.0%	$922.84
3270	Leanne	Weaver	5/13/2010	7.64	6/5/1959	57	Portland	FT	$98,219	2.0%	$1,964.38
1150	Linda	Brown	3/13/2012	5.81	7/8/1949	67	Portland	FT	$96,944	2.0%	$1,938.88
3294	Nicole	Wright	10/17/2009	8.21	3/20/1967	49	Portland	FT	$111,426	2.0%	$2,228.52
1162	Patricia	Butler	1/8/2015	2.98	3/6/1970	46	Portland	FT	$81,536	2.0%	$1,630.72
1154	Santina	Bryant	8/8/2015	2.40	12/1/1956	60	Portland	FT	$92,091	2.0%	$1,841.82
2022	Shannon	Merrill	3/5/2015	2.83	5/2/1958	58	Portland	FT	$93,248	2.0%	$1,864.96
3002	Stephen	Rodriguez	6/3/2010	7.58	6/16/1953	63	Portland	FT	$89,391	2.0%	$1,787.82
3286	Thomas	Wilt	4/10/2015	2.73	7/15/1982	34	Portland	FT	$77,468	2.0%	$1,549.36
1102	Vanesa	Allen	7/10/2012	3.48	4/11/1961	55	Portland	FT	$106,010	2.0%	$2,120.20

EmployeeData | COLA | AdvancedFilter

Figure 5.17 Calculated Columns

TOTAL ROW

A useful table tool for data analysis is the Total Row. You can quickly total data in an Excel table by enabling the **Total Row** option, and then use one of several built-in functions provided in a drop-down list, per column. The Total row, which is added to the end of the table after the last data record can calculate summary statistics, including the average, sum, minimum, and maximum of select fields within the table. The Total row is formatted with values displayed in bold, the double border line option is separating the data records from the Total row.

Apply a Total Row, and follow the below steps to sum three columns of data:

1. Click anywhere in the table and choose the Table Tools Design tab, and click on the Total Row option.

Mac Users: just click the Table tab and click on the Total Row option

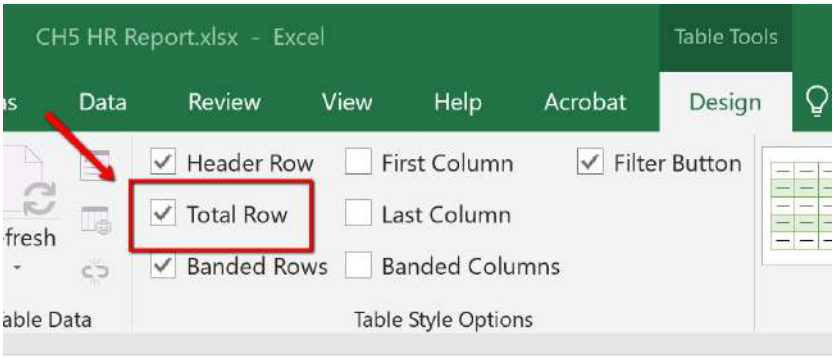


Figure 5.18 Total Row

2. Excel redirects you to the bottom of the table to view the total row, where a SUM defaulted in

the Projected Salary Increase column. Click cell **J108**, select choose the Total Row menu arrow. Choose **SUM** to total the Current Salary column.

 Figure 5.19 Total Row, Current Salary Column, Sum Function

3. Click cell **K108**, from the total row menu select Average. The average COLA increase will display.

 Figure 5.20 Total Row, COLA Column, Average Function

CENTER ACROSS SELECTION

Follow the below steps to center the title in cell A1:L2 using the Center Across Selection tool located in the Format Cells dialog box. In prior chapters, we used the 'Merge & Center' button to center text across a range. The Merge & Center tool centers the title but removes access to individual cells. This restriction can present a problem when trying to autofit column widths in a table. The Center Across Selection format centers text across multiple cells but does not merge the selected cell range into one cell making it a better formatting choice when working with tables.

1. Select cell **A1:L2,** and right-click to access the short cut menu.

   Mac Users: hold down **CTRL key** and click the selected cells to access the short cut menu
2. Choose Format Cells.
3. In the Format Cells dialogue box, choose the Alignment tab.
4. From the Horizontal alignment menu, choose **Center Across Selection**. Click **OK** to return to the table.

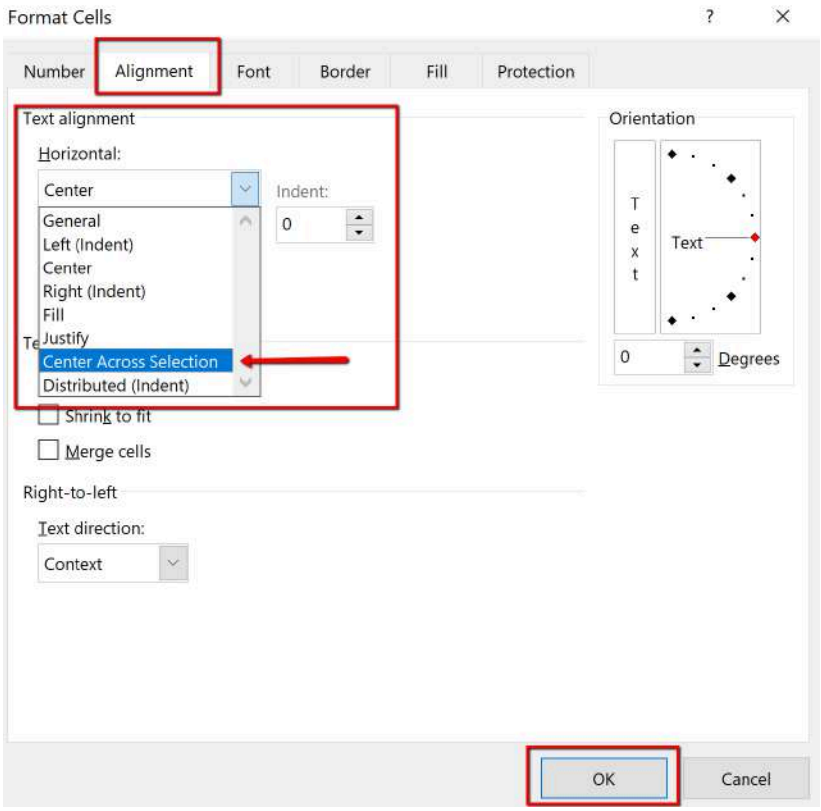
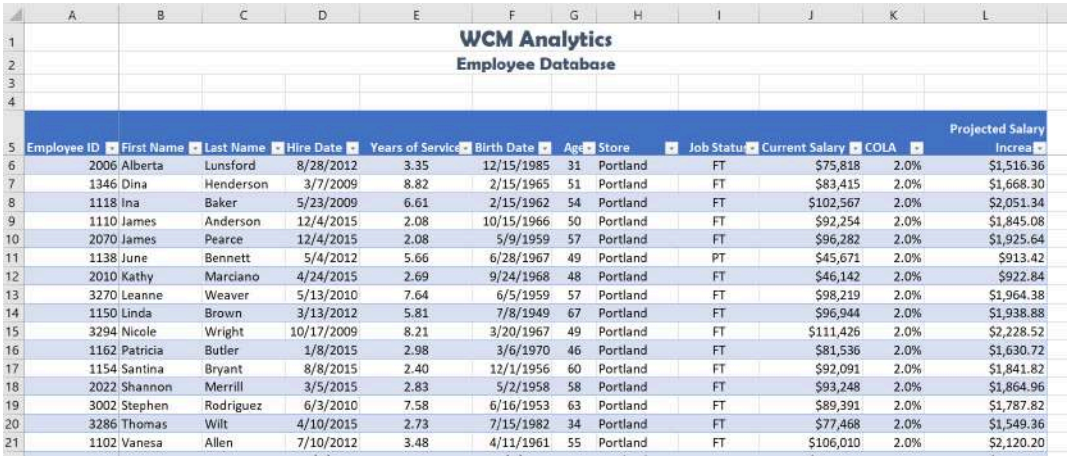


Figure 5.21 Format Cells Dialogue Box



WCM Analytics
Employee Database

Employee ID	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary	COLA	Projected Salary Increa
2006	Alberta	Lunsford	8/28/2012	3.35	12/15/1985	31	Portland	FT	$75,818	2.0%	$1,516.36
1346	Dina	Henderson	3/7/2009	8.82	2/15/1965	51	Portland	FT	$83,415	2.0%	$1,668.30
1118	Ina	Baker	5/23/2009	6.61	2/15/1962	54	Portland	FT	$102,567	2.0%	$2,051.34
1110	James	Anderson	12/4/2015	2.08	10/15/1966	50	Portland	FT	$92,254	2.0%	$1,845.08
2070	James	Pearce	12/4/2015	2.08	5/9/1959	57	Portland	FT	$96,282	2.0%	$1,925.64
1138	June	Bennett	5/4/2012	5.66	6/28/1967	49	Portland	PT	$45,671	2.0%	$913.42
2010	Kathy	Marciano	4/24/2015	2.69	9/24/1968	48	Portland	FT	$46,142	2.0%	$922.84
3270	Leanne	Weaver	5/13/2010	7.64	6/5/1959	57	Portland	FT	$98,219	2.0%	$1,964.38
1150	Linda	Brown	3/13/2012	5.81	7/8/1949	67	Portland	FT	$96,944	2.0%	$1,938.88
3294	Nicole	Wright	10/17/2009	8.21	3/20/1967	49	Portland	FT	$111,426	2.0%	$2,228.52
1162	Patricia	Butler	1/8/2015	2.98	3/6/1970	46	Portland	FT	$81,536	2.0%	$1,630.72
1154	Santina	Bryant	8/8/2015	2.40	12/1/1956	60	Portland	FT	$92,091	2.0%	$1,841.82
2022	Shannon	Merrill	3/5/2015	2.83	5/2/1958	58	Portland	FT	$93,248	2.0%	$1,864.96
3002	Stephen	Rodriguez	6/3/2010	7.58	6/16/1953	63	Portland	FT	$89,391	2.0%	$1,787.82
3286	Thomas	Wilt	4/10/2015	2.73	7/15/1982	34	Portland	FT	$77,468	2.0%	$1,549.36
1102	Vanesa	Allen	7/10/2012	3.48	4/11/1961	55	Portland	FT	$106,010	2.0%	$2,120.20

Figure 5.22 Center Across Selection

Attribution

7.XLSX.2 INTERMEDIATE TABLE SKILLS

Noreen Brown; Barbara Lave; Hallie Puncochar; Julie Romey; Mary Schatz; Art Schneider; Diane Shingledecker; and Jennifer Evans

Learning Objectives

- Sort table data.
- Custom Sort table data.
- Apply Custom List sort options.
- Filter table data using criteria filters.
- Use the Advanced Filter option to filter table data.
- Analyze data with PivotTables & Subtotals.

INTERMEDIATE TABLE SKILLS

SORT, FILTER, AND ANALYZE DATA WITH PIVOT TABLES & SUBTOTALS

SORTING

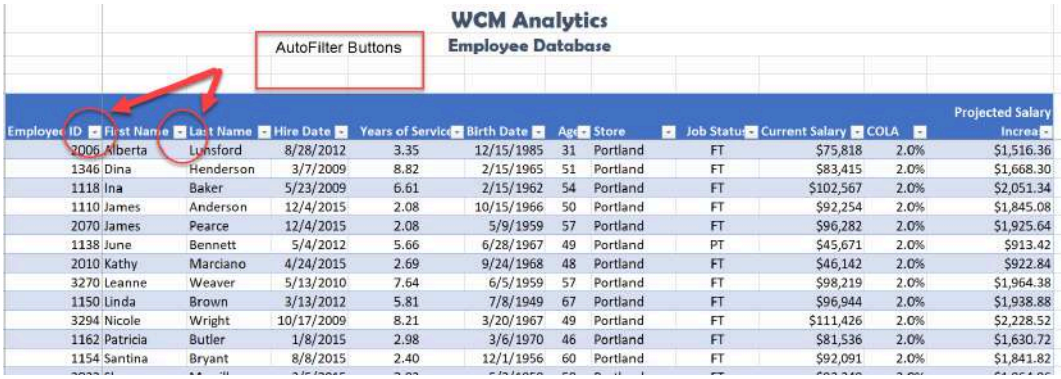
Sorting is one of the most common tools for data management. By arranging data sequentially the information becomes more meaningful. Arranging records in a specific sequence is called sorting. If you sort by one column this is considered a single sort. If you need to sort by more than one column, this is considered a custom sort.

The field or fields you select to sort are called sort keys. In Excel, you can sort your table by ascending or descending order. Data in ascending order appears lowest to highest, earliest to most recent, or alphabetically from A to Z. Data in descending order in arranged by highest to lowest, most recent to earliest, or alphabetically from Z to A.

Excel will sort a range of data that is not in a table. However, when working with large sets of information it is wise to make the data a table for integrity. Excel locks the row of information creating a record, thus when sorted, the record remains intact, just reorganized. For example, when you sort the table by last name, all of the records in each row move together. It is always a good idea to save a copy of your worksheet before applying sorts.

There are multiple places you can find and use sorting tools:

- When you first create a table, Excel automatically enables AutoFilter buttons; a tool used to sort, query, and filter the records in a table. The filter buttons appear to the right of the column headings. When you click the filter button sorting options appear on the menu options.



Employee ID	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary	COLA	Projected Salary Increa
2006	Alberta	Lunsford	8/28/2012	3.35	12/15/1985	31	Portland	FT	$75,818	2.0%	$1,516.36
1346	Dina	Henderson	3/7/2009	8.82	2/15/1965	51	Portland	FT	$83,415	2.0%	$1,668.30
1118	Ina	Baker	5/23/2009	6.61	2/15/1962	54	Portland	FT	$102,567	2.0%	$2,051.34
1110	James	Anderson	12/4/2015	2.08	10/15/1966	50	Portland	FT	$92,254	2.0%	$1,845.08
2070	James	Pearce	12/4/2015	2.08	5/9/1959	57	Portland	FT	$96,282	2.0%	$1,925.64
1138	June	Bennett	5/4/2012	5.66	6/28/1967	49	Portland	PT	$45,671	2.0%	$913.42
2010	Kathy	Marciano	4/24/2015	2.69	9/24/1968	48	Portland	FT	$46,142	2.0%	$922.84
3270	Leanne	Weaver	5/13/2010	7.64	6/5/1959	57	Portland	FT	$98,219	2.0%	$1,964.38
1150	Linda	Brown	3/13/2012	5.81	7/8/1949	67	Portland	FT	$96,944	2.0%	$1,938.88
3294	Nicole	Wright	10/17/2009	8.21	3/20/1967	49	Portland	FT	$111,426	2.0%	$2,228.52
1162	Patricia	Butler	1/8/2015	2.98	3/6/1970	46	Portland	FT	$81,536	2.0%	$1,630.72
1154	Santina	Bryant	8/8/2015	2.40	12/1/1956	60	Portland	FT	$92,091	2.0%	$1,841.82

Figure 5.23 AutoFilter Buttons

- From the Home tab, in the Editing group, click the 'Sort & Filter' button, and then click one of the sorting options on the Sort & Filter menu.

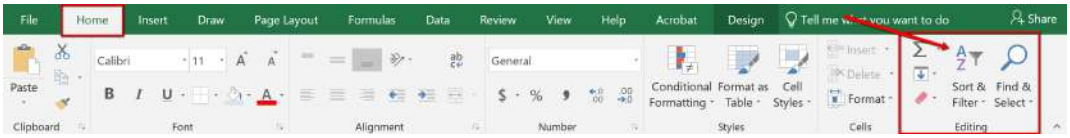


Figure 5.24 Sort and Filter Menu

- From the Data tab, use the 'Sort A to Z' or 'Sort Z to A' buttons or for multiple levels select the Sort button to open the Custom Sort dialogue.

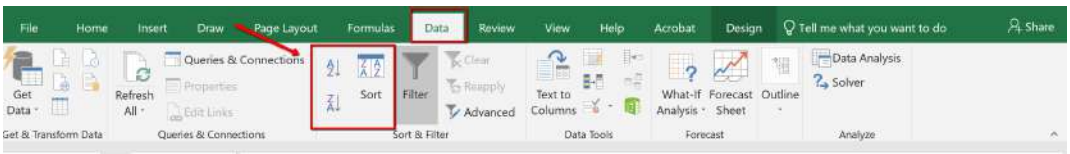


Figure 5.25 Data Tab Sort options

- Right-click anywhere in a table and then point to Sort on the shortcut menu to display the Sort sub-menu.

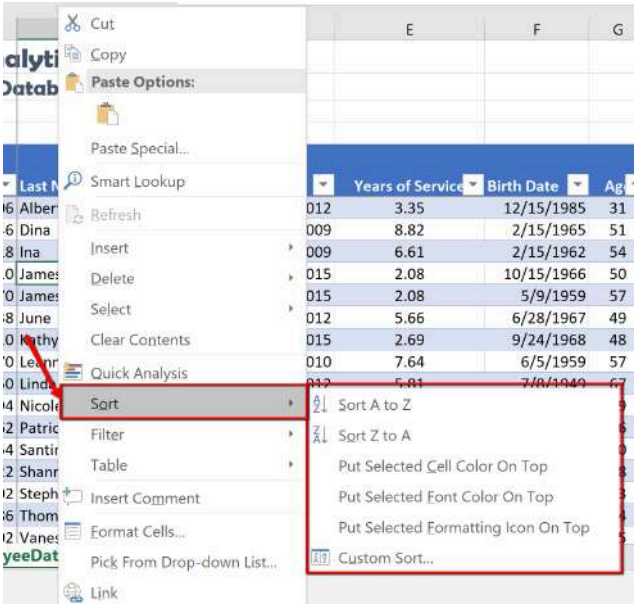
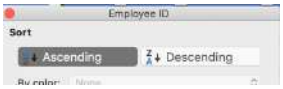


Figure 5.26
Right-Click Menu

Complete a single level sort by following the steps:

1. In the EmployeeID heading, click the filter button.
2. Choose to Sort Smallest to Largest.

Mac Users: Click the A-Z Ascending button 

Notice Excel arranges in chronological order all the employee data based on the EmployeeID number, however keeping each record together. You will also notice the filter button now displays an up arrow denoting an ascending sort.

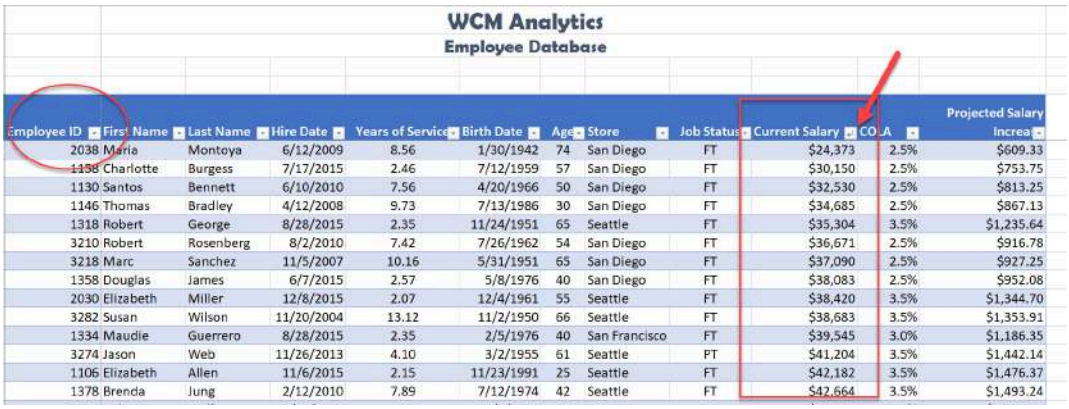
 Figure 5.27 EmployeeID Sort

The following steps will sort the records in descending order by Current Salary using the 'Sort Largest to Smallest' option form the filter button.

1. Click the filter button located in the Current Salary heading.
2. Choose Sort Largest to Smallest option from the menu.

Mac Users: click the "**Descending**" button

Notice the original sort has been overridden, and the information is now organized based on the largest Current Salary. You will see the small arrow on the EmployeeID filter is gone, and an arrow pointing down for Descending Order is visible on the Current Salary filter button.



WCM Analytics
Employee Database

Employee ID	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary	COLA	Projected Salary Increa
2038	Maria	Montoya	6/12/2009	8.56	1/30/1942	74	San Diego	FT	$24,373	2.5%	$609.33
1138	Charlotte	Burgess	7/17/2015	2.46	7/12/1959	57	San Diego	FT	$30,150	2.5%	$753.75
1130	Santos	Bennett	6/10/2010	7.56	4/20/1966	50	San Diego	FT	$32,530	2.5%	$813.25
1146	Thomas	Bradley	4/12/2008	9.73	7/13/1986	30	San Diego	FT	$34,685	2.5%	$867.13
1318	Robert	George	8/28/2015	2.35	11/24/1951	65	Seattle	FT	$35,304	3.5%	$1,235.64
3210	Robert	Rosenberg	8/2/2010	7.42	7/26/1962	54	San Diego	FT	$36,671	2.5%	$916.78
3218	Marc	Sanchez	11/5/2007	10.16	5/31/1951	65	San Diego	FT	$37,090	2.5%	$927.25
1358	Douglas	James	6/7/2015	2.57	5/8/1976	40	San Diego	FT	$38,083	2.5%	$952.08
2030	Elizabeth	Miller	12/8/2015	2.07	12/4/1961	55	Seattle	FT	$38,420	3.5%	$1,344.70
3282	Susan	Wilson	11/20/2004	13.12	11/2/1950	66	Seattle	FT	$38,683	3.5%	$1,353.91
1334	Maudie	Guerrero	8/28/2015	2.35	2/5/1976	40	San Francisco	FT	$39,545	3.0%	$1,186.35
3274	Jason	Web	11/26/2013	4.10	3/2/1955	61	Seattle	PT	$41,204	3.5%	$1,442.14
1106	Elizabeth	Allen	11/6/2015	2.15	11/23/1991	25	Seattle	FT	$42,182	3.5%	$1,476.37
1378	Brenda	Jung	2/12/2010	7.89	7/12/1974	42	Seattle	FT	$42,664	3.5%	$1,493.24

Figure 5.28 Current Salary Sort

Skill Refreshed

Sort a Column

1. Click on the filter Click arrow to the right of the header in the column you want to sort.
2. Click on the choice AZ↑ or ZA↓ to sort your data by that column.

CUSTOM SORT

When you need to sort by more than one level, you must use the Custom Sort option. Complete the following steps to organize the data by Store, Last Name, Current Salary, all in Ascending Order (A-Z).

1. Select the Data tab, and click the Sort button. Notice the last column sorted by is listed. Change the column heading name by dropping down the **Sort by** menu and select **Store.**
2. Click Add Level.

Mac Users: click the + symbol 

3. Click the down arrow in the **Then by** section, and choose the column heading names **as shown below in Figure 5.29**. Note to click Add Level to add the next column heading. The order you select the headings will determine how the table information is sorted.

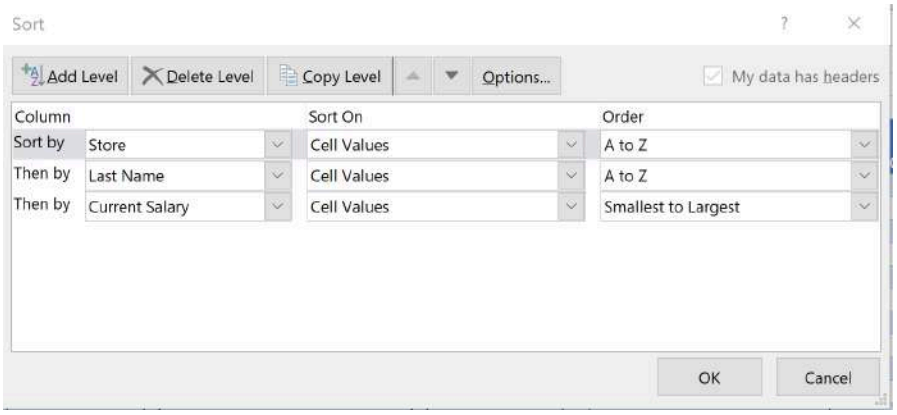


Figure 5.29 Sort Dialogue Box

4. Once you select to Sort by column headings, choose the Order by selecting to sort in ascending order (A-Z) for the Store and Last name fields, and Smallest to Largest, for the Current Salary field.
5. Click OK.

Notice the information is now sorted by three levels, per **Store**, each employee is organized by **Last Name**, and **Current Salary** in ascending order (smallest to largest). Each of the filter buttons indicates the sort with the up arrow.

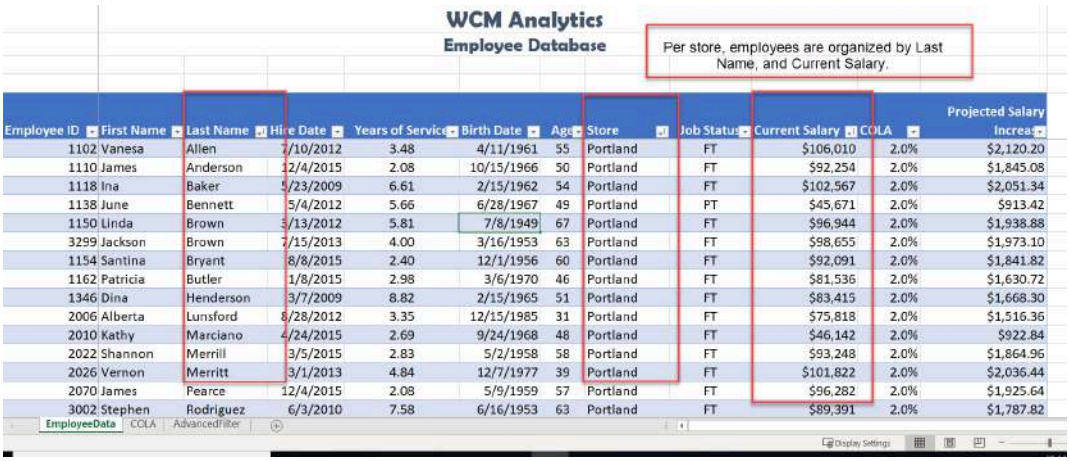


Figure 5.30
Custom Sort Visual

Skill Refresher

Custom Sort (Multiple Level Sort)

1. Select the Data tab, and click the Sort button.
2. Choose Add Level.
3. Click the down arrow in the Column field and choose the column heading to sort by.
4. Repeat the above steps to add another level and select the next column heading to sort by.
5. The order you select the headings will determine how the table information is sorted.

CUSTOM LIST SORT

When sorting you can create custom lists that allow sorting by characteristics that do not sort alphabetically. Example, text items such as high, medium, and low—or S, M, L, XL. Dates commonly require custom lists so you can vary in the way data is sorted by days of the week or months of the year.

In our case, we want to create a custom list that sorts our stores, which is not, in ascending or descending order. The human resources office likes to order the stores based on the location size. The company headquarters is in Seattle and employs the most people. The next biggest location is San Diego etc. Follow the below steps to create a custom list ordering the stores as shown below:

Seattle
San Diego
Portland
San Francisco

Mac Users: The steps to create a **custom sort list are different for Excel for Mac. Please skip the below steps and follow the alternate steps below Figure 5.34.**

Follow the below steps to create a custom list ordering:

1. While clicked in the table, choose the Data tab and click the **Sort** button.
2. In the Sort by row, click the drop-down menu in the Order Column for the **Store** heading. Choose **Custom List**.

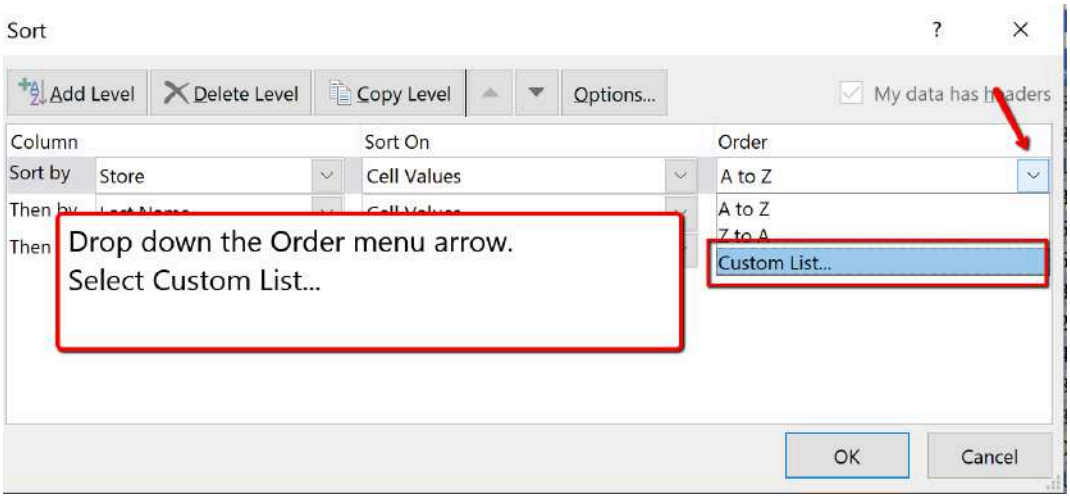


Figure 5.31 Custom List Dialogue Box

3. Click in the **List entries:** box and type **Seattle**, and press enter. Type the remainder of the locations shown in Figure 5.32, pressing enter after each store location typed. Once all locations are entered, click **Add**. Then choose **Ok**.

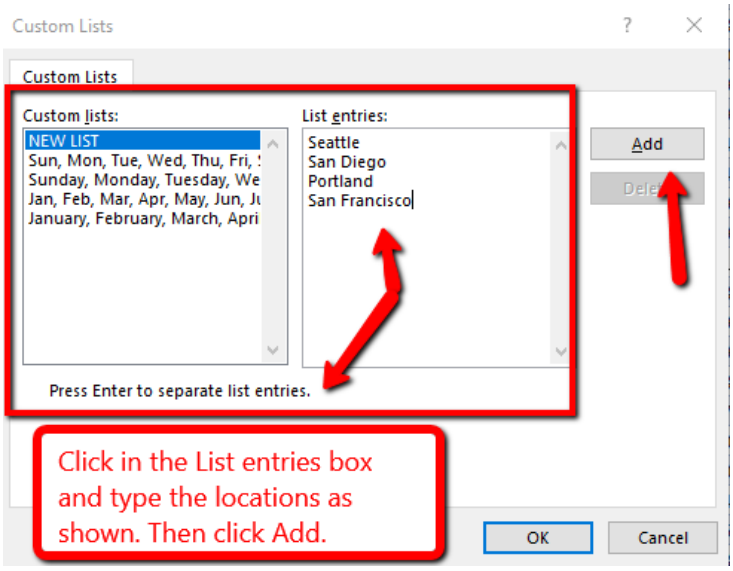


Figure 5.32 Custom List Entries dialogue box

4. You will see the Order of the Store sort update. Click **OK** to close the Sort dialogue box.

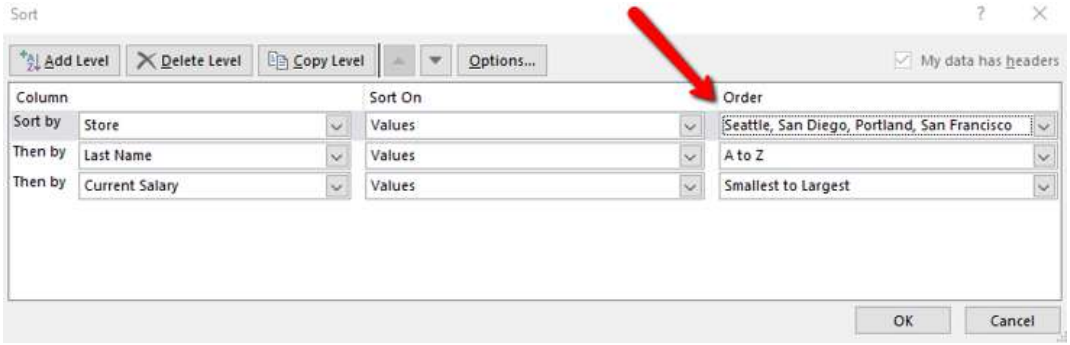
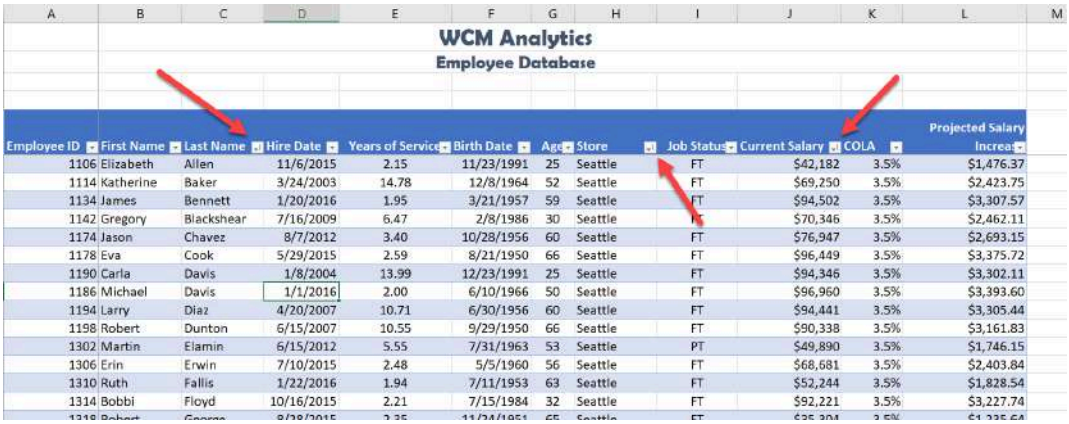


Figure 5.33
Sort_Dialogue Box
Custom List Order

The custom sort is applied and the table is now sorted by Store, using the custom order, then the Last Name of the employee and then by the Current Salary column.

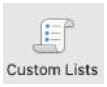


WCM Analytics
Employee Database

Employee ID	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary	COLA	Projected Salary Increas
1106	Elizabeth	Allen	11/6/2015	2.15	11/23/1991	25	Seattle	FT	$42,182	3.5%	$1,476.37
1114	Katherine	Baker	3/24/2003	14.78	12/8/1964	52	Seattle	FT	$69,250	3.5%	$2,423.75
1134	James	Bennett	1/20/2016	1.95	3/21/1957	59	Seattle	FT	$94,502	3.5%	$3,307.57
1142	Gregory	Blackshear	7/16/2009	6.47	2/8/1986	30	Seattle	FT	$70,346	3.5%	$2,462.11
1174	Jason	Chavez	8/7/2012	3.40	10/28/1956	60	Seattle	FT	$76,947	3.5%	$2,693.15
1178	Eva	Cook	5/29/2015	2.59	8/21/1950	66	Seattle	FT	$96,449	3.5%	$3,375.72
1190	Carla	Davis	1/8/2004	13.99	12/23/1991	25	Seattle	FT	$94,346	3.5%	$3,302.11
1186	Michael	Davis	1/1/2016	2.00	6/10/1966	50	Seattle	FT	$96,960	3.5%	$3,393.60
1194	Larry	Diaz	4/20/2007	10.71	6/30/1956	60	Seattle	FT	$94,441	3.5%	$3,305.44
1198	Robert	Dunton	6/15/2007	10.55	9/29/1950	66	Seattle	FT	$90,338	3.5%	$3,161.83
1302	Martin	Elamin	6/15/2012	5.55	7/31/1963	53	Seattle	PT	$49,890	3.5%	$1,746.15
1306	Erin	Erwin	7/10/2015	2.48	5/5/1960	56	Seattle	FT	$68,681	3.5%	$2,403.84
1310	Ruth	Fallis	1/22/2016	1.94	7/11/1953	63	Seattle	FT	$52,244	3.5%	$1,828.54
1314	Bobbi	Floyd	10/16/2015	2.21	7/15/1984	32	Seattle	FT	$92,221	3.5%	$3,227.74

Figure 5.34
Custom List Sort
Visual

Mac Users alternate steps for creating a custom sort list:

1. Click the **Excel** menu option and choose **Preferences**
2. Click on the **Custom List** button 

3. Type the list of cities in the "List entries" box as shown in **Figure 5.32 above** then click the **Add** button and close the Custom List dialog box
4. Click anywhere in the table, and then click the **Data tab** and click the **Sort** button
5. Click the drop-down menu in the **Order** Column for the **Store** heading. Choose **Custom List**
6. Click on the custom list of cities that you just created and then click the **OK button twice**
7. The custom sort is applied and the table is now sorted by Store, using the custom order, then the Last Name of the employee and then by the Current Salary column. See **Figure 5.34 above.**

Skill Refresher

Custom List Sort

1. Select the Data tab, and click the Sort button.
2. Click the drop-down menu in the Order Column of the field needing a custom list created.
3. Choose Custom List.
4. Click in the List entries box and type the custom list desired.
5. Then click Add.
6. Click Ok.

FILTER DATA

If your worksheet contains a lot of data, it can be difficult to find information quickly. Applying **Filters** is an efficient and effective way to only show the information needed. Typically when filtering you are searching the data for specific information. Generally speaking, you are searching the data based on a question, or in other words, querying the data, and returning only the information that satisfies the question. The process of filtering records based on one or more filter criteria is called a query. Filtering data hides the rows whose values do not match the search criteria. The information that does not display is not deleted, it is just hidden, and will be redisplayed by removing the filter or applying a new filter.

