Adult Basic Education
Advanced Level
MATHEMATICS

Data Analysis

BRITISH COLUMBIA
Ministry of Advanced Education,
Training and Technology

CENTRE FOR CURRICULUM,
TRANSFER & TECHNOLOGY

Adult Basic Education
Advanced Level Mathematics

Data Analysis

Prepared by
Paul Grinder, Okanagan University College
with
Pat Corbett-Labatt, North Island College
Bob Darling, Malaspina University-College
Peter Robbins, Kwantlen University College
Ada Sarsiat, Northwest Community College
for the
Province of British Columbia
Ministry of Advanced Education, Training and Technology
and the
Centre for Curriculum, Transfer and Technology

Contents

Learning outcomes ii
Glossary iii
Unit 1: The uses and abuses of statistics 1
Unit 2: Introduction: Mean, median, mode, range and graphs 5
Unit 3: Measures of position: quartiles and percentiles 16
Unit 4: The standard deviation 26
Unit 5: The normal distribution 33
Unit 6: The normal curve 44
Unit 7: Analysing survey data 55
Unit 8: A statistics project 62
Appendix A 66
Appendix B 67
Answers 69

Learning outcomes

The word statistics is derived from the Latin word status which means “state”. Governments were the first to use statistics. They used statistics to collect and interpret data about their countries. Today, statistics are used in almost every major field of study.

Upon completion of this Module, you should be able to:

- explain the uses and misuses of statistics
- demonstrate an understanding of mean, median, mode, range, quartiles, percentiles, standard deviation, the normal curve, z scores, sampling error and confidence intervals
- graphically present data in the form of frequency tables, line graphs, bar graphs and stem and leaf plots
- design and conduct a statistics project, analyze the data and communicate your observations about the data

Procedure for independent study

1. Read each of the units in order and complete all of the exercises. If you need assistance, contact your instructor.

2. Complete the Activity Exercises wherever possible.

3. Study the terminology in the Glossary to become familiar with the definitions.

4. If recommended by your instructor, complete additional problem sets.

5. Complete the Project for this Module.

Glossary

Bar graph
A graph that uses side by side bars of different lengths to represent ranked data.

Confidence interval
The interval in which a statistic will likely fall, a certain percent of the time, after repeated experimentation.

Data
The information collected for statistical analysis.

Deviation
The difference between one data value and the mean.

Frequency
The number of times that a particular value occurs in a set of data.

Frequency graph
Sometimes called a broken line graph. A graph with a horizontal axis representing data values and a vertical axis representing frequency values.

Frequency histogram
Also known as a bar graph.

Measures of central tendency
Statistics that describe where the data is centred. The mean, median and mode are measures of central tendency.

Measures of position
Statistics that describe how one data value compares to another. Percentiles, quartiles and z scores are measures of position.

Measures of variation
Statistics that describe how the data is spread out or dispersed. The range, deviation and standard deviation are measures of variation.

Mean
The average. The mean is obtained by finding the sum of the data values and dividing by the number of data values.

Median
The middle value, or the average of the two middle values, of a set of ranked data.

Mode
The data value that occurs most frequently.

Normal curve
Also called a bell curve. Data that is distributed symmetrically about the mean so that most of the data is close to the mean.

Normal distribution
A distribution that takes the shape of a normal curve when graphed. Approximately 68% of the data values will fall within one standard deviation of the mean, 95.5% will fall within two standard deviations of the mean and 99.7% of the data will fall within three standard deviations of the mean.

Percentile
One of the 100 values that divide a set of ranked data into 100 equal intervals. The 48^{th} percentile is a value that is higher than 48% of all the data values.

Population
A large group from which samples are taken for statistical analysis.

Quartile
One of four values that divide a set of ranked data into four equal intervals. The first quartile is equal to the 25^{th} percentile.

Random
A value is random if it has an equal chance of occurring as any other value from the same set.

Random sample
A sample that has the same probability of being chosen as any other sample of the same size.

Range
The difference between the largest data value and the smallest data value.

Ranked data
Data that is listed from highest to lowest or lowest to highest.

Sample
A small set of data chosen from a larger set of data.

Sampling error
The amount of error associated with a calculated value as determined by the size of the sample.

Standard deviation
The square root of the average squared deviation of a set of data.

Statistic
A value calculated from a set of data. The mean and z scores are statistics.

Statistics
A branch of mathematics that collects, organizes and analyzes data.

Stem and leaf plot
A table of data values where the last digits of data values (leaves) are strung out behind their first digits (or stem values).

Survey
Information derived from a sampling of a certain population.

Tally
A method of counting data using “tic” marks.

Yes population
A 40% yes population is one that has responded yes to a particular question 40% of the time.

z score
Also known as a standard score. The value obtained by dividing the deviation by the standard deviation.

Unit 1: The uses and abuses of statistics

The word *statistics* has two meanings. A statistic is a numerical measurement describing some characteristic of a set of *data*. For example, a statistic like 290 pounds could be used to describe the average or mean weight of a football team. *Statistics* is also a collection of methods for planning experiments, collecting data, analyzing the data and drawing conclusions.

The uses of statistics

It is hard to read a magazine or newspaper without coming across some statistical survey or analysis. Sportscasts, TV documentaries and newscasts also have their share of statistics. The uses of statistics include applications in business, sports, medicine, agriculture, psychology, sociology, education and political science. Governments use statistics to monitor everything from life style preferences to crime rates. New drugs are statistically analyzed to determine their effectiveness on patients. The statistical technique of *random* selection is employed to guarantee that a small sample of a larger population group is actually an unbiased representation of the whole population. Statistics, such as plus-minus records, can even be used to determine whether a certain hockey player should be given more or less ice time.

The abuses of statistics

Just as statistics can be used to provide a solid quantitative analysis of a set of data, statistics can be misused to distort data. The abuse of statistics is what Benjamin Disraeli (nineteenth century British prime minister) was referring to when he made the famous comment, "There are three kinds of lies – lies, damned lies and statistics."

Statistics can be used to misrepresent a situation. Suppose a small store employs 6 people who earn an average, or *mean*, wage of \$8.50 per hour as calculated below,

$$\frac{\$8+\$8+\$8+\$8+\$9+\$10}{6}=\$8.50$$

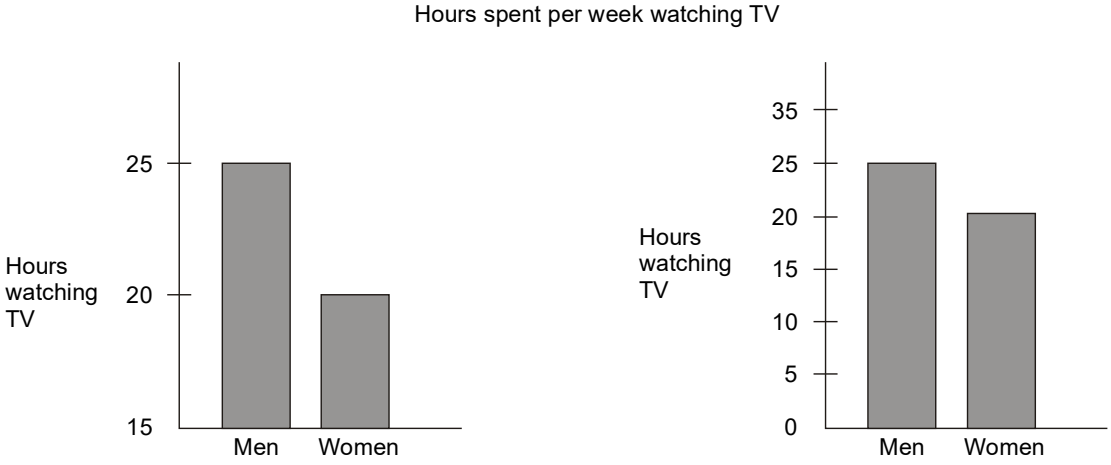
Now suppose the store owner, who earns \$40 per hour, includes his wages in the calculation,

$$\frac{\$8+\$8+\$8+\$8+\$9+\$10+\$40}{7}=\$13$$

If the store owner reports that the average wage earned at the store is \$13, he or she is misrepresenting the situation since the store owner is the only person making \$13 per hour or more.