Like sorting, Filter options are located in the filter button alongside each field name. By clicking the filter button, you can choose which values in that field to display, hiding the rows or records that do not match that value. The filter lets you choose to display only those records that meet specified criteria such as color, number, or text. In this situation, criteria is defined as; a logical rule by which data is tested and chosen.

For example, you can filter the table to display a specific name or item by typing it in a Search box. The name you selected acts as the criterion for filtering the table, which results in Excel displaying only those records that match the criterion. The selected checkboxes indicate which items will appear in the table. By default, all of the items are selected. If you deselect an item from the filter menu, it is removed from the filter criterion. Excel will not display any record that contains the unchecked item. As with the previous sort techniques, you can include more than one column when you filter by clicking a second filter button and making choices. After you filter data, you can copy, find, edit, format, chart, or print the filtered data without rearranging or moving it.

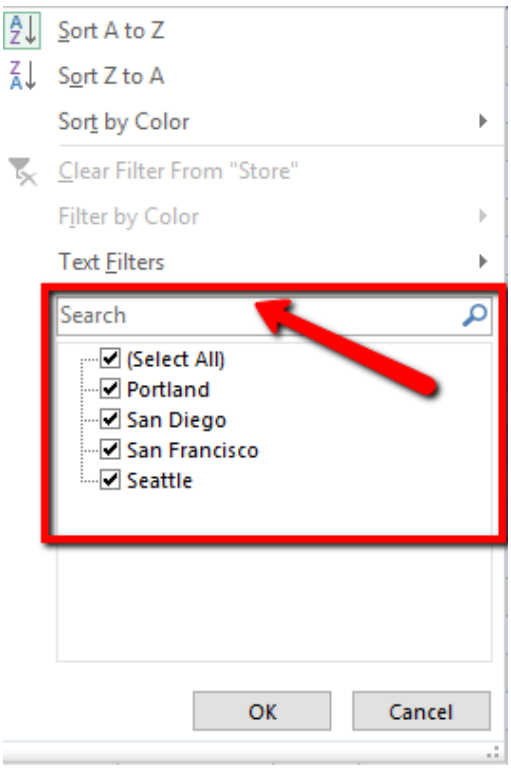


Figure 5.35 Filter Search Menu

Complete the following steps and filter data according to each query.

How many employees are at a Part-Time (PT) status?

1. Click the filter button on the **Job Status** column heading.
2. Click **Select All**, to deselect options.
3. Click the **PT** box to only display the part-time employees.
4. From the total row, in cell **I108**, choose the **Count function** count the number of employees at a PT status.

The answer to the question is there are currently are 11 employees at a PT time status. The total row will display the part-time total current salaries, and what the projected salary increase for part-time help will be after COLA adjustments.

WCM Analytics
Employee Database

Employee ID	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary	COLA	Projected Salary Increa
2034	Billy	Miller	8/11/2009	6.39	1/10/1959	57	Seattle	PT	$53,582	3.5%	$1,875.37
3234	Sherri	Spaulding	3/13/2009	6.81	5/12/1969	47	Seattle	PT	$64,598	3.5%	$2,260.93
3274	Jason	Web	11/26/2013	4.10	3/2/1955	61	Seattle	PT	$41,204	3.5%	$1,442.14
1330	Robert	Griffin	3/11/2010	7.81	3/3/1958	58	San Diego	PT	$45,657	2.5%	$1,141.43
3222	Mary	Smith	7/17/2015	2.46	10/15/1984	32	San Diego	PT	$51,639	2.5%	$1,290.98
1138	June	Bennett	5/4/2012	5.66	6/28/1967	49	Portland	PT	$45,671	2.0%	$913.42
2090	James	Price	11/27/2012	3.10	6/8/1952	64	San Francisco	PT	$64,826	3.0%	$1,944.78
3278	Jeffery	Whiting	8/25/2004	13.36	5/13/1964	52	San Francisco	PT	$65,978	3.0%	$1,979.34
3280	Jennifer	Wiliams	9/6/2008	9.32	6/26/1967	49	San Francisco	PT	$96,354	3.0%	$2,890.62
Total								11	$628,997	3.0%	$19,221.08

Figure 5.36 PT Filter Visual

USING CRITERIA FILTERS

The filters created are limited to selecting records for fields matching a specific value or set of values. For more general criteria, you can use criteria filters, which are expression involving dates and times, numeric values, and text strings. Excel will identify what criteria filter to display based on the information in the column. For example, you can filter the employee data to show only those employees hired within a specific date range. Notice the criteria filter changes to **Date Filters.** If we were looking at the Current Salary column, the filter would be a Numbers Filter.

Using criteria filters, follow the below steps to search for employees who have been with the company for a specific time period.

Identify employees who have been with the company between 2013-2016.

1. While clicked in the table, clear any sort or filter applied by clicking the Data tab. In the Sort & Filter group choose the **Clear** button. 

2. Click the Filter button in the **Hire Date** column. Select Date Filters, and choose the Between criteria.

Mac Users: uncheck the **Select All** checkbox **before** choosing the **Between** option.

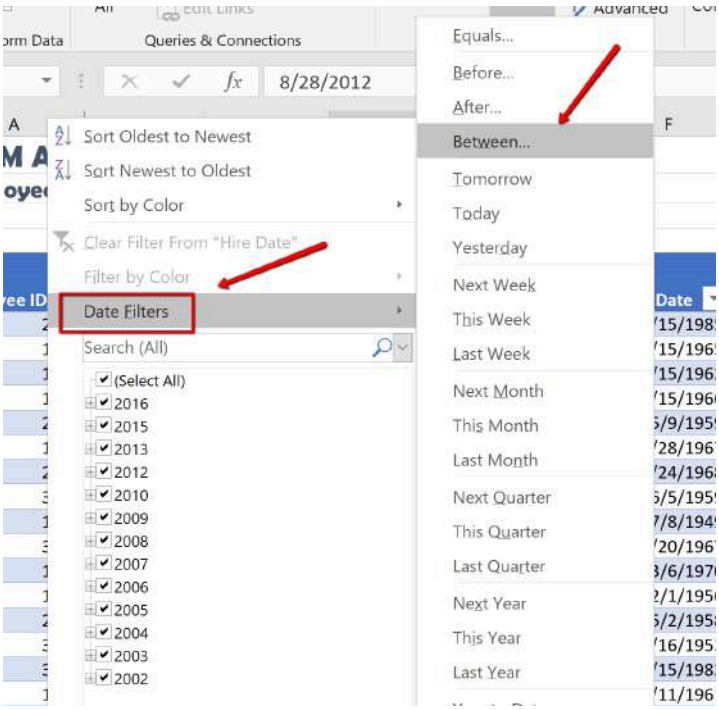


Figure 5.37 Date Filter Menu

3. Search for employees with a hire date between 2013, and 2016. In the "**is after or equal to"** section type **1/01/2013**, and typing in the "**is before or equal to"** section type **12/31/2016**. Then click **OK**.

Mac Users: Excel for Mac sections simply say "**After"** and **"Before"**

Figure 5.38 Date Filter Between Dialogue Box

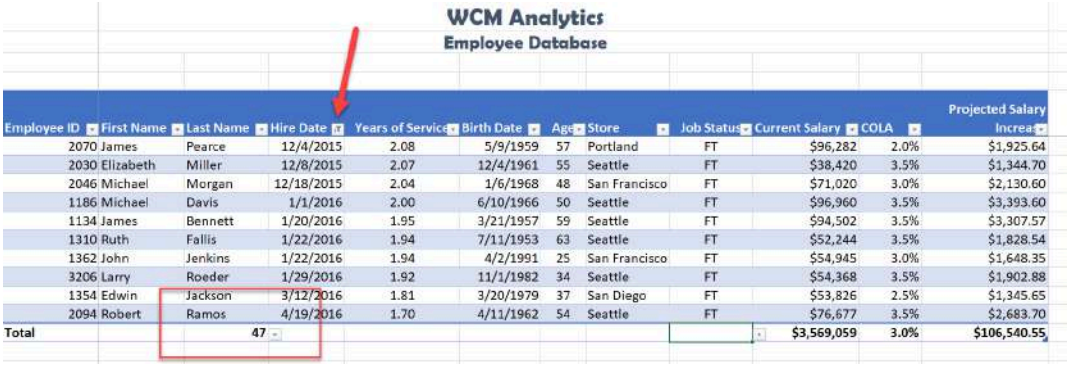
4. Sort the filtered table from **Oldest to Newest** by Date Hired.

5. In the total row section, count the last name names of the employees by applying the count function in cell B108.

6. In the total row, select cell **I108**, and choose **None** to turn off the count function in the Job Status Column.

Notice the table total row show 47 employees hired between the specified dates. These employees will be evaluated for a COLA adjustment.

Notice the filter button displays a filter symbol and an up arrow indicating the column is filtered and sorted in ascending order.



WCM Analytics
Employee Database

Employee ID	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary	COLA	Projected Salary Increa
2070	James	Pearce	12/4/2015	2.08	5/9/1959	57	Portland	FT	$96,282	2.0%	$1,925.64
2030	Elizabeth	Miller	12/8/2015	2.07	12/4/1961	55	Seattle	FT	$38,420	3.5%	$1,344.70
2046	Michael	Morgan	12/18/2015	2.04	1/6/1968	48	San Francisco	FT	$71,020	3.0%	$2,130.60
1186	Michael	Davis	1/1/2016	2.00	6/10/1966	50	Seattle	FT	$96,960	3.5%	$3,393.60
1134	James	Bennett	1/20/2016	1.95	3/21/1957	59	Seattle	FT	$94,502	3.5%	$3,307.57
1310	Ruth	Fallis	1/22/2016	1.94	7/11/1953	63	Seattle	FT	$52,244	3.5%	$1,828.54
1362	John	Jenkins	1/22/2016	1.94	4/2/1991	25	San Francisco	FT	$54,945	3.0%	$1,648.35
3206	Larry	Roeder	1/29/2016	1.92	11/1/1982	34	Seattle	FT	$54,368	3.5%	$1,902.88
1354	Edwin	Jackson	3/12/2016	1.81	3/20/1979	37	San Diego	FT	$53,826	2.5%	$1,345.65
2094	Robert	Ramos	4/19/2016	1.70	4/11/1962	54	Seattle	FT	$76,677	3.5%	$2,683.70
Total			47						$3,569,059	3.0%	$106,540.55

Figure 5.39 Date Filter and Sort

SLICERS

Another way to filter an Excel table is with slicers. Slicers, generally speaking, are visual filter buttons you can click to filter the table data. Slicers show the current filtered category, which makes it easy to understand what exactly is displayed. For example, a slicer for the Store field would have buttons for the Seattle, San Diego, Portland, and San Francisco locations.

When slicer buttons are selected, the data is filtered to show only those records that match the criteria. Multiple buttons can be selected at the same time, and a table can have multiple slicers, each linked to a different field. When multiple slicers are used, Excel uses the AND logical operator so filtered records must meet all of the criteria indicated in the slicer. When selecting multiple buttons in a Slicer, use the shift key to select adjacent field names. If the field names are not adjacent, use the non-adjacent selection method, pressing the CTL button, and selecting the field names needed.

Follow the below steps to filter the table using visual Slicer buttons.

1. Click in the table area. From the Data tab, choose **Clear** to remove the current sort and filter applied to the data.
2. To make room for the **Slicer** buttons at the top of the table, we will add 4 rows between the title and the table area. Right-click cell A3. Choose Insert. Select **Entire Row**. Repeat these steps until the table heading starts in row **A9**.

Mac users should hold down **CTRL key** and click cell A3. Then repeat until the table heading starts in row **A9.**

 Figure 5.40 Added Rows

3. Click back into the table area. Choose the **Insert** tab. Click **Slicer**. When the Insert Slicers dialogue box opens, click the **Store** and **Job Status** field names to display as slicers. Click **OK**.

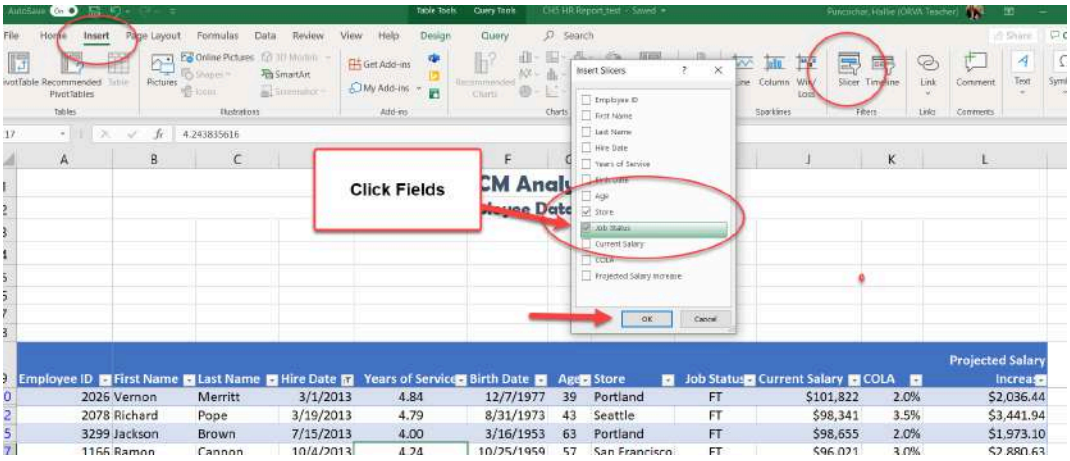


Figure 5.41 Slicer Dialogue Box

4. Move, and re-size the Slicer boxes to fit in the **approximate** area of **I1:J8** and **K1:L8**. Make sure the buttons remain visible. Below is a visual example.

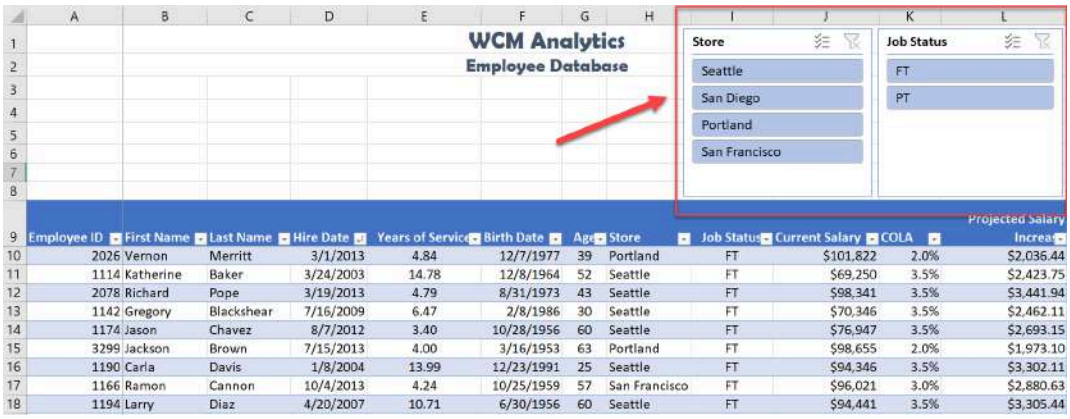
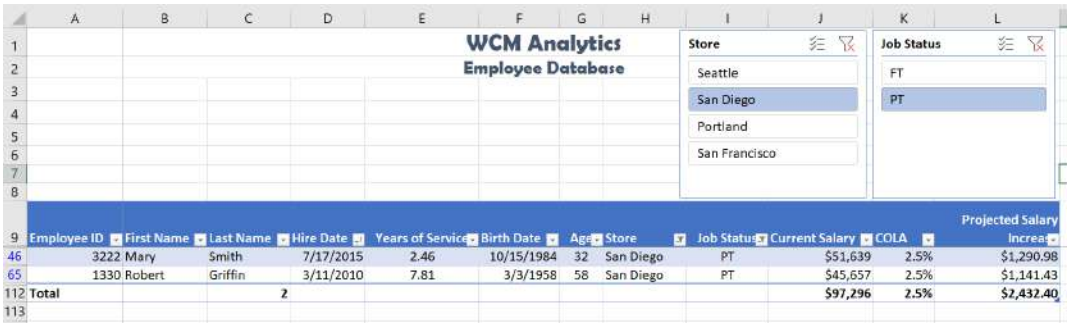


Figure 5.42 Slicer Layout Example

5. From the **Store** slicer, click the **San Diego** button. Notice the data filters to only show the data for San Diego.

6. From the **Job Status** slicer click **PT**. Notice the data filters to only show the data for PT employees in San Diego.



Employee ID	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary	COLA	Projected Salary Increa
3222	Mary	Smith	7/17/2015	2.46	10/15/1984	32	San Diego	PT	$51,639	2.5%	$1,290.98
1330	Robert	Griffin	3/11/2010	7.81	3/3/1958	58	San Diego	PT	$45,657	2.5%	$1,141.43
Total			2						$97,296	2.5%	$2,432.40

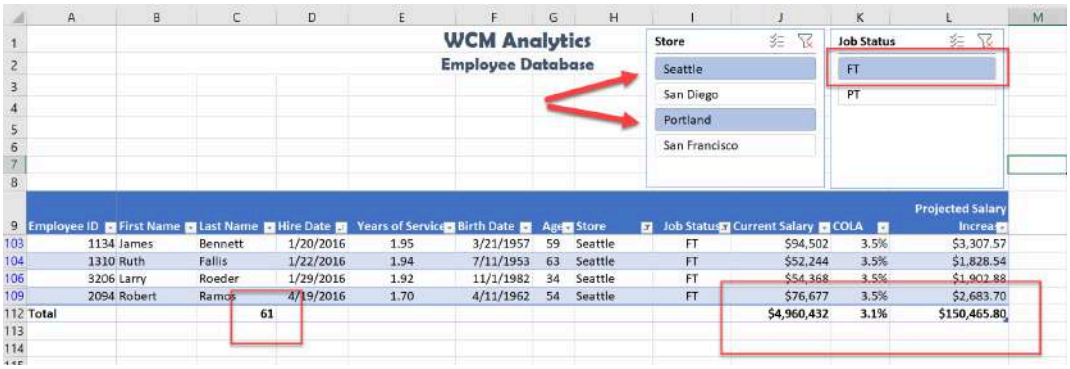
Figure 5.43 Slicer Solution

7. Return to the Store slicer and choose **Seattle** and **Portland**. Note the non-adjacent selection method is needed. Select Seattle first, then press and hold the Ctrl button on the keyboard, and then select **Portland**.

Mac Users: hold down the Command key not the Ctrl key before you click on Portland.

8. Change the Job Status slicer selection to **FT**.

The table results show there are **61** FT employees in Seattle and Portland. The Projected Salary Increase after the COLA adjustment for the Northwest region is $150,465.80.



Employee ID	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary	COLA	Projected Salary Increa
1134	James	Bennett	1/20/2016	1.95	3/21/1957	59	Seattle	FT	$94,502	3.5%	$3,307.57
1310	Ruth	Fallis	1/22/2016	1.94	7/11/1953	63	Seattle	FT	$52,244	3.5%	$1,828.54
3206	Larry	Roeder	1/29/2016	1.92	11/1/1982	34	Seattle	FT	$54,368	3.5%	$1,902.88
2094	Robert	Ramos	4/19/2016	1.70	4/11/1962	54	Seattle	FT	$76,677	3.5%	$2,683.70
Total			61						$4,960,432	3.1%	$150,465.80

Figure 5.44 Non-Adjacent Slicer Solution

ADVANCED FILTERS

Filter buttons are limited to combining fields using advanced logic or complex criteria. If the data you want to filter requires complex criteria, you can use the **Advanced Filter** dialog box. The Advanced Filter works differently from the **Filter** command in several important ways:

- It displays the **Advanced Filter** dialog box instead of the AutoFilter menu.
- You type the advanced criteria in a separate criteria range in a worksheet and above the range of cells or table that you want to filter. Excel uses the separate criteria range in the **Advanced Filter** dialog box as the source for the advanced criteria.

For example, you searched records for employees in the Seattle and San Diego offices AND for employees working at full-time bases, AND have a base salary between the below Salary Ranges:

Store	Job Status	Current Salary	Current Salary
• Seattle • San Diego	• Full Time • Full Time	• >=$50,000 • >=$70,000	• <=$60,000 • <=$80,000

Figure 5.45 Advanced Filter Criteria

To run the above complex criteria mentioned above follow the below steps:

1. From the **EmployeeData** sheet, click in the table, then select the Data tab and clear the current filters by selecting the Clear button.
2. Select the Table Tools Design tab and turn off the Total Row.
   Mac Users: just click the Table tab and turn off the Total Row
3. Select the **Advanced Filter** sheet. Click cell A10. The criteria mentioned in the above example has already been entered for this advanced filter exercise. Next, you will use an advanced filter to copy the records that match these criteria.
4. From the **Data** tab, click the **Advanced** button. The Advanced Filter dialog box opens.
5. Click the **Copy to another location** option button to copy matching records from the data range.
6. Click in the List range box to make it active, and then navigate to the EmployeesData sheet, click **cell A9**, and then press and hold the **CTRL** and **SHIFT** and **SPACEBAR** to select the entire table. In the List range box, you will see **Employee_DB[#All]** in the list range box.
   Mac Users: The keyboard shortcut of "CTRL, SHIFT, SPACEBAR" **does not work** in Exel for Mac. You should click in **Cell A9**, scroll down to the end of the data, hold down the **Shift key** and click in C**ell L112** to select the entire table
7. Click, or press the tab key, to move to the Criteria Range box.
8. From the Advanced Filter sheet, select A6:D8. You will see **'Advanced Filter'!Criteria** populate in the criteria range box.
9. Click, or press the tab key, to move to the Copy to box, and then click cell A10 to specify the location for inserting the copied records. You will see **'Advanced Filter'!$A$10** in the Copy to criteria range box.

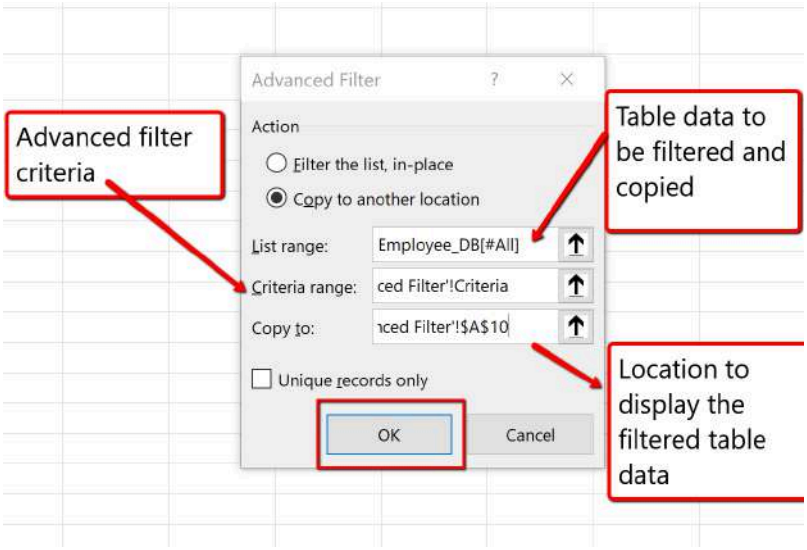


Figure 5.46
Advanced Filter
Dialogue Box

9. Click **OK** to copy the records that match the advanced filter criteria. Save your work.

The advanced search results list 7 employees that meet the criteria. Of these 7 employees, only 1 full-time employee in San Diego has a current salary between $70,000 and $80,000 dollars, and 6 full-time Seattle employees have a current salary between $50,000 and $60,000 dollars.

WCM Analytics
Advanced Filter Criteria
Seattle & San Diego Current Salaries

Store	Job Status	Current Salary	Current Salary
Seattle	FT	>=50000	<=60000
San Diego	FT	>=70000	<=80000

Employee	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary	COLA	Projected Salary Increase
1394	Shanika	Lloyd	8/24/2004	13.36	8/11/1966	50	Seattle	FT	$53,186	3.5%	$1,861.51
3242	Charles	Taylor	2/5/2012	3.90	1/18/1947	69	Seattle	FT	$58,291	3.5%	$2,040.19
3258	Johnny	Vazquez	9/18/2006	11.29	6/16/1953	63	Seattle	FT	$52,125	3.5%	$1,824.38
3214	Robert	Ross	8/28/2015	2.35	9/24/1974	42	Seattle	FT	$58,309	3.5%	$2,040.82
1350	Patrice	Hutton	12/29/2009	6.01	4/4/1953	63	San Diego	FT	$75,037	2.5%	$1,875.93
1310	Ruth	Fallis	1/22/2016	1.94	7/11/1953	63	Seattle	FT	$52,244	3.5%	$1,828.54
3206	Larry	Roeder	1/29/2016	1.92	11/1/1982	34	Seattle	FT	$54,368	3.5%	$1,902.88

Figure 5.47
Advanced Filter
Results

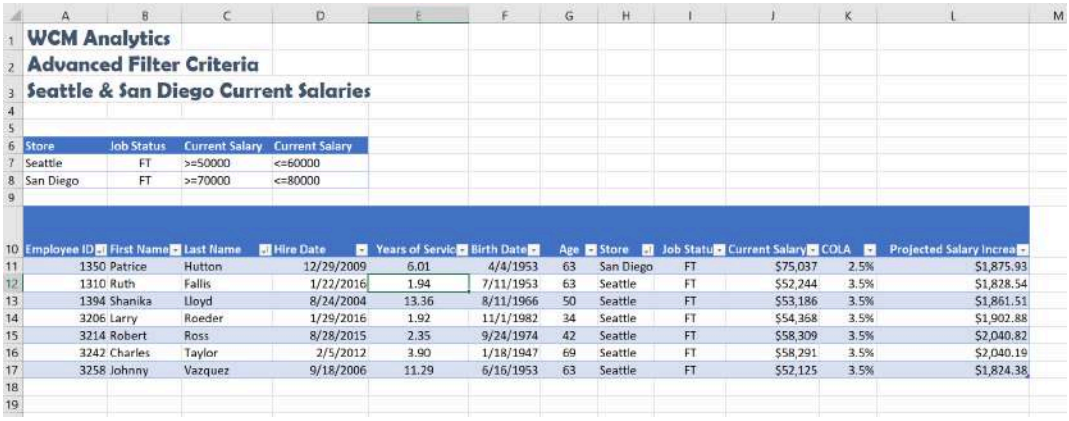
INSERT TABLE

Let's review another away to turn a range of data into a table.

1. Select the **Advanced Filter** sheet, and click cell A10.
2. From the Insert tab, choose **Table**.
3. The Create Table dialogue box will appear.
4. Make sure "My table has headers" is selected so Excel recognizes the column headings.

5. Click **OK**. Excel turns our advance search data into a table.
6. Sort the table in ascending order (A-Z), by Store, and Employee ID, then Last Name. **Hint: Click the Data tab, Click the Sort button, add levels for the three fields.**
7. Autofit the column widths and row height to make sure the heading row is visible.
8. Save your work.

Excel turns the information into a table and sorts accordingly:



WCM Analytics
Advanced Filter Criteria
Seattle & San Diego Current Salaries

Store	Job Status	Current Salary	Current Salary
Seattle	FT	>=50000	<=60000
San Diego	FT	>=70000	<=80000

Employee ID	First Name	Last Name	Hire Date	Years of Servic	Birth Date	Age	Store	Job Statu	Current Salary	COLA	Projected Salary Increa
1350	Patrice	Hutton	12/29/2009	6.01	4/4/1953	63	San Diego	FT	$75,037	2.5%	$1,875.93
1310	Ruth	Fallis	1/22/2016	1.94	7/11/1953	63	Seattle	FT	$52,244	3.5%	$1,828.54
1394	Shanika	Lloyd	8/24/2004	13.36	8/11/1966	50	Seattle	FT	$53,186	3.5%	$1,861.51
3206	Larry	Roeder	1/29/2016	1.92	11/1/1982	34	Seattle	FT	$54,368	3.5%	$1,902.88
3214	Robert	Ross	8/28/2015	2.35	9/24/1974	42	Seattle	FT	$58,309	3.5%	$2,040.82
3242	Charles	Taylor	2/5/2012	3.90	1/18/1947	69	Seattle	FT	$58,291	3.5%	$2,040.19
3258	Johnny	Vazquez	9/18/2006	11.29	6/16/1953	63	Seattle	FT	$52,125	3.5%	$1,824.38

Figure 5.48 Advanced Filter Table

ANALYZING WORKSHEET DATA

INTRODUCTION TO PIVOT TABLES

Another way to analyze table information is with PivotTables. A PivotTable is a powerful tool that calculates, summarizes, and analyzes table data to compare, patterns, and trends. PivotTables are inserted directly from a table, linking the table data. Generally speaking, when you pivot on the table data you are reorganizing the table information to reveal different levels of detail that allow you to analyze specific subgroups of information and summarize data quickly and easily without having to change the structure or layout of the original table area.

When you pull table data into a PivotTable there are four main area fields: Rows, Columns, Values, and Filters. The Rows and Columns fields can interchange quickly to summarize the data in different ways or to run new reports based on the question or criteria being asked. The Value field is data from the table that can be calculated, or that contain values that the PivotTable will summarize. The Values field has multiple settings to choose how you want to calculate the data; SUM, COUNT, AVERAGE, MIN, MAX, and can even show the displayed values as a percentage of the total, column total, grand total, and so on. Lastly, is the Filters area, which restricts the PivotTable to only show the values matching specified criteria.

Four Primary PivotTable Areas:

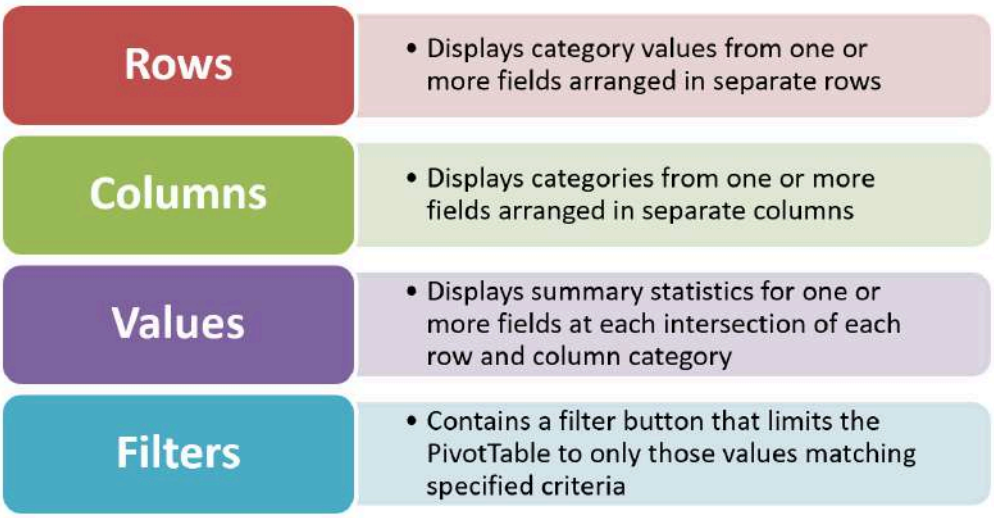


Figure 5.49 Four Primary PivotTable Areas

In our situation, shown below, we will create a PivotTable to summarize employee data to show Projected Salary Increases, for both Part-Time (PT) and Full -Time (FT) employees for all store locations.

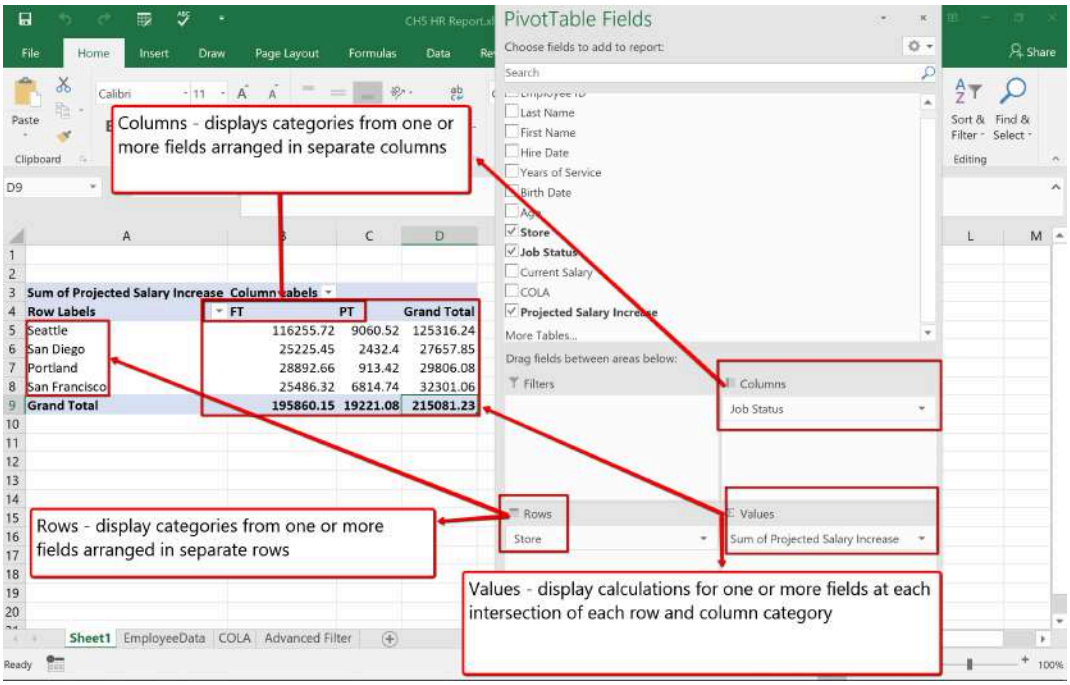


Figure 5.50 Parts Of A PivotTable

Follow the below steps to explore and build a PivotTable report.

1. Click the **EmployeeData** sheet. Click anywhere in the table area.
2. From the **Insert** tab, choose **PivotTable**.

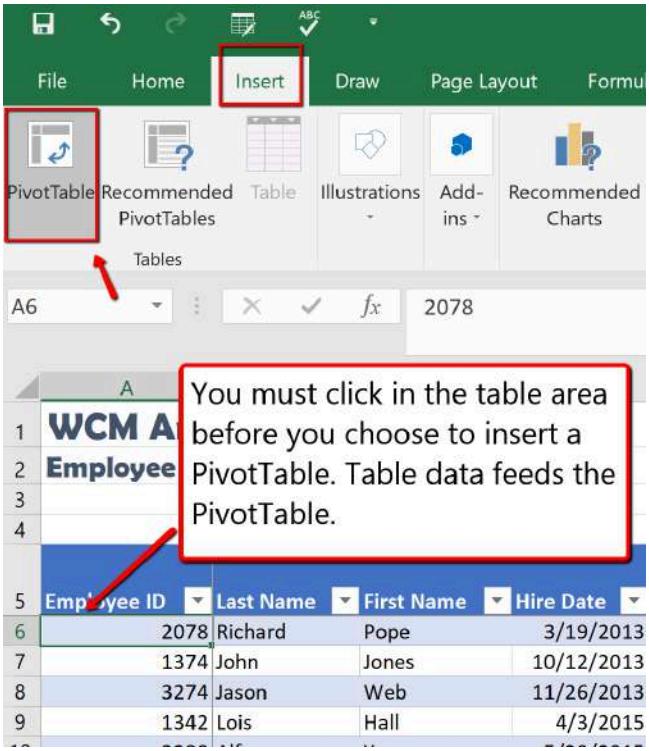


Figure 5.51 Insert PivotTable

3. From the Create PivotTable dialogue box, make sure the PivotTable report will be placed in a **New Worksheet**, and click **OK**.

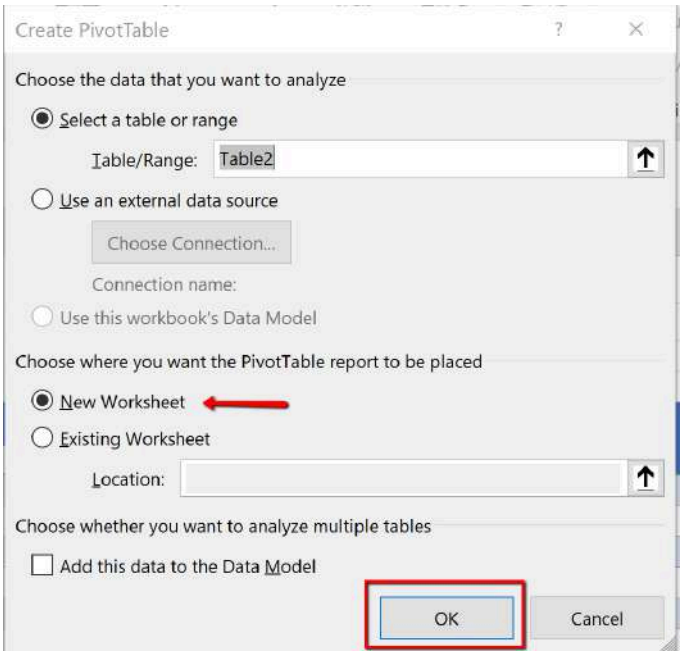


Figure 5.52 Create PivotTable Dialogue Box

4. Notice a new sheet (Sheet1) is inserted, at the bottom of the workbook, that contains the PivotTable1 area and fields dialogue box. Rename the default name (Sheet 1) to **StorePT**.