Another source of deceptive statistics results from the faulty collection of data. Companies that conduct public opinion polls have to be extremely careful that they survey a large enough sample of the population and also an unbiased segment of the population. For example, suppose a poll was conducted in BC to determine whether a luxury tax should be imposed on buyers of new pick-up trucks. The citizens of Prince George might respond quite differently to the poll than the residents of Victoria. The poll could be quite biased if it was only conducted in Victoria, or only conducted in Prince George.

Statistical graphs can be presented in a deceptive manner. Consider the two *bar graphs* below depicting the same data.



Without a close inspection of the vertical scale, the first bar graph creates the impression that men watch twice as much TV as women do. In the second graph, the vertical scale starts at 0, and the length of the bars are proportional to the actual hour of TV watching.

The above examples illustrate only a few of the abuses of statistics. To avoid the "lies and damned lies", every step of the statistical process must be scrupulously carried out; from the collection of the data, to the calculation of a statistic, to the presentation of conclusions.

Now complete Exercise 1 and check your answers.

Exercise 1

1. Why is the following bar graph misleading?

2. What factor or factors might cause the following surveys to be biased?

   a. TV news watchers are asked to phone in their opinion on whether marijuana smoking should be legalized.

   b. A questionnaire asking family members to list the number of books they read in the last year is mailed to 1000 homes in the city of Vancouver.

   c. To determine how many college students are smokers, Butler asks the first 20 students he sees standing outside the main entrance to the college, “Are you a smoker?”

Answers are on page 69.

Activity 1: Watching TV

Ask every student in the room to write down, on a small piece of paper, an estimate of the number of minutes they spent watching TV yesterday. Collect the data (pieces of paper) in some sort of container.

1. a. Draw one piece of paper and record the number. _____

   b. Do you think that this one piece of data is a good representation of the actual (yet to be calculated) average? _____

2. Replace the first piece of paper and draw two pieces of data. Find the mean of these two. _____

3. Replace the two pieces of data and now draw four pieces of paper. What is the average time spent watching TV based on just these four pieces of data? _____

4. Replace the four pieces of paper and draw one half (or one half plus one) of the data. Find the average time for one half the data. _____

5. Now find the mean using all the data. _____

   a. How do the previous calculations of the mean, using smaller samples of the total data, compare to the actual mean?

   _____

   b. Some of the students may have recorded 0 minutes for the time they spent watching TV yesterday. How did these zeros affect the mean time?

   _____

   c. Now calculate the mean for only those students who actually watched some TV yesterday.

   _____

Unit 2: Introduction: Mean, median, mode, range and graphs

Statistics is the science of collecting, classifying, presenting and interpreting numerical data. The *data* are numbers or measurements collected by a statistician. For example, the data below are scores obtained by 12 students on a math quiz out of 40 marks.

32, 39, 32, 27, 30, 34, 32, 35, 40, 36, 32, 36

In order to statistically describe the above data, we might ask the following questions.

1. What is the average, or *mean*, score?

2. What is the middle, or *median*, score?

3. What score occurs most often, or what is the *mode*?

4. What is the difference between the highest and lowest score, or what is the *range*?

6. How can the data be represented graphically, with a *line graph*, *bar graph* or *stem and leaf plot*?

The mean, median, mode, and range are four statistics which can be used to describe a set of data. The mean, median, and mode are called *measures of central tendency* because they tell us where the data is centered. The range is a *measure of variation* because it tells us how much the data is spread out.

The mean is the most important measure of central tendency. It is calculated as follows:

The *mean* is the sum of all the data values divided by the number of data values, or

$$\bar{x} = \frac{\Sigma x}{n}$$

where $\bar{x}$ = mean, x is a data value, and n is the number of data values

The symbol "Σ" is the Greek letter "sigma" and means "the sum of all". Here "Σx" means the sum of all x (or data) values.

Example 1

Find the mean score of the following 12 math test scores;

32, 39, 32, 27, 30, 34, 32, 35, 40, 36, 32 and 36.

Solution

Using the formula, (note that n = 12),

$$\bar{x} = \frac{\Sigma x}{n} = \frac{32 + 39 + 32 + 27 + 30 + 34 + 32 + 35 + 40 + 36 + 32 + 36}{12}$$

$$= \frac{405}{12} = 33.75$$

The mean score is 33.75.

The *median* is the middle value when the data is arranged from highest to lowest. If there are two middle values, then the median is the mean of these two values.

Example 2

Find the median score of the above 12 math scores.

Solution

Arrange the data from the highest to lowest.

40	35		32
39	**34**	**middle**	32
36	**32**	**values**	30
36	32		27

Note that because there are an *even number* (twelve) of data values, we have two middle values. The mean of these two values is

$$\frac{34 + 32}{2} = \frac{66}{2} = 33$$

Hence the median math score is 33.

The *mode* is the data value which occurs most often (or with the greatest *frequency*). The mode may not exist.

Example 3

Determine the mode of the above twelve math scores.

Solution

Again, arrange the data from highest to lowest.

40	35	32
39	34	32
36	32	30
36	32	27

Notice that the score of 32 occurs most often. Hence the modal score is 32.

The *range* is the difference between the largest data value and the smallest data value.

Example 4

Determine the range of the 12 math scores above.

Solution

The data is,

40	35	32
39	34	32
36	32	30
36	32	27

The highest value is 40. The lowest is 27. The range is,

$$40 - 27 = 13.$$

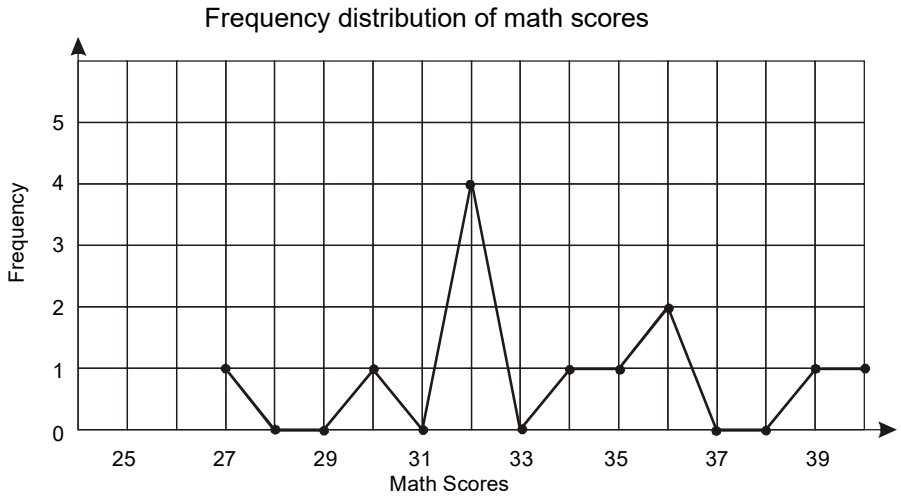
Also known as *frequency distributions*, the line graph plots data values (horizontal axis) against the frequency of those data values (vertical axis).

Example 5

Plot the frequency distribution of the 12 math scores.

Solution

Label each axis. Plot points and connect points with a straight line.



Bar graphs or, *frequency histograms*, use bars to represent frequencies for certain data intervals.

Example 6

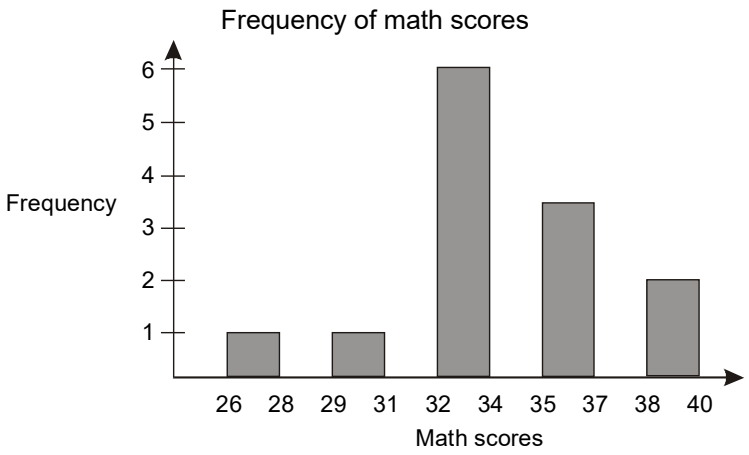
Prepare a frequency histogram for the 12 math scores.