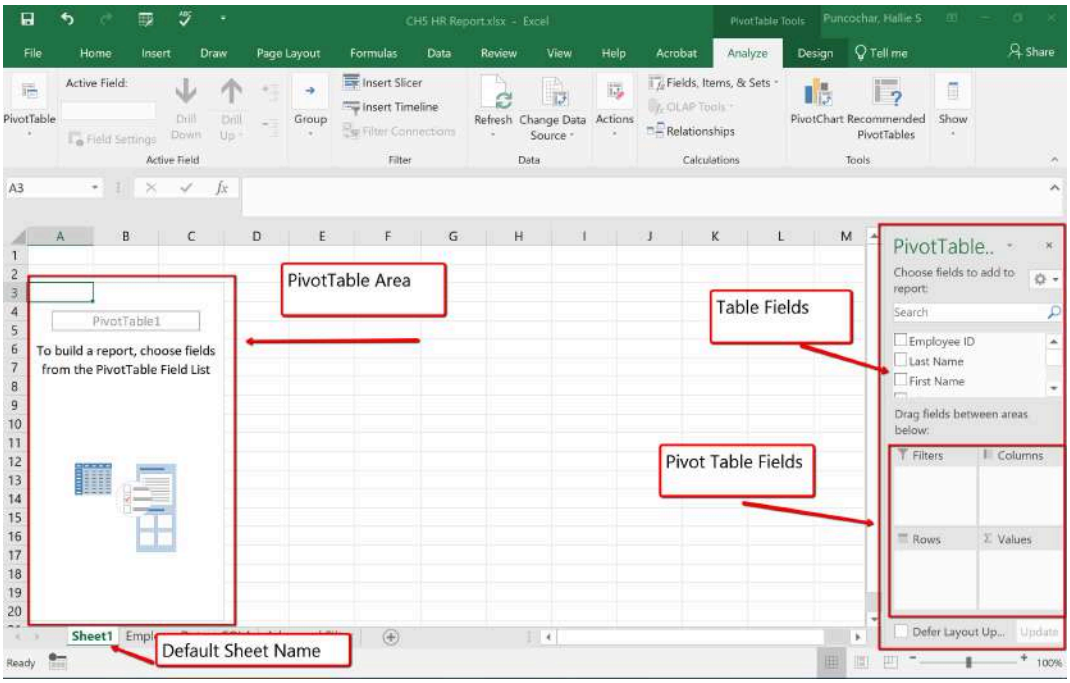


Figure 5.53 PivotTable Window

5. From the PivotTable pane, drag and drop the **Store** heading to the **Rows** section of PivotTable field area.

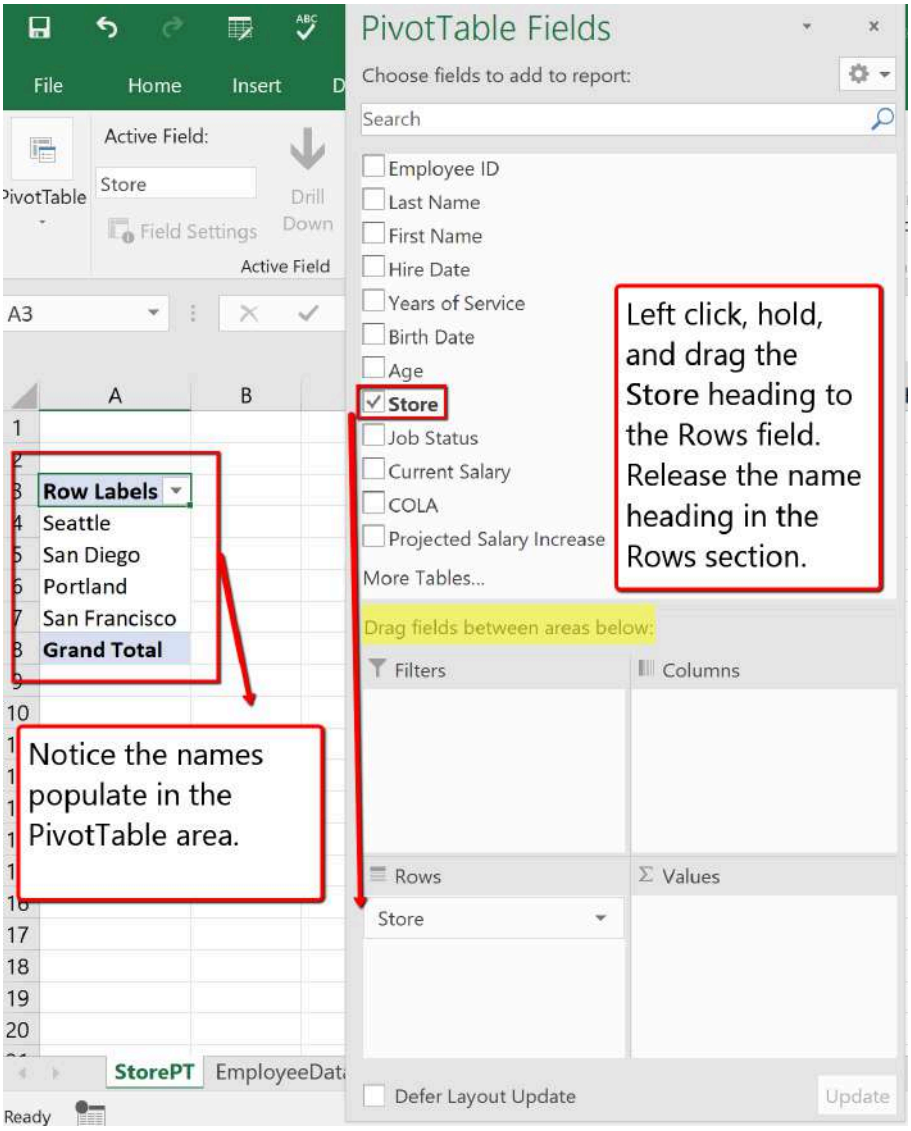


Figure 5.54
PivotTable Row Selection

6. From the PivotTable fields list drag and drop the **Projected Salary Increase** heading to the **Values** section.

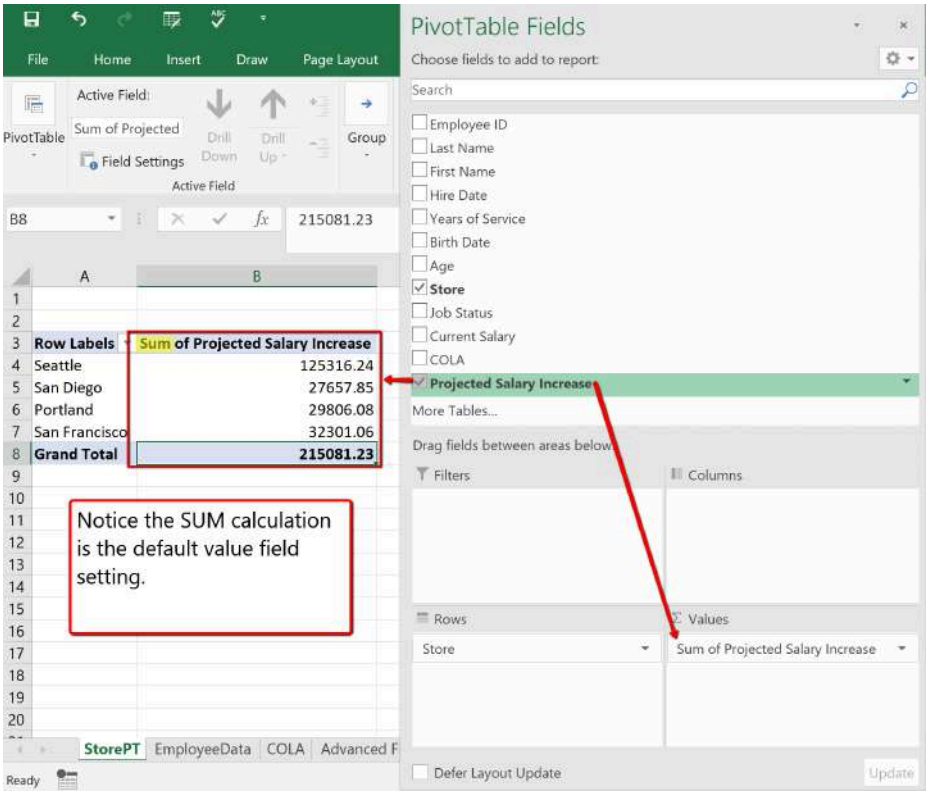


Figure 5.55
PivotTable Value Selection

7. Drag and Drop the **Job Status** heading to the **Columns** field section. Notice the Job Status categories display. In this case, displaying Full-Time (FT) and Part-Time (PT) employees.

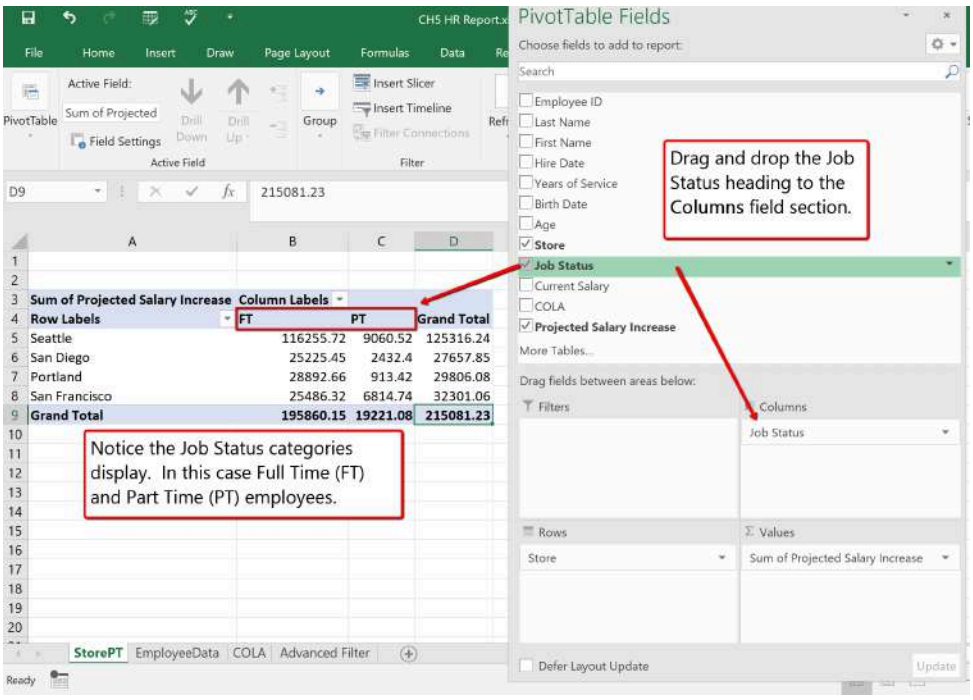


Figure 5.56
PivotTable Columns Selection

FORMATTING PIVOT TABLES

After creating a PivotTable and adding the fields that you want to analyze, you may want to enhance the report to include slicers, or graphs and or format the data to make it easier to read and scan for details. When clicked in the PivotTable area you will see a contextual tab appear on the ribbon, containing PivotTable Tools and two specific tabs; Analyze and Design. **Mac Users: there is not a "PivotTable Tools" tab but you will see two tabs named: PivotTable Analyze and Design. They are only visible when you have clicked inside the PivotTable area.**

The Analyze tab contains tools specifically for examining data, for example, the ability to insert Slicers, or PivotCharts. The Design tab contains tools that specifically tie to how the table and data visibly display. For example, when you have a lot of data in your PivotTable, it may help to show banded rows or columns for easy scanning or to highlight important data to make it stand out.

Follow the below steps to add format the PivotTable, and add a PivotChart.

1. Click in the PivotTable. From the PivotTable Tools choose the Design tab.
2. In the PivotTable Styles gallery select the Light Blue, Pivot Style Medium Style 2 format.

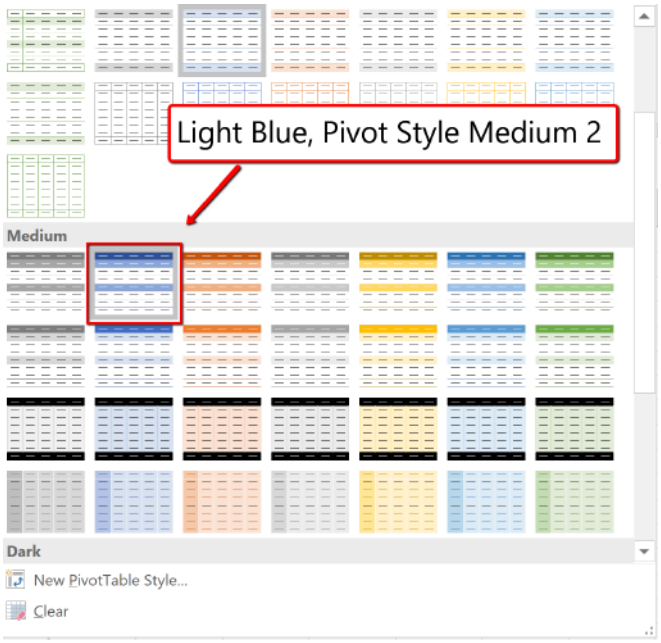


Figure 5.57 Light_Blue Medium Style Pivot 2

3. To format the PivotTable numbers, select B5: D9. Click the Home tab. Apply the Currency number format and decrease the decimal place to zero decimals.

(The alternative method to number formatting in a PivotTable is to expand the menu on value field; Sum of Projected Salary Increase. Click the Value Field Settings. Choose Number Format and apply the desired number

format option. *Mac Users should click the small circle with an "i" next to "Sum Projected Salary Increase" in the Values section then click the Number button to change the Number Format.* 

)

NOW LET'S CREATE A PIVOTCHART!

4. Click in the PivotTable. Click the Analyze tab. Choose the PivotChart button on the Ribbon.
5. From the listed chart types, choose Column. And select the 3D Clustered Column option. Click OK.

Mac Users: Only a basic, 2D column chart is available when clicking the **Pivot Chart** button. In order to select a different chart type, such as the 3D clustered column option, you must do the following:

- Click on the 2D chart that was just inserted
- Click the **Design tab** on the Ribbon
- Click the **Change Chart Type** button
- Select the 3D Clustered Column option

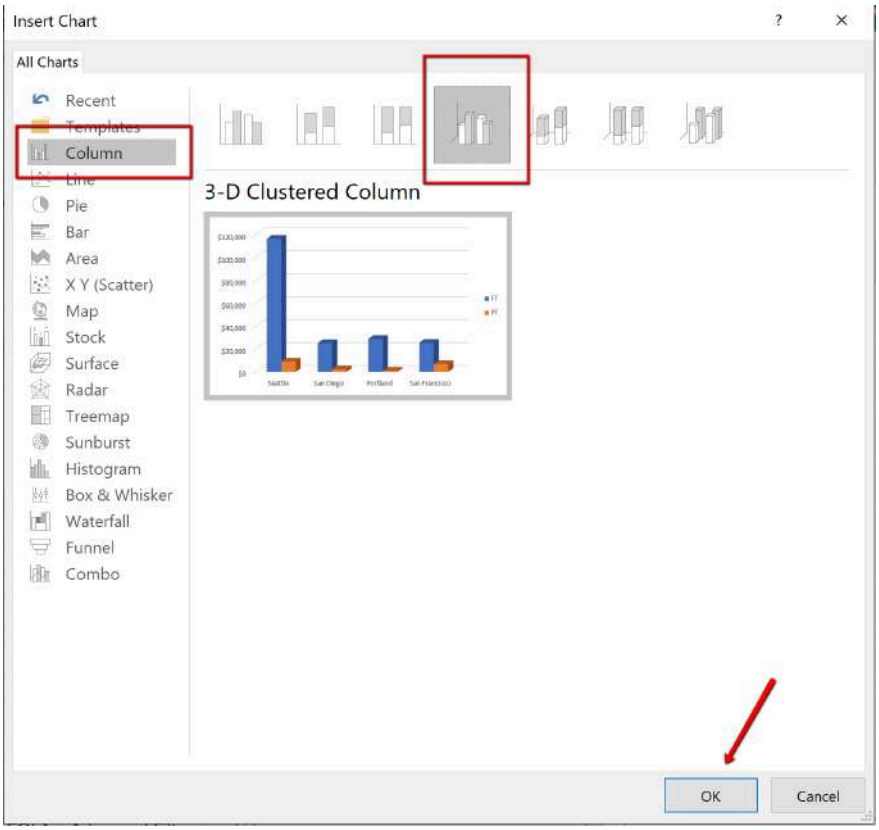


Figure 5.58 3D Clustered_Column Chart

6. Move the PivotChart under the PivotTable area. Resize accordingly. Save your work.

Note the formatting changes in the new chart below. The "Job Status" and "Store" buttons are column and row "filters" for the Pivot Chart.

Mac Users: Excel for Mac does not insert these formatting changes within a Pivot Chart. You can add a chart title by clicking the "Add Chart Element" button from the Design tab. It is **not possible** to add the "chart filter" buttons as shown in Figure 5.59. The filters on the pivot table can be used to also filter the columns and rows in the Pivot Chart.

Sum of Projected Salary Increase	Column Labels		
Row Labels	FT	PT	Grand Total
Seattle	$119,600	$9,061	$128,661
San Diego	$25,225	$2,432	$27,658
Portland	$30,866	$913	$31,779
San Francisco	$25,486	$6,815	$32,301
Grand Total	**$201,178**	**$19,221**	**$220,399**

Figure 5.59 PivotTable Solution

SUBTOTALS

Another way to summarize data is by using subtotals. Analyzing a large data range usually includes making calculations on the data. You can summarize the data by applying summary functions such as COUNT, SUM, and AVERAGE to the entire organized range of information. Subtotals, in general, are summary functions applied to parts of an organized data range.

For example, you can SUM Current Salaries for employees from each Store location. To subtotal the information the data must first be sorted by the Store field. For subtotals, the field that you sort is referred to as the control field. For example, if you choose the Store location as your control field, all of the Seattle, San

Diego, Portland, and San Francisco entries will be grouped together within the data range. The SUM function then can be applied to SUM the Current Salary fields for each Store location. Excel calculates and displays the subtotal each time the Store location changes.

A new row containing a subtotal of that particular location will be inserted, and wherever the field changes a value will display; a subtotal group of records. Excel updates the subtotal automatically when the control field is changed. In theory, when subtotaling, you are adding a calculation row to the set of data. Adding rows that total information in the middle of a table would compromise the integrity of data in the table. The table tools would look at the total as a record, not a calculation. Therefore the Subtotal feature cannot be used in tabling, and can only be applied to a normal range of data. You must convert all tables to a range prior to subtotaling.

Multiple functions can be applied within the same Subtotal. For example, we will explore how you can SUM Current Salary's and also provide the AVERAGE Current Salary for each Store location within the same Subtotal. Note Subtotal data can also be filtered.

The best practice when subtotaling is to follow four rules:

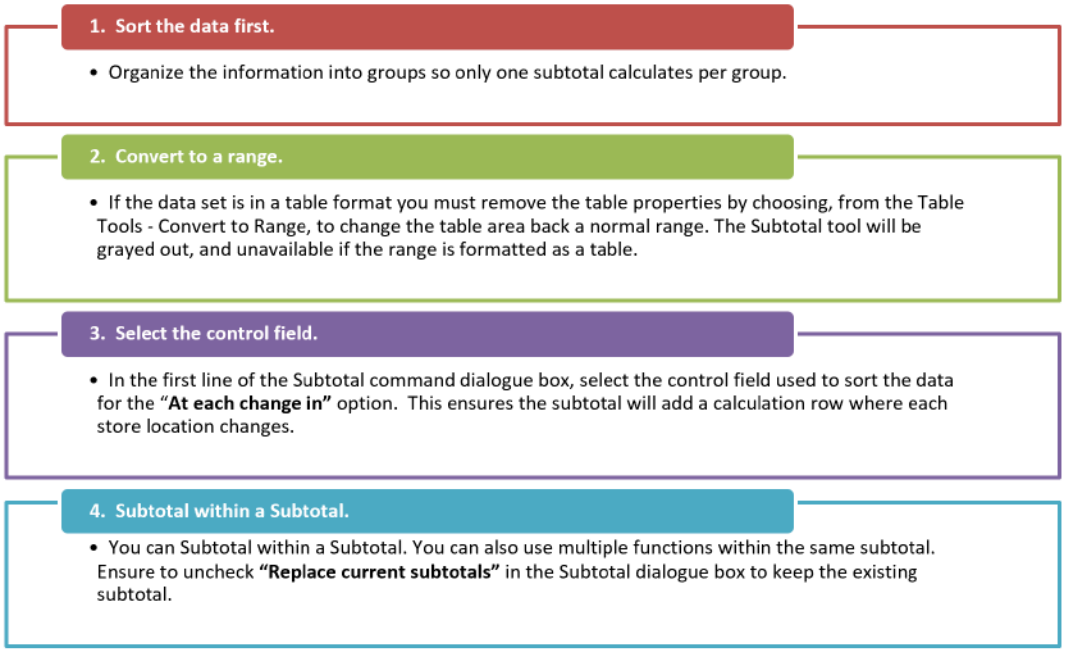


Figure 5.60
Subtotal Rules

Follow the below steps to Subtotal the Employee Data and provide a total Current Salary per Store.

1. Select the **Employee Data** sheet. If necessary clear any filters applied to the data by clicking the Data tab and choosing the Clear filter option.
2. From the Data tab, choose Sort button. Sort the Store Location, using the preferred Custom List order of Seattle, San Diego, Portland, and San Francisco. If the list we set up previously is not available type the

entries in the List entries area. Choose Add, and then OK.

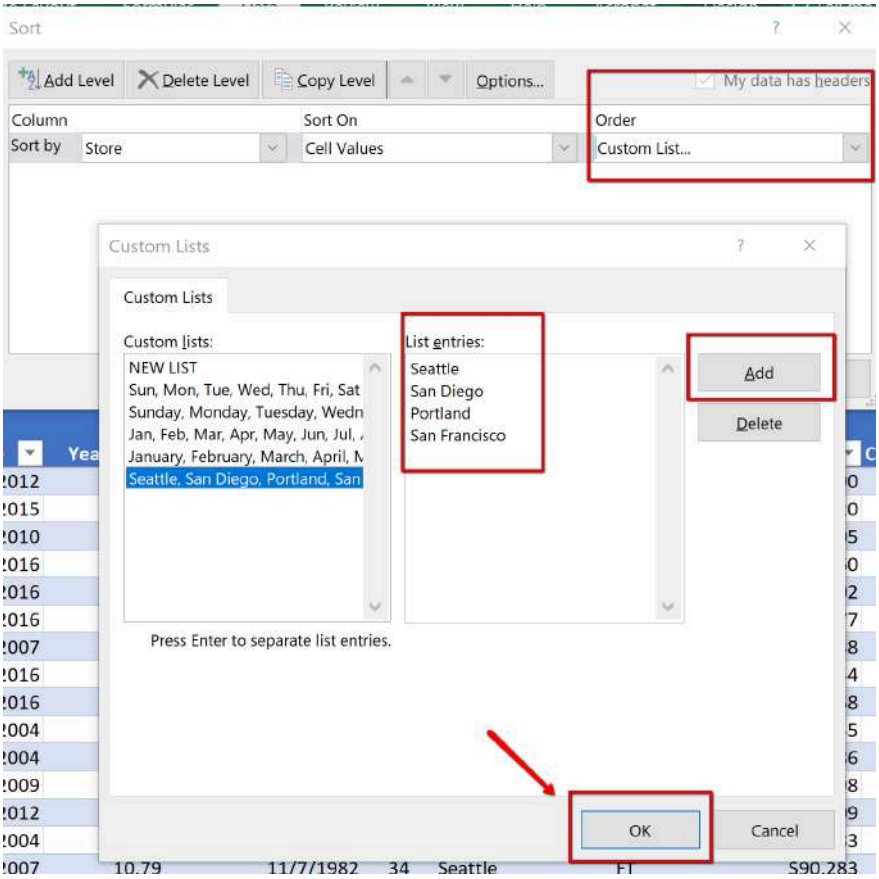


Figure 5.61 Custom Sort, Custom List Dialogue Box.

3. Choose the Table Tools Design tab. **Mac Users:** just click the **"Table" tab.**

Select "**Convert to Range**." Excel will display a message asking if you really want to convert the table back to a normal range. Choose **Yes.**

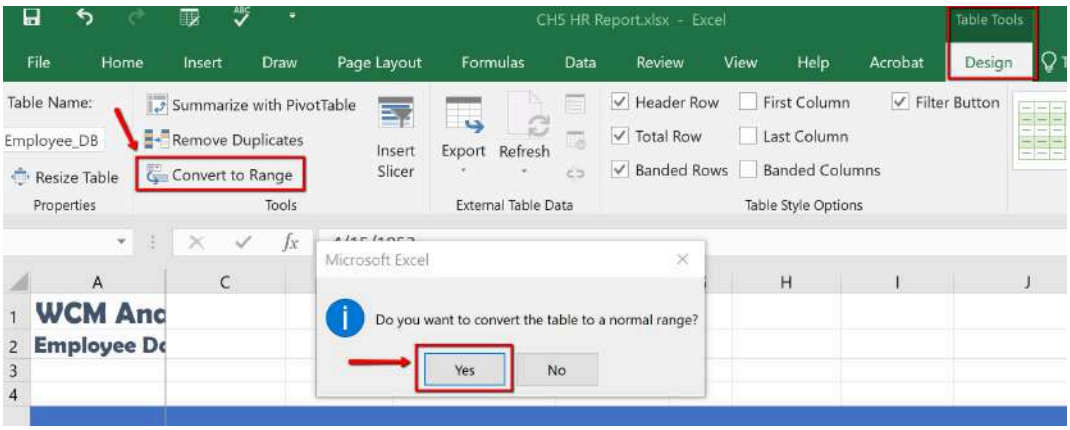


Figure 5.62 Convert To A Range

4. Click the Data tab, in the Outline group find and select the Subtotal Command. (Notice the heading row no longer has filters buttons. The data looks like a table but is not a table. The table tools are not active, and the information is a normal range.)

5. In the Subtotal dialogue box, choose the Store field in the "**At each change in**." For the "**Use Function**," choose **Sum**, and only check **Current Salary**. Click **OK.**

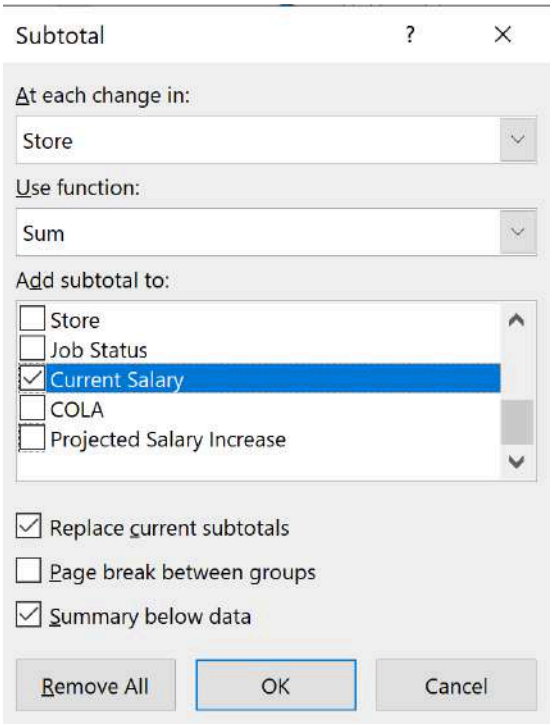
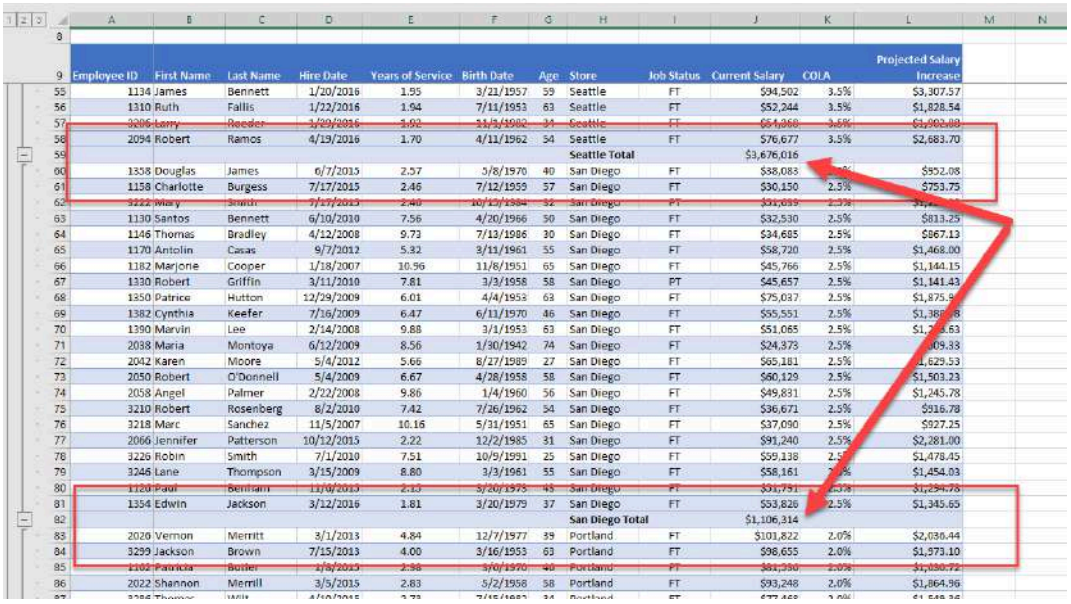


Figure 5.64
Subtotal Dialogue Box

6. Notice the Current Salary column is totaled, per location. Save your work.



Employee ID	First Name	Last Name	Hire Date	Years of Service	Birth Date	Age	Store	Job Status	Current Salary	COLA	Projected Salary Increase
1134	James	Bennett	1/20/2016	1.95	3/21/1957	59	Seattle	FT	$94,502	3.5%	$3,307.57
1310	Ruth	Fallis	1/22/2016	1.94	7/11/1953	63	Seattle	FT	$52,244	3.5%	$1,828.54
2094	Robert	Ramos	4/19/2016	1.70	4/11/1962	54	Seattle	FT	$76,677	3.5%	$2,683.70
							Seattle Total		$3,676,016		
1338	Douglas	James	6/7/2015	2.57	5/8/1970	40	San Diego	FT	$38,083	2.5%	$952.08
1158	Charlotte	Burgess	7/17/2015	2.46	7/12/1959	57	San Diego	FT	$30,150	2.5%	$753.75
1130	Santos	Bennett	6/10/2010	7.56	4/20/1966	50	San Diego	FT	$32,530	2.5%	$813.25
1146	Thomas	Bradley	4/12/2008	9.73	7/13/1986	30	San Diego	FT	$34,685	2.5%	$867.13
1170	Antolin	Casas	9/7/2012	5.32	3/11/1961	55	San Diego	FT	$58,720	2.5%	$1,468.00
1182	Marjorie	Cooper	1/18/2007	10.96	11/8/1951	65	San Diego	FT	$45,766	2.5%	$1,144.15
1330	Robert	Griffin	3/11/2010	7.81	3/3/1958	58	San Diego	PT	$45,657	2.5%	$1,141.43
1350	Patrice	Hutton	12/29/2009	6.01	4/4/1953	63	San Diego	FT	$75,037	2.5%	$1,875.9
1382	Cynthia	Keefer	7/16/2009	6.47	6/11/1970	46	San Diego	FT	$55,551	2.5%	$1,38
1390	Marvin	Lee	2/14/2008	9.88	3/1/1953	63	San Diego	FT	$51,065	2.5%	$1,2
2038	Maria	Montoya	6/12/2009	8.56	1/30/1942	74	San Diego	FT	$24,373	2.5%	09.33
2042	Karen	Moore	5/4/2012	5.66	8/27/1989	27	San Diego	FT	$65,181	2.5%	1,629.53
2050	Robert	O'Donnell	5/4/2009	6.67	4/28/1958	58	San Diego	FT	$60,129	2.5%	$1,503.23
2058	Angel	Palmer	2/22/2008	9.86	1/4/1960	56	San Diego	FT	$49,831	2.5%	$1,245.78
3210	Robert	Rosenberg	8/2/2010	7.42	7/26/1962	54	San Diego	FT	$36,671	2.5%	$916.78
3218	Marc	Sanchez	11/5/2007	10.16	5/31/1951	65	San Diego	FT	$37,090	2.5%	$927.25
2066	Jennifer	Patterson	10/12/2015	2.22	12/2/1985	31	San Diego	FT	$91,240	2.5%	$2,281.00
3226	Robin	Smith	7/1/2010	7.51	10/9/1991	25	San Diego	FT	$59,138	2.5%	$1,478.45
3246	Lane	Thompson	3/15/2009	8.80	3/3/1961	55	San Diego	FT	$58,161	2.5%	$1,454.03
1354	Edwin	Jackson	3/12/2016	1.81	3/20/1979	37	San Diego	FT	$53,826	2.5%	$1,345.65
							San Diego Total		$1,106,314		
2026	Vernon	Merritt	3/1/2013	4.84	12/7/1977	39	Portland	FT	$101,822	2.0%	$2,036.44
3299	Jackson	Brown	7/15/2013	4.00	3/16/1953	63	Portland	FT	$98,655	2.0%	$1,973.10
2022	Shannon	Merrill	3/5/2015	2.83	5/2/1958	58	Portland	FT	$93,248	2.0%	$1,864.96

Figure 5.65
Subtotal Solution

SUBTOTAL OUTLINE VIEW

The Outline views, located on the left side panel, show summary statistics. The Outline tool, with levels, allows you to control the expanse of detail displayed in the worksheet. The EmployeeData worksheet has three levels in the outline of its data range:

- Level 1, displays only the grand totals.
- Level 2, displays the total spent at each Store.
- Level 3, displays the total Salary.

Figure 5.66 above shows the Level 3 Outline, all the employee detail per store location. Clicking the outline buttons located to the left of the row numbers lets you choose how much detail you want to see in the worksheet. (Note that the three level numbers are at the top left side of the worksheet, just below the Name box.)

You will use the outline buttons to expand and collapse different sections of the data range.

1. Click level 1. Notice it displays the Grand Total.
2. Click level 2. Notice the totals for all store locations are displayed.

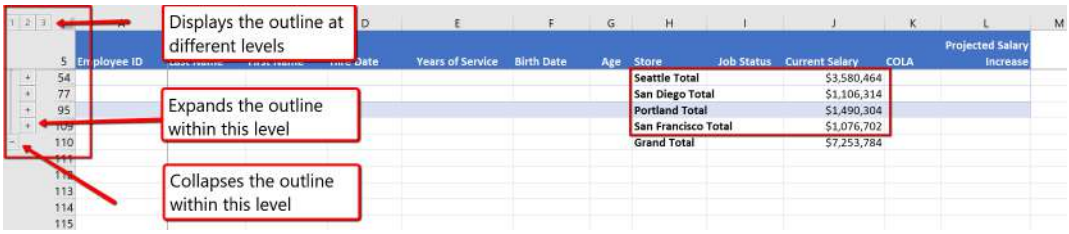


Figure 5.66
Subtotal Outline
Level 2

ADDING A SUBTOTAL WITHIN A SUBTOTAL

As mentioned at the beginning of the section, you can use multiple functions within the same subtotal. We will now explore how you can SUM Current Salary's and also provide the Average Current Salary for each Store location within the same Subtotal.

1. Click within the Subtotal data, go to the Outline, click Level 3, to display all the subtotal data.
2. From the Data tab, and click Subtotal.
3. In the Subtotal dialogue box, select the Store field for the "**At each change in:"** option.

4. In the "**Use function:"** section select to display the **Average**.
5. Only check the **Current Salary** field in the **"Add subtotal to section:"**. (Note Excel will default check something in this area. Uncheck any other fields.)
6. Uncheck **"Replace current subtotals"**; we do not want to replace the current subtotal summing the Current Salary.
7. Click OK

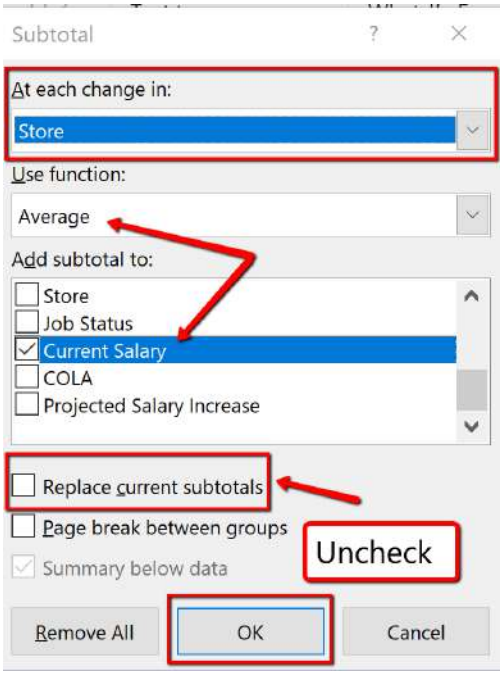


Figure 5.67
Subtotal Within A Subtotal

8. Notice each location is now subtotaled showing the Average and Total Current Salary. Excel has also added 4th level to the Outline, accounting for the Averages. Save your work.