27, 30, 32, 32, 32, 32, 34, 35, 36, 39, and 40.

Solution

The range of scores is 13. If we want 5 bars, each data interval should be 3 scores wide.

Interval	Frequency
38 – 40	38, 40 (2 scores)
35 – 37	35, 36, 36 (3 scores)
32 – 34	32, 32, 32, 32, 34 (5 scores)
29 – 31	30 (1 score)
26 – 28	27 (1 score)



A *stem and leaf plot* is similar to a bar graph, except the bars are replaced by digits. These leaf digits are the last digits of data having the same first digit(s), called the stem.

Example 7

Construct a stem and leaf plot for the following data (minutes taken to run 15 km).

49, 52, 53, 53, 57, 58, 60, 63, 64, 66, 66, 66, 69, 70, 70, 72, 75, 77, 77, 79, 83, 88, 89, 94, 106

Solution

The stems are the first digits of the data numbers (4, 5, 6, 7, 8, 9 and 10). The leaf digits are strung out beside their stems as displayed below.

	Time (min.) taken to run 15 km
4	9
5	2 3 3 7 8
6	0 3 4 6 6 6 9
7	0 0 2 5 7 7 9
8	3 8 9
9	4
10	6

Now complete Exercise 2 and check your answers.

Exercise 2

1. A small company employs 15 workers. Their annual salaries are as follows:

$15 000	$ 25 000	$ 25 000
17 000	25 000	25 000
17 000	25 000	25 000
22 000	25 000	30 000
22 000	25 000	65 000

a. Determine the mean, median, mode, and range of the above data.

$\bar{x}$ = _____ median = _____ mode = _____

range = _____

b. Which statistic (mean, median, mode, or range) best describes the annual salary *most* of the workers receive?

_____

c. Which statistic best describes the gap that exists between annual salaries?

_____

2. a. Find the daily mean temperature and daily range for the temperatures below.

DAY	°C HIGH	°C LOW	DAILY MEAN	DAILY RANGE
1	6	0		
2	3	2		
3	12	4		
4	13	10		
5	15	7		
6	13	10		
7	12	8		

b. Determine the mean of the daily means.

_____

c. Determine the total range of temperature over this 7 day period.

_____

d. Determine the mean range over this 7 day period.

_____

3. Below are the statistics for a final exam given to two different math classes. (The exam was worth 100 marks.)

CLASS A (n = 25)	CLASS B (n = 25)
$\bar{x}$ = 80	$\bar{x}$ = 72
range = 40	range = 10

a. Which class (overall) seemed to do better on the exam?

_____

b. Which class probably had the student with the highest mark?

_____

c. Which class probably had the student with the lowest mark?

_____

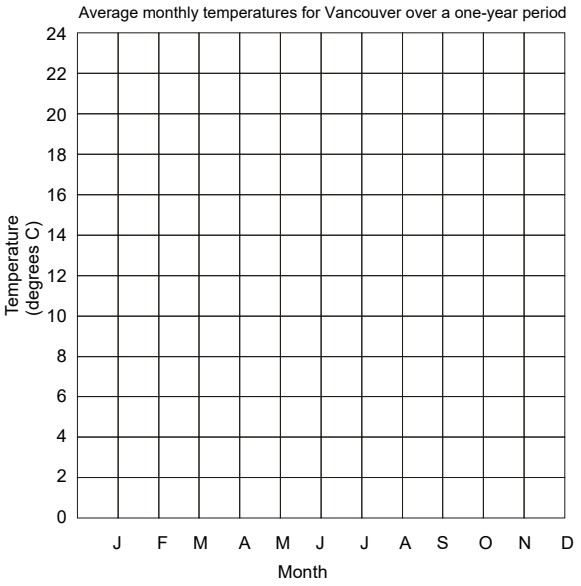
d. Which class had students with similar abilities? Explain.

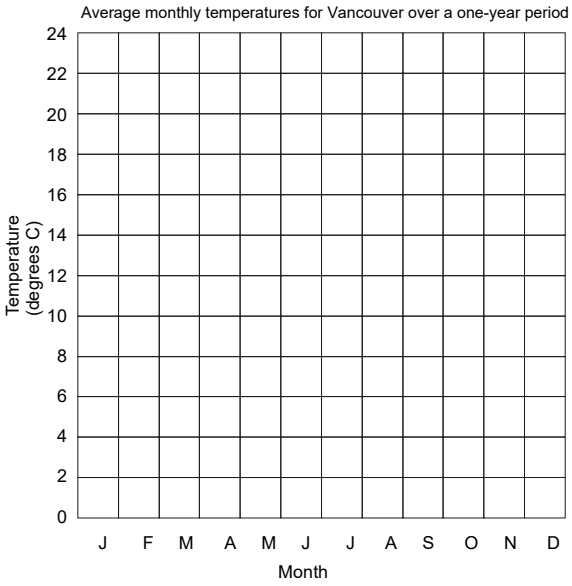
_____

4. The following data represents the average monthly temperature for Vancouver over a one year period.

Month	J	F	M	A	M	J	J	A	S	O	N	D
Average Temperature (°C)	4	7	10	12	15	20	22	19	14	8	3	0

Draw a frequency distribution graph and a histogram on the following page.





5. The following data represents final grades for a computer course. Construct a stem and leaf plot for the data.

   40%, 42%, 42%, 50%, 54%, 56%, 58%, 66%, 66%, 68%, 69%, 70%, 73%, 80%, 84%, 85%, 88%, 89%, 93%

6. The following stem and leaf plot represents the time in seconds taken to type 60 words by a class of business students.

4	4	6	8						
5	0	1	6	8	8	9			
6	0	0	2	3	3	5	7	7	7
7	1	3	3	4	4				
8	4								
9	5								

   a. How many students were tested? _____

   b. Find the mean and median for this data.

   $\bar{x}$ = _____ median = _____

7. Jody wants to receive a grade of 80% for the laboratory part of her Chemistry course. So far, she has a 78% average on her last 5 labs. What grade does she need on her sixth lab to earn an 80% average?

8. Neil has 182 out of 260 marks thus far in his math course. If the final exam is worth 100 marks, what mark does Neil need on the final exam to earn a final grade of 75% for the course? How do you feel about Neil's chances, and why?

Answers are on pages 6969.

Activity 2: Shoe size

Ask at least 15 male students and 15 female students to write their shoe size on a piece of paper. Include M for male and F for female as well, on the slip of paper.

1. Below, arrange the shoe sizes in order from smallest to largest.

MALE	FEMALE

2. Determine the following.

	MALE	FEMALE
mean	_____	_____
median	_____	_____
mode	_____	_____
range	_____	_____

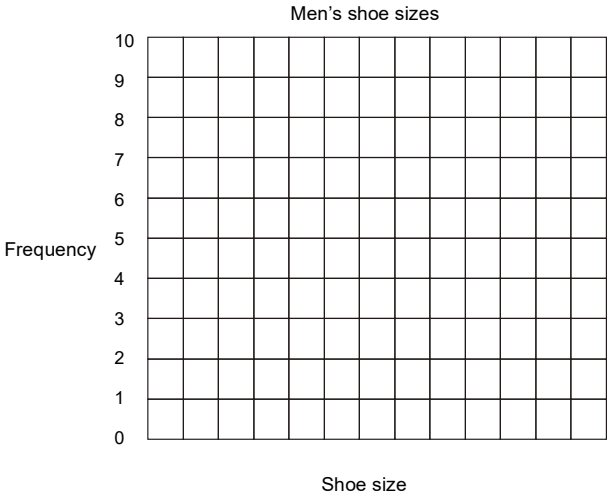
3. Imagine you are a shoe manufacturer.