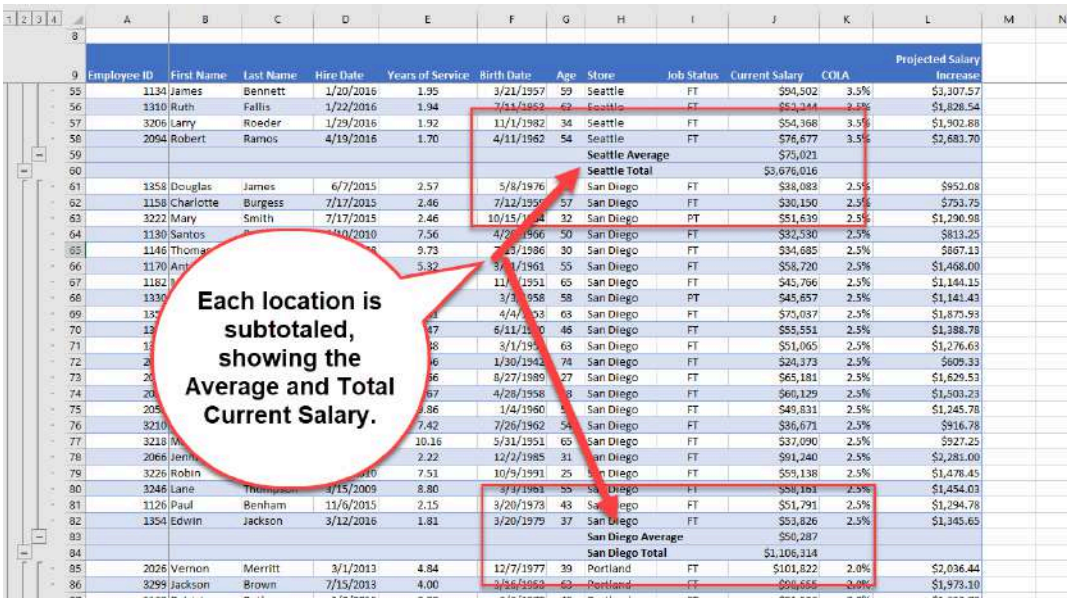


Figure 5.68
Solution, Subtotal Within A Subtotal

Key Takeaways

- A table is made up of a data set that is organized into columns and rows representing fields and records, such as employee information.
- You can create a table by clicking formatting the data set as a table, or using the Insert Table feature.
- Excel offers pre-built table styles, and options to choose from to format a table.
- You can add records (rows) and our fields (columns) to a table. You can then sort to reorganize your data.
- Freezing heading keeps your column headings displayed while you scroll through your table data.
- You can use the filter arrows in the table headings to sort by a single column. When sorting by more than one field, use the Custom Sort option.
- Custom List Sorts can be used when a field needs to be sorted in a special way.
- A slicer is a visual filter button (object) used to filter data in an Excel table. Each unique value in the field is a button.
- A PivotTable is an interactive table that summarizes data from a data source such as a data range or an Excel table.
- The Subtotal tool includes summary statistics for each group of records. Excel organizes subtotals using an outline that can be expanded or contracted to view or hide details about the data.

7.XLSX.3 PREPARING TO PRINT

Noreen Brown; Barbara Lave; Hallie Puncochar; Julie Romey; Mary Schatz; Art Schneider; Diane Shingledecker; and Jennifer Evans

Learning Objectives

1. Adjust page settings for printing.
2. Add footer information for user integrity.
3. Preview a worksheet, adjust settings to print in a professional manner.
4. Insert a 3D Model to enhance the visual appearance of a worksheet.

Previewing a Worksheet

Although printing large data sets is uncommon, it is an industry curiosity to set up Excel workbooks to print correctly, and to also add documentation as to when data was revised. Follow the below steps to prepare the worksheets to print.

1. Click on the **AdvancedFilter** worksheet. At the bottom of the screen choose the Page Layout option.

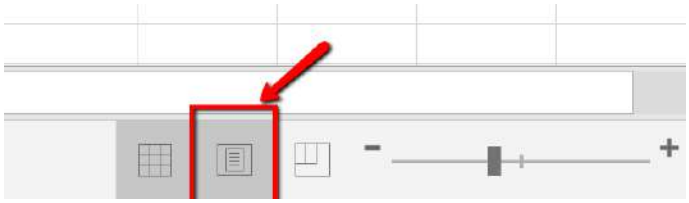


Figure 5.67 Page Layout View

2. At the bottom of the page, click into the left section, of the Add Footer panel.

Figure 5.68 Footer Area Left Section

3. From the Header and Footer Design tab, choose to insert the Current Date field.

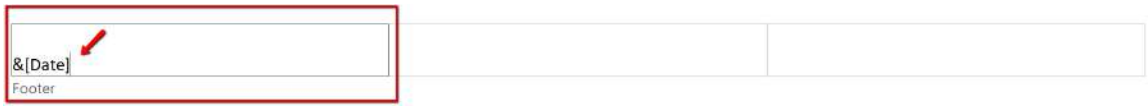


Figure 5.69 Date field in the Footer Area Left Section

4. Click in the right panel section, insert the File Name field.

Figure 5.70 File Name Footer Area Right Section

5. Click back into the spreadsheet to close the Header and Footer section, and choose the Normal page layout.

6. From the File tab, select Print. Change the Orientation to Landscape. In the Scaling section, choose **Fit Sheet on One Page.**

Mac Users: click the **"Scale to Fit" option**

7. Save your work. You don't have to actually print this sheet. Go back to your worksheet.

Follow the below steps to add a footer to indicate when the last update was made and apply settings to the **EmployeeData** worksheet to ensure it will print correctly if needed.

1. Click the EmployeeData worksheet. At the bottom of the screen choose the Page Layout option. You may get a message telling you that Page Layout and Freeze Panes are not compatible. You should click **OK** to remove the Freeze Panes setting.

2. At the bottom of the page, click into the left section, of the Add Footer panel type **Revision Date:** add a space, then click the **Current Date button from the Ribbon**. Example: Revision Date: 1/01/2020.

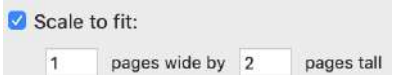
3. Click in the center panel, add the page number field.

4. Click in the right panel section, type **Revised by:** then type Y**our Name**. Example: Revised by: Jane Doe

5. Click back into the spreadsheet to close the Header and Footer section, and choose the Normal page layout.

6. From the File tab, select Print. Change the Margins to **Narrow**. In the Scaling section, choose **Fit All Columns on One Page.**

Mac Users: set the "**Scale to Fit**" option to **1 page wide by 2 pages tall**.



7. Save your work. Again, you do not have to print this sheet. Go back to the worksheet.

2054 Ruth	Olson	11/26/2004	13.10	10/4/1971	45	Seattle	FT	$64,845	3.5%	$2,269.58
2058 Angel	Palmer	2/22/2008	9.86	1/4/1960	56	San Diego	FT	$49,831	2.5%	$1,245.78
2062 Jose	Parham	12/4/2015	2.08	9/16/1970	46	Seattle	FT	$76,706	3.5%	$2,684.71
2066 Jennifer	Patterson	10/12/2015	2.22	12/2/1985	31	San Diego	FT	$91,240	2.5%	$2,281.00
2070 James	Pearce	12/4/2015	2.08	5/9/1959	57	Portland	FT	$96,282	2.0%	$1,925.64

Revision Date: (Type the Current Date) 1 Revised by: Student Name

Figure 5.65 Footer EmployeeData

INSERTING A 3D MODEL TO ENHANCE A WORKSHEET

Insert a 3D Model to the worksheet to enhance its appearance. In Excel, you can either insert Pictures, Shapes, Icons, SmartArt, Screenshots or 3D Models.

Figure 5.66 Illustrations Group

In this example, we will insert (from online) a 3D Model that looks like the Seattle Space Needle.

1. Click the Advanced Filter sheet tab, then click the Insert tab on the ribbon.
2. Click 3D Models button from the Illustrations group. (If necessary choose **From Online Sources or Stock 3D Models**.)

Mac Users: click the 3D Model icon button and then choose "**Stock 3D Models...**".

3. In the Search box type **Tower,** and hit **Enter** from the keyboard.
4. From the results window, choose a model that looks like the Space Needle. And click Insert. **Again, if the Space Needle is not available in the gallery, click the Back arrow and find an alternate building or tower from the 3D Model "Buildings" category.**

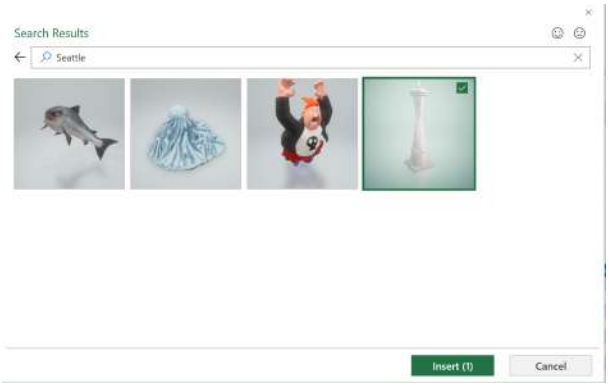


Figure 5.67 3D Model Search Box

5. Notice the model can be manipulated 360 degrees tilted up and down to show a specific feature of the object. Adjust based on your preference.

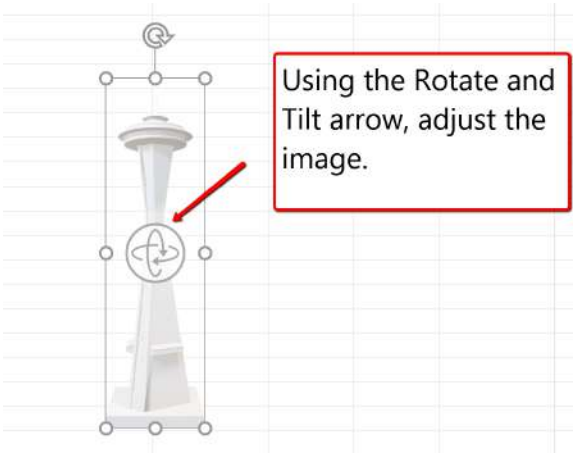
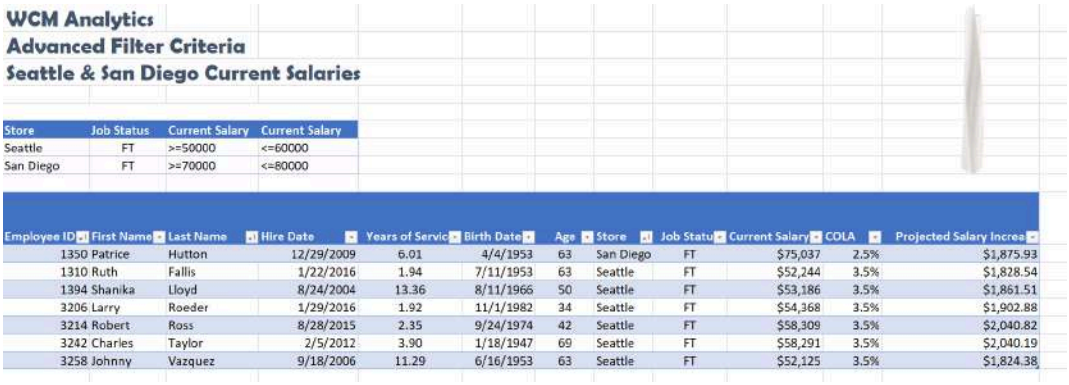


Figure 5.68 3D Model Image

6. Place, and resize the image to the upper left-hand corner of the sheet, above the last column of data. Make sure it does not overlap on the table.

7. Check the spelling on all of the worksheets and make any necessary changes. Save your work. Submit **CH5 HR Report** as directed by your instructor.



WCM Analytics
Advanced Filter Criteria
Seattle & San Diego Current Salaries

Store	Job Status	Current Salary	Current Salary
Seattle	FT	>=50000	<=60000
San Diego	FT	>=70000	<=80000

Employee ID	First Name	Last Name	Hire Date	Years of Servic	Birth Date	Age	Store	Job Statu	Current Salary	COLA	Projected Salary Increa
1350	Patrice	Hutton	12/29/2009	6.01	4/4/1953	63	San Diego	FT	$75,037	2.5%	$1,875.93
1310	Ruth	Fallis	1/22/2016	1.94	7/11/1953	63	Seattle	FT	$52,244	3.5%	$1,828.54
1394	Shanika	Lloyd	8/24/2004	13.36	8/11/1966	50	Seattle	FT	$53,186	3.5%	$1,861.51
3206	Larry	Roeder	1/29/2016	1.92	11/1/1982	34	Seattle	FT	$54,368	3.5%	$1,902.88
3214	Robert	Ross	8/28/2015	2.35	9/24/1974	42	Seattle	FT	$58,309	3.5%	$2,040.82
3242	Charles	Taylor	2/5/2012	3.90	1/18/1947	69	Seattle	FT	$58,291	3.5%	$2,040.19
3258	Johnny	Vazquez	9/18/2006	11.29	6/16/1953	63	Seattle	FT	$52,125	3.5%	$1,824.38

Figure 5.69 3D Model Solution

Key Takeaways

- When working with Excel workbooks, the final step should always be to review the worksheets in Print Preview to make sure they are printing appropriately.
- You can add images you have saved, or images you find online, to a worksheet to enhance its appearance. Be sure to resize and move them appropriately so they do not detract from the data.

Attribution

7.XLSX.4 CHAPTER PRACTICE

Noreen Brown; Barbara Lave; Hallie Puncochar; Julie Romey; Mary Schatz; Art Schneider; Diane Shingledecker; and Jennifer Evans

---

Tables for a Tourism Company

Download Data File: PR5 Data

Travel and tour companies need to keep track of client data, as well as, travel/tour options and tour guides. Keeping up-to-date, accurate records is essential to their bottom line. To run a tour company, employees must be able to manipulate their data quickly and easily. This exercise illustrates how to use the skills presented in this chapter to generate the data needed on a daily basis by a tourism company.

1. Open the data file **PR5 Data** and save the file to your computer as **PR5 Canyon Trails**.
2. Click Sheet 1. Choose cell B3.
3. From the Home tab, choose Format as Table. Choose the Orange, Table Style Medium 3.
4. In J4, calculate Total Cost (number of Guests *Per Person Cost). Note Excel will add the formula to the entire column. (If prompted, choose to overwrite the formula to the cells below.)
5. Format Columns I and J with Accounting format, no decimal places.
6. Center all headings in Row 3.
7. Adjust column widths within the table so that all the headings are completely visible.
8. Rename Sheet 1 **Current Tours**. Sort this sheet alphabetically (A to Z) by Last Name.

Tours By Canyon

First Name	Last Name	Guests	Average Age	Home Country	Tour Canyon	Tour State	Days	Per Person Cost	Total Cost
Santos	Albert	2	58	Brazil	Grand Canyon National Park	Arizona	5	$ 550	$ 1,100
Ian	Armstrong	2	62	United States	Cedar Breaks National Monument	Utah	5	$ 550	$ 1,100
Laurie	Black	2	66	Canada	Fall Canyon Death Valley National Park	California	7	$ 900	$ 1,800
Richard	Branson	2	65	United Kingdom	Zion National Park	Utah	7	$ 900	$ 1,800
Vanessa	Burleigh	4	30	United Kingdom	Grand Canyon National Park	Arizona	7	$ 900	$ 3,600
Jim	Carrey	1	54	Canada	Canyon de Chelly National Monument	Arizona	5	$ 550	$ 550
Jennifer	Connelly	2	45	United States	Canyon de Chelly National Monument	Arizona	5	$ 550	$ 1,100
Ellen	Cronan	2	68	Canada	Bryce Canyon National Park	Utah	7	$ 900	$ 1,800
James	Doug	2	50	United Kingdom	Zion National Park	Utah	7	$ 900	$ 1,800
Marianne	Eliot	2	50	United States	Yellowstone National Park	Wyoming	5	$ 550	$ 1,100
Jonas	Finamore	2	45	United Kingdom	Yellowstone National Park	Wyoming	3	$ 450	$ 900
James	Gardipee	1	46	United States	Grand Canyon National Park	Arizona	5	$ 550	$ 550
Sharon	Glatz	2	63	Australia	Cedar Breaks National Monument	Utah	3	$ 450	$ 900
Sofia	Guthenberg	1	60	Canada	Glen Canyon National Recreation Area	Arizona	3	$ 450	$ 450
Yoko	Hanamoto	2	29	Japan	Black Canyon of the Gunnison National Park	Colorado	7	$ 900	$ 1,800
Charles	Hector	2	56	United States	Zion National Park	Utah	7	$ 900	$ 1,800
Scarlett	Johansson	3	31	United States	Fall Canyon Death Valley National Park	California	7	$ 900	$ 2,700
Sally	Kee	4	57	Canada	Glen Canyon National Recreation Area	Arizona	3	$ 450	$ 1,800
Frank	Kee	3	58	United States	Zion National Park	Utah	7	$ 900	$ 2,700
Rosella	Kim	2	40	United States	Zion National Park	Utah	7	$ 900	$ 1,800
Deborah	Klein	2	65	Germany	Fall Canyon Death Valley National Park	California	5	$ 550	$ 1,100
Nick	Kopec	2	65	United Kingdom	Black Canyon of the Gunnison National Park	Colorado	5	$ 550	$ 1,100
Grace	Kruger	2	56	Germany	Black Canyon of the Gunnison National Park	Colorado	7	$ 900	$ 1,800
Erik	Laarson	2	63	Canada	Black Canyon of the Gunnison National Park	Colorado	5	$ 550	$ 1,100
Samuel	Larocca	2	59	Brazil	Cedar Breaks National Monument	Utah	7	$ 900	$ 1,800

Figure 5.70 Current Tours

9. Make a copy of the Current Tours sheet and rename it **Tours by Canyon**. One way to make a copy of a worksheet is to right-click on the worksheet tab (**Mac Users**: Ctrl+click) and select Move or Copy. Be sure to check the **Create a Copy** box. Place the Tours by Canyon sheet to the right of the Current Tours sheet.

10. Sort the Tours by Canyon sheet by **Tour Canyon, Home Country,** and then **Last Name** all in Ascending order (A to Z).

Tours By Canyon

First Name	Last Name	Guests	Average Age	Home Country	Tour Canyon	Tour State	Days	Per Person Cost	Total Cost
Erik	Laarson	2	63	Canada	Black Canyon of the Gunnison National Park	Colorado	5	550	$ 1,100
Grace	Kruger	2	56	Germany	Black Canyon of the Gunnison National Park	Colorado	7	900	$ 1,800
Ingrid	Schultz	2	57	Germany	Black Canyon of the Gunnison National Park	Colorado	7	900	$ 1,800
Yoko	Hanamoto	2	29	Japan	Black Canyon of the Gunnison National Park	Colorado	7	900	$ 1,800
Nick	Kopec	2	65	United Kingdom	Black Canyon of the Gunnison National Park	Colorado	5	550	$ 1,100
Analise	Wendle	2	58	United States	Black Canyon of the Gunnison National Park	Colorado	7	900	$ 1,800
Brian	Lawson	2	66	Australia	Bryce Canyon National Park	Utah	5	550	$ 1,100
Ellen	Cronan	2	68	Canada	Bryce Canyon National Park	Utah	7	900	$ 1,800
Miguel	Piggott	2	70	United Kingdom	Bryce Canyon National Park	Utah	7	900	$ 1,800
Skye	Quillin	2	45	United States	Bryce Canyon National Park	Utah	7	900	$ 1,800
Gil	Thompson	1	62	United States	Bryce Canyon National Park	Utah	7	900	$ 900
Darlene	Welsh	2	63	United States	Bryce Canyon National Park	Utah	5	550	$ 1,100
Jim	Carrey	1	54	Canada	Canyon de Chelly National Monument	Arizona	5	550	$ 550
Jennifer	Connelly	2	45	United States	Canyon de Chelly National Monument	Arizona	5	550	$ 1,100
Alfred	Yankovic	2	56	United States	Canyon de Chelly National Monument	Arizona	7	900	$ 1,800
Sharon	Glatz	2	63	Australia	Cedar Breaks National Monument	Utah	3	450	$ 900
Samuel	Larocca	2	59	Brazil	Cedar Breaks National Monument	Utah	7	900	$ 1,800
John	Lawrence	2	52	Canada	Cedar Breaks National Monument	Utah	7	900	$ 1,800
Omar	Rafael	2	59	United Kingdom	Cedar Breaks National Monument	Utah	3	450	$ 900
Ian	Armstrong	2	62	United States	Cedar Breaks National Monument	Utah	5	550	$ 1,100
Laurie	Black	2	66	Canada	Fall Canyon Death Valley National Park	California	7	900	$ 1,800
Jolene	Terry	2	67	Canada	Fall Canyon Death Valley National Park	California	7	900	$ 1,800
Deborah	Klein	2	65	Germany	Fall Canyon Death Valley National Park	California	5	550	$ 1,100
Scarlett	Johansson	3	31	United States	Fall Canyon Death Valley National Park	California	7	900	$ 2,700
Sofia	Guthenberg	1	60	Canada	Glen Canyon National Recreation Area	Arizona	3	450	$ 450
Sally	Kee	4	57	Canada	Glen Canyon National Recreation Area	Arizona	3	450	$ 1,800

Figure 5.71 Tours by Canyon

11. Make another copy of the Current Tours sheet and rename it **US Guests**. Place the US Guests sheet to the right of the Tours by Canyon sheet.

12. Filter the **US Guests** sheet to display customers who live in the United States. Sort the filtered data alphabetically (A to Z) by Tour State. Add a Total Row that sums the Guests and Total Cost columns.

Tours By Canyon

First Name	Last Name	Guests	Average Age	Home Country	Tour Canyon	Tour State	Days	Per Person Cost	Total Cost
Jennifer	Connelly	2	45	United States	Canyon de Chelly National Monument	Arizona	5	550	$ 1,100
James	Gardipee	1	46	United States	Grand Canyon National Park	Arizona	5	550	$ 550
Vince	Shad	3	69	United States	Grand Canyon National Park	Arizona	7	900	$ 2,700
Rod	Vanderzee	2	58	United States	Grand Canyon National Park	Arizona	7	900	$ 1,800
Alex	Wigham	2	70	United States	Grand Canyon National Park	Arizona	7	900	$ 1,800
Alfred	Yankovic	2	56	United States	Canyon de Chelly National Monument	Arizona	7	900	$ 1,800
Scarlett	Johansson	3	31	United States	Fall Canyon Death Valley National Park	California	7	900	$ 2,700
Analise	Wendle	2	58	United States	Black Canyon of the Gunnison National Park	Colorado	7	900	$ 1,800
Ian	Armstrong	2	62	United States	Cedar Breaks National Monument	Utah	5	550	$ 1,100
Charles	Hector	2	56	United States	Zion National Park	Utah	7	900	$ 1,800
Frank	Kee	3	58	United States	Zion National Park	Utah	7	900	$ 2,700
Rosella	Kim	2	40	United States	Zion National Park	Utah	7	900	$ 1,800
Skye	Quillin	2	45	United States	Bryce Canyon National Park	Utah	7	900	$ 1,800
Gil	Thompson	1	62	United States	Bryce Canyon National Park	Utah	7	900	$ 900
Darlene	Welsh	2	63	United States	Bryce Canyon National Park	Utah	5	550	$ 1,100
Marianne	Eliot	2	50	United States	Yellowstone National Park	Wyoming	5	550	$ 1,100
Lucas	Lee	2	42	United States	Yellowstone National Park	Wyoming	3	450	$ 900
Sierra	Sloane	2	33	United States	Yellowstone National Park	Wyoming	7	900	$ 1,800
Total		37							$ 29,250

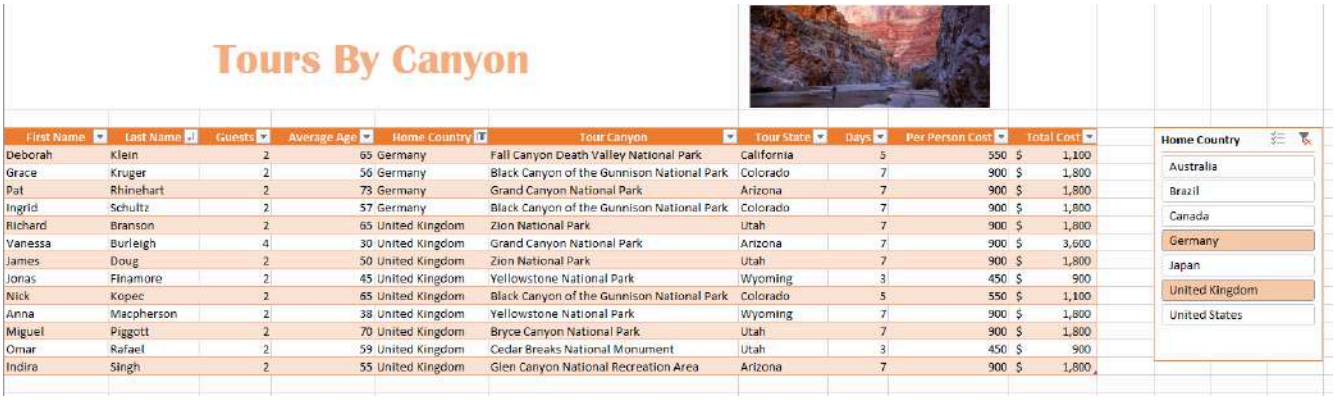
Figure 5.72 US Guests

13. Make another copy of the Current Tours sheet and rename it, **European Guests**. Place the European Guests sheet to the right of the US Guests sheet.

14. Insert a slicer in the **European Guests** sheet for Home Country. Move the top left corner of the slicer to the top left-hand corner of cell L3. Resize the slicer so all buttons display. Format the slicer to match the table.

15. Using the slicer, filter the data to display customers from Germany and the United Kingdom.

16. Sort the filtered data by the Home Country, and Last Name fields displaying both in Ascending order (A to Z).



Tours By Canyon

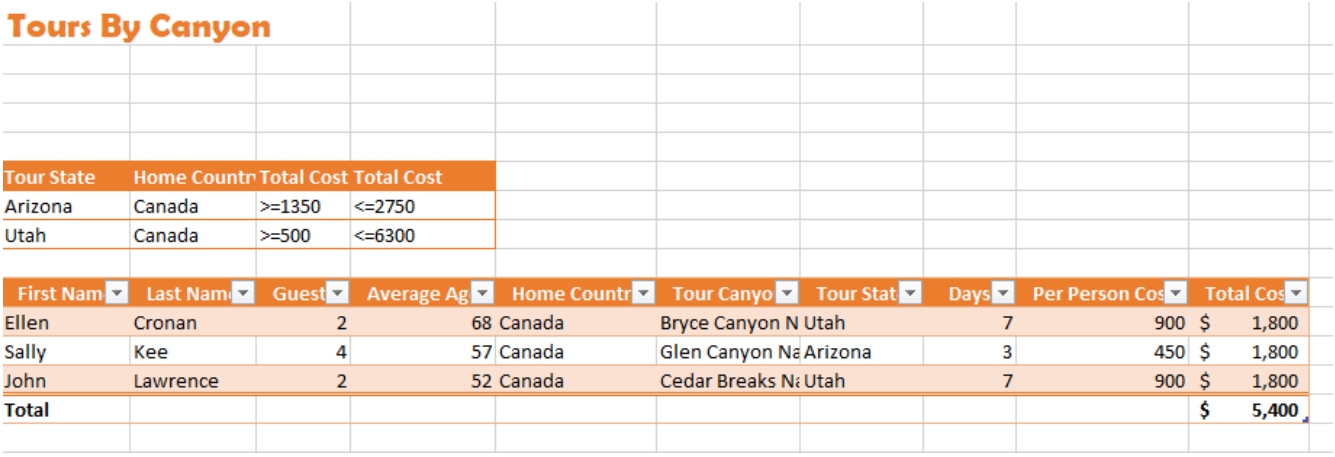
First Name	Last Name	Guests	Average Age	Home Country	Tour Canyon	Tour State	Days	Per Person Cost	Total Cost
Deborah	Klein	2	65	Germany	Fall Canyon Death Valley National Park	California	5	550	$ 1,100
Grace	Kruger	2	56	Germany	Black Canyon of the Gunnison National Park	Colorado	7	900	$ 1,800
Pat	Rhinehart	2	73	Germany	Grand Canyon National Park	Arizona	7	900	$ 1,800
Ingrid	Schultz	2	57	Germany	Black Canyon of the Gunnison National Park	Colorado	7	900	$ 1,800
Richard	Branson	2	65	United Kingdom	Zion National Park	Utah	7	900	$ 1,800
Vanessa	Burleigh	4	30	United Kingdom	Grand Canyon National Park	Arizona	7	900	$ 3,600
James	Doug	2	50	United Kingdom	Zion National Park	Utah	7	900	$ 1,800
Jonas	Finamore	2	45	United Kingdom	Yellowstone National Park	Wyoming	3	450	$ 900
Nick	Kopec	2	65	United Kingdom	Black Canyon of the Gunnison National Park	Colorado	5	550	$ 1,100
Anna	Macpherson	2	38	United Kingdom	Yellowstone National Park	Wyoming	7	900	$ 1,800
Miguel	Piggott	2	70	United Kingdom	Bryce Canyon National Park	Utah	7	900	$ 1,800
Omar	Rafael	2	59	United Kingdom	Cedar Breaks National Monument	Utah	3	450	$ 900
Indira	Singh	2	55	United Kingdom	Glen Canyon National Recreation Area	Arizona	7	900	$ 1,800

Home Country
Australia
Brazil
Canada
Germany
Japan
United Kingdom
United States

Figure 5.73 European Guests

17. Click the Advanced Filter sheet. Using the Advanced Filter option, filter the Current Tours table based on the criteria given. Determine how many guests from Canada are taking tours in Arizona and Utah between the costs indicated in the criteria table. Place the results in A10.

18. Turn the results into a table. Format the table to match the criteria area. Turn on the total row and show the Sum of the Total Cost column.



Tours By Canyon

Tour State	Home Countr	Total Cost	Total Cost
Arizona	Canada	>=1350	<=2750
Utah	Canada	>=500	<=6300

First Nam	Last Nam	Guest	Average Ag	Home Countr	Tour Canyo	Tour Stat	Days	Per Person Cos	Total Cos
Ellen	Cronan	2	68	Canada	Bryce Canyon N	Utah	7	900	$ 1,800
Sally	Kee	4	57	Canada	Glen Canyon Na	Arizona	3	450	$ 1,800
John	Lawrence	2	52	Canada	Cedar Breaks Na	Utah	7	900	$ 1,800
Total									**$ 5,400**

Figure 5.74 Advanced Filter

19. Select the Current Tours sheet. Click in the table area and insert a PivotTable as a new sheet. Name the sheet **ToursPT.** Run a report to show the Total Cost per Home Country, for each available Tour States. Format the numbers in currency format, zero decimal places. Choose a PivotStyle format to match the current orange theme.

Sum of Total Cost	Column Labels					
Row Labels	Arizona	California	Colorado	Utah	Wyoming	Grand Total
Australia	$1,800			$2,000	$2,200	$6,000
Brazil	$2,900			$1,800		$4,700
Canada	$3,900	$3,600	$1,100	$3,600	$2,200	$14,400
Germany	$1,800	$1,100	$3,600			$6,500
Japan			$1,800	$1,800		$3,600
United Kingdom	$5,400		$1,100	$6,300	$2,700	$15,500
United States	$9,750	$2,700	$1,800	$11,200	$3,800	$29,250
Grand Total	**$25,550**	**$7,400**	**$9,400**	**$26,700**	**$10,900**	**$79,950**

Figure 5.75 ToursPT

20. Make one more copy of the Current Tours sheet and rename it **Tours by State**. Place the Tours by State sheet to the right of the European Guests sheet. Go to the Table Tools and turn off the Banded Rows.

21. Subtotal the data by State, summing the Total Cost column. (**Note:** Remember to follow the four rules of subtotaling!)

22. After you subtotal, turn on filters and filter out 3-day tours in the table.

Tours By Canyon

First Name	Last Name	Guests	Average Age	Home Country	Tour Canyon	Tour State	Days	Per Person Cost	Total Cost
Santos	Albert	2	58	Brazil	Grand Canyon National Park	Arizona	5	550	$ 1,100
Vanessa	Burleigh	4	30	United Kingdom	Grand Canyon National Park	Arizona	7	900	$ 3,600
Jim	Carrey	1	54	Canada	Canyon de Chelly National Monument	Arizona	5	550	$ 550
Jennifer	Connelly	2	45	United States	Canyon de Chelly National Monument	Arizona	5	550	$ 1,100
James	Gardipee	1	46	United States	Grand Canyon National Park	Arizona	5	550	$ 550
Leonora	Maag	2	45	Brazil	Grand Canyon National Park	Arizona	7	900	$ 1,800
Raymond	Mah	2	55	Canada	Grand Canyon National Park	Arizona	5	550	$ 1,100
Pat	Rhinehart	2	73	Germany	Grand Canyon National Park	Arizona	7	900	$ 1,800
Vince	Shad	3	69	United States	Grand Canyon National Park	Arizona	7	900	$ 2,700
Indira	Singh	2	55	United Kingdom	Glen Canyon National Recreation Area	Arizona	7	900	$ 1,800
Rod	Vanderzee	2	58	United States	Grand Canyon National Park	Arizona	7	900	$ 1,800
Alex	Wigham	2	70	United States	Grand Canyon National Park	Arizona	7	900	$ 1,800
Alfred	Yankovic	2	56	United States	Canyon de Chelly National Monument	Arizona	7	900	$ 1,800
						Arizona Total			$ 21,500
Laurie	Black	2	66	Canada	Fall Canyon Death Valley National Park	California	7	900	$ 1,800
Scarlett	Johansson	3	31	United States	Fall Canyon Death Valley National Park	California	7	900	$ 2,700
Deborah	Klein	2	65	Germany	Fall Canyon Death Valley National Park	California	5	550	$ 1,100
Jolene	Terry	2	67	Canada	Fall Canyon Death Valley National Park	California	7	900	$ 1,800
						California Total			$ 7,400
Yoko	Hanamoto	2	29	Japan	Black Canyon of the Gunnison National Park	Colorado	7	900	$ 1,800
Nick	Kopec	2	65	United Kingdom	Black Canyon of the Gunnison National Park	Colorado	5	550	$ 1,100
Grace	Kruger	2	56	Germany	Black Canyon of the Gunnison National Park	Colorado	7	900	$ 1,800
Erik	Laarson	2	63	Canada	Black Canyon of the Gunnison National Park	Colorado	5	550	$ 1,100
Ingrid	Schultz	2	57	Germany	Black Canyon of the Gunnison National Park	Colorado	7	900	$ 1,800
Analise	Wendle	2	58	United States	Black Canyon of the Gunnison National Park	Colorado	7	900	$ 1,800
						Colorado Total			$ 9,400

Figure 5.76 Subtotal

23. On each worksheet, make the following print setup changes:

    a) Add a footer with the **current date, worksheet name**, and your name.

    b) Change to Landscape Orientation

    c) Set the scaling to Fit All Columns on One Page

    d) For any worksheets that print on more than one page, add **Print Titles** to repeat the first three rows at the top of each page.

24. Check the spelling on all of the worksheets and make any necessary changes. Save the **PR5 Canyon Trails** workbook. Submit the **PR5 Canyon Trails** workbook as directed by your instructor.

Attribution

SECTION XIII

8. PROBABILITY

8.1 What Are the Chances?
8.2 Probability Rules
8.3 More About Chance

8.1 WHAT ARE THE CHANCES?

8.1: What Are the Chances?

8.1.1: Fundamentals of Probability

Probability is the branch of mathematics that deals with the likelihood that certain outcomes will occur. There are five basic rules, or axioms, that one must understand while studying the fundamentals of probability.

Learning Objective

Explain the most basic and most important rules in determining the probability of an event

Key Takeaways

Key Points

- Probability is a number that can be assigned to outcomes and events. It always is greater than or equal to zero, and less than or equal to one.
- The sum of the probabilities of all outcomes must equal 1">1.
- If two events have no outcomes in common, the probability that one or the other occurs is the sum of their individual probabilities.

- The probability that an event does not occur is 1">1 minus the probability that the event does occur.
- Two events A">A and B">B are independent if knowing that one occurs does not change the probability that the other occurs.

Key Terms

experiment
: Something that is done that produces measurable results, called outcomes.

outcome
: One of the individual results that can occur in an experiment.

event
: A subset of the sample space.

sample space
: The set of all outcomes of an experiment.

In discrete probability, we assume a well-defined experiment, such as flipping a coin or rolling a die. Each individual result which could occur is called an outcome. The set of all outcomes is called the sample space, and any subset of the sample space is called an event.

For example, consider the experiment of flipping a coin two times. There are four individual outcomes, namely HH,HT,TH,TT">HH,HT,TH,TT. The sample space is thus {HH,HT,TH,TT}">{HH,HT,TH,TT}. The event “at least one heads occurs” would be the set {HH,HT,TH}">{HH,HT,TH}. If the coin were a normal coin, we would assign the probability of 1/4 to each outcome.

In probability theory, the probability P">P of some event E">E, denoted P(E)">P(E), is usually defined in such a way that P">P satisfies a number of axioms, or rules. The most basic and most important rules are listed below.

Probability Rules

Probability is a number. It is always greater than or equal to zero, and less than or equal to one. This can be written as 0≤P(A)≥1">0≤P(A)≥1. An impossible event, or an event that never occurs, has a probability of 0">0. An event that always occurs has a probability of 1">1. An event with a probability of 0.5">0.5 will occur half of the time.

The sum of the probabilities of all possibilities must equal 1″>1. Some outcome must occur on every trial, and the sum of all probabilities is 100%, or in this case, 1″>1. This can be written as P(S)=1″>P(S)=1, where S″>S represents the entire sample space.