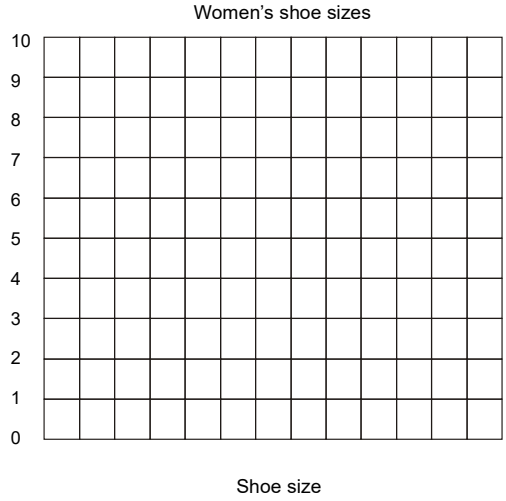
a. Why would it be important to know the range of shoe sizes for males and females?

b. Why would it be important to know the modal shoe size for males and females?

c. Why would knowing the mean and median shoe sizes be important?

4. Organize the data into bar graphs.





Unit 3: Measures of position: quartiles and percentiles

The mean, median and mode are measures of central tendency. These statistics tell us where the data is centered. Quartiles and percentiles are *measures of position*. Quartiles and percentiles can be used to compare one particular data value to all the rest of the data. These statistics enable us to answer questions such as, "Is a certain value unusually high, unusually low or just average?"

Quartiles divide ranked data into four equal parts. Ranked data is data arranged from highest to lowest, or lowest to highest. The quartile values are denoted as Q_1, Q_2 and Q_3. Q_1 separates the bottom 25% of the data from the top 75%. Q_2 is the same as the median and separates the top 50% from the bottom 50% of the data. Q_3 separates the top 25% of the data from the bottom 75%.

Example 1

Find Q_1, Q_2 and Q_3 for the following set of data (resting heart rates of 30 college males).

55	94	80	68	78	61
60	55	88	60	70	70
70	60	86	42	65	74
72	68	80	100	58	84
81	72	71	85	57	96

Solution

Rank the data from lowest to highest.

42	60	68	71	80	86
55	60	68	72	80	88
55	60	70	72	81	94
57	61	70	74	84	96
58	65	70	78	85	100

Find Q_2 or the median, first. Since there are 30 values, there are two middle values, 70 and 71.

$$Q_2 = \frac{70 + 71}{2} = 70.5$$

To find Q_1, find the middle value of the bottom 50% of the data. The bottom half of the data has 15 values and ranges from 42 to 70. Counting to the 8th value,

$Q_1 = 60$

Q_3 can be found in a similar fashion. Q_3 is the middle value of the upper 50% of the data or the 8th value from 71 to 100.

$Q_3 = 81$

Finding Q_1, Q_2 and Q_3 in the above example was simply a matter of ranking and counting. But there was an even number (n = 30) of data values. When there is an odd number of data values, a different method of calculating Q_1 and Q_3 will have to be used.

The three quartiles Q_1, Q_2 and Q_3 divide the ranked data into 4 equal parts. There are 99 *percentiles* P_1, P_2, P_3, …, P_{99} that divide the ranked data into 100 equal parts. For example, P_{80} is called the "80th percentile" and P_{80} is a value that is higher than 80% of the rest of the data values. P_{10} is a value that is higher than 10% of the data values.

Each *data value*, x, corresponds to a particular percentile, P_k, where k is given by the expression,

$$k = \frac{\text{number of data values less than x}}{n} \times 100\%$$

where n is the total number of data values.

Example 2

Below are the number of chin ups completed in one minute by 70 male college students. The data is ranked lowest to highest.

0	3	7	8	10	13	20
0	3	7	9	10	13	20
1	4	7	9	10	13	22
1	4	7	9	10	14	22
2	6	7	9	11	14	23
2	6	7	9	11	15	25
2	6	7	9	11	15	28
3	6	8	10	11	15	30
3	6	8	10	12	18	30
3	7	8	10	13	20	33

Find the percentiles associated with the data values 0, 7 and 25.

Solution

a. For data value 0, $k = \frac{0}{70} \times 100\% = 0\%$

So, 0 is the 0^{th} percentile or $P_0 = 0$. This makes sense because 0 is not higher than any other value.

b. There are 19 data values that are less than the value 7.

For data value 7, $k = \frac{19}{70} \times 100\% = 27\% \text{ (rounded)}$.

So, 7 is the 27^{th} percentile or $P_{27} = 7$.

c. There are 65 data values less than 25.

For data value 25, $k = \frac{65}{70} \times 100\% = 93\% \text{ (rounded)}$.

So, $P_{93} = 25$. The person who did 25 chin ups in one minute did better than 93% of the other college men.

The reverse procedure, finding what data value corresponds to a given percentile, is rather involved.

To find the *data value* associated with a certain *percentile*, P_k, follow the steps below:

Step 1 Rank the data from lowest to highest.

Step 2 Calculate $C = \left(\frac{k}{100}\right)n$ where k is the percentile in question and n is the number of data values.

Step 3 If C is a whole number, then $P_k = \frac{C^{th} \text{ data value} + (C+1)^{th} \text{ data value}}{2}$.

Step 4 If C is not a whole number, round C up to the next larger whole number and P_k = the C^{th} (rounded up) data value, counting from the lowest value.

Example 3

The following stem and leaf plot depicts the number of words typed in 2 minutes by 125 office administration students.

7	3	7	8	9	9																	
8	0	1	2	3	5	5	5	8	8	9												
9	0	0	0	2	3	3	7	8	8	8	9	9										
10	0	0	0	1	1	1	1	1	3	3	7	8	8	8	8	8						
11	0	1	1	2	2	2	2	3	4	4	4	5	5	7	7	7	7	8	8	9	9	9
12	0	0	1	1	1	1	1	5	5	5	6	6	6	6	6	7	7	8	9			
13	0	0	0	0	0	2	3	3	3	6	6	6	6	7	7	8	8	9				
14	0	0	1	1	1	1	3	3	4	4	6	7	8	8	8							
15	0	0	3	8	9	9																
16	1	7																				

Find P_{40}, P_{25} and P_{95}.

Solution

The data is already ranked.

a. To find P_{40}, calculate $C = \left(\frac{k}{100}\right)n$ where k = 40 and n = 125,

$$C = \left(\frac{40}{100}\right)125 = 50$$

Since C = 50 is a whole number, calculate

$$P_{40} = \frac{C^{th} \text{ data value} + (C+1)^{th} \text{ data value}}{2}$$

where the $C = 50^{th}$ data value (from the stem and leaf plot) is 112 and the $C + 1 = 51^{st}$ data value is 113.

$$P_{40} = \frac{112 + 113}{2} = 112.5$$

b. Finding P_{25} is the same as finding Q_1, the first quartile.

Calculate $C = \left(\frac{k}{100}\right)n$ where K = 25 and n = 125

$$C = \left(\frac{25}{100}\right)125 = 31.25$$

The C = 31.25 is not a whole number.

Round C = 31.25 to 32 and P_{25} is the 32^{nd} data value from the lowest value.

$$P_{25} = 98$$

c. To find P_{95}, calculate

$$C = \left(\frac{95}{100}\right)125 = 118.75$$

Round C up to 119.

The 119^{th} score is 150.

So, $P_{95} = 150$.

Now complete Exercise 3 and check your answers.

Exercise 3

1. The following stem and leaf plot depicts the number of words typed in 2 minutes by 125 office administration students.

7	3	7	8	9	9																
8	0	1	2	3	5	5	5	8	8	9											
9	0	0	0	2	3	3	7	8	8	8	9	9									
10	0	0	0	1	1	1	1	1	3	3	7	8	8	8	8	8					
11	0	1	1	2	2	2	2	3	4	4	4	5	5	7	7	7	7	8	8	9	9	9
12	0	0	1	1	1	1	1	5	5	5	6	6	6	6	6	7	7	8	9			
13	0	0	0	0	0	2	3	3	3	6	6	6	6	7	7	8	8	9				
14	0	0	1	1	1	1	3	3	4	4	6	7	8	8	8							
15	0	0	3	8	9	9																
16	1	7																				

   a. Find P_{50} (the median) and P_{80}.

   b. Jill can type 140 words in 2 minutes. What percentile score is this? Jill can type faster than what percent of the other 124 students?

   c. A student needs to type 90 words in 2 minutes in order to pass the test. What percentile is associated with this value? How many students failed the test?

d. The instructor decided that those students who achieved less than the 20^{th} percentile would have to retake the test. What data value is represented by P_{20}?