If two events have no outcomes in common, the probability that one or the other occurs is the sum of their individual probabilities. If one event occurs in 30% of the trials, a different event occurs in 20% of the trials, and the two cannot occur together (if they are disjoint), then the probability that one or the other occurs is 30%+20%=50%. This is sometimes referred to as the addition rule, and can be simplified with the following: P(A or B)=P(A)+P(B)″>P(A or B)=P(A)+P(B). The word "or" means the same thing in mathematics as the union, which uses the following symbol: ∪″>∪. Thus when A″>A and B″>B are disjoint, we have P(A∪B)=P(A)+P(B)″>P(A∪B)=P(A)+P(B). The probability that an event does not occur is 1″>1 minus the probability that the event does occur. If an event occurs in 60% of all trials, it fails to occur in the other 40%, because 100%−60%=40%. The probability that an event occurs and the probability that it does not occur always add up to 100%, or 1″>1. These events are called complementary events, and this rule is sometimes called the complement rule. It can be simplified with P(Ac)=1−P(A)″>P(Ac)=1−P(A), where Ac″>Ac is the complement of A″>A.

Two events A″>A and B″>B are independent if knowing that one occurs does not change the probability that the other occurs. This is often called the multiplication rule. If A″>A and B″>B are independent, then P(A and B)=P(A)P(B)″>P(A and B)=P(A)P(B). The word "and" in mathematics means the same thing in mathematics as the intersection, which uses the following symbol: ∩″>∩. Therefore when A″>A and B″>B are independent, we have P(A∩B)=P(A)P(B)″>P(A∩B)=P(A)P(B).

Extension of the Example

Elaborating on our example above of flipping two coins, assign the probability 1/4″>1/4 to each of the 4″>4 outcomes. We consider each of the five rules above in the context of this example.

1. Note that each probability is 1/4″>1/4, which is between 0″>0 and 1″>1.
2. Note that the sum of all the probabilities is 1″>1, since 14+14+14+14=1″>$\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+=1$.
3. Suppose A″>A is the event exactly one head occurs, and B is the event exactly two tails occur. Then A={HT,TH}″>A={HT,TH} and B={TT}″>B={TT} are disjoint. Also, $P(A \cup B) = \frac{3}{4} = \frac{2}{4} + \frac{1}{4} = P(A) + P(B)$.
4. The probability that no heads occurs is 1/4″>1/4, which is

Die

Dice are often used when

learning the rules of probability.

equal to 1−3/4">1−3/4. So if A={HT,TH,HH}">A={HT,TH,HH} is the event that a head occurs, we have P(Ac)=14=1−34=1−P(A)">$P(A^C) = \frac{1}{4} = 1 - \frac{3}{4} = 1 - P(A)$.

5. If A">A is the event that the first flip is a heads and B">B is the event that the second flip is a heads, then A">A andB">B are independent. We have A={HT,HH}">A={HT,HH} and B=TH,HH">B=TH,HH and A∩B=HH">A∩B=HH. Note that P(A∩B)=14=12⋅12=P(A)P(B)">$P(A \cap B) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A)P(B)$.

8.1.2: Conditional Probability

The conditional probability of an event is the probability that an event will occur given that another event has occurred.

Learning Objective

Explain the significance of Bayes' theorem in manipulating conditional probabilities

Key Takeaways

Key Points

- The conditional probability P(B∣A)">P(B ∣ A) of an event B">B, given an event A">A, is defined by: $P(B|A) = \frac{P(A \cap B)}{P(A)}$, when P(A)>0">P(A)>0.

- If the knowledge that event A">A occurs does not change the probability that event B">B occurs, then A">A and B">B are independent events, and thus, P(B∣A)=P(B)">P(B | A)=P(B).
- Mathematically, Bayes' theorem gives the relationship between the probabilities of A">A and B,P(A)">B,P(A) and P(B)">P(B), and the conditional probabilities of A">A given B">B and B">B given A,P(A∩B)">A,P(A∩B) and P(B∩A)">P(B∩A). In its most common form, it is: P(A∩B)=P(B∣A)P(A)P(B)">$P(A \cap B) = \frac{P(B|A)P(A)}{P(B)}$.

Key Terms

conditional probability
: The probability that an event will take place given the restrictive assumption that another event has taken place, or that a combination of other events has taken place

independent
: Not dependent; not contingent or depending on something else; free.

Probability of B Given That A Has Occurred

Our estimation of the likelihood of an event can change if we know that some other event has occurred. For example, the probability that a rolled die shows a 2">2 is 1/6">1/6 without any other information, but if someone looks at the die and tells you that is is an even number, the probability is now 1/3">1/3 that it is a 2">2. The notation P(B∣A)">P(B|A) indicates a conditional probability, meaning it indicates the probability of one event under the condition that we know another event has happened. The bar "∣">|" can be read as "given", so that P(B∣A)">P(B|A) is read as "the probability of B">B given that A">A has occurred".

The conditional probability P(B∣A)">P(B|A) of an event B">B, given an event A">A, is defined by:

P(B∣A)=P(A∩B)P(A)">$P(B|A) = \frac{P(A \cap B)}{P(A)}$

When P(A)>0">P(A)>0. Be sure to remember the distinct roles of B">B and A">A in this formula. The set after the bar is the one we are assuming has occurred, and its probability occurs in the denominator of the formula.

Example

Suppose that a coin is flipped 3 times giving the sample space:

S={HHH,HHT,HTH,THH,TTH,THT,HTT,TTT}">S={HHH,HHT,HTH,THH,TTH,THT,HTT,TTT}

Each individual outcome has probability 1/8">1/8. Suppose that B">B is the event that at least one heads occurs and A">A is the event that all 3 coins are the same. Then the probability of B">B given A">A is 1/2">1/2, since A∩B={HHH}">A∩B={HHH} which has probability 1/8">1/8 and A={HHH,TTT}">A={HHH,TTT} which has probability 2/8">2/8, and 1/82/8=12">$\frac{\frac{1}{8}}{\frac{2}{8}} = \frac{1}{2}$.

Independence

The conditional probability P(B∣A)">P(B|A) is not always equal to the unconditional probability P(B)">P(B). The reason behind this is that the occurrence of event A">A may provide extra information that can change the probability that event B">B occurs. If the knowledge that event A">A occurs does not change the probability that event B">B occurs, then A">A and B">B are independent events, and thus, P(B∣A)=P(B)">P(B|A)=P(B).

Bayes' Theorem

In probability theory and statistics, Bayes' theorem (alternatively Bayes' law or Bayes' rule) is a result that is of importance in the mathematical manipulation of conditional probabilities. It can be derived from the basic axioms of probability.

Mathematically, Bayes' theorem gives the relationship between the probabilities of A">A and B">B, P(A)">P(A) and P(B)">P(B), and the conditional probabilities of A">A given B">B and B">B given A">A. In its most common form, it is:

P(A∣B)=P(B∣A)P(A)P(B)">$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

This may be easier to remember in this alternate symmetric form:

P(A∣B)P(B∣A)=P(A)P(B)">$\frac{P(A|B)}{P(B|A)} = \frac{P(A)}{P(B)}$

Example

Suppose someone told you they had a nice conversation with someone on the train. Not knowing anything else about this conversation, the probability that they were speaking to a woman is 50%. Now suppose they also told you that this person had long hair. It is now more likely they were speaking to a woman, since women in in this city are more likely to have long hair than men. Bayes's theorem can be used to calculate the probability that the person is a woman.

To see how this is done, let W">W represent the event that the conversation was held with a woman, and L">L denote the event that the conversation was held with a long-haired person. It can be assumed that women constitute half the population for this example. So, not knowing anything else, the probability that W">W occurs is P(W)=0.5">P(W)=0.5.

Suppose it is also known that 75% of women in this city have long hair, which we denote as P(L∣W)=0.75">P(L | W)=0.75. Likewise, suppose it is known that 25% of men in this city have long hair, or P(L∣M)=0.25">P(L | M)=0.25, where M">M is the complementary event of W">W, i.e., the event that the conversation was held with a man (assuming that every human is either a man or a woman).

Our goal is to calculate the probability that the conversation was held with a woman, given the fact that the person had long hair, or, in our notation, P(W∣L)">P(W | L). Using the formula for Bayes's theorem, we have:

P(W∣L)=P(L∣W)P(W)P(L)=P(L∣W)P(W)P(L∣W)P(W)+P(L∣M)P(M)=0.75⋅0.50.75⋅0.5+0.25⋅0.5=0.75">
$$P(W|L) = \frac{P(L|W)P(W)}{P(L)} = \frac{P(L|W)P(W)}{P(L|W)P(W)+P(L|M)P(M)} = \frac{0.75 \cdot 0.5}{0.75 \cdot 0.5+0.25 \cdot 0.5} = 0.75$$

8.1.3: Unions and Intersections

Union and intersection are two key concepts in set theory and probability.

Learning Objective

Give examples of the intersection and the union of two or more sets

Key Takeaways

Key Points

- The union of two or more sets is the set that contains all the elements of the two or more sets. Union is denoted by the symbol ∪"> ∪ .
- The general probability addition rule for the union of two events states that P(A∪B)=P(A)+P(B)−P(A∩B)">P(A ∪ B)=P(A)+P(B)−P(A∩B), where A∩B">A∩Bis the intersection of the two sets.
- The addition rule can be shortened if the sets are disjoint: P(A∪B)=P(A)+P(B)">P(A ∪ B)=P(A)+P(B). This can even be extended to more sets if they are all disjoint: P(A∪B∪C)=P(A)+P(B)+P(C)">P(A ∪ B ∪ C)=P(A)+P(B)+P(C).
- The intersection of two or more sets is the set of elements that are common to every set. The symbol ∩">∩ is used to denote the intersection.
- When events are independent, we can use the multiplication rule for independent events, which states that P(A∩B)=P(A)P(B)">P(A∩B)=P(A)P(B).

Key Terms

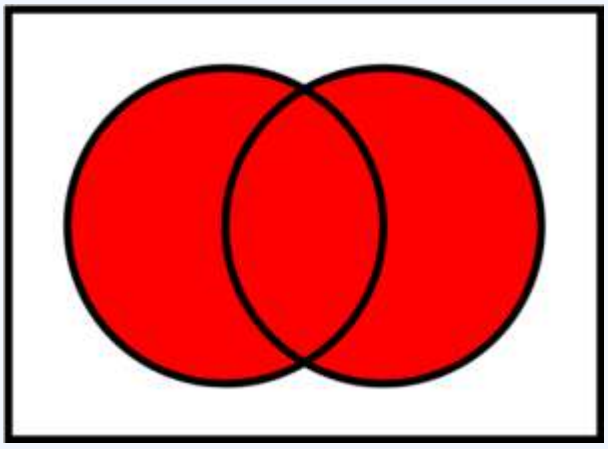
independent
: Not contingent or dependent on something else.

disjoint
: Having no members in common; having an intersection equal to the empty set.

Probability uses the mathematical ideas of sets, as we have seen in the definition of both the sample space of an experiment and in the definition of an event. In order to perform basic probability calculations, we need to review the ideas from set theory related to the set operations of union, intersection, and complement.

Union

The union of two or more sets is the set that contains all the elements of each of the sets; an element is in the union if it belongs to at least one of the sets. The symbol for union is ∪"> ∪ , and is associated with the word "or", because A∪B">A ∪ B is the set of all elements that are in A">A or B">B (or both.) To find the union of two sets, list the elements that are in either (or both) sets. In terms of a Venn Diagram, the union of sets A">A and B">B can be shown as two completely shaded interlocking circles.



Union of Two Sets

Union of Two Sets: The shaded Venn Diagram shows the union of set A">A (the circle on left) with set B">B (the circle on the right). It can be written shorthand as A∪B">A ∪ B.

In symbols, since the union of A″>A and B″>B contains all the points that are in A″>A or B″>B or both, the definition of the union is:

A∪B={x:x∈A or x∈B}″>A ∪ B={x:x∈A or x∈B}

For example, if A={1,3,5,7}″>A={1,3,5,7} and B={1,2,4,6}″>B={1,2,4,6} , then A∪B={1,2,3,4,5,6,7}″>A ∪ B={1,2,3,4,5,6,7}. Notice that the element 1 is not listed twice in the union, even though it appears in both sets A″>A and B″>B. This leads us to the general addition rule for the union of two events:

P(A∪B)=P(A)+P(B)−P(A∩B)″>P(A ∪ B)=P(A)+P(B)−P(A∩B)

Where P(A∩B)″>P(A∩B) is the intersection of the two sets. We must subtract this out to avoid double counting of the inclusion of an element.

If sets A″>A and B″>B are disjoint, however, the event A∩B″>A∩B has no outcomes in it, and is an empty set denoted as ∅, which has a probability of zero. So, the above rule can be shortened for disjoint sets only:

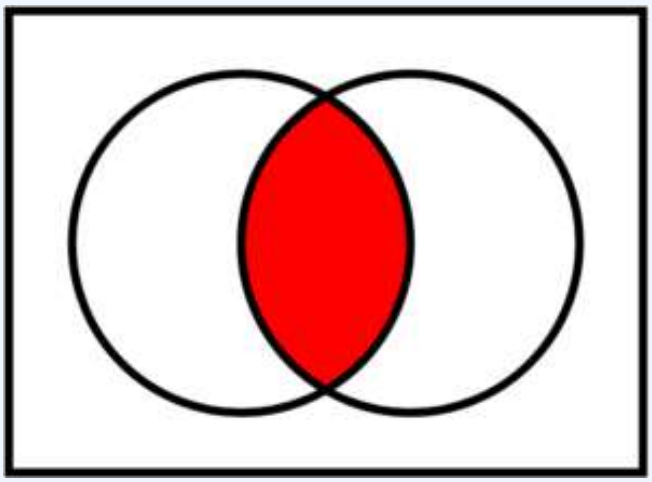
P(A∪B)=P(A)+P(B)″>P(A ∪ B)=P(A)+P(B)

This can even be extended to more sets if they are all disjoint:

P(A∪B∪C)=P(A)+P(B)+P(C)″>P(A ∪ B ∪ C)=P(A)+P(B)+P(C)

Intersection

The intersection of two or more sets is the set of elements that are common to each of the sets. An element is in the intersection if it belongs to all of the sets. The symbol for intersection is ∩″>∩, and is associated with the word “and”, because A∩B″>A∩B is the set of elements that are in A″>A and B″>B simultaneously. To find the intersection of two (or more) sets, include only those elements that are listed in both (or all) of the sets. In terms of a Venn Diagram, the intersection of two sets A″>A and B″>B can be shown at the shaded region in the middle of two interlocking circles .



Intersection of Two Sets

Set A is the circle on the left, set B is the circle on the right, and the intersection of A and B, or A∩B">A∩B, is the shaded portion in the middle.

In mathematical notation, the intersection of A">A and B">B is written asA∩B={x:x∈A">A∩B={x:x∈A and x∈B}">x∈B}. For example, if A={1,3,5,7}">A={1,3,5,7} and B={1,2,4,6}">B={1,2,4,6}, then A∩B={1}">A∩B={1} because 1">1 is the only element that appears in both sets A">A and B">B.

When events are independent, meaning that the outcome of one event doesn't affect the outcome of another event, we can use the multiplication rule for independent events, which states:

P(A∩B)=P(A)P(B)">P(A∩B)=P(A)P(B)

For example, let's say we were tossing a coin twice, and we want to know the probability of tossing two heads. Since the first toss doesn't affect the second toss, the events are independent. Say is the event that the first toss is a heads and B">B is the event that the second toss is a heads, then $P(A \cap B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

8.1.4: Complementary Events

The complement of A is the event in which A does not occur.

Learning Objectives

Explain an example of a complementary event

Key Takeaways

Key Points

- The complement of an event A">A is usually denoted as A′">A‘, Ac">A^C or $\bar{A}$ A¯">.
- An event and its complement are mutually exclusive, meaning that if one of the two events occurs, the other event cannot occur.
- An event and its complement are exhaustive, meaning that both events cover all possibilities.

Key Terms

exhaustive
: including every possible element

mutually exclusive
: describing multiple events or states of being such that the occurrence of any one implies the non-occurrence of all the others

What are Complementary Events?

In probability theory, the complement of any event A">A is the event [not A]">[not A], i.e. the event in which A">A does not occur. The event A">A and its complement [not A]">[not A] are mutually exclusive

and exhaustive, meaning that if one occurs, the other does not, and that both groups cover all possibilities. Generally, there is only one event B″>B such that A″>A and B″>B are both mutually exclusive and exhaustive; that event is the complement of A″>A. The complement of an event A″>A is usually denoted as A′″>A‘, Ac″>A^c or $\bar{A}$A¯″>.

Simple Examples

A common example used to demonstrate complementary events is the flip of a coin. Let's say a coin is flipped and one assumes it cannot land on its edge. It can either land on heads or on tails. There are no other possibilities (exhaustive), and both events cannot occur at the same time (mutually exclusive). Because these two events are complementary, we know that P(heads) + P(tails)=1.

Another simple example of complementary events is picking a ball out of a bag. Let's say there are three plastic balls in a bag. One is blue and two are red. Assuming that each ball has an equal chance of being pulled out of the bag, we know that P(blue)=13″>P(blue)=1/3 and P(red)=23″>P(red)=2/3. Since we can only either chose blue or red (exhaustive) and we cannot choose both at the same time (mutually exclusive), choosing blue and choosing red are complementary events, and P(blue)+P(red)=1″>P(blue)+P(red)=1.

Coin Flip

Often in sports games, such as tennis, a coin flip is used to determine who will serve first because heads and tails are complementary events.

Finally, let's examine a non-example of complementary events. If you were asked to choose any number, you might think that that number could either be prime or composite. Clearly, a number cannot be both prime and composite, so that takes care of the mutually exclusive property. However, being prime or being composite are not exhaustive because the number 1 in mathematics is designated as "unique. "

Attributions

- Fundamentals of Probability

- "Boundless."
  http://www.boundless.com/.
  Boundless Learning
  CC BY-SA 3.0.
- "Probability axioms."
  http://en.wikipedia.org/wiki/Probability_axioms.
  Wikipedia
  CC BY-SA 3.0.
- "sample space."
  http://en.wiktionary.org/wiki/sample_space.
  Wiktionary
  CC BY-SA 3.0.
- "disjoint."
  http://en.wiktionary.org/wiki/disjoint.
  Wiktionary
  CC BY-SA 3.0.
- "independent."
  http://en.wiktionary.org/wiki/independent.
  Wiktionary
  CC BY-SA 3.0.
- "Nuvola apps atlantik."
  http://en.wikipedia.org/wiki/File:Nuvola_apps_atlantik.png.
  Wikipedia
  CC BY-SA.

- Conditional Probability
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Bayes' theorem."
    http://en.wikipedia.org/wiki/Bayes'_theorem.
    Wikipedia
    CC BY-SA 3.0.
  - "Conditional probability."
    http://en.wikipedia.org/wiki/Conditional_probability.
    Wikipedia
    CC BY-SA 3.0.

- "conditional probability."
  http://en.wiktionary.org/wiki/conditional_probability.
  Wiktionary
  CC BY-SA 3.0.
- "independent."
  http://en.wiktionary.org/wiki/independent.
  Wiktionary
  CC BY-SA 3.0.

- Unions and Intersections
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Union (set theory)."
    http://en.wikipedia.org/wiki/Union_(set_theory).
    Wikipedia
    CC BY-SA 3.0.
  - "disjoint."
    http://en.wiktionary.org/wiki/disjoint.
    Wiktionary
    CC BY-SA 3.0.
  - "Intersection (set theory)."
    http://en.wikipedia.org/wiki/Intersection_(set_theory).
    Wikipedia
    CC BY-SA 3.0.
  - "independent."
    http://en.wiktionary.org/wiki/independent.
    Wiktionary
    CC BY-SA 3.0.
  - "Venn0111."
    http://en.wikipedia.org/wiki/File:Venn0111.svg.
    Wikipedia
    CC BY-SA.
  - "Venn0001."
    http://en.wikipedia.org/wiki/File:Venn0001.svg.
    Wikipedia
    CC BY-SA.

- Complementary Events
  - “Boundless.” http://www.boundless.com/. Boundless Learning CC BY-SA 3.0.
  - “Complementary event.” http://en.wikipedia.org/wiki/Complementary_event. Wikipedia CC BY-SA 3.0.
  - “mutually exclusive.” http://en.wiktionary.org/wiki/mutually_exclusive. Wiktionary CC BY-SA 3.0.
  - “exhaustive.” http://en.wiktionary.org/wiki/exhaustive. Wiktionary CC BY-SA 3.0.
  - “Heads or tails.” http://commons.wikimedia.org/wiki/File:Heads_or_tails.jpg. Wikimedia CC BY-SA.

8.2 PROBABILITY RULES

8.2: Probability Rules

8.2.1: The Addition Rule

The addition rule states the probability of two events is the sum of the probability that either will happen minus the probability that both will happen.

Learning Objective

Calculate the probability of an event using the addition rule

Key Takeaways

Key Points

- The addition rule is: P(A∪B)=P(A)+P(B)−P(A∩B).">P(A ∪ B)=P(A)+P(B)−P(A∩B).
- The last term has been accounted for twice, once in P(A)">P(A) and once in P(B)">P(B), so it must be subtracted once so that it is not double-counted.
- If A">A and B">B are disjoint, then P(A∩B)=0">P(A∩B)=0, so the formula

becomes P(A∪B)=P(A)+P(B).">P(A ∪ B)=P(A)+P(B).

Key Term

probability
: The relative likelihood of an event happening.

Addition Law

The addition law of probability (sometimes referred to as the addition rule or sum rule), states that the probability that A">A or B">B will occur is the sum of the probabilities that A">A will happen and that B">B will happen, minus the probability that both A">A and B">B will happen. The addition rule is summarized by the formula:

P(A∪B)=P(A)+P(B)−P(A∩B)">P(A ∪ B)=P(A)+P(B)−P(A∩B)

Consider the following example. When drawing one card out of a deck of 52">52 playing cards, what is the probability of getting heart or a face card (king, queen, or jack)? Let H">H denote drawing a heart and F">F denote drawing a face card. Since there are 13">13 hearts and a total of 12">12 face cards (3">3 of each suit: spades, hearts, diamonds and clubs), but only 3">3 face cards of hearts, we obtain:

P(H)=1352">P(H)=13/52

P(F)=1252">P(F)=12/52

P(F∩H)=352">P(F∩H)=3/52

Using the addition rule, we get:

P(H∪F)=P(H)+P(F)−P(H∩F)=1352+1252−352">P(H ∪ F)=P(H)+P(F)−P(H∩F)

P(H∪F)=P(H)+P(F)−P(H∩F)=1352+1252−352">=(13/52)+(12/52)−3/52

The reason for subtracting the last term is that otherwise we would be counting the middle section twice (since H">H and F">F overlap).

Addition Rule for Disjoint Events

Suppose A">A and B">B are disjoint, their intersection is empty. Then the probability of their intersection is zero. In symbols: P(A∩B)=0">P(A∩B)=0. The addition law then simplifies to:

P(A∪B)=P(A)+P(B)whenA∩B=∅">P(A ∪ B)=P(A)+P(B) when A∩B=∅

The symbol ∅">∅ represents the empty set, which indicates that in this case A">A and B">B do not have any elements in common (they do not overlap).

Example

Suppose a card is drawn from a deck of 52 playing cards: what is the probability of getting a king or a queen? Let A">A represent the event that a king is drawn and B">B represent the event that a queen is drawn. These two events are disjoint, since there are no kings that are also queens. Thus:

P(A∪B)=P(A)+P(B)=452+452=852=213">P(A ∪ B)=P(A)+P(B)

P(A∪B)=P(A)+P(B)=452+452=852=213">$= \frac{4}{52} + \frac{4}{52}$

P(A∪B)=P(A)+P(B)=452+452=852=213">$= \frac{8}{52}$

P(A∪B)=P(A)+P(B)=452+452=852=213">$= \frac{2}{13}$

8.2.2: The Multiplication Rule

The multiplication rule states that the probability that A and B both occur is equal to the probability that B occurs times the conditional probability that A occurs given that B occurs.

Learning Objective

Apply the multiplication rule to calculate the probability of both A and B occurring

Key Takeaways

Key Points

- The multiplication rule can be written as: P(A∩B)=P(B)⋅P(A|B)">P(A∩B)=P(B)·P(A|B).
- We obtain the general multiplication rule by multiplying both sides of the definition of conditional probability by the denominator.

Key Term

sample space
: The set of all possible outcomes of a game, experiment or other situation.

The Multiplication Rule

In probability theory, the Multiplication Rule states that the probability that A">A and B">B occur is equal to the probability that A">A occurs times the conditional probability that B">B occurs, given that we know A">A has already occurred. This rule can be written:

P(A∩B)=P(B)⋅P(A|B)">P(A∩B)=P(B)·P(A|B)

Switching the role of A">A and B">B, we can also write the rule as:

P(A∩B)=P(A)⋅P(B|A)">P(A∩B)=P(A)·P(B|A)

We obtain the general multiplication rule by multiplying both sides of the definition of conditional probability by the denominator. That is, in the equation $P(A|B) = \frac{P(A \cap B)}{P(B)}$ P(A|B)=P(A∩B)P(B)">, if we multiply both sides by P(B)">P(B), we obtain the Multiplication Rule.

The rule is useful when we know both P(B)">P(B) and P(A|B)">P(A|B), or both P(A)">P(A) and P(B|A)."">P(B|A).

Example

Suppose that we draw two cards out of a deck of cards and let A">A be the event the the first card is an ace, and B">B be the event that the second card is an ace, then:

P(A)=452">P(A)=4/52

And:

P(B|A)=351">P(B|A)=3/51

The denominator in the second equation is 51">51 since we know a card has already been drawn. Therefore, there are 51">51 left in total. We also know the first card was an ace, therefore:

P(A∩B)=P(A)⋅P(B|A)=452⋅351=0.0045">P(A∩B)=P(A)⋅P(B|A)

P(A∩B)=P(A)⋅P(B|A)=452⋅351=0.0045">=(4/52)⋅(3/51)

P(A∩B)=P(A)⋅P(B|A)=452⋅351=0.0045">=0.0045

Independent Event

Note that when A">A and B">B are independent, we have that P(B|A)=P(B)">P(B|A)=P(B), so the formula becomes P(A∩B)=P(A)P(B)">P(A∩B)=P(A)P(B), which we encountered in a previous section. As an example, consider the experiment of rolling a die and flipping a coin. The probability that we get a 2">2 on the die and a tails on the coin is (16⋅12=112">1/6)⋅(1/2)=1/12, since the two events are independent.

8.2.3: Independence

To say that two events are independent means that the occurrence of one does not affect the probability of the other.

Learning Objective

Explain the concept of independence in relation to probability theory

Key Takeaways

Key Points

- Two events are independent if the following are true: P(A|B)=P(A)">P(A|B)=P(A),P(B|A)=P(B)">P(B|A)=P(B), and P(A and B)=P(A)⋅P(B)">P(A and B)=P(A)⋅P(B).
- If any one of these conditions is true, then all of them are true.
- If events A">A and B">B are independent, then the chance of A">A occurring does not affect the chance of B">B occurring and vice versa.

Key Terms

independence
: The occurrence of one event does not affect the probability of the occurrence of another.

probability theory
: The mathematical study of probability (the likelihood of occurrence of random events in order to predict the behavior of defined systems).

Independent Events

In probability theory, to say that two events are independent means that the occurrence of one does not affect

the probability that the other will occur. In other words, if events A″>A and B″>B are independent, then the chance of A″>A occurring does not affect the chance of B″>B occurring and vice versa. The concept of independence extends to dealing with collections of more than two events.

Two events are independent if any of the following are true:

1. P(A|B)=P(A)″>P(A|B)=P(A)
2. P(B|A)=P(B)″>P(B|A)=P(B)

Independent Events

Selecting two cards from a deck by first selecting one, then replacing it in the deck before selecting a second is an example of independent events.

3. P(A and B)=P(A)⋅P(B)″>P(A and B)=P(A)·P(B)

To show that two events are independent, you must show only one of the conditions listed above. If any one of these conditions is true, then all of them are true.

Translating the symbols into words, the first two mathematical statements listed above say that the probability for the event with the condition is the same as the probability for the event without the condition. For independent events, the condition does not change the probability for the event. The third statement says that the probability of both independent events A″>A and B″>B occurring is the same as the probability of A″>A occurring, multiplied by the probability of B″>B occurring.

As an example, imagine you select two cards consecutively from a complete deck of playing cards. The two selections are not independent. The result of the first selection changes the remaining deck and affects the probabilities for the second selection. This is referred to as selecting “without replacement” because the first card has not been replaced into the deck before the second card is selected.

However, suppose you were to select two cards “with replacement” by returning your first card to the deck and shuffling the deck before selecting the second card. Because the deck of cards is complete for both selections, the first selection does not affect the probability of the second selection. When selecting cards with replacement, the selections are independent.

Consider a fair die role, which provides another example of independent events. If a person roles two die, the outcome of the first roll does not change the probability for the outcome of the second roll.

Example

Two friends are playing billiards, and decide to flip a coin to determine who will play first during each round. For the first two rounds, the coin lands on heads. They decide to play a third round, and flip the coin again. What is the probability that the coin will land on heads again?

First, note that each coin flip is an independent event. The side that a coin lands on does not depend on what occurred previously.

For any coin flip, there is a 1/2 chance that the coin will land on heads. Thus, the probability that the coin will land on heads during the third round is 1/2.

Example

When flipping a coin, what is the probability of getting tails 5 times in a row?

Recall that each coin flip is independent, and the probability of getting tails is 1/2 for any flip. Also recall that the following statement holds true for any two independent events A and B:

P(A and B)=P(A)⋅P(B)">P(A and B)=P(A)·P(B)

Finally, the concept of independence extends to collections of more than 2 events.

Therefore, the probability of getting tails 4 times in a row is:

12⋅12⋅12⋅12=116">$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$

8.2.4: Counting Rules and Techniques

Combinatorics is a branch of mathematics concerning the study of finite or countable discrete structures.

Learning Objective

Describe the different rules and properties for combinatorics

Key Takeaways

Key Points

- The rule of sum (addition rule), rule of product (multiplication rule), and inclusion-exclusion principle are often used for enumerative purposes.
- Bijective proofs are utilized to demonstrate that two sets have the same number of elements.
- Double counting is a technique used to demonstrate that two expressions are equal. The pigeonhole principle often ascertains the existence of something or is used to determine the minimum or maximum number of something in a discrete context.
- Generating functions and recurrence relations are powerful tools that can be used to manipulate sequences, and can describe if not resolve many combinatorial situations.
- Double counting is a technique used to demonstrate that two expressions are equal.

Key Terms

polynomial
: An expression consisting of a sum of a finite number of terms: each term being the product of a constant coefficient and one or more variables raised to a non-negative integer power.

combinatorics
: A branch of mathematics that studies (usually finite) collections of objects that satisfy specified criteria.

Combinatorics is a branch of mathematics concerning the study of finite or countable discrete structures. Combinatorial techniques are applicable to many areas of mathematics, and a knowledge of combinatorics is necessary to build a solid command of statistics. It involves the enumeration, combination, and permutation of sets of elements and the mathematical relations that characterize their properties.

Aspects of combinatorics include: counting the structures of a given kind and size, deciding when certain criteria can be met, and constructing and analyzing objects meeting the criteria. Aspects also include finding "largest," "smallest," or "optimal" objects, studying combinatorial structures arising in an algebraic context, or applying algebraic techniques to combinatorial problems.

Combinatorial Rules and Techniques

Several useful combinatorial rules or combinatorial principles are commonly recognized and used. Each of these principles is used for a specific purpose. The rule of sum (addition rule), rule of product (multiplication rule), and inclusion-exclusion principle are often used for enumerative purposes. Bijective proofs are utilized to demonstrate that two sets have the same number of elements. Double counting is a method of showing that two expressions are equal. The pigeonhole principle often ascertains the existence of something or is used to determine the minimum or maximum number of something in a discrete context. Generating functions and recurrence relations are powerful tools that can be used to manipulate sequences, and can describe if not resolve many combinatorial situations. Each of these techniques is described in greater detail below.

Rule of Sum

The rule of sum is an intuitive principle stating that if there are a">a possible ways to do something, and b">b possible ways to do another thing, and the two things can't both be done, then there are a+b">a+b total possible ways to do one of the things. More formally, the sum of the sizes of two disjoint sets is equal to the size of the union of these sets.

Rule of Product

The rule of product is another intuitive principle stating that if there are a">a ways to do something and b">b ways to do another thing, then there are a⋅b">a·b ways to do both things.

Inclusion-Exclusion Principle

The inclusion-exclusion principle is a counting technique that is used to obtain the number of elements in a union of multiple sets. This counting method ensures that elements that are present in more than one set in the union are not counted more than once. It considers the size of each set and the size of the intersections of

the sets. The smallest example is when there are two sets: the number of elements in the union of A″>A and B″>B is equal to the sum of the number of elements in A″>A and B″>B, minus the number of elements in their intersection. See the diagram below for an example with three sets.

Bijective Proof

A bijective proof is a proof technique that finds a bijective function f:A→B″>f:A→B between two finite sets A″>A and B″>B, which proves that they have the same number of elements, |A|=|B|″>|A|=|B|. A bijective function is one in which there is a one-to-one correspondence between the elements of two sets. In other words, each element in set B″>B is paired with exactly one element in set A″>A. This technique is useful if we wish to know the size of A″>A, but can find no direct way of counting its elements. If B″>B is more easily countable, establishing a bijection from A″>A to B″>B solves the problem.

Double Counting

Double counting is a combinatorial proof technique for showing that two expressions are equal. This is done by demonstrating that the two expressions are two different ways of counting the size of one set. In this technique, a finite set X″>X is described from two perspectives, leading to two distinct expressions for the size of the set. Since both expressions equal the size of the same set, they equal each other.

Pigeonhole Principle

The pigeonhole principle states that if a″>a items are each put into one of b″>b boxes, where a>b″>a>b, then at least one of the boxes contains more than one item. This principle allows one to demonstrate the existence of some element in a set with some specific properties. For example, consider a set of three gloves. In such a set, there must be either two left gloves or two right gloves (or three of left or right). This is an application of the pigeonhole principle that yields information about the properties of the gloves in the set.

Generating Function

Generating functions can be thought of as polynomials with infinitely many terms whose coefficients correspond to the terms of a sequence. The (ordinary) generating function of a sequence an″>an is given by:

f(x)=∑n=0∞anxn″>$f(x) = \sum_{n=0}^{\infty} a_n x^n$

whose coefficients give the sequence {a0,a1,a2,…}″>{a_0,a_1,a_2,...}.

Recurrence Relation

A recurrence relation defines each term of a sequence in terms of the preceding terms. In other words, once one or more initial terms are given, each of the following terms of the sequence is a function of the preceding terms.

The Fibonacci sequence is one example of a recurrence relation. Each term of the Fibonacci sequence is given by Fn=Fn−1+Fn−2">$F_n=F_{n-1}+F_{n-2}$, with initial values F0=0">$F_0=0$ and F1=1">$F_1=1$. Thus, the sequence of Fibonacci numbers begins:

0,1,1,2,3,5,8,13,21,34,55,89,…">0,1,1,2,3,5,8,13,21,34,55,89,...

8.2.5: Bayes' Rule

Bayes' rule expresses how a subjective degree of belief should rationally change to account for evidence.

Learning Objective

Explain the importance of Bayes's theorem in mathematical manipulation of conditional probabilities

Key Takeaways

Key Points

- Bayes' rule relates the odds of event A1">A_1 to event A2">A_2, before (prior to) and after (posterior to) conditioning on another event B">B.
- More specifically, given events A1">A1">A_1, A2">A_2, and B">B, Bayes' rule states that the

conditional odds of A1:A2">A_1:A_2 given B">B are equal to the marginal odds A1:A2">A1">A_1:A_2 if multiplied by the Bayes' factor.

- Bayes' rule shows how one's judgement on whether A1">A_1or A2">A_2 is true should be updated based on observing the evidence.
- Bayesian inference is a method of inference in which Bayes' rule is used to update the probability estimate for a hypothesis as additional evidence is learned.

Key Term

Bayes' factor

The ratio of the conditional probabilities of the event B given that A_1 is the case or that A_2 is the case, respectively.

In probability theory and statistics, Bayes' theorem (or Bayes' rule) is a result that is of importance in the mathematical manipulation of conditional probabilities. It is a result that derives from the more basic axioms of probability. When applied, the probabilities involved in Bayes' theorem may have any of a number of probability interpretations. In one of these interpretations, the theorem is used directly as part of a particular approach to statistical inference. In particular, with the Bayesian interpretation of probability, the theorem expresses how a subjective degree of belief should rationally change to account for evidence. This is known as Bayesian inference, which is fundamental to Bayesian statistics.

Bayes' rule relates the odds of event A1">A1 to event A2">A2, before (prior to) and after (posterior to) conditioning on another event B">B. The odds on A1">A1 to event A2">A2 is simply the ratio of the probabilities of the two events. The relationship is expressed in terms of the likelihood ratio, or Bayes' factor. By definition, this is the ratio of the conditional probabilities of the event B">B given that A1">A1 is the case or that A2">A2 is the case, respectively. The rule simply states:

Posterior odds equals prior odds times Bayes' factor.

More specifically, given events A1">A1, A2">A2 and B">B, Bayes' rule states that the conditional odds of A1:A2">A1:A2 given B">B are equal to the marginal odds A1:A2">A1:A2 multiplied by the Bayes factor or likelihood ratio. This is shown in the following formulas:

O(A1:A2|B)=Λ(A1:A2|B)⋅O(A1:A2)">$O(A_1:A_2|B)=\Lambda(A_1:A_2|B)\cdot O(A_1:A_2)$

Where the likelihood ratio Λ">Λ is the ratio of the conditional probabilities of the event B">B given that A1">A1 is the case or that A2">A2 is the case, respectively:

Λ(A1:A2|B)=P(B|A1)P(B|A2)">$\Lambda(A_1:A_2|B)=P(B|A_1)P(B|A_2)$

Bayes' rule is widely used in statistics, science and engineering, such as in: model selection, probabilistic expert systems based on Bayes' networks, statistical proof in legal proceedings, email spam filters, etc. Bayes' rule tells us how unconditional and conditional probabilities are related whether we work with a frequentist

or a Bayesian interpretation of probability. Under the Bayesian interpretation it is frequently applied in the situation where A1">A_1 and A2">A_2 are competing hypotheses, and B">B is some observed evidence. The rule shows how one's judgement on whether A1">A_1 or A2">A_2 is true should be updated on observing the evidence.

Bayesian Inference

Bayesian inference is a method of inference in which Bayes' rule is used to update the probability estimate for a hypothesis as additional evidence is learned. Bayesian updating is an important technique throughout statistics, and especially in mathematical statistics. Bayesian updating is especially important in the dynamic analysis of a sequence of data. Bayesian inference has found application in a range of fields including science, engineering, philosophy, medicine, and law.

Informal Definition

Rationally, Bayes' rule makes a great deal of sense. If the evidence does not match up with a hypothesis, one should reject the hypothesis. But if a hypothesis is extremely unlikely *a priori*, one should also reject it, even if the evidence does appear to match up.

For example, imagine that we have various hypotheses about the nature of a newborn baby of a friend, including:

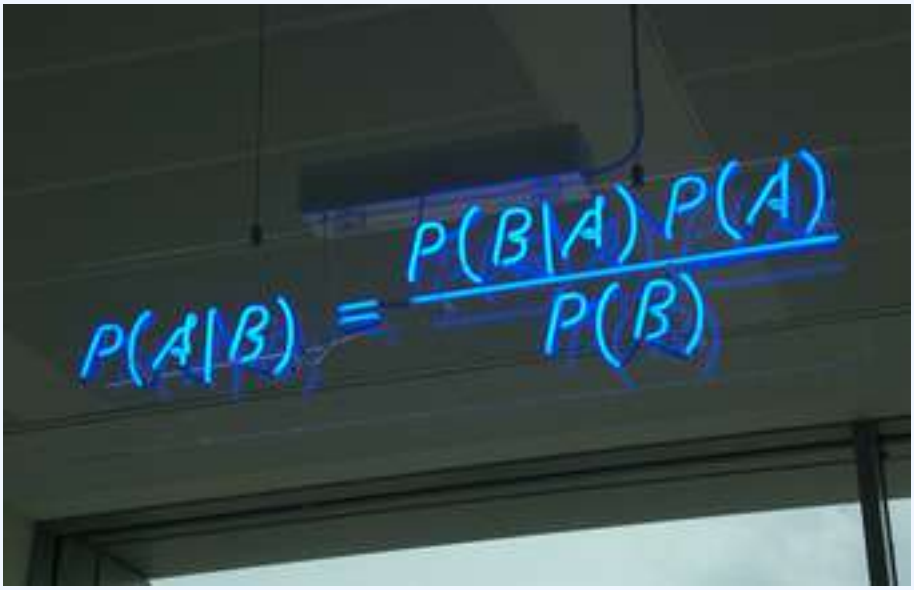
- H_1: The baby is a brown-haired boy.
- H_2: The baby is a blond-haired girl.
- H_3: The baby is a dog.

Then, consider two scenarios:

1. We're presented with evidence in the form of a picture of a blond-haired baby girl. We find this evidence supports H2">H_2 and opposes H1">H_1 and H3">H_3.
2. We're presented with evidence in the form of a picture of a baby dog. Although this evidence, treated in isolation, supports H3">H_3, my prior belief in this hypothesis (that a human can give birth to a dog) is extremely small. Therefore, the posterior probability is nevertheless small.

The critical point about Bayesian inference, then, is that it provides a principled way of combining new evidence with prior beliefs, through the application of Bayes' rule. Furthermore, Bayes' rule can be applied iteratively. After observing some evidence, the resulting posterior probability can then be treated as a prior probability, and a new posterior probability computed from new evidence. This allows for Bayesian principles

to be applied to various kinds of evidence, whether viewed all at once or over time. This procedure is termed Bayesian updating.



Bayes' Theorem

A blue neon sign at the Autonomy Corporation in Cambridge, showing the simple statement of Bayes' theorem.

8.2.6: The Collins Case

The People of the State of California v. Collins was a 1968 jury trial in California that made notorious forensic use of statistics and probability.

Learning Objective

Argue what causes prosecutor's fallacy

Key Takeaways

Key Points

- Bystanders to a robbery in Los Angeles testified that the perpetrators had been a black male, with a beard and moustache, and a caucasian female with blonde hair tied in a ponytail. They had escaped in a yellow motor car.
- A witness of the prosecution, an instructor in mathematics, explained the multiplication rule to the jury, but failed to give weight to independence, or the difference between conditional and unconditional probabilities.
- The Collins case is a prime example of a phenomenon known as the prosecutor's fallacy.

Key Terms

multiplication rule
: The probability that A and B occur is equal to the probability that A occurs times the probability that B occurs, given that we know A has already occurred.

prosecutor's fallacy
: A fallacy of statistical reasoning when used as an argument in legal proceedings.

The People of the State of California v. Collins was a 1968 jury trial in California. It made notorious forensic use of statistics and probability. Bystanders to a robbery in Los Angeles testified that the perpetrators had been a black male, with a beard and moustache, and a caucasian female with blonde hair tied in a ponytail. They had escaped in a yellow motor car.

The prosecutor called upon for testimony an instructor in mathematics from a local state college. The instructor explained the multiplication rule to the jury, but failed to give weight to independence, or the difference between conditional and unconditional probabilities. The prosecutor then suggested that the jury would be safe in estimating the following probabilities:

- Black man with beard: 1 in 10
- Man with moustache: 1 in 4
- White woman with pony tail: 1 in 10

- White woman with blonde hair: 1 in 3
- Yellow motor car: 1 in 10
- Interracial couple in car: 1 in 1000

These probabilities, when considered together, result in a 1 in 12,000,000 chance that any other couple with similar characteristics had committed the crime – according to the prosecutor, that is. The jury returned a verdict of guilty.

As seen in , upon appeal, the Supreme Court of California set aside the conviction, criticizing the statistical reasoning and disallowing the way the decision was put to the jury. In their judgment, the justices observed that mathematics:

... while assisting the trier of fact in the search of truth, must not cast a spell over him.

Prosecutor's Fallacy

The Collins' case is a prime example of a phenomenon known as the prosecutor's fallacy—a fallacy of statistical reasoning when used as an argument in legal proceedings. At its heart, the fallacy involves assuming that the prior probability of a random match is equal to the probability that the defendant is innocent. For example, if a perpetrator is known to have the same blood type as a defendant (and 10% of the population share that blood type), to argue solely on that basis that the probability of the defendant being guilty is 90% makes the prosecutors's fallacy (in a very simple form).

The basic fallacy results from misunderstanding conditional probability, and neglecting the prior odds of a defendant being guilty before that evidence was introduced. When a prosecutor has collected some evidence (for instance, a DNA match) and has an expert testify that the probability of finding this evidence if the accused were innocent is tiny, the fallacy occurs if it is concluded that the probability of the accused being innocent must be comparably tiny. If the DNA match is used to confirm guilt that is otherwise suspected, then it is indeed strong evidence. However, if the DNA evidence is the sole evidence against the accused, and the accused was picked out of a large database of DNA profiles, then the odds of the match being made at random may be reduced. Therefore, it is less damaging to the defendant. The odds in this scenario do not relate to the odds of being guilty; they relate to the odds of being picked at random.

The Collins Case

The Collins case is a classic example of the prosecutor's fallacy. The guilty verdict was reversed upon appeal to the Supreme Court of California in 1968.

Attributions

- The Addition Rule
  - “Boundless.” http://www.boundless.com/. Boundless Learning CC BY-SA 3.0.
  - “Probability axioms.” http://en.wikipedia.org/wiki/Probability_axioms. Wikipedia CC BY-SA 3.0.
  - “addition rule.” http://en.wikipedia.org/wiki/addition%20rule. Wikipedia CC BY-SA 3.0.
  - “probability.” http://en.wiktionary.org/wiki/probability. Wiktionary CC BY-SA 3.0.
  - “Probability.” https://en.wikipedia.org/wiki/Probability. Wikipedia CC BY-SA 3.0.
  - “Some rules of probability – Statistics.” http://statistics.wikidot.com/ch5. Wikidot CC BY-SA.
- The Multiplication Rule
  - “Boundless.” http://www.boundless.com/. Boundless Learning CC BY-SA 3.0.
  - “sample space.” http://en.wiktionary.org/wiki/sample_space. Wiktionary CC BY-SA 3.0.
  - “Some rules of probability – Statistics.”

http://statistics.wikidot.com/ch5.
Wikidot
CC BY-SA.

- Independence
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Independence (probability theory)."
    http://en.wikipedia.org/wiki/Independence_(probability_theory).
    Wikipedia
    CC BY-SA 3.0.
  - "independence."
    http://en.wikipedia.org/wiki/independence.
    Wikipedia
    CC BY-SA 3.0.
  - "probability theory."
    http://en.wiktionary.org/wiki/probability_theory.
    Wiktionary
    CC BY-SA 3.0.
  - "Roberta Bloom, Probability Topics: Independent & Mutually Exclusive Events (modified R. Bloom). September 17, 2013."
    http://cnx.org/content/m18868/latest/.
    OpenStax CNX
    CC BY 3.0.
  - "All sizes | Ace of Spades Card Deck Trick Magic Macro 10-19-09 2 | Flickr – Photo Sharing!."
    http://www.flickr.com/photos/stevendepolo/4028160820/sizes/o/in/photostream/.
    Flickr
    CC BY.
- Counting Rules and Techniques
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Combinatorics."
    http://en.wikipedia.org/wiki/Combinatorics.
    Wikipedia

CC BY-SA 3.0.

- "Pigeonhole principle."
  https://en.wikipedia.org/wiki/Pigeonhole_principle.
  Wikipedi
  CC BY-SA 3.0.
- "Bijective proof."
  http://en.wikipedia.org/wiki/Bijective_proof.
  Wikipedia
  CC BY-SA 3.0.
- "Double counting (proof technique)."
  http://en.wikipedia.org/wiki/Double_counting_(proof_technique).
  Wikipedia
  CC BY-SA 3.0.
- "Bijection."
  https://en.wikipedia.org/wiki/Bijection.
  Wikipedia
  CC BY-SA 3.0.
- "Inclusion-exclusion principle."
  https://en.wikipedia.org/wiki/Inclusion%E2%80%93exclusion_principle.
  Wikipedia
  CC BY-SA 3.0.
- "polynomial."
  http://en.wiktionary.org/wiki/polynomial.
  Wiktionary
  CC BY-SA 3.0.
- "combinatorics."
  http://en.wiktionary.org/wiki/combinatorics.
  Wiktionary
  CC BY-SA 3.0.
- "Combinatorial principles."
  http://en.wikipedia.org/wiki/Combinatorial_principles.
  Wikipedia
  CC BY-SA 3.0.

- Bayes' Rule
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning

CC BY-SA 3.0.

- "Bayes' rule."
  http://en.wikipedia.org/wiki/Bayes'_rule.
  Wikipedia
  CC BY-SA 3.0.
- "Bayesian inference."
  http://en.wikipedia.org/wiki/Bayesian_inference.
  Wikipedia
  CC BY-SA 3.0.
- "Bayes' theorem."
  http://en.wikipedia.org/wiki/Bayes'_theorem.
  Wikipedia
  CC BY-SA 3.0.
- "Bayes' factor."
  http://en.wikipedia.org/wiki/Bayes'%20factor.
  Wikipedia
  CC BY-SA 3.0.
- "Bayes' Theorem MMB 01."
  http://commons.wikimedia.org/wiki/File:Bayes'_Theorem_MMB_01.jpg.
  Wikimedia
  CC BY-SA.

- The Collins Case
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Prosecutor's fallacy."
    http://en.wikipedia.org/wiki/Prosecutor's_fallacy.
    Wikipedia
    CC BY-SA 3.0.
  - "People v. Collins."
    http://en.wikipedia.org/wiki/People_v._Collins.
    Wikipedia
    CC BY-SA 3.0.
  - "prosecutor's fallacy."
    http://en.wikipedia.org/wiki/prosecutor's%20fallacy.
    Wikipedia

CC BY-SA 3.0.

- "CA SC seal." http://commons.wikimedia.org/wiki/File:CA_SC_seal.png. Wikimedia Public domain.

8.3 MORE ABOUT CHANCE

8.3: More About Chance

8.3.1: The Paradox of the Chevalier De Méré

de Méré observed that getting at least one 6 with 4 throws of a die was more probable than getting double 6's with 24 throws of a pair of dice.

Learning Objective

Explain Chevalier de Méré's Paradox when rolling a die

Key Takeaways

Key Points

- Chevalier de Méré originally thought that rolling a 6 in 4 throws of a die was equiprobable to rolling a pair of 6's in 24 throws of a pair of dice.
- In practice, he would win the first bet more than half the time, but lose the second bet more than half the time.
- de Méré asked his mathematician friend, Pascal, to help him solve the problem.

- The probability of rolling a 6 in 4 throws is $1 - \frac{5}{6}^4$, which turns out to be just over 50%.
- The probability of rolling two 6's in 24 throws of a pair of dice is $1 - \frac{35}{36}^2 4$, which turns out to be just under 50%.

Key Terms

veridical paradox
: a situation in which a result appears absurd but is demonstrated to be true nevertheless

independent event
: the fact that A occurs does not affect the probability that B occurs

equiprobable
: having an equal chance of occurring mathematically

Chevalier de Méré

Antoine Gombaud, Chevalier de Méré (1607 – 1684) was a French writer, born in Poitou. Although he was not a nobleman, he adopted the title Chevalier (Knight) for the character in his dialogues who represented his own views (Chevalier de Méré because he was educated at Méré). Later, his friends began calling him by that name.

Méré was an important Salon theorist. Like many 17th century liberal thinkers, he distrusted both hereditary power and democracy. He believed that questions are best resolved in open discussions among witty, fashionable, intelligent people.

He is most well known for his contribution to probability. One of the problems he was interested in was called the *problem of points*. Suppose two players agree to play a certain number of games — say, a best-of-seven series — and are interrupted before they can finish. How should the stake be divided among them if, say, one has won three games and the other has won one?

Another one of his problems has come to be called "De Méré's Paradox," and it is explained below.

De Mere's Paradox

Which of these two is more probable:

1. Getting at least one six with four throws of a die or
2. Getting at least one double six with 24 throws of a pair of dice?

The self-styled Chevalier de Méré believed the two to be equiprobable, based on the following reasoning:

1. Getting a pair of sixes on a single roll of two dice is the same probability of rolling two sixes on two rolls of one die.
2. The probability of rolling two sixes on two rolls is 1/6 as likely as one six in one roll.
3. To make up for this, a pair of dice should be rolled six times for every one roll of a single die in order to get the same chance of a pair of sixes.
4. Therefore, rolling a pair of dice six times as often as rolling one die should equal the probabilities.
5. So, rolling 2 dice 24 times should result in as many double sixes as getting one six with throwing one die four times.

However, when betting on getting two sixes when rolling 24 times, Chevalier de Méré lost consistently. He posed this problem to his friend, mathematician Blaise Pascal, who solved it.

Explanation

Throwing a die is an experiment with a finite number of equiprobable outcomes. There are 6 sides to a die, so there is
probability for a 6 to turn up in 1 throw. That is, there is a $\frac{1}{6} - \frac{1}{6} = \frac{5}{6}$ probability for a 6 not to turn up. When you throw a die 4 times, the probability of a 6 not turning up at all is $(1 - \frac{1}{6})^4 = (\frac{5}{6})^4$. So, there is a probability of $(\frac{6}{6}) - (\frac{5}{6})^4$ of getting at least one 6 with 4 rolls of a die. If you do the arithmetic, this gives you a probability of approximately 0.5177, or a favorable probability of a 6 appearing in 4 rolls.

Now, when you throw a pair of dice, from the definition of independent events, there is a $(\frac{1}{6})^2 = \frac{1}{36}$ probability of a pair of 6's appearing. That is the same as saying the probability for a pair of 6's not showing is 35/36. Therefore, there is a probability of $(\frac{36}{36}) - (\frac{35}{36})^{24}$ of getting at least one pair of 6's with 24 rolls of a pair of dice. If you do the arithmetic, this gives you a probability of approximately 0.4914, or a favorable probability of a pair of 6's *not* appearing in 24 rolls.

This is a veridical paradox. Counter-intuitively, the odds are distributed differently from how they would be expected to be.

de Méré's Paradox

de Méré observed that getting at least one 6 with 4 throws of a die was more probable than getting double 6's with 24 throws of a pair of dice.

8.3.2: Are Real Dice Fair?

A fair die has an equal probability of landing face-up on each number.

Learning Objective

Infer how dice act as a random number generator

Key Takeaways

Key Points

- Regardless of what it is made out of, the angle at which the sides connect, and the spin and speed of the roll, a fair die gives each number an equal probability of landing face-up. Every side must be equal, and every set of sides must be equal.
- The result of a die roll is determined by the way it is thrown; they are made random by uncertainty due to factors like movements in the thrower's hand. Thus, they are a type of hardware random number generator.
- Precision casino dice have their pips drilled, then filled flush with a paint of the same density as the material used for the dice, such that the center of gravity of the dice is as close to the geometric center as possible.
- A loaded, weighted, or crooked die is one that has been tampered with to land with a specific side facing upwards more often than it normally would.

Key Terms

random number
: number allotted randomly using suitable generator (electronic machine or as simple "generator" as die)

pip
: one of the spots or symbols on a playing card, domino, die, etc.

Platonic solid
: any one of the following five polyhedra: the regular tetrahedron, the cube, the regular octahedron, the regular dodecahedron and the regular icosahedron

A die (plural dice) is a small throw-able object with multiple resting positions, used for generating random numbers. This makes dice suitable as gambling devices for games like craps, or for use in non-gambling tabletop games.

An example of a traditional die is a rounded cube, with each of its six faces showing a different number of dots (pips) from one to six. When thrown or rolled, the die comes to rest showing on its upper surface a

random integer from one to six, each value being equally likely. A variety of similar devices are also described as dice; such specialized dice may have polyhedral or irregular shapes and may have faces marked with symbols instead of numbers. They may be used to produce results other than one through six. Loaded and crooked dice are designed to favor some results over others for purposes of cheating or amusement.

What Makes Dice Fair?

A fair die is a shape that is labelled so that each side has an equal probability of facing upwards when rolled onto a flat surface, regardless of what it is made out of, the angle at which the sides connect, and the spin and speed of the roll. Every side must be equal, and every set of sides must be equal.

The result of a die roll is determined by the way it is thrown, according to the laws of classical mechanics; they are made random by uncertainty due to factors like movements in the thrower's hand. Thus, they are a type of hardware random number generator. Perhaps to mitigate concerns that the pips on the faces of certain styles of dice cause a small bias, casinos use precision dice with flush markings.

Precision casino dice may have a polished or sand finish, making them transparent or translucent, respectively. Casino dice have their pips drilled, then filled flush with a paint of the same density as the material used for the dice, such that the center of gravity of the dice is as close to the geometric center as possible. All such dice are stamped with a serial number to prevent potential cheaters from substituting a die.

The most common fair die used is the cube, but there are many other types of fair dice. The other four Platonic solids are the most common non-cubical dice; these can make for 4, 8, 12, and 20 faces . The only other common non-cubical die is the 10-sided die.

Platonic Solids as Dice

A Platonic solids set of five dice; tetrahedron (four faces), cube/hexahedron (six faces), octahedron (eight faces), dodecahedron (twelve faces), and icosahedron (twenty faces).

Loaded Dice

A loaded, weighted, or crooked die is one that has been tampered with to land with a specific side facing

upwards more often than it normally would. There are several methods for creating loaded dice; these include round and off-square faces and (if not transparent) weights. Tappers have a mercury drop in a reservoir at the center, with a capillary tube leading to another reservoir at a side; the load is activated by tapping the die so that the mercury travels to the side.

Attributions

- The Paradox of the Chevalier De Méré
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Antoine Gombaud."
    http://en.wikipedia.org/wiki/Antoine_Gombaud.
    Wikipedia
    CC BY-SA 3.0.
  - "equiprobable."
    http://en.wiktionary.org/wiki/equiprobable.
    Wiktionary
    CC BY-SA 3.0.
  - Proof Wiki.

    http://www.proofwiki.org/wiki/De_M%C3%A9r%C3%A9's_Paradox.
    GNU FDL.
  - "6sided dice."
    http://en.wikipedia.org/wiki/File:6sided_dice.jpg.
    Wikipedia
    CC BY-SA.
- Are Real Dice Fair?
  - "Boundless."
    http://www.boundless.com/.
    Boundless Learning
    CC BY-SA 3.0.
  - "Dice."
    http://en.wikipedia.org/wiki/Dice.
    Wikipedia

- "pip."
  http://en.wiktionary.org/wiki/pip.
  Wiktionary

- "random number."
  http://en.wiktionary.org/wiki/random_number.
  Wiktionary

- "Platonic solid."
  http://en.wiktionary.org/wiki/Platonic_solid.
  Wiktionary

- "BluePlatonicDice."
  http://en.wikipedia.org/wiki/File:BluePlatonicDice.jpg.
  Wikipedia

SECTION XIV

8.XLSX - EXCEL CHALLENGE - INTEGRATING MICROSOFT OFFICE SUITE APPLICATIONS

Microsoft Excel is just one of many programs you will need to communicate your data analysis findings. Additional applications from the Office suite of products, including PowePoint and Word, are not only necessary, but can integrate easily with Excel (and vice-versa). With the newest version of Microsoft's Office, 365, you can even hyperlink spreadsheets into your documents and presentations so that they can update automatically when the source file is changed. The following lessons and quiz will help you understand how these programs work together.

8.XLSX.1 CLEANING AND RESTRUCTURING DATA IN EXCEL

Emese Felvegi; Noreen Brown; Barbara Lave; Julie Romey; Mary Schatz; Diane Shingledecker; and Robert McCarn

---

Before we can work with our data, we need to make sure it's valid, accurate, and reliable. In the age of Big Data, companies may spend just as much or more on maintaining the health and cleaning their data as they spend on collecting or purchasing it in the first place. Consider the issues that can stem from missing or wrong values, duplicates, and typos. The validity, accuracy, and reliability of your calculations depend on your ability to keep your data up-to-date. Many estimates show that about 30% of your data may become inaccurate over time (JD Supra, 2019; Strategic DB, 2019) and even small data sets can be costly to clean, let alone files that are tens or hundreds of thousands of records deep – or much more if you are using large scale databases.

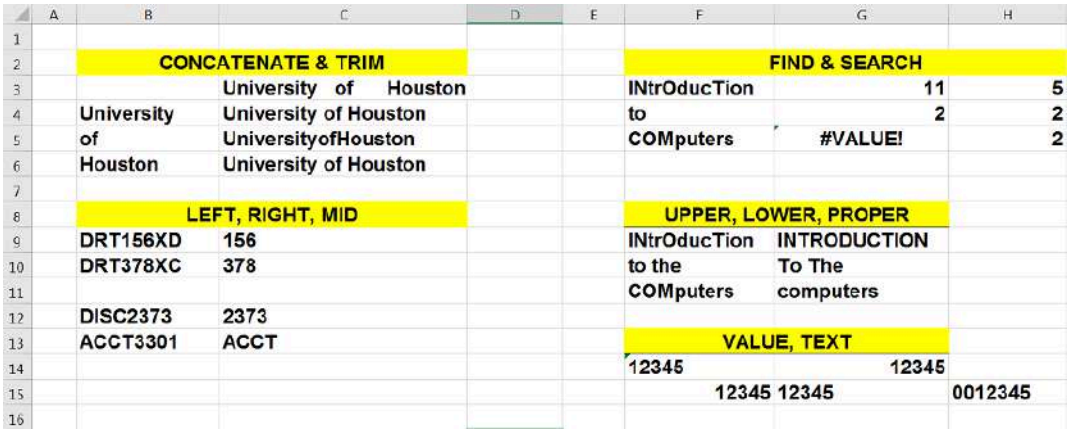
There are many data cleaning solutions out there for a wide range of file formats, data volumes, or budgets. However, there are many things we can accomplish using Excel functions and features so that you can process our data quickly and effectively. Instead of purchasing an application, assigning data cleaning to an employee, or hiring a service to scrub your data, for records under a million per sheet, Excel can save you a great deal of time and funds using a variety of functions and features. Table 10.1 shows you some important functions that can help you clean up your data.

CLEAN	Removes all nonprintable characters from text.
TRIM	Removes all spaces from text except for single spaces between words.
CONCATENATE	Join two or more text strings into one string.
LEFT	Returns a string containing a specified number of characters from the left side of a string.
RIGHT	Returns a string containing a specified number of characters from the right side of a string.
MID	Returns a specific number of characters from a text string.
SEARCH	SEARCH returns the number of the character at which a specific character or text string is first found.
FIND and FINDB	Locate one text string within a second text string.
UPPER	Converts text to uppercase.
LOWER	Converts text to lowercase.
PROPER	Capitalizes the first letter in a text string and any other letters in text that follow any character other than a letter. Converts all other letters to lowercase letters.
TEXT	Change the way a number appears by applying formatting to it with format codes.
VALUE	Converts a text string that represents a number to a number.

Table 10.1 A sample of text and data cleaning functions in Excel.

The following sections show the functions above in action. The Ch10_Data_File contains four sheets. The *Documentation* sheet notes the sources of our data. *Text_FUNC* sheet features a variety of common errors you may see in a data set, including line breaks in the wrong place, extra spaces or no spaces in between words, non-printing characters, improperly capitalized or all upper case, all lower case text, ill-formatted data values. The *DataGen_Companies* sheet contains a set of "dummy" (plausible, but not real) data about companies generated at https://www.generatedata.com/ that the author of this chapter intentionally injected with common errors seen in data in order to unfold and process it for the sake of practicing Excel functions for the Chapter Practice section. The *Mockaroo_Cars* sheet is a "dummy" dataset about consumers and their addresses generated at https://mockaroo.com/, this data set will be used for the Mail Merge section. Both of these "dummy" data sets are archived here for educational purposes.

Figure 10.1.1 below shows the *Text_FUNC* sheet with a variety of common errors seen in data you import from other sources. The CONCATENATE & TRIM range is an example of how a single line of text can be created from the contents of three rows by nesting two Excel functions. CONCATENATE on its own will merge the three cells into one, but alone, it does nothing about the extra spaces we see in the text. TRIM will remove all spaces, which means we need to add " " in order for Excel to add the needed blank cells in between words.

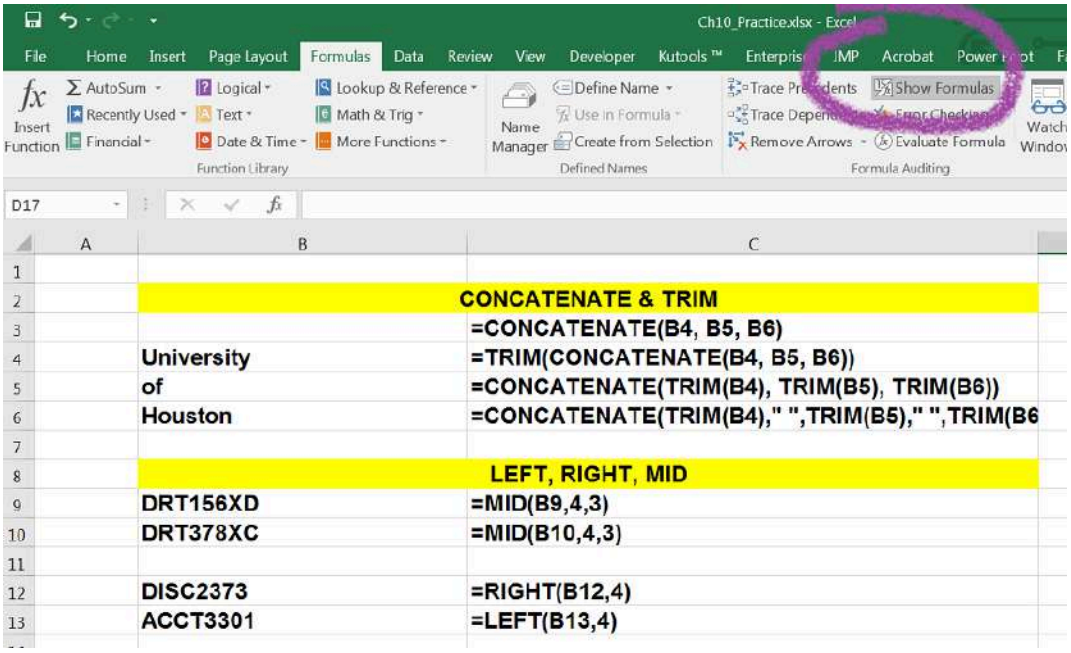


	A	B	C	D	E	F	G	H
1								
2		CONCATENATE & TRIM				FIND & SEARCH		
3			University of Houston			INtrOducTion	11	5
4		University	University of Houston			to	2	2
5		of	UniversityofHouston			COMputers	#VALUE!	2
6		Houston	University of Houston					
7								
8		LEFT, RIGHT, MID				UPPER, LOWER, PROPER		
9		DRT156XD	156			INtrOducTion	INTRODUCTION	
10		DRT378XC	378			to the	To The	
11						COMputers	computers	
12		DISC2373	2373					
13		ACCT3301	ACCT			VALUE, TEXT		
14						12345	12345	
15						12345	12345	0012345
16								

Figure 10.1.1 The Text_FUNC sheet with original and cleaned content side by side.

The LEFT, RIGHT, MID range in columns A:C illustrate another common set of functions used to process data. Oftentimes data comes in large chunks merged together. While we can use the Data > Text to Columns feature with delimiters to tell Excel where we want our data split, the LEFT, RIGHT, MID functions will process data from certain directions depending on where in the string is the text or number we wish to extract. B9 and B10 show a part number we can extract portions of using the MID function into C9, C10. B12 and B13 show course numbers we can extract portions of using the RIGHT and LEFT functions into C12, C13.

Figure 10.1.2 shows the formulas in columns A:C to illustrate the combination of CONCATENATE and TRIM nested in a variety of ways to find the best configuration to output the way we want our text to appear with the syntax for LEFT, RIGHT, and MID showing underneath.



	A	B	C
1			
2			CONCATENATE & TRIM
3			=CONCATENATE(B4, B5, B6)
4		University	=TRIM(CONCATENATE(B4, B5, B6))
5		of	=CONCATENATE(TRIM(B4), TRIM(B5), TRIM(B6))
6		Houston	=CONCATENATE(TRIM(B4)," ",TRIM(B5)," ",TRIM(B6
7			
8			LEFT, RIGHT, MID
9		DRT156XD	=MID(B9,4,3)
10		DRT378XC	=MID(B10,4,3)
11			
12		DISC2373	=RIGHT(B12,4)
13		ACCT3301	=LEFT(B13,4)

Figure 10.1.2 The Text_FUNC sheet with the "Show Formulas" option enabled for columns A:C.

Figure 10.1.3 below shows the formulas in columns F:H to illustrate the different between FIND and

SEARCH, as well as show the UPPER, LOWER, PROPER, VALUE and TEXT functions used to produce the contents for data in those ranges.

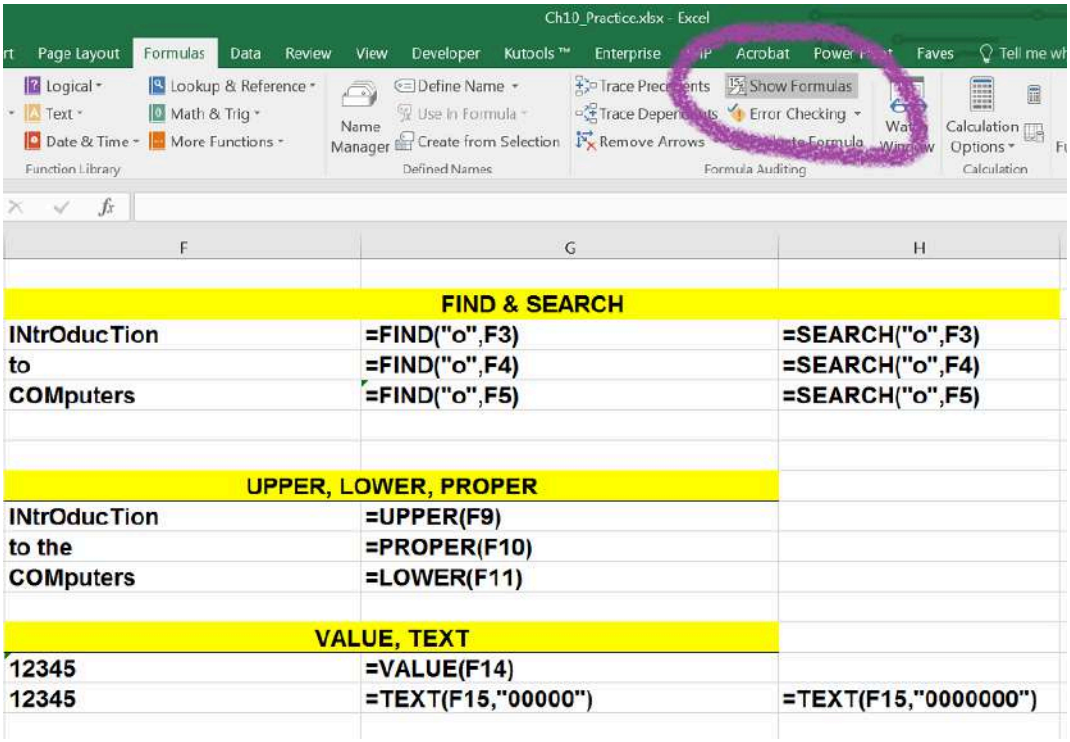


Figure 10.1.3 The Text_FUNC with the "Show Formulas" option enabled for columns F:H.

More Examples

Visit the Official Microsoft site for a list of common text functions in Excel.

Observe the variety of tasks you can achieve by using relatively simple formulas and nested alternatives.

> **"Note:** Although you can use the TEXT function to change formatting, it's not the only way. You can change the format without a formula by pressing **CTRL+1** (or ⌘ **+1** on the Mac), then pick the format you want from the **Format Cells** > **Number** dialog (Source)."

Consider possible uses of these functions in order to clean your data. We will revisit these functions and the use of delimiters in the Chapter Practice.

ATTRIBUTION

8.XLSX.2 MAIL MERGE

Emese Felvegi; Noreen Brown; Barbara Lave; Julie Romey; Mary Schatz; Diane Shingledecker; and Robert McCarn

Everyday communications between colleagues, business partners, a business and a customer, a non-profit and its donors can take many shapes or forms. Thank you notes, reminders, product updates, invoices, and many other topics may require an individual to send identical documents with small changes to each document such as the recipient's name, address, donation amount, product number, purchase date, or more. Mail merge automates the tedious task of copy-pasting a large number of data from one application to another one field at a time a hundred or a thousand times over. We can use mail merge in Word or Outlook while depending on a data source from Excel or Access and allow employees to process hundreds or thousands (or more, depending on your processing speed or patience) of records to populate fields (name, address, donation amount, etc.) in a pre-written document or email.

> "With the combination of your letter or email and a mailing list, you can create a mail merge document that sends out bulk mail to specific people or to all people on your mailing list. You also can create and print mailing labels and envelopes by using mail merge (support.office.com)."

We will use the *Mockaroo_Cars* sheet in the Ch10_Data_File in combination with a Word document to create a letter to mail to our clients regarding an extended warranty offer for their vehicle. The *Mockaroo_Cars* sheet is a "dummy" dataset about fictional consumers, their addresses, and their vehicles generated at https://mockaroo.com/. The data set generated online is archived here for educational purposes.

1. Open the Ch10_Data_File and go to the *Mockaroo_Cars* sheet.
2. Observe the field names in the header row.
    - How would these appear on a mailing label?
    - How would you add them as an address block in a letter?
    - How much time would it take for you to manually copy and paste the car_model (Column K) and the car_year (Column L) into a letter that wishes to personalize content for its recipients?
3. Open a blank Word document. Highlight, copy and then paste the following text into it. "Address|Dear [name],|We are pleased to inform you about our ongoing special regarding your [car_model, car_year].|Please contact us regarding this limited time offer and schedule a meeting with your service advisor.|Sincerely,|Mockaroo Cars." Replace the | symbols with hard line breaks using the ENTER key

to format your document to match Figure 10.2.1 below.