2. Below are the number of chin ups completed in one minute by 70 male college students. The data is ranked lowest to highest.

0	3	7	8	10	13	20
0	3	7	9	10	13	20
1	4	7	9	10	13	22
1	4	7	9	10	14	22
2	6	7	9	11	14	23
2	6	7	9	11	15	25
2	6	7	9	11	15	28
3	6	8	10	11	15	30
3	6	8	10	12	18	30
3	7	8	10	13	20	33

a. Find Q_3 (or P_{75}) for the above data.

b. Now find the percentile associated with 13 chin ups.

c. In a. you found that $P_{75} = 13$ but when the process was reversed, in b. you found that $13 = P_{70}$. Explain this discrepancy.

Answers are on page 69.

Activity 3: Mutual funds

Work in groups of 3 or 4.

1. On the following page, thirteen Canadian "Asia ex-Japan" mutual funds are listed with their current value and percentage gains over 1 day, 1 week, 30 days and 1 year. On page 66 in the spaces provided, rank each fund from best to worst based on their percentage gain for the given time interval. For example, for the "1 day %" ranking column, Fund B would rank 1st and Fund H would rank last or 13th.

2. a) Which, if any, funds always ranked in the top half of the group (above Q_2) in all four categories? (These would be the best funds with low ranks.)

   b) Are there any fund(s) that ranked above Q_3 in all categories? Which one(s)?

3. a) Find the sum of the four rankings for each fund in the last column.

   b) What does a "low" sum indicate?

   c) What does a "high" sum indicate?

4. a) Is there any fund that has a consistently high ranking? Which one?

   b) Which fund has the worst performance based on overall ranking?

5. a) Rank what your group thinks are the three best performing "Asia ex-Japan" mutual funds.

   b) Also rank the three worst performing funds.

Activity 3: Asia ex-Japan Canadian mutual funds

		As of December 23, 1999										
Fund Letter Code	**Fund name**	**Price $**	**1 day $ Chg**	**1 day %**	**Rank**	**1 week %**	**Rank**	**30 day %**	**Rank**	**YTD %**	**Rank**	**Sum of ranks**
A	AGF Asian Growth Class	13.880	.620	4.68		10.86		16.35		56.31		
B	AGF Asian Growth Class (US$)	9.420	.440	4.90		11.22		15.72		62.98		
C	Clarington Asia Pacific	16.548	.214	1.31		4.88		16.72		67.37		
D	Fidelity Far East	39.630	.340	.87		4.92		11.20		53.72		
E	Fidelity Far East (US$)	26.970	.360	1.35		5.39		10.99		60.15		
F	First Canadian Far East	10.412	.021	.20		2.21		8.52		33.86		
G	Green Line Asian Growth	12.450	.500	4.18		12.77		19.37		82.28		
H	Investors Pacific International	8.970	.050	-.55		1.24		4.30		25.88		
I	National Bank Far East Equity	10.540	.040	.38		2.73		6.90		27.76		
J	Navigator Asia Pacific	10.044	.003	.03		-.21		-.39		19.49		
K	Royal Asian Growth	12.463	.088	.71		4.51		5.80		58.24		
L	Universal Far East	4.735	.062	1.34		6.45		10.84		50.60		
M	Universal Far East (US$)	3.214	.049	1.54		6.81		10.18		56.43		

Source: http://globefund.com

Unit 4: The standard deviation

Consider the data in set A and B below.

Set A: 30, 50, 70 Set B: 40, 50, 60

Notice that the mean and median for both sets are exactly the same. Their *measures of central tendency* are the same, but the range is very different for both sets. The range for Set A is 40 and for Set B it is 20. The range is a very simple *measure of variation*. Set A and Set B are dispersed, or spread out, quite differently.

One way of measuring the variation or dispersion of specific data values is to calculate the *deviation*, $x - \bar{x}$, where x is a data value and $\bar{x}$ is the mean. For example, in Set A above, the amount that 70 deviates from the mean is 20.

A very precise way of measuring the variation of an entire set of data is to calculate the *standard deviation*. The standard deviation is a statistic that indicates a kind of average deviation from the mean of the data. The larger the standard deviation number, the more spread out the data is. The Greek letter sigma, σ, is the symbol for standard deviation.

The *standard deviation*[1] formula is as follows,

$$\sigma = \sqrt{\frac{\Sigma\left(x - \bar{x}\right)^2}{n}}$$

where σ is the standard deviation, x is a data value, $\bar{x}$ is the mean of the data, and n is the number of data values.

[1] The formula given above is for finding the standard deviation of a given population. The formula $\sigma_{n-1} = \sqrt{\frac{\Sigma\left(x - \bar{x}\right)^2}{n-1}}$ is used to find the standard deviation of smaller sample of a given population. Throughout this module, we will use the formula for σ (not σ_{n-1}).

Example 1

Find the mean and standard deviation of the 12 science quiz scores below.

5, 6, 8, 8, 9, 10, 10, 11, 11, 12, 15 and 15

Solution

Find the mean, or $\bar{x}$.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{5+6+8+8+9+10+10+11+11+12+15+15}{12} = \underline{10}$$

The following table can be used to determine the standard deviation.

x	$x - \bar{x}$	$\left(x-\bar{x}\right)^2$
5	5-10= -5	$(-5)^2 = 25$
6	-4	16
8	-2	4
8	-2	4
9	-1	1
10	0	0
10	0	0
11	1	1
11	1	1
12	2	4
15	5	25
15	5	25
$= \Sigma\left(x-\bar{x}\right)^2$		
		106

The standard deviation, or σ, is

$$\sigma = \sqrt{\frac{\Sigma\left(x-\bar{x}\right)^2}{n}} = \sqrt{\frac{106}{12}} = \sqrt{8.83} \approx \underline{2.97}$$

The reason we square the deviations before finding the average is to avoid adding positive and negative values. Notice above that the sum of the deviations, $x - \bar{x}$ is zero.

The paper and pencil method of calculating the standard deviation can be extremely lengthy, especially when *n*, the *population size*, is large.

Example 2

Calculate the mean and standard deviation of the data in Example 1 using a calculator with a statistics mode. (As many calculators function differently, please bring your calculator manual to class.)

Solution

Put your calculator in the statistics mode (if necessary), enter the data and find $\bar{x}$ and σ.

1. To operate in statistics mode, press: _____.

2. Enter the data (Find your data button. FRQ can be used to input repeated data values.):

   5 DATA (or Σ+)
   6 DATA
   8 DATA
   8 DATA
   9 DATA
   10 DATA
   10 DATA
   11 DATA
   11 DATA
   12 DATA
   15 DATA
   15 DATA

3. Find $\bar{x}$ and σ (you may have to press SHIFT or 2^{nd} or some other key).

   To find $\bar{x}$, press: _____
   To find σ, press: _____

4. To return to calculation mode, press: _____

Now complete Exercise 4 and check your answers.

Exercise 4

1. The following data represents quiz scores on a test out of 10 by 21 math students.

   0, 0, 1, 2, 4, 5, 5, 5, 6, 6, 6, 7, 8, 8, 8, 8, 9, 9, 9, 10, 10

   a. What is the range? _____

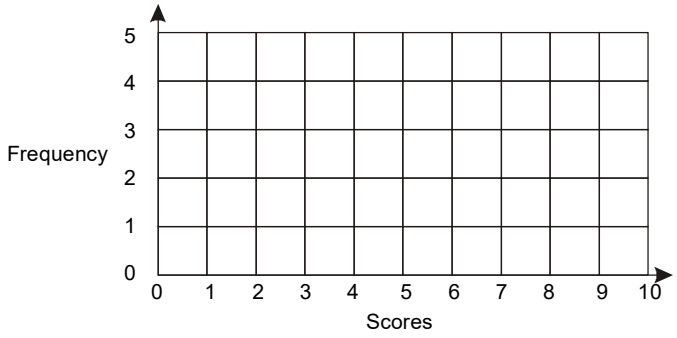
   b. What is the mode? _____

   c. What is the median? _____

   d. What is the mean? _____

   e. Calculate the standard deviation. _____

   f. Plot a frequency distribution graph for the data below.



   g. Calculate $\bar{x} - \sigma =$ _____ and $\bar{x} + \sigma =$ _____.

   h. How many scores are there between 2.9 and 9.1, or, how many scores lie within one standard deviation of the mean?