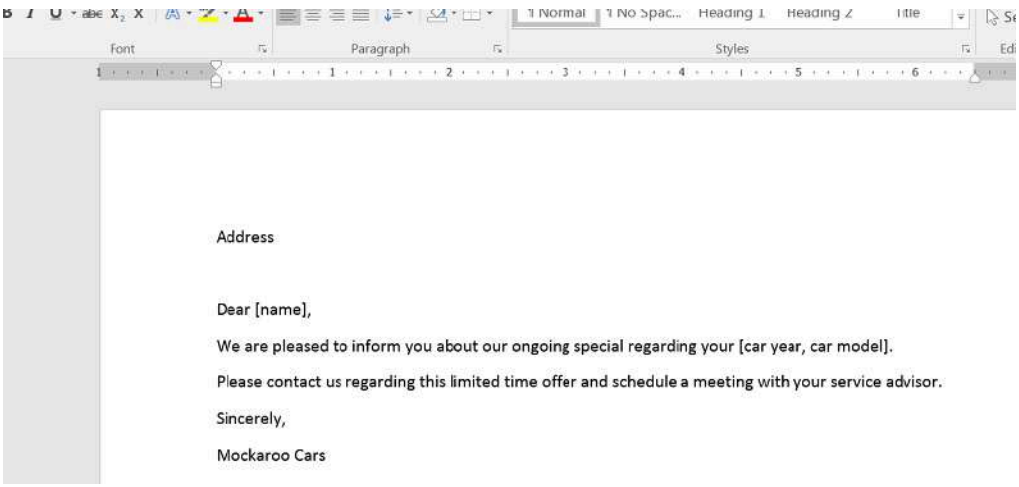


Figure 10.2.1 Word document with pre-written content.

4. Save your document as Mail_Merge_Sample.docx in the folder where you have been saving your course-related documents in a subfolder under Chapter 10.
5. Click into the Mailings Tab > Start Mail Merge > Step-by-Step Mail Merge Wizard as shown in Figure 10.2.2.

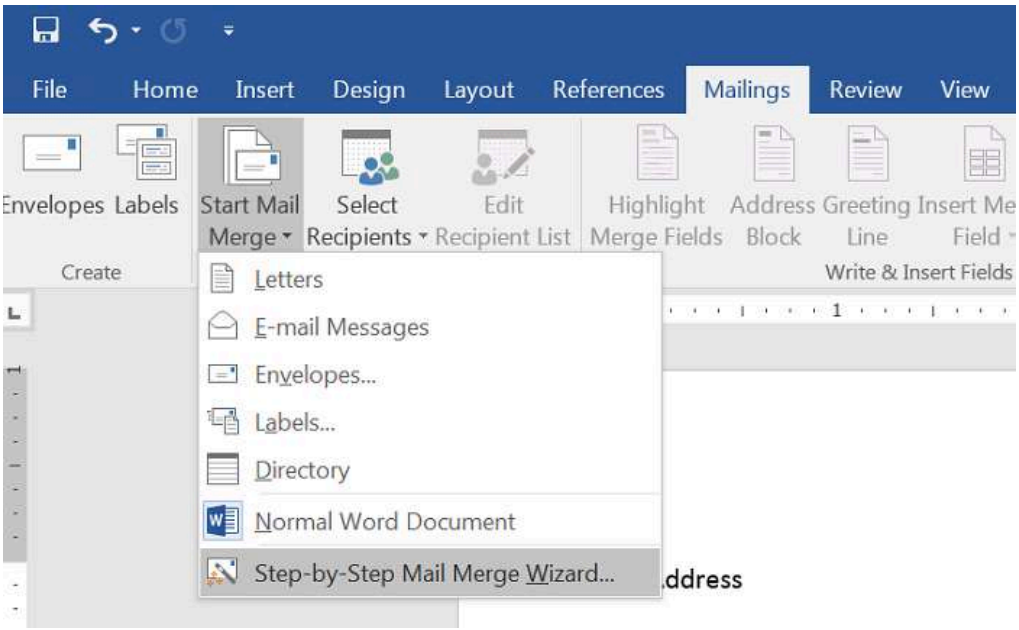


Figure 10.2.2 Starting a Mail Merge Wizard.

6. You will be asked to confirm the type of mail merge you wish to complete. In the navigation pane that appears on the right side of your window, make sure the *Letters* option is selected as the document type. Click ->*Next: Starting document* as shown in Figure 10.2.3.

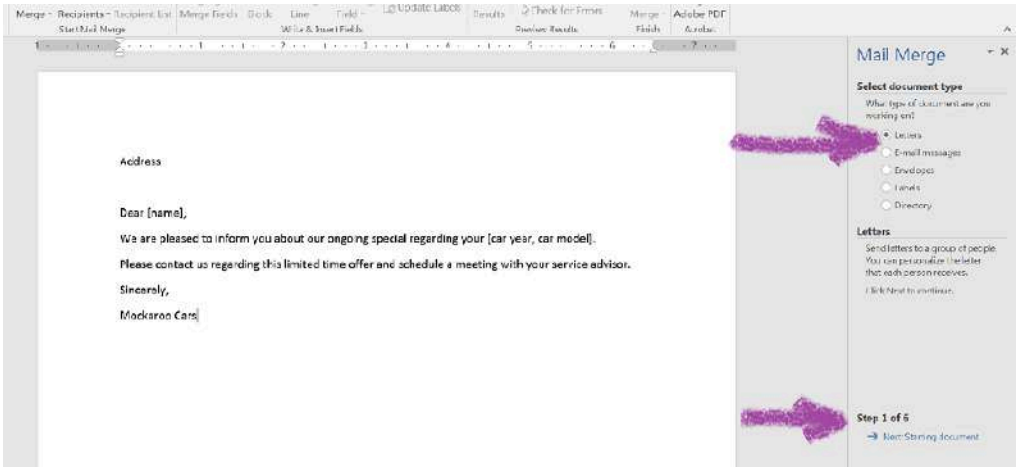


Figure 10.2.3 Selecting document type.

7. You will be asked to confirm whether you wish to use the document you have open or other sources. Select the *Use current document* option for this practice at the top of the navigation pane. At the bottom, click *Next: Select recipients* as shown in Figure 10.2.4.

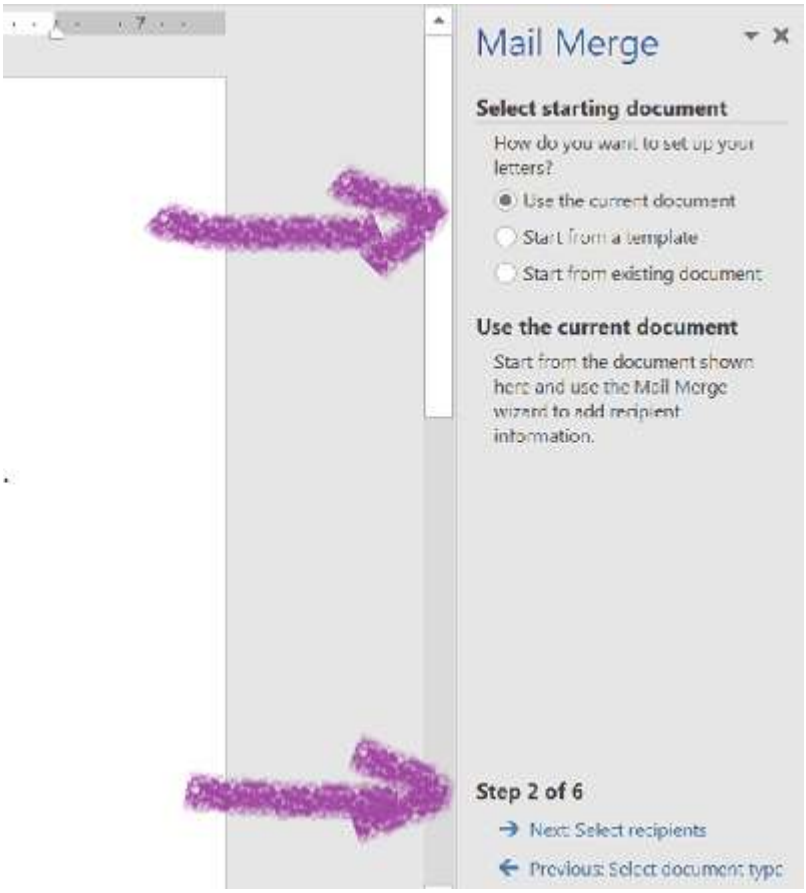


Figure 10.2.4 select your starting document.

8. Next, you will select the fields you want to use from your *Mockaroo_Cars* sheet Excel file. Under the Mail Marge pane on the right, under *Use an existing list*, (1) click *Browse* to select names and addresses from a file or database. Navigate to the folder where you downloaded the Ch10_Data_File and (2) select

the *Mockaroo_Cars* sheet. Make sure that the checkbox is selected next to the *First row of data contains column headers* option as shown in Figure 10.2.5. (3) Click Next: Write your letter.

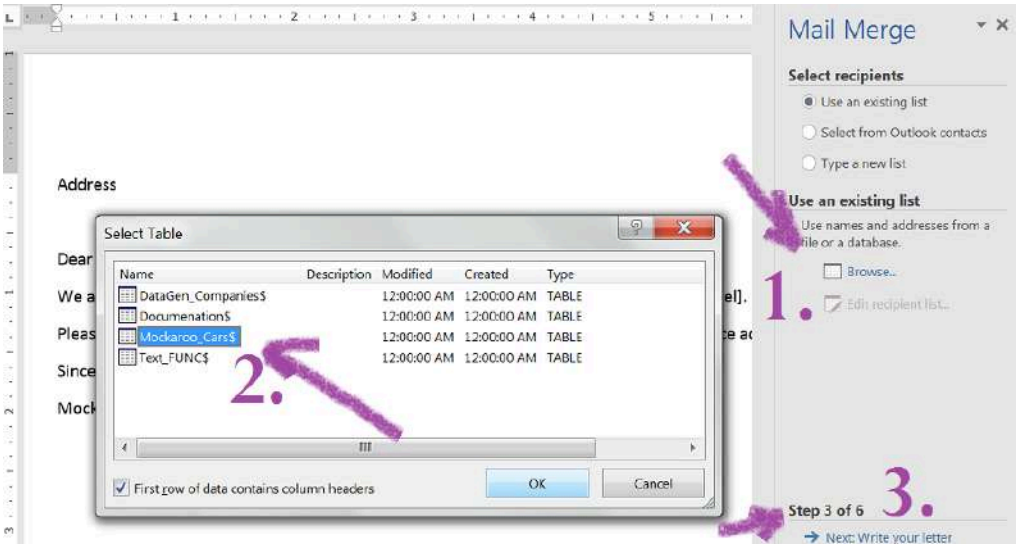


Figure 10.2.5 Selecting your data source.

9. A dialogue box will allow you to confirm the Data Source, correct sheet/fields and to make edits as needed. Click OK upon approval of the contents shown.
10. In the pane on the right, you will see the *Write your letter* options. This is the time for you to update your letter with the data from your Excel sheet to populate fields like *Address*, Dear *[name]*, *[car year, car model]* in your Word document. Click *Address block...* and preview how the default selection appears based on your data. You can use the *Match Fields...* button to call up more fields from your Excel file as shown in Figure 10.2.6.

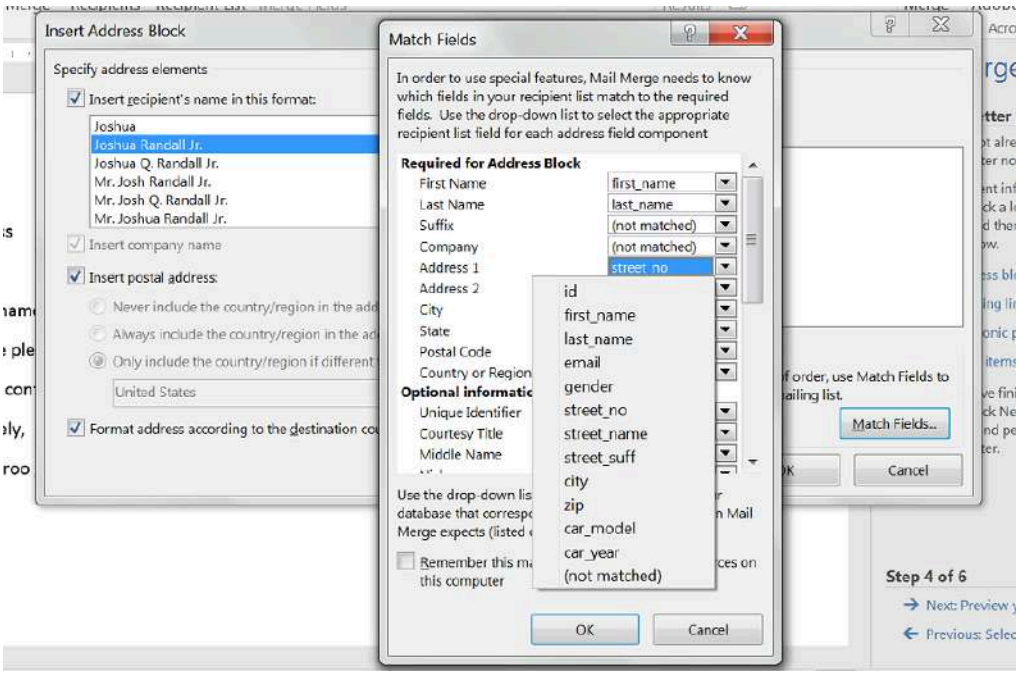


Figure 10.2.6 Matching fields from your Excel list to create the desired address block.

11. You have two options as your Excel field names don't necessarily map out onto an Address Block exactly as you would like. (1) You can insert all the fields through the Match Fields... button and dialogue box as is and then edit the line breaks manually on the next step. (2) You can also go back to your Excel sheet and use the CONCATENATE function to merge the customers' addresses into single lines (street_no, street_name, street_suff), save your Excel file, then *Browse* to select your source again. For option (2), your field name for the combined address line will show up as your *Address Block*. You can now delete the Addresses word from before *Address Block* in your Word document we used as a placeholder.
12. Click *Greeting Line...* and replace Dear [name] with an option of your choice from the available presets as shown in Figure 10.2.7.

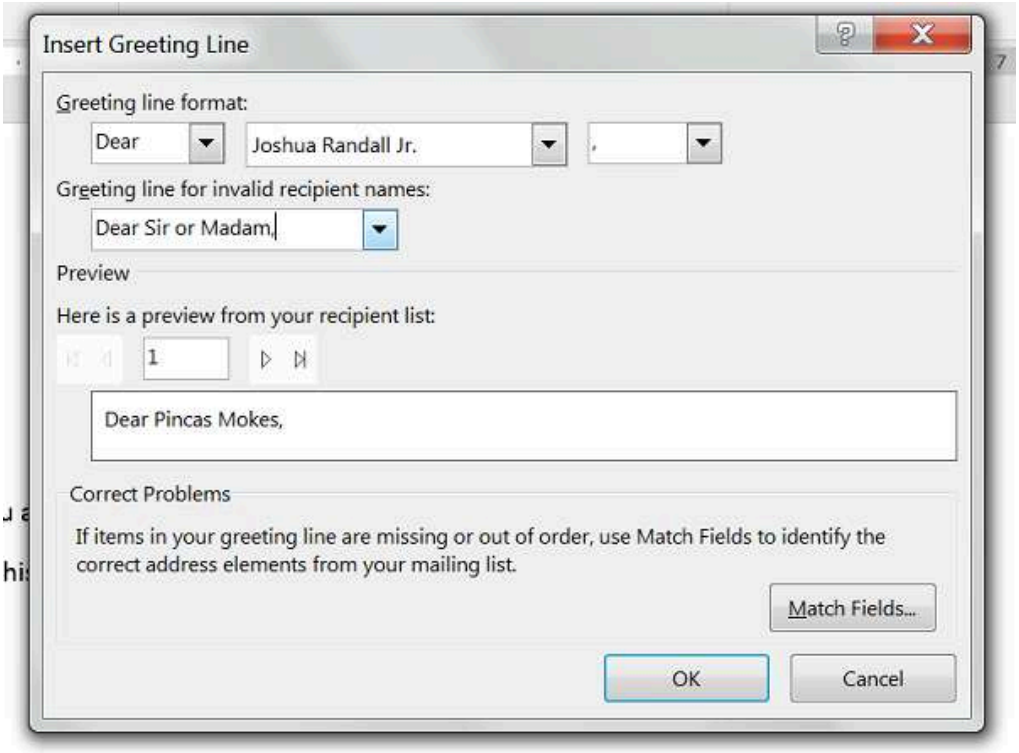


Figure 10.2.7 Greeting line options.

13. Go to More items... and insert the fields for car year and model. Your Word document should match what is shown in Figure 10.2.8 (without the yellow highlighting).

«AddressBlock»

«GreetingLine»

We are pleased to inform you about our ongoing special regarding your «car_year»«car_model».

Please contact us regarding this limited time offer and schedule a meeting with your service advisor.

Sincerely,

Mockaroo Cars

Pincas Mokes

245 Graedel Crossing

Pasirlimus

Dear Pincas Mokes,

We are pleased to inform you about our ongoing special regarding your 1990 Electra.

Please contact us regarding this limited time offer and schedule a meeting with your service advisor.

Sincerely,

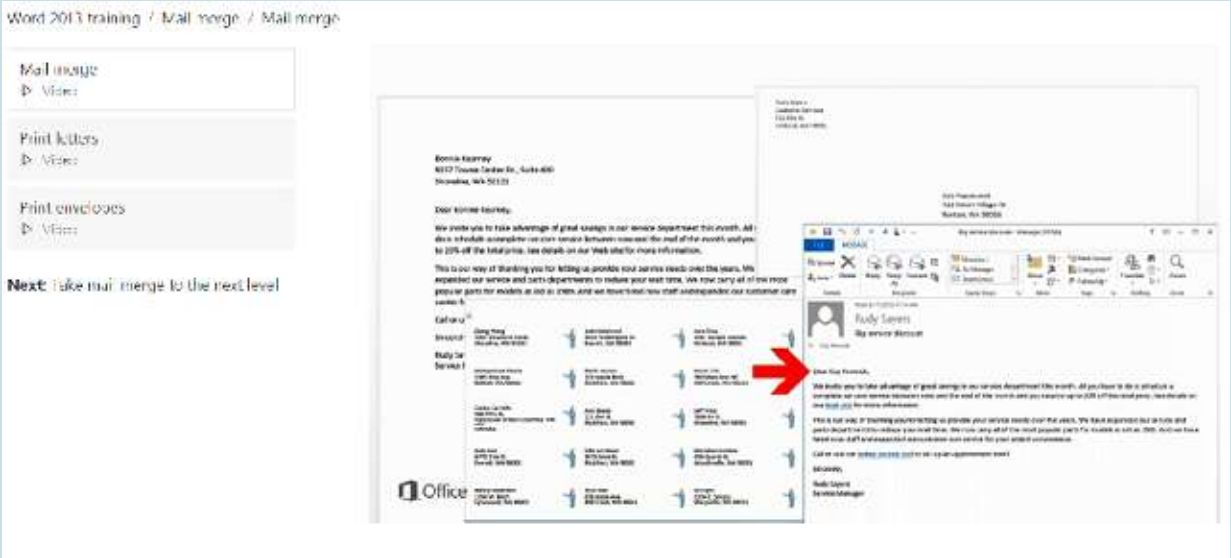
Mockaroo Cars

Figure 10.2.8 Preview of your fields inserted into your Word document.

14. You can now *Complete the merge*. It is best to save all letters you are creating as a separate file, that way you can have a record of all mailers in a New Document. You will as many pages in Word for your customers as the number of records in your Excel sheet. Depending on your processing speed and working memory, this process may take a few minutes.

Mail Merge e-Mail Exercise

Complete this 10-minute training on support.office.com to practice other forms of mail merge at the official Microsoft Office website.



ATTRIBUTION

8.XLSX.3 INTEGRATING EXCEL® WITH WORD® AND POWERPOINT®

Emese Felvegi; Noreen Brown; Barbara Lave; Julie Romey; Mary Schatz; Diane Shingledecker; and Robert McCarn

Learning Objectives

1. Learn how to paste an image of an Excel chart into a Word document.
2. Learn how to paste a link to an Excel chart into a PowerPoint slide.

Charts that are created in Excel are commonly used in Microsoft Word documents or for presentations that use Microsoft PowerPoint slides. Excel provides options for pasting an image of a chart into either a Word document or a PowerPoint slide. You can also establish a link to your Excel charts so that if you change the data in your Excel file, it is automatically reflected in your Word or PowerPoint files. We will demonstrate both methods in this section.

Pasting a Chart Image into Word

For this exercise you will need two files:

- *The Excel spreadsheet you have been working with during the Charts & Graphs chapter:* ***CH4 Charting****.*
- *A Word document data file —* *CH4 Diversity*

Excel charts can be valuable tools for explaining quantitative data in a written report. Reports that address business plans, public policies, budgets, and so on all involve quantitative data. For this example, we will assume that the Change in Enrollment Statistics Spend Source stacked column chart is being used in a student's written report (see **Figure 10.3.1**).

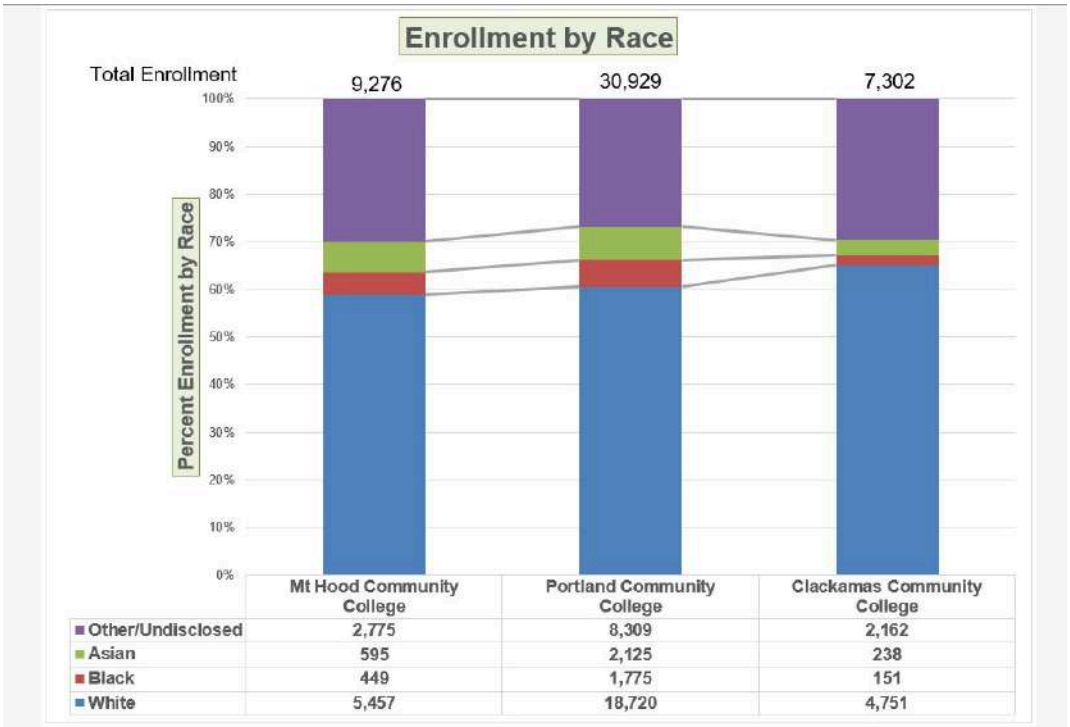


Figure 10.3.1 Completed Stacked Column Chart

The following steps demonstrate how to paste an image, or picture, of this chart into a Word document:

1. Open *CH4 Diversity*. Save it as *CH4 Diversity in Enrollment in Community Colleges*
2. Click below the figure heading in the Word document that reads: **Figure 1: Enrollment by Race**. The image of the stacked column chart will be placed below this heading.
3. If needed, open the Excel file you have been working with (*CH4 Charting)*. Activate the **Enrollment by Race** chart in the **Enrollment by Race Chart** sheet.
4. Click the down arrow on the **Copy** button in the **Home** tab of the ribbon. Select **Copy as Picture**
5. Select **OK** — Accepting the Copy Pictures defaults:

    - As shown on Screen
    - Picture

6. Go back to the *CH4 Diversity in Enrollment in Community Colleges* Word document by clicking the file in the taskbar.
7. Confirm that the insertion point is below the Figure 1: Enrollment by Race heading (see Figure 10.3.2) and click the **Paste** button in the Home tab of the ribbon (or press **Crtl-V**).

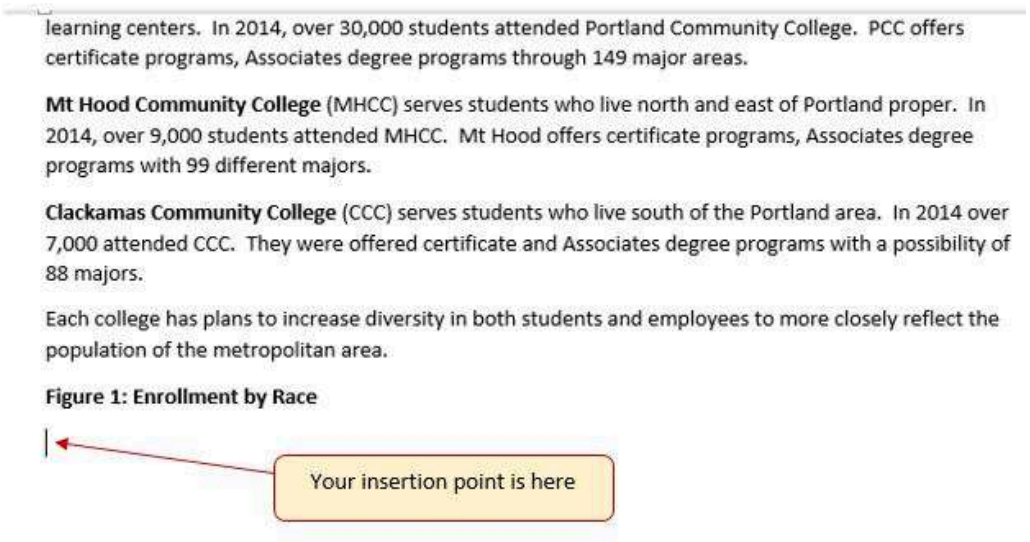


Figure 10.3.2 Paste Picture in Word

Oh no!! The picture is so big that it falls on to the next page. We will need to change its size.

1. Click anywhere on the picture of the chart to activate it.
2. Click the **Format** tab under the **Picture Tools** section of the ribbon (see **Figure 10.3.2**).
3. Click the down arrow on the **Shape Width** button in the **Size** group of commands. Continue to click the down arrow until the width of the picture is **5.4**." As you reduce the width of the picture, the height is automatically reduced as well. (The height should be ~ **3.92**″)
4. To center the chart on the page, make sure the chart is activated. Then go to the **Home** tab, to the **Paragraph** group, and select **Center.**
5. Save your work.

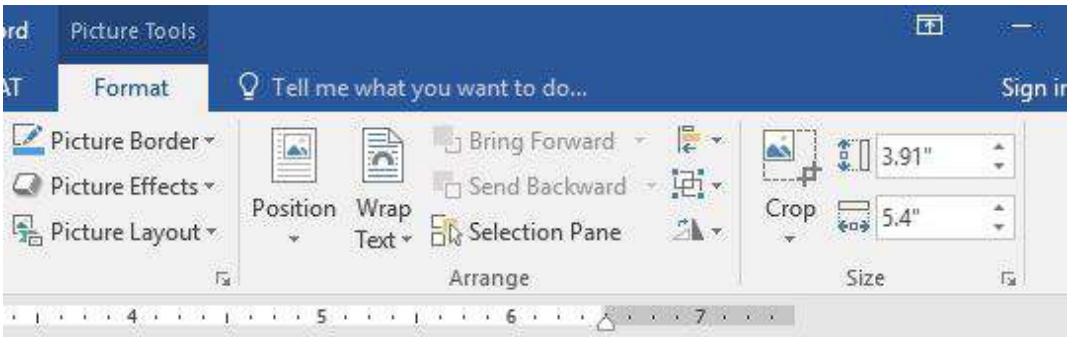


Figure 10.3.3 Changing the Size of a Picture in Word

Figure 10.3.4 shows the final appearance of the Enrollment by Race Source chart pasted into a Word document. It is best to use either the Shape Width or Shape Height buttons to reduce the size of the chart. Using either button automatically reduces the height and width of the chart in proper proportion. If you choose to use the sizing handles to resize the chart, holding the SHIFT key while clicking and dragging on a corner sizing handle will also keep the chart in proper proportion.

Figure 10.3.4 Final Appearance of Pasting a Chart Image into Word

Diversity in Enrollment in Community Colleges in the Portland Metropolitan Area.

The Portland metropolitan area benefits from a wide array of public and private colleges. By far, most students are enrolled at one of the local community colleges.

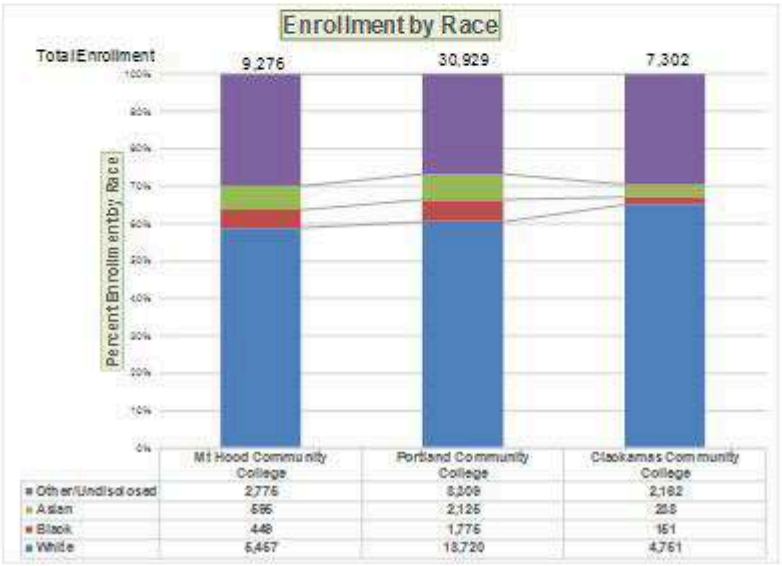
Portland Community College (PCC) is the largest, with four full fledges campuses and several smaller learning centers. In 2014, over 30,000 students attended Portland Community College. PCC offers certificate programs, Associates degree programs through 149 major areas.

Mt Hood Community College (MHCC) serves students who live north and east of Portland proper. In 2014, over 9,000 students attended MHCC. Mt Hood offers certificate programs, Associates degree programs with 99 different majors.

Clackamas Community College (CCC) serves students who live south of the Portland area. In 2014 over 7,000 attended CCC. They were offered certificate and Associates degree programs with a possibility of 88 majors.

Each college has plans to increase diversity in both students and employees to more closely reflect the population of the metropolitan area.

Figure 1: Enrollment by Race



Skill Refresher

Pasting a Chart Image into Word

1. Activate an Excel chart and click the Copy button in the Home tab of the ribbon.

2. Click on the location in the Word document where the Excel chart will be pasted.
3. Click the down arrow of the Paste button in the Home tab of the ribbon.
4. Click the Picture option from the drop-down list.
5. Click the Format tab in the Picture Tools section of the ribbon.
6. Resize the picture by clicking the up or down arrow on the Shape Width or Shape Height buttons.

Pasting a Linked Chart Image into PowerPoint

For this exercise you will need two files:

- *The Excel spreadsheet you have been working in your Charts & Graphs chapter:* ***CH4 Charting****.*
- *A PowerPoint data file –* *CH4 Diversity*.

Microsoft PowerPoint is perhaps the most commonly used tool for delivering live presentations. The charts used in a live presentation are critical for efficiently delivering your ideas to an audience. Similar to written documents, a wide range of presentations may require the explanation of quantitative data. This demonstration includes a PowerPoint slide that could be used in a presentation. We will paste the Enrollment by Race chart into this PowerPoint slide. However, instead of pasting an image, as demonstrated in the Word document, we will establish a link to the Excel file. As a result, if we change the chart in the Excel file, the change will be reflected in the PowerPoint file. The following steps explain how to accomplish this:

1. Open *CH4 Diversity.pptx*. Save it as *CH4 Diversity in Enrollment in Community Colleges*.
2. Navigate to **Slide 6 – Diversity in Enrollment**. This is the slide where you will place the linked chart.
3. If needed, open the Excel file you have been working with (*CH4 Charting)*. Activate the **Enrollment by Race** chart in the **Enrollment by Race Chart** sheet.
4. Click the down arrow on the **Copy** button in the **Home** tab of the ribbon. Select **Copy** (not Copy as Picture.)
5. Go back to the *CH4 Diversity in Enrollment in Community Colleges* presenation by clicking the file in the taskbar.
6. Make sure you are still on Slide 6 – Diversity in Enrollment. Click on the **outside edge** of the empty prompt box on the right.
7. Click the down arrow below the **Paste** button in the **Home** tab of the ribbon in the PowerPoint file.
8. Hover over each of the Paste Options until you find **Keep Source Formatting & Link Data** (see

Figure **10.3.5**). Select this option. This pastes an image of the Excel chart into the PowerPoint slide. In addition, a link is created so that any changes made to the chart (in Excel) appear on the PowerPoint slide.

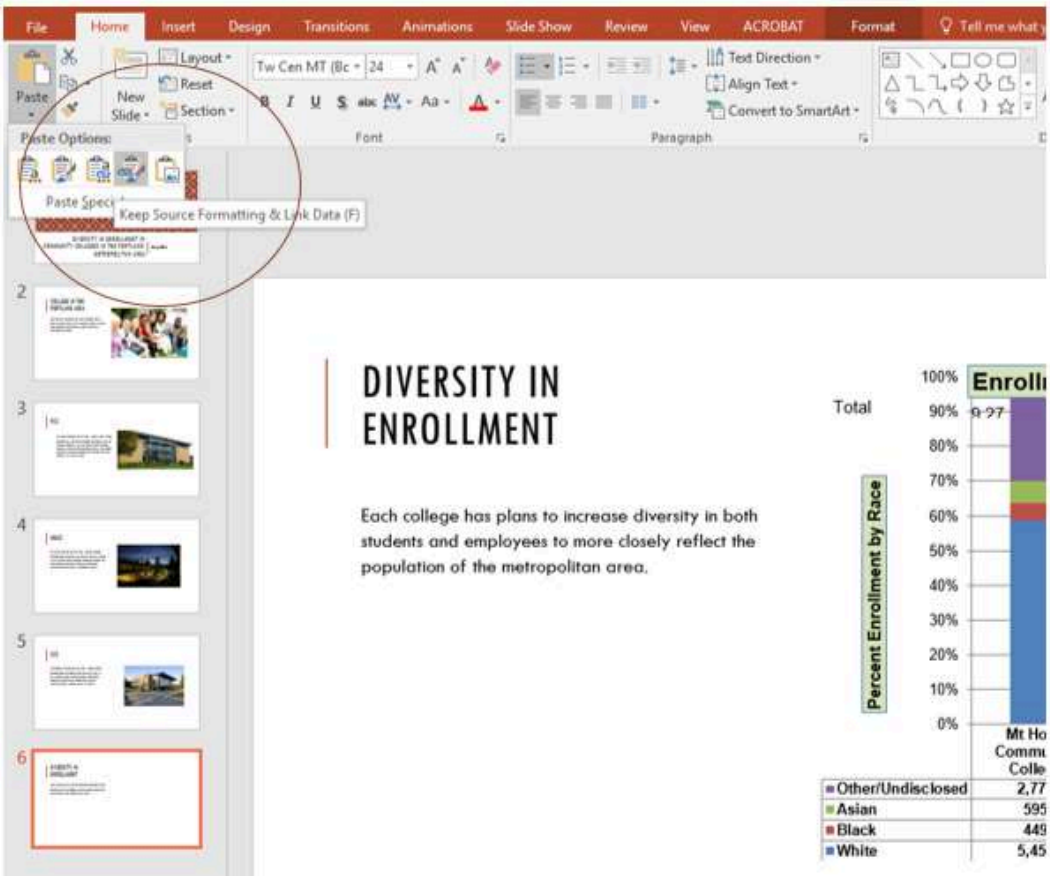


Figure 10.3.5 Creating a Link to an Excel Chart in PowerPoint

Next we need to make some changes to clean up the chart a bit. First, we are going to apply a different chart style.

1. Click anywhere in the plot area of the column chart pasted into the PowerPoint slide. You will see the same **Excel Chart Tools** tabs added to the ribbon (see **Figure 10.3.6**).
2. On the **Design** tab, select **Style 8** in the **Chart Style** group.