   _____

   i. What percent of the scores is this?

   _____

2. a. In Example 1 of this section, the mean of the 12 science quiz scores was 10 and the standard deviation was 2.97. What percent of the scores lie within one standard deviation of the mean? The scores were 5, 6, 8, 8, 9, 10, 10, 11, 11, 12, 15, and 15.

_____

b. What percent of the scores lie within two standard deviations of the mean?

_____

3. Suppose the standard deviation of a set of numbers is 0. What does this tell you about the data?

_____

4. Two classes, each with 100 students, wrote an examination with a possible maximum score of 100. In the first class the mean score was 75 and the standard deviation was 5. In the second class, the mean score was 70 and the standard deviation was 15. Which of the two classes do you think had more scores of 85 or better? Why?

_____

5. The following data represents the weights (in kg) of a small class of students:

78, 42, 72, 88, 86, 97, 91, 79, 82, 86, 91, and 74

a. Calculate the range. _____

b. Calculate the mean weight. _____

c. Calculate the standard deviation _____

d. What percentage of the weights fall within one standard deviation of the mean weight? _____

6. It is found that the time taken by a bank teller to serve 7 people is 3, 3, 4, 5, 6, 6, and 7 minutes.

a. Find the mean time. _____

b. Find the standard deviation. _____

Answers are on pages 69.

Activity 4: Pop quiz

1. Your instructor will ask you to write the quiz in Appendix B (pages 67) along with the rest of the class. You will only have 5 minutes to complete the test. Do not look at it until your instructor says "Go!"

2. After all the tests are marked out of 20, record the marks below, ranked from highest to lowest. (See page 75 for the answers.)

3. Calculate the following.

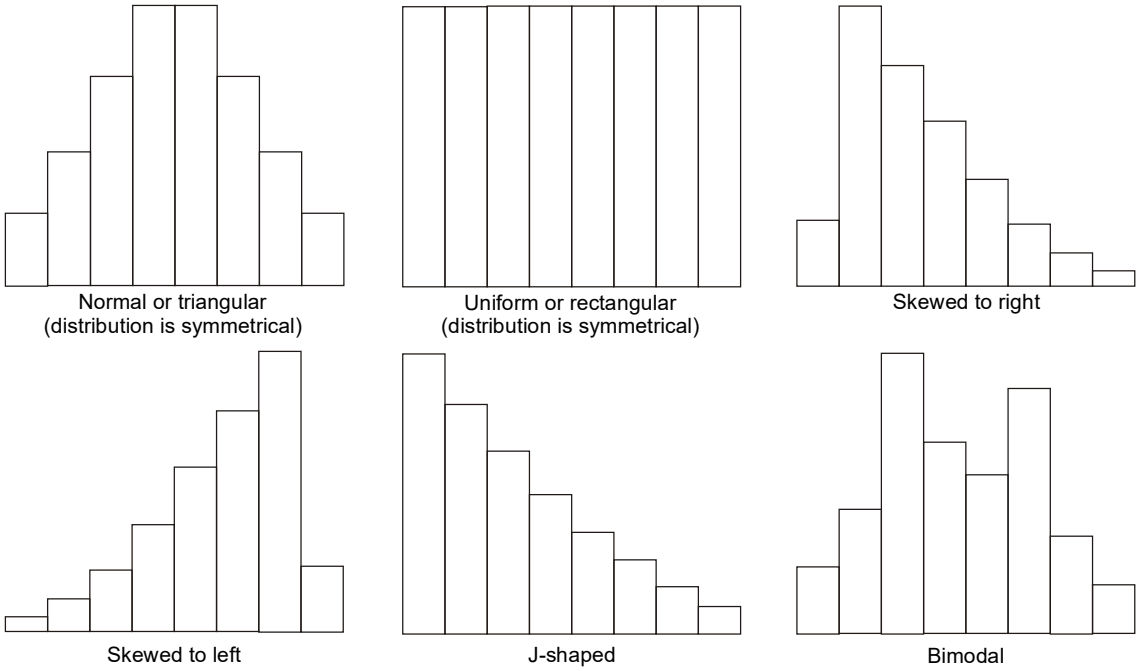
   mode = _____
   $\bar{x}$ = _____
   σ = _____
   Q_1 = _____
   Q_2 = _____
   Q_3 = _____

4. What percent of the scores lie within

   a. one standard deviation of the mean? _____
   b. two standard deviations of the mean? _____
   c. three standard deviations of the mean? _____

Unit 5: The normal distribution

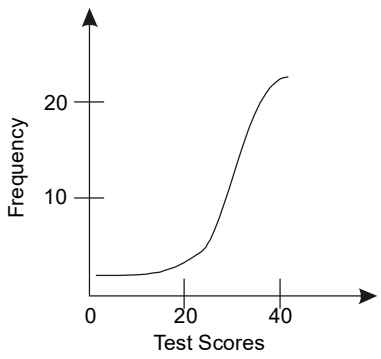
Data can be distributed in quite a variety of ways. Consider the frequency histograms below,



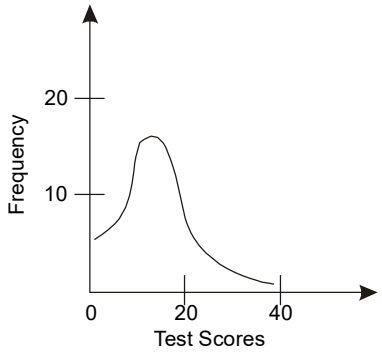
- *Symmetrical:* Both sides of this distribution are identical.
- *Uniform (rectangular):* Every value appears with equal frequency.
- *Skewed:* One tail is stretched out longer than the other. The direction of skewness is on the side of the longer tail.
- *J-shaped:* There is no tail on the side of the class with the highest frequency.
- *Bimodal:* The two most populous classes are separated by one or more classes. This situation often implies that two populations are being sampled.

Some examples of the above distributions follow. Imagine that a group of math students were given a math test out of 40. The difficulty of the test can have quite an influence on the shape of a frequency distribution.

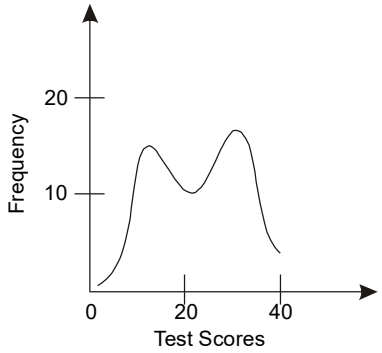
Case 1 If the test was so easy that most of the students received a perfect mark, this distribution would be considered J-shaped.



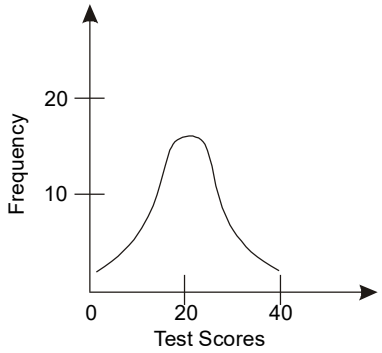
Case 2 If the test was quite difficult and most of the students received a mark of less than 50%, this distribution would be considered skewed to the right.



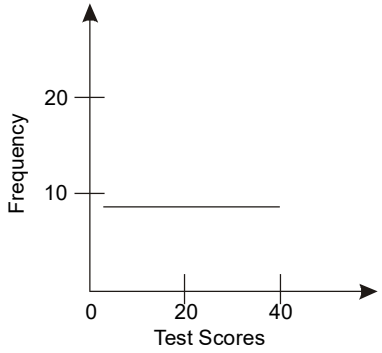
Case 3 If the test was given to two different math classes, one of which had not been taught half the material, this distribution would be bimodal.



Case 4 If the test was "fair", at least from the instructors point of view, this distribution would be normal.



Case 5 What kind of test would produce a uniform distribution?



When a population is measured for some attribute or ability, the most frequently occurring distribution is the normal distribution. When a product is tested for some characteristic the result is most often a normal distribution[2]. For example, if a sample population of men (or women) is tested for physical strength (or blood pressure or intelligence or shoe size), most of the people will be close to average strength with a small minority either much stronger than or much weaker than the majority.

In the previous assignment you were often asked, "what percent of the data lie within one standard deviation of the mean?" Knowing how a population bunches around its mean value can be quite useful.

Pafnuty Chebyshev (1821-1894) was a Russian mathematician who worked on probability, theory of prime numbers, and problems in mechanics.

Chebyshev's Theorem states that the proportion of any distribution that lies within k standard deviations of the mean is at least $1-\left(\frac{1}{k}\right)^2$, where $k > 1$. This applies to *any* distribution of data.

For example, when $k = 2$, Chebyshev's Theorem states that, $1-\left(\frac{1}{2}\right)^2 = 1-\frac{1}{4} = \frac{3}{4}$ or more of the data will lie within 2 standard deviations of the mean.

Chebyshev's Theorem applies to *any* set of data.

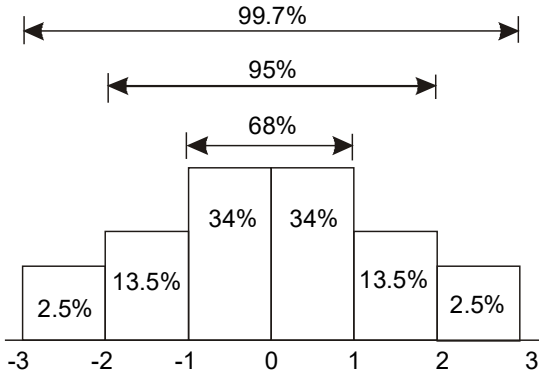
When data is distributed *normally*, (see top left histogram) Chebyshev's proportion can be "improved" on dramatically.

When data is *distributed normally*, then approximately 68% of the data is within one standard deviation of the mean, 95% of the data is within 2 standard deviations of the mean and 99.7% is within 3 standard deviations of the mean. This is the Empirical Rule.

The vast majority of statistical analysis is done on normally distributed data; from biology to psychology to economics to medicine to sports.

[2] Normal distribution is not appropriate for all kinds of distribution. For example, small sample sizes or biased populations would not necessarily be normally distributed and could be statistically analyzed by different methods.

The histogram below is a representation of an “ideal” normal population, where the mean is 0 and the standard deviation is 1.



Now complete Exercise 5 and check your answers.

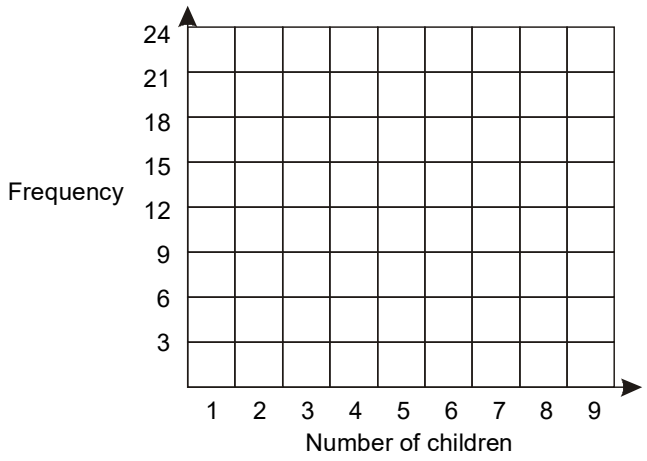
Exercise 5

1. Sixty college students were asked for the total number of children in their family. The data collected follows:

1	6	3	5	5	3	4	1	2	7	3	2
3	4	5	3	1	3	2	1	4	4	2	2
3	9	4	3	3	5	3	5	7	3	1	1
3	5	2	6	4	3	3	3	3	3	2	3
4	3	5	7	3	2	1	2	3	2	4	3

a. What is the range for this data? _____

b. Draw a histogram for the data below.



c. Find the mean for this data. $\bar{x}$ = _____

d. Find the standard deviation. σ = _____

e. Find the values, $\bar{x} - \sigma$ and $\bar{x} + \sigma$. _____ and _____.

f. What percent of the data lies between $\bar{x} - \sigma$ and $\bar{x} + \sigma$?

_____

g. What are the values, $\bar{x} - 2\sigma$ and $\bar{x} + 2\sigma$?

_____ and _____

h. What percent of the data lies within two standard deviations of the mean?

_____

i. What are the $\bar{x} - 3\sigma$ and $\bar{x} + 3\sigma$ values?

_____ and _____

j. What percent of the data lies within three standard deviations of the mean?

_____

k. Compare your answers for f., h., and j. to the results predicted by the Empirical Rule. Does the result suggest an approximately normal distribution?

_____

2. The following table tallies the number of hour of TV watched in one day by 75 high school students. Only those who watched some TV yesterday were included in the tally.

Hours	Tally	Frequency
0.5	|||	
1.0	𝍸 𝍸 |	
1.5	𝍸 𝍸 ||	
2.0	𝍸 𝍸 𝍸 ||	
2.5	𝍸 𝍸 𝍸	
3.0	𝍸 𝍸 |||	
3.5	||||	

a. Convert the tally counts to frequency numbers above.

b. Calculate the mean. $\bar{x}$ = _____

c. Calculate the standard deviation. σ = _____

d. What percent of the data lies within one standard deviation of the mean?

_____

e. What percent of the data lies within two standard deviations of the mean?

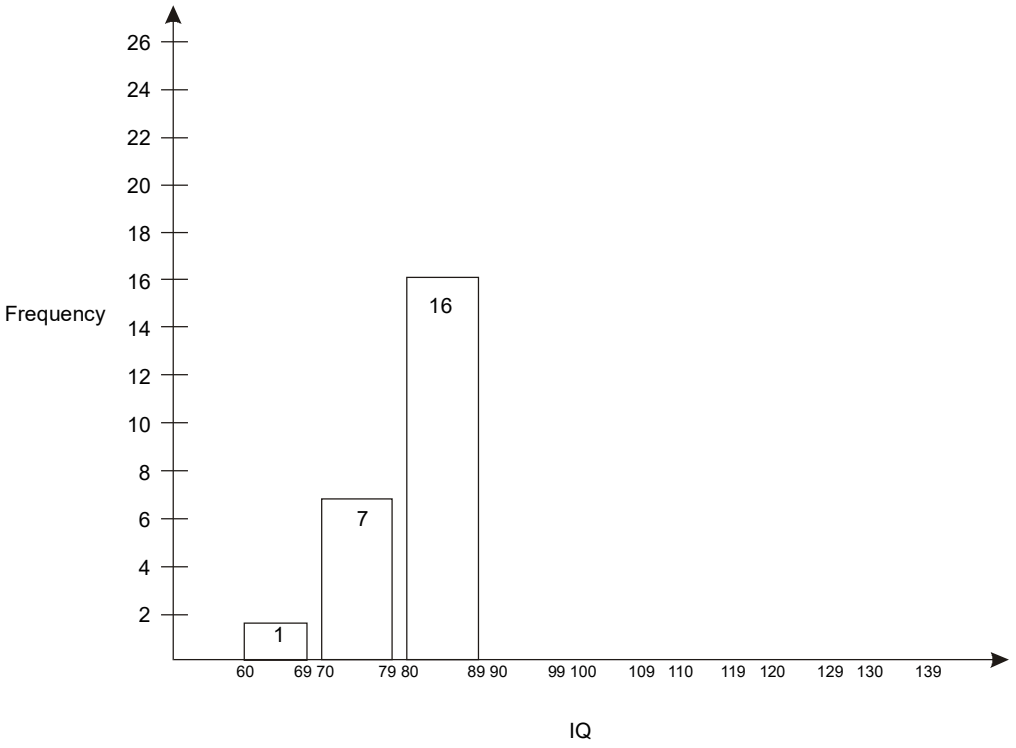
_____

f. What percent lies within three standard deviations of the mean? _____

3. The following is a collection of IQ scores: IQ stands for "intelligence quotient".

66	81	88	93	97	100	102	106	112	119
71	83	89	93	98	100	102	107	112	119
71	83	89	95	98	100	102	107	113	121
72	84	89	95	98	100	102	107	113	122
73	85	90	96	98	100	103	108	114	123
74	85	91	96	98	100	103	110	114	126
76	85	92	96	99	100	103	110	115	126
77	86	92	97	99	101	104	111	117	127
80	86	92	97	99	101	105	111	118	130
81	88	92	97	99	101	106	112	118	136

a. Complete the histogram for the data below.



b. Find the range. _____

c. Find the mean.