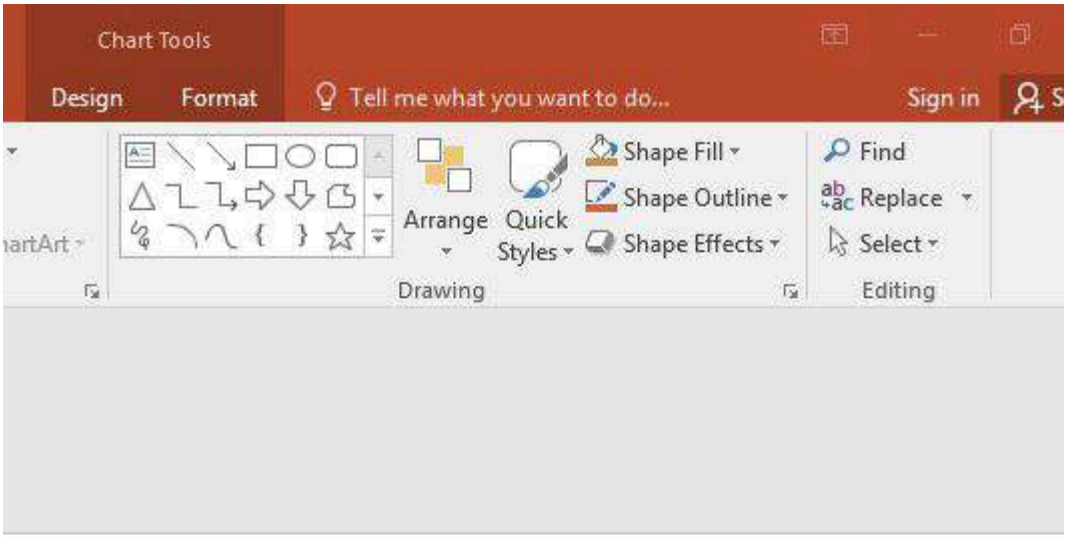
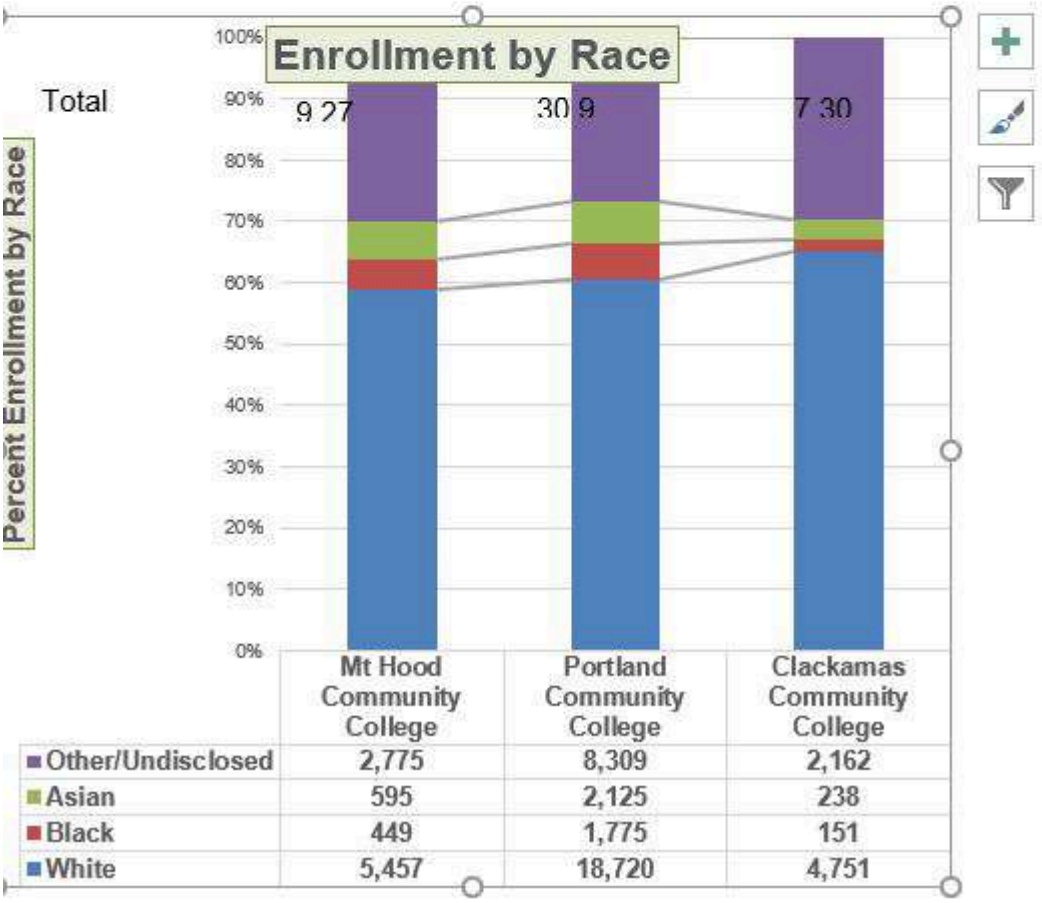


Figure 10.3.6 Modifying and Excel Chart Pasted into a PowerPoint Slide



Paste linking this chart caused trouble with the text boxes we added, so next, we are going to delete them.

1. Select each text box by clicking on the outside edge of the text box with the four-headed arrow. Press the

delete key on your keyboard. Be sure that the insertion point is NOT blinking inside the text box. If it is, you will be editing the contents of the text box instead of deleting the actual text box.

The benefit of adding this chart to the presentation as a link is that it will automatically update when you change the data in the linked spreadsheet file.

1. Return to your *CH4 Charting* Excel file.
2. Select the **Enrollment Statistics** worksheet (the one with the Enrollment data.) Change the value in cell D3 to **1000**. You have just changed the number of white students at Clackamas Community College to **1000**. This isn't true, but you want to change the data enough to see the effect in the charts.
3. Select the **Enrollment by Race Chart** worksheet. Notice how the chart has changed.
4. Return to the Diversity in Enrollment in Community Colleges PowerPoint file by clicking the file in the taskbar.
5. On Slide 6, you should see the updated chart.
6. If the chart has not changed; be sure that your chart is selected, click the **Design** tab in the **Chart Tools** section of the ribbon. Click the **Refresh Data** button. The change made in the Excel workbook is now reflected on the PowerPoint slide.
7. If that still doesn't work, you may have created a "normal" link — instead of a **Paste Link**. Delete the chart and follow the steps again. Start from the beginning of this section.
8. Save your work. You will submit both the Word and PowerPoint files, along with the Excel file, at the end of the next section.

Figure 10.3.7 shows the appearance of the column chart after the change was made in the Enrollment Statistics worksheet in the Excel file. Note that the Data Chart at the bottom reflects the new number, too. The change that was made in the Excel file will appear in the PowerPoint file after clicking the Refresh Data button.

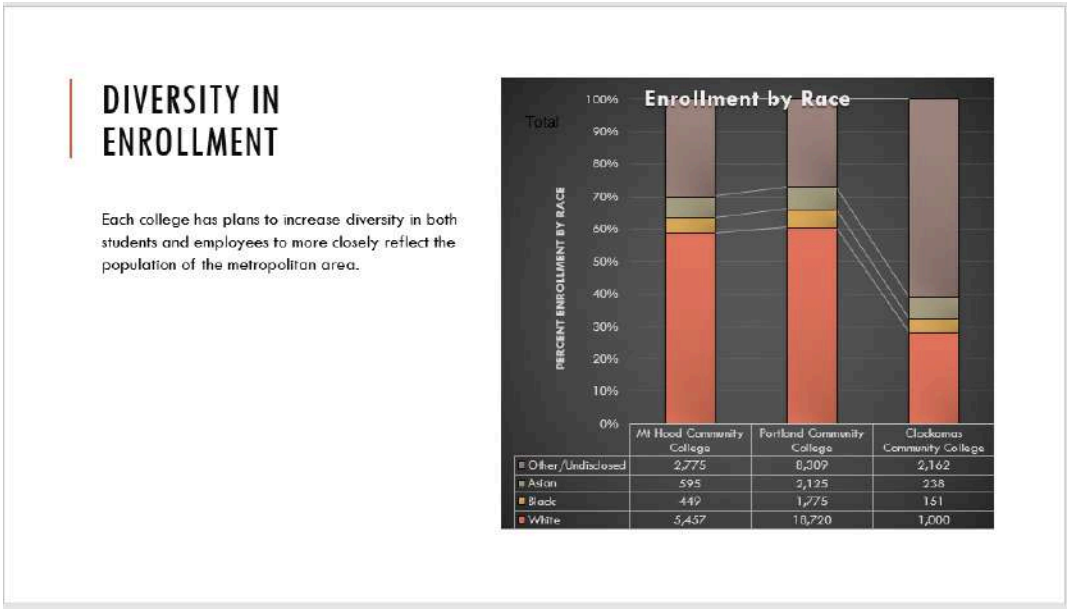


Figure 10.3.7 Styled and Updated Chart.

Integrity Check

Refreshing Linked Charts in PowerPoint and Word

When creating a link to a chart in Word or PowerPoint, you must refresh the data if you make any changes in the Excel workbook. This is especially true if you make changes in the Excel file prior to opening the Word or PowerPoint file that contains a link to a chart. To refresh the chart, make sure it is activated, then click the Refresh Data button in the Design tab of the ribbon. Forgetting this step can result in old or erroneous data being displayed on the chart.

Integrity Check

Severed Link?

When creating a link to an Excel chart in Word or PowerPoint, you must keep the Excel

workbook in its original location on your computer or network. If you move or delete the Excel workbook, you will get an error message when you try to update the link in your Word or PowerPoint file. You will also get an error if the Excel workbook is saved on a network drive that your computer cannot access. These errors occur because the link to the Excel workbook has been severed. Therefore, if you know in advance that you will be using a USB drive to pull up your documents or presentation, move the Excel workbook to your USB drive before you establish the link in your Word or PowerPoint file.

Skill Refresher:

Pasting a Linked Chart Image into PowerPoint

1. Activate an Excel chart and click the Copy button in the Home tab of the ribbon.
2. Click in the PowerPoint slide where the Excel chart will be pasted.
3. Click the down arrow of the Paste button in the Home tab of the ribbon.
4. Click the Keep Source Formatting & Link Data option from the drop-down list.
5. Click the Refresh Data button in the Design tab of the ribbon to ensure any changes in the Excel file are reflected in the chart.

Key Takeaways

- When pasting an image of an Excel chart into a Word document or PowerPoint file, use the **Picture** option from the Paste drop-down list of options – if you want the image to act as an image. You will not be able to make any changes to the content of the picture.
- When creating a link to a chart in Word or PowerPoint, you may need to refresh the data if you make any changes in the originating spreadsheet. You should not use the **Picture** option.

Attribution

8.XLSX.4 CHAPTER PRACTICE

Emese Felvegi; Noreen Brown; Barbara Lave; Julie Romey; Mary Schatz; Diane Shingledecker; and Robert McCarn

---

To expand your understanding of the material covered in the chapter, complete the following assignment. You will be working with the *DataGen_Companies* sheet in your Ch10_Data_File workbook. As noted before, the *DataGen_Companies* sheet contains a set of "dummy" (plausible, but not real) data about companies generated at https://www.generatedata.com/ that the author of this chapter intentionally injected with common errors seen in data in order to unfold and process it for the sake of practicing Excel functions for the Chapter Practice section. Our goal is to clean and restructure that data using functions and features discussed earlier in this chapter.

1. Open the Ch10_Data_File workbook and examine the data in the *DataGen_Companies* sheet.
   – What issues do you see with this data?
   – What is present, what is missing: what do we need to delete and what do we need to add?
   – Currently, all our data is in a single cell for each company. We want to have the company name in one column, their street address in another, their city, their ZIP code in others. Altogether we wish to have the data chopped up into segments that correspond with how we may want to use them in the future and align with categories generally associated with mailing addresses.
2. Highlight column A, where all your data is, then go to the Text to Column feature under Data > Data Tools on the ribbon.
3. The Convert Text to Columns Wizard pops up and will guide you through the process of converting a single cell into multiple ones based on where commas or any other recurring characters or patterns may be in your data. Each category in your data (company name, street address, city, ZIP code) is separated from one another using a comma. Click the Delimited checkbox, then Click next.
4. On Step 2 of 3 of the wizard, you are asked to select the delimited present in your data Excel can use to process the conversion. Your text has a comma in between the categories you want to display in individual cells, so select the Comma option by checking the box next to that option. The data preview will show vertical lines where your columns will be inserted (Figure 10.4.1). Click next after confirming that the text would convert as you like.

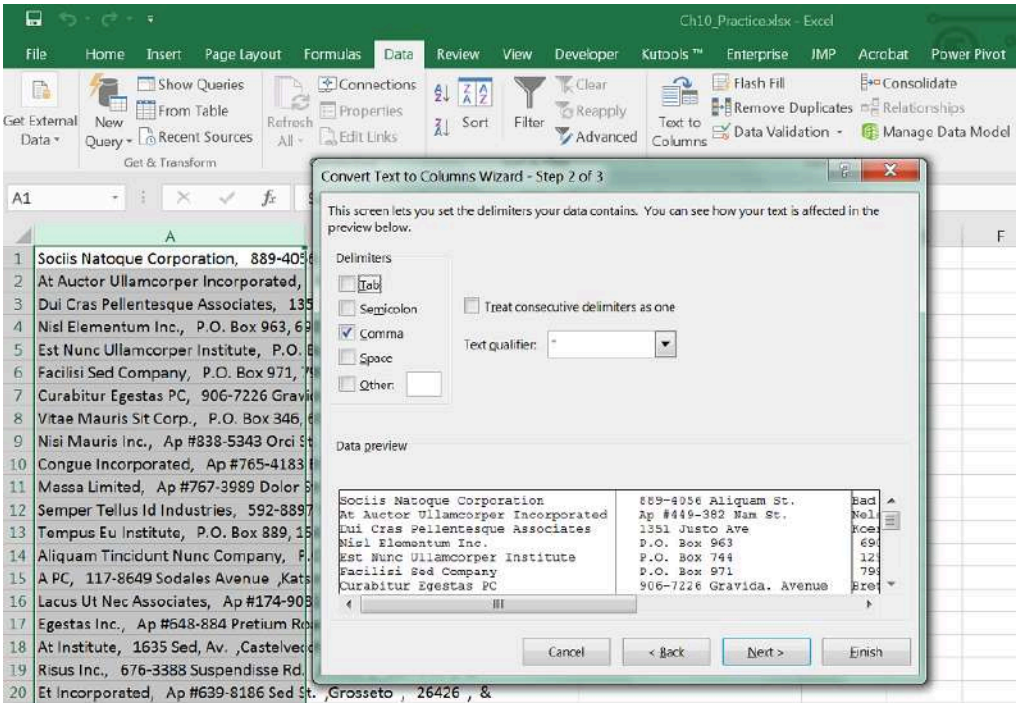


Figure 10.4.1 Setting Comma as a Delimiter for your Convert Text to Columns Wizard.

5. Your data will now display over 6 columns instead of one, with company names in column A and with & in column F. Even though we used the feature correctly, our conversion is not perfect because of the different types of addresses present. Some businesses have a street or apartment number, some have a P.O. Box number in a cell preceding their street address. Businesses with a P.O. Box number have one more cell's worth of data than others (Figure 10.4.2).

	A	B	C	D	E	F
1	Sociis Natoque Corporation	889-4056 Aliquam St.	Bad Nauheim	3084	&	
2	At Auctor Ullamcorper Incorporated	Ap #449-382 Nam St.	Nelson	5572 GW	&	
3	Dui Cras Pellentesque Associates	1351 Justo Ave	Koersel	4162	&	
4	Nisl Elementum Inc.	P.O. Box 963	6905 Convallis Road	Shippagan	52670-3	&
5	Est Nunc Ullamcorper Institute	P.O. Box 744	1298 Lobortis Av.	Spokane	J7B 3S6	&
6	Facilisi Sed Company	P.O. Box 971	7999 In Rd.	Huntly	6806	&
7	Curabitur Egestas PC	906-7226 Gravida. Avenue	Breton	5694	&	
8	Vitae Mauris Sit Corp.	P.O. Box 346	6593 Vulputate Rd.	Reading	958	&
9	Nisi Mauris Inc.	Ap #838-5343 Orci St.	Ramara	51650	&	

Figure 10.4.2 Our data is mismatched because of different types of business addresses.

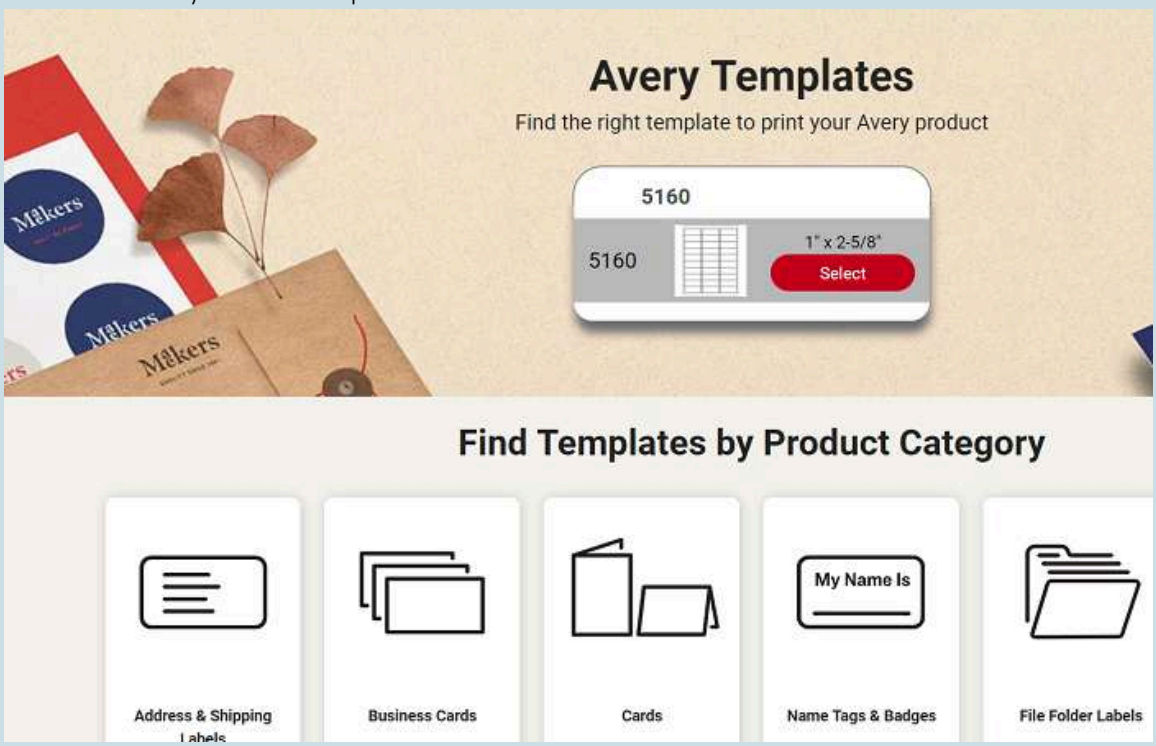
6. Let us consolidate the addresses into a single cell for the sake of consistency. Sort column B in Ascending Order to have all the street, apartment and P.O.Box addresses below one another by type.
7. Select the range that contains the P.O. Box numbers. Insert cells to Shift Cells Right.
8. In the blank range, use CONCATENATE to merge the P.O. Box numbers with the street address.
9. Move cell contents to ensure that the City and Zip codes are in the adjacent range without blanks in between.
10. Sort your data to resolve issues with street names with periods or other issues you may see with the data set.
11. Delete Column E with the superfluous & symbol.

12. Save your work for your records.

Mail Merge: Printing Mailing Labels Exercise

"One of the most popular Avery label sizes is 2.625in x 1in which is the white label 5160. It is available as 30 labels per page and is used for addressing and mailing purposes. It is one of the most important mailing labels and its layout has been copied by many other manufacturers (Streetdirectory.com). "

- Go to avery.com and examine the wide range of labels available for purchase at one of the most commonly used office products.



- Observe all the other types of labels or mailers available from Avery.
- Search for and download 5160.
- Use this template to create mailing labels from your address lists.

ATTRIBUTION

Chapter by Emese Felvégi. CC BY-NC-SA 3.0. Dummy data set from https://mockaroo.com archived here for educational purposes.

8.XLSX.5 CHAPTER ASSESSMENT

Emese Felvegi; Noreen Brown; Barbara Lave; Julie Romey; Mary Schatz; Diane Shingledecker; and Robert McCarn

The following are **sample questions** to test whether you know, understand, and are able to apply your learning from this chapter.

1. A common process to create mailing labels or marketing materials combining Excel and Word is:
a) Charts
b) Macros
c) Mail Merge
d) PivotTables
e) Templates

2. The fastest way to create mailing labels is through using the built-in:
a) PivotTable
b) Script
c) Warlock
d) Wizard

3. This function removes all spaces from text except for single spaces between words.
a) CONCATENATE
b) CLEAN
c) TRIM

d) VALUE

4. This function capitalizes the first letter in a text string and any other letters in text that follow any character other than a letter. Converts all other letters to lowercase letters.
a) CAPITAL
b) LOWER
c) PROPER
d) UPPER

5. Which function produces C10 from B10?

	A	B	C
9			
10		DISC2373	DISC
11			2373
12			
13		ACCT3301	ACCT
14			3301

a) LEFT
b) RIGHT
c) MID
d) EXTRACT

6. Which function produces C11 from B10?

	A	B	C
9			
10		DISC2373	DISC
11			2373
12			
13		ACCT3301	ACCT
14			3301

a) LEFT
b) RIGHT
c) MID
d) EXTRACT

7. The second argument in the MID function specifies where the extraction should start.
TRUE/FALSE

8. =CONCATENATE(TRIM(B3)," ",TRIM(B4)," ",TRIM(B5)) does what?
a) Merges cells B3, B4, B5
b) Clears spaces from B3, B4, B5
c) Extracts characters from B3, B4, B5
d) Chops up B3, B4, B5
e) Both a and b
f) Both c and d
g) Both a and d

9. If an Excel chart is linked into a Word document, what file(s) must be transferred in order for another person to work with the information completely?
a) Word document file only
b) Excel spreadsheet file only
c) Both the Word document and Excel spreadsheet files
d) Neither the Word document or the Excel spreadsheet file

10. What happens when a standard copy/paste procedure is used to insert an Excel chart into a Word document?
a) it creates a table in Word which can only be revised using Word capabilities
b) it creates a separate object in Word which can only be revised by double-clicking and opening Excel capabilities
c) it causes updates in one location to update the other location
d) it creates a permanent link between the spreadsheet and the document
e) all of the above

Solutions: [redacted]

---

you can also test your knowledge of the functions using the quizlet below.

https://quizlet.com/414448350/flashcards/embed?i=24veoc&x=1jj1[/embed]

ATTRIBUTION

GLOSSARY OF KEY TERMS

aggregate

a mass, assemblage, or sum of particulars; something consisting of elements but considered as a whole

arithmetic mean

the measure of central tendency of a set of values computed by dividing the sum of the values by their number; commonly called the mean or the average

average

any measure of central tendency, especially any mean, the median, or the mode

Bayes' factor

The ratio of the conditional probabilities of the event B given that A_1 is the case or that A_2 is the case, respectively.

bell curve

In mathematics, the bell-shaped curve that is typical of the normal distribution. A symmetrical bell-shaped curve that represents the distribution of values, frequencies, or probabilities of a set of data. It slopes downward from a point in the middle corresponding to the mean value, or the maximum probability. Data that reflect the aggregate outcome of large numbers of unrelated events tend to result in bell curve distributions. (Dictionary.com, 2021)

bellwether

anything that indicates future trends

bias

(Uncountable) Inclination towards something; predisposition, partiality, prejudice, preference, predilection.

bivariate

Having or involving exactly two variables.

box plot

A graphical summary of a numerical data sample through five statistics: median, lower quartile, upper quartile, and some indication of more extreme upper and lower values.

box-and-whisker plot

a convenient way of graphically depicting groups of numerical data through their quartiles

breakdown point

the number or proportion of arbitrarily large or small extreme values that must be introduced into a batch or sample to cause the estimator to yield an arbitrarily large result

breeding

the process through which propagation, growth, or development occurs

causality

the relationship between an event (the cause) and a second event (the effect), where the second event is understood as a consequence of the first

census

an official count of members of a population (not necessarily human), usually residents or citizens in a particular region, often done at regular intervals

central limit theorem

The theorem that states: If the sum of independent identically distributed random variables has a finite variance, then it will be (approximately) normally distributed.

central tendency

a term that relates the way in which quantitative data tend to cluster around some value

chance variation

the presence of chance in determining the variation in experimental results

chi-squared test

In probability theory and statistics, refers to a test in which the chi-squared distribution (also chi-square or χ-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.

chromosome

A structure in the cell nucleus that contains DNA, histone protein, and other structural proteins.

cluster

a significant subset within a population

coefficient of variation

The ratio of the standard deviation to the mean.

combinatorics

A branch of mathematics that studies (usually finite) collections of objects that satisfy specified criteria.

conditional probability

The probability that an event will take place given the restrictive assumption that another event has taken place, or that a combination of other events has taken place

confidence interval

A type of interval estimate of a population parameter used to indicate the reliability of an estimate.

confounding variable

an extraneous variable in a statistical model that correlates (positively or negatively) with both the dependent variable and the independent variable

contingency table

a table presenting the joint distribution of two categorical variables

continuous random variable

obtained from data that can take infinitely many values

continuous variable

a variable that has a continuous distribution function, such as temperature

control

a separate group or subject in an experiment against which the results are compared where the primary variable is low or nonexistence

control group

the group of test subjects left untreated or unexposed to some procedure and then compared with treated subjects in order to validate the results of the test

correlation

One of the several measures of the linear statistical relationship between two random variables, indicating both the strength and direction of the relationship.

critical thinking

the application of logical principles, rigorous standards of evidence, and careful reasoning to the analysis and discussion of claims, beliefs, and issues

cross tabulation

a presentation of data in a tabular form to aid in identifying a relationship between variables

cumulative relative frequency

the accumulation of the previous relative frequencies

data mining

a technique for searching large-scale databases for patterns; used mainly to find previously unknown correlations between variables that may be commercially useful

density

the probability that an event will occur, as a function of some observed variable

dependent variable

in an equation, the variable whose value depends on one or more variables in the equation

descriptive statistics

A branch of mathematics dealing with summarization and description of collections of data sets, including the concepts of arithmetic mean, median, and mode.

deviation

For interval variables and ratio variables, a measure of difference between the observed value and the mean.

dichotomous

dividing or branching into two pieces

discrete random variable

obtained by counting values for which there are no in-between values, such as the integers 0, 1, 2, ...

discrete variable

a variable that takes values from a finite or countable set, such as the number of legs of an animal

disjoint

Having no members in common; having an intersection equal to the empty set.

disparity

the state of being unequal; difference

dispersion

the degree of scatter of data

distribution

the set of relative likelihoods that a variable will have a value in a given interval

ellipsis

a mark consisting of three periods, historically with spaces in between, before, and after them "... ", nowadays a single character … (used in printing to indicate an omission)

empirical

verifiable by means of scientific experimentation

empirical rule

That a normal distribution has 68% of its observations within one standard deviation of the mean, 95% within two, and 99.7% within three.

equiprobable

having an equal chance of occurring mathematically

event

A subset of the sample space.

evolution

a gradual directional change, especially one leading to a more advanced or complex form; growth; development

exhaustive

including every possible element

expected value

of a discrete random variable, the sum of the probability of each possible outcome of the experiment multiplied by the value itself

experiment

A test under controlled conditions made to either demonstrate a known truth, examine the validity of a hypothesis, or determine the efficacy of something previously untried.

exploratory data analysis

an approach to analyzing data sets that is concerned with uncovering underlying structure, extracting important variables, detecting outliers and anomalies, testing underlying assumptions, and developing models

finite

limited, constrained by bounds, having an end

frequency

number of times an event occurred in an experiment (absolute frequency)

frequency distribution

a representation, either in a graphical or tabular format, which displays the number of observations within a given interval

gene

a unit of heredity; a segment of DNA or RNA that is transmitted from one generation to the next, and that carries genetic information such as the sequence of amino acids for a protein

gradient

of a function $y = f(x)$ or the graph of such a function, the rate of change of y with respect to x, that is, the amount by which y changes for a certain (often unit) change in x

graph

A diagram displaying data; in particular one showing the relationship between two or more quantities, measurements or indicative numbers that may or may not have a specific mathematical formula relating them to each other.

heterogeneous

diverse in kind or nature; composed of diverse parts

histogram

a representation of tabulated frequencies, shown as adjacent rectangles, erected over discrete intervals (bins), with an area equal to the frequency of the observations in the interval

independence

The occurrence of one event does not affect the probability of the occurrence of another.

independent

Not dependent; not contingent or depending on something else; free.

independent event

the fact that A occurs does not affect the probability that B occurs

independent variable

in an equation, any variable whose value is not dependent on any other in the equation

inferential statistics

A branch of mathematics that involves drawing conclusions about a population based on sample data drawn from it.

integral

the limit of the sums computed in a process in which the domain of a function is divided into small subsets and a possibly nominal value of the function on each subset is multiplied by the measure of that subset, all these products then being summed

intercept

the coordinate of the point at which a curve intersects an axis

interquartile range

The difference between the first and third quartiles; a robust measure of sample dispersion.

labor force

The collective group of people who are available for employment, whether currently employed or unemployed (though sometimes only those unemployed people who are seeking work are included).

line

a path through two or more points (compare 'segment'); a continuous mark, including as made by a pen; any path, curved or straight

linear regression

an approach to modeling the relationship between a scalar dependent variable y and one or more explanatory variables denoted x.

logarithm

for a number x, the power to which a given base number must be raised in order to obtain x

margin of error

An expression of the lack of precision in the results obtained from a sample.

mean squared error

A measure of the average of the squares of the "errors"; the amount by which the value implied by the estimator differs from the quantity to be estimated.

median

the numerical value separating the higher half of a data sample, a population, or a probability distribution, from the lower half

mode

the most frequently occurring value in a distribution

Monte Carlo simulation

a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results–i.e., by running simulations many times over in order to calculate those same probabilities

multiplication rule

The probability that A and B occur is equal to the probability that A occurs times the probability that B occurs, given that we know A has already occurred.

mutually exclusive

describing multiple events or states of being such that the occurrence of any one implies the non-occurrence of all the others

nominal

Having values whose order is insignificant.

non-response

the absence of a response

non-response bias

Occurs when the sample becomes biased because some of those initially selected refuse to respond.

normal distribution

A family of continuous probability distributions such that the probability density function is the normal (or Gaussian) function.

nuisance parameters

any parameter that is not of immediate interest but which must be accounted for in the analysis of those parameters which are of interest; the classic example of a nuisance parameter is the variance $sigma^2$, of a normal distribution, when the mean, mu, is of primary interest

null hypothesis

A hypothesis set up to be refuted in order to support an alternative hypothesis; presumed true until statistical evidence in the form of a hypothesis test indicates otherwise.

objective

not influenced by the emotions or prejudices

observational study

a study drawing inferences about the possible effect of a treatment on subjects, where the assignment of subjects into a treated group versus a control group is outside the control of the investigator

odds

the ratio of the probabilities of an event happening to that of it not happening

ordinal

Of a number, indicating position in a sequence.

outcome

One of the individual results that can occur in an experiment.

outlier

a value in a statistical sample which does not fit a pattern that describes most other data points; specifically, a value that lies 1.5 IQR beyond the upper or lower quartile

Pareto chart

a type of bar graph where where the bars are drawn in decreasing order of frequency or relative frequency

Pareto distribution

The Pareto distribution, named after the Italian economist Vilfredo Pareto, is a power law probability distribution that is used in description of social, scientific, geophysical, actuarial, and many other types of observable phenomena.

partition

a part of something that had been divided, each of its results

peer review

the scholarly process whereby manuscripts intended to be published in an academic journal are reviewed by independent researchers (referees) to evaluate the contribution, i.e. the importance, novelty and accuracy of the manuscript's contents

percentile

any of the ninety-nine points that divide an ordered distribution into one hundred parts, each containing one per cent of the population

pictogram

a picture that represents a word or an idea by illustration; used often in graphs

pip

one of the spots or symbols on a playing card, domino, die, etc.

placebo

an inactive substance or preparation used as a control in an experiment or test to determine the effectiveness of a medicinal drug

placebo effect

the tendency of any medication or treatment, even an inert or ineffective one, to exhibit results simply because the recipient believes that it will work

Platonic solid

any one of the following five polyhedra: the regular tetrahedron, the cube, the regular octahedron, the regular dodecahedron and the regular icosahedron

plot

a graph or diagram drawn by hand or produced by a mechanical or electronic device

polynomial

An expression consisting of a sum of a finite number of terms: each term being the product of a constant coefficient and one or more variables raised to a non-negative integer power.

population

a group of units (persons, objects, or other items) enumerated in a census or from which a sample is drawn

probability

The relative likelihood of an event happening.

probability density function

any function whose integral over a set gives the probability that a random variable has a value in that set

probability distribution

A function of a discrete random variable yielding the probability that the variable will have a given value.

probability sample

a sample in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined

probability theory

The mathematical study of probability (the likelihood of occurrence of random events in order to predict the behavior of defined systems).

prognostic

a sign by which a future event may be known or foretold

prosecutor's fallacy

A fallacy of statistical reasoning when used as an argument in legal proceedings.

public opinion polls

surveys designed to represent the beliefs of a population by conducting a series of questions and then extrapolating generalities in ratio or within confidence intervals

purposive sampling

occurs when the researchers choose the sample based on who they think would be appropriate for the study; used primarily when there is a limited number of people that have expertise in the area being researched

quadrennial

happening every four years

qualitative

of descriptions or distinctions based on some quality rather than on some quantity

qualitative analysis

The numerical examination and interpretation of observations for the purpose of discovering underlying meanings and patterns of relationships.

qualitative data

data centered around descriptions or distinctions based on some quality or characteristic rather than on some quantity or measured value

quantitative

of a measurement based on some quantity or number rather than on some quality

quantity

of a measurement based on some quantity or number rather than on some quality

quartile

any of the three points that divide an ordered distribution into four parts, each containing a quarter of the population

quota sampling

a sampling method that chooses a representative cross-section of the population by taking into consideration each important characteristic of the population proportionally, such as income, sex, race, age, etc.

R

A free software programming language and a software environment for statistical computing and graphics.

random assignment

an experimental technique for assigning subjects to different treatments (or no treatment)

random number

number allotted randomly using suitable generator (electronic machine or as simple "generator" as die)

random sample

a sample randomly taken from an investigated population

random variable

a quantity whose value is random and to which a probability distribution is assigned, such as the possible outcome of a roll of a die

random walk

a stochastic path consisting of a series of sequential movements, the direction (and sometime length) of which is chosen at random

range

the length of the smallest interval which contains all the data in a sample; the difference between the largest and smallest observations in the sample

raw score

an original observation that has not been transformed to a z-score

regression

An analytic method to measure the association of one or more independent variables with a dependent variable.

regression to the mean

the phenomenon by which extreme examples from any set of data are likely to be followed by examples which are less extreme; a tendency towards the average of any sample

relative frequency

the fraction or proportion of times a value occurs

relative frequency distribution

a representation, either in graphical or tabular format, which displays the fraction of observations in a certain category

residual

The difference between the observed value and the estimated function value.

response bias

Occurs when the answers given by respondents do not reflect their true beliefs.

root mean square

the square root of the arithmetic mean of the squares

sample

a subset of a population selected for measurement, observation, or questioning to provide statistical information about the population

sample mean

the mean of a sample of random variables taken from the entire population of those variables

sample space

The set of all outcomes of an experiment.

sampling

the process or technique of obtaining a representative sample

sampling distribution

The probability distribution of a given statistic based on a random sample.

scatter plot

A type of display using Cartesian coordinates to display values for two variables for a set of data.

scientific control

an experiment or observation designed to minimize the effects of variables other than the single independent variable

shunt

a passage between body channels constructed surgically as a bypass

Simpson's paradox

a paradox in which a trend that appears in different groups of data disappears when these groups are combined, and the reverse trend appears for the aggregate data

skewed

Biased or distorted (pertaining to statistics or information).

skewness

A measure of the asymmetry of the probability distribution of a real-valued random variable; is the third standardized moment, defined as where is the third moment about the mean and is the standard deviation.

slope

the ratio of the vertical and horizontal distances between two points on a line; zero if the line is horizontal, undefined if it is vertical.

spread

A numerical difference.

standard deviation

a measure of how spread out data values are around the mean, defined as the square root of the variance

standard error

A measure of how spread out data values are around the mean, defined as the square root of the variance.

statistical literacy

the ability to understand statistics, necessary for citizens to understand material presented in publications such as newspapers, television, and the Internet

statistics

a mathematical science concerned with data collection, presentation, analysis, and interpretation

stem-and-leaf display

a means of displaying data used especially in exploratory data analysis; another name for stemplot

stemplot

a means of displaying data used especially in exploratory data analysis; another name for stem-and-leaf display

stochastic

random; randomly determined

stratum

a category composed of people with certain similarities, such as gender, race, religion, or even grade level

straw poll

a survey of opinion which is unofficial, casual, or ad hoc

Student's t-distribution

A distribution that arises when the population standard deviation is unknown and has to be estimated from the data; originally derived by William Sealy Gosset (who wrote under the pseudonym “Student”).

Student's t-statistic

a ratio of the departure of an estimated parameter from its notional value and its standard error

summation notation

a notation, given by the Greek letter sigma, that denotes the operation of adding a sequence of numbers

TI-83

A calculator manufactured by Texas Instruments that is one of the most popular graphing calculators for statistical purposes.

truncate

To shorten something as if by cutting off part of it.

unbiased

impartial or without prejudice

undercoverage

Occurs when a survey fails to reach a certain portion of the population.

unemployment

The level of joblessness in an economy, often measured as a percentage of the workforce.

variable

a quantity that may assume any one of a set of values

variation ratio

the proportion of cases not in the mode

vector

in statistics, a set of real-valued random variables that may be correlated

veridical paradox

a situation in which a result appears absurd but is demonstrated to be true nevertheless

volatility

the state of sharp and regular fluctuation

weighted average

an arithmetic mean of values biased according to agreed weightings

z-score

The standardized value of observation x from a distribution that has mean mu and standard deviation $sigma$.

z-value

the standardized value of an observation found by subtracting the mean from the observed value, and then dividing that value by the standard deviation; also called z-score