$\bar{x}$ = _____

d. Find the standard deviation.

σ = _____

e. What percent of the IQ's lie within one standard deviation of the mean?

_____

f. What percent lie within 2 standard deviations of the mean?

_____

g. What percent lie within 3 standard deviations of the mean?

_____

h. Does IQ seem to be normally distributed?

_____

i. What IQ score would be three standard deviations above the mean?

_____

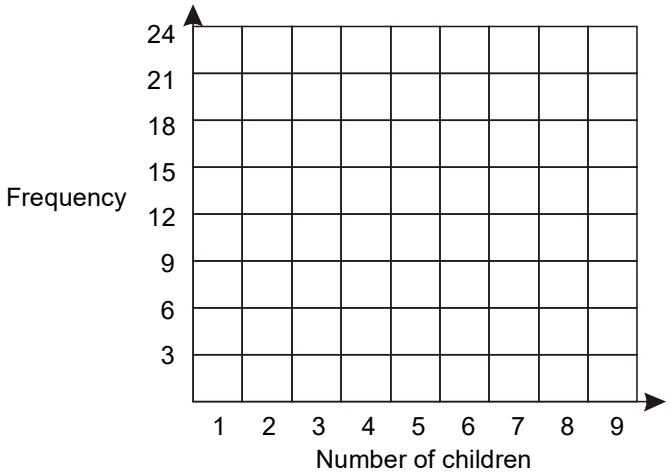
Answers are on pages 69.

Activity 5: Brothers and sisters

Duplicate the survey conducted in Question 1 of Exercise 5. Ask 40 people how many siblings (brothers and sisters) they have. Record the answers. Add 1 to each number so that the person asked is included.

a. What is the range for this data? _____

b. Draw a histogram for the data below.



c. Find the mean for this data. $\bar{x}$ = _____

d. Find the standard deviation. σ = _____

e. Find the values, $\bar{x} - \sigma$ and $\bar{x} + \sigma$. _____ and _____.

f. What percent of the data lies between $\bar{x} - \sigma$ and $\bar{x} + \sigma$? _____

g. What are the values, $\bar{x} - 2\sigma$ and $\bar{x} + 2\sigma$? _____ and _____

h. What percent of the data lies within two standard deviations of the mean?

_____

i. What are the $\bar{x} - 3\sigma$ and $\bar{x} + 3\sigma$ values? _____ and _____

j. What percent of the data lies within three standard deviations of the mean?

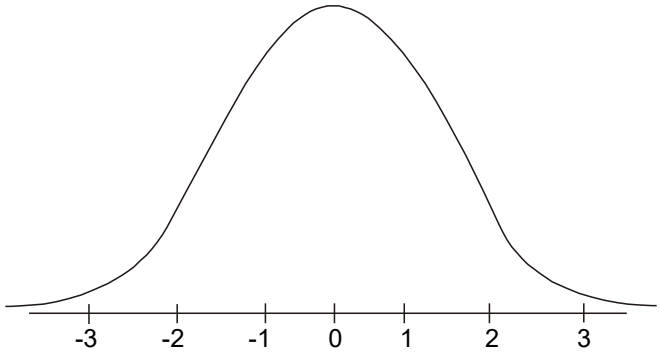
_____

k. Compare your answers for f., h., and j. to the results predicted by the Empirical Rule. Does the result suggest an approximately normal distribution?

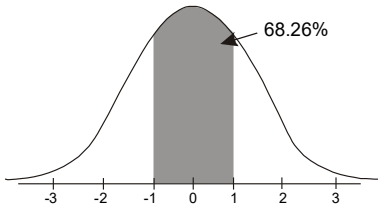
_____

Unit 6: The normal curve

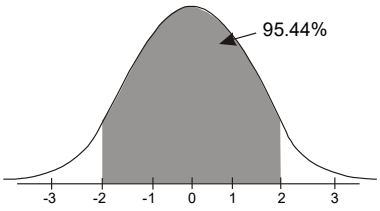
The *normal curve* is an idealized representation of a normally distributed population. The normal curve, also called a bell-shaped curve, is drawn below, where the mean score is 0 and the standard deviation is 1.



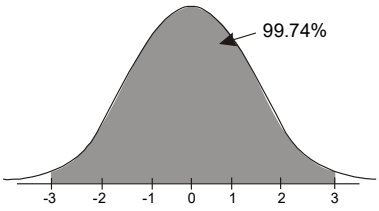
The *area under the curve* represents 100% (or 1.00) of the data (or population). By the empirical rule, the area under the curve and within one standard deviation of the mean is 68.26%, within two standard deviations is 95.44%, and within three standard deviations is 99.74%, as shown below.



The area under the curve is 0.6826



The area under the curve is 0.9544



The area under the curve is 0.9974

The normal curve is the graph of the exponential equation;

$$y = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \quad \text{where } e \approx 2.718$$

By using *z scores*, the area of any region under the curve can be determined.

The *z score* or *standard score* represents the number of standard deviations a data value is from the mean value. The formula for z is,

$$z = \frac{x - \bar{x}}{\sigma}$$

Where x is a data value, $\bar{x}$ is the mean, and σ is the standard deviation.

Example 1

The mean IQ is 100 and the standard deviation is 15. If Frank has an IQ of 127, find his z score.

Solution

Here, $\bar{x} = 100$, $\sigma = 15$, and x = 127.

$$z = \frac{127 - 100}{15} = 1.8$$

A z score is similar to a percentile in that it is a *measure of position*. As a rule, z scores above 2.0 (or below –2.0) are considered "unusual" values. In a normal population such scores would occur less than 5% of the time. Z scores between –2.0 and 2.0 are considered "ordinary" values.

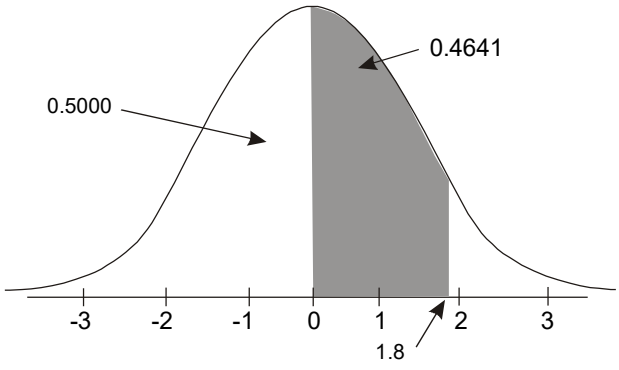
Example 2

Frank has an IQ of 127, or a z score of 1.8. What percent of the population have IQ scores less than (or equal to) 127 and what percent have IQ scores higher than 127?

Solution

Refer to Appendix A (see page 66).
Locate 1.8 under the z column and read across to the value under the 0.00 column.

A z score of 1.8 relates to the area under the curve from 0 to 1.80. The area is 0.4641. The area under the curve from –∞ to 0 is 0.5000.



The percent of the population that have IQ scores less than (or equal to) 127 is,

$$0.5000 + 0.4641 = 0.9641 \text{ or } 96.41\%.$$

A z score of 1.8 can be considered as equivalent to a percentile of 96, since it is higher than 96.41% of the population. In other words, an IQ of 127 has a 96th percentile ranking.

The percent of the population with IQ scores above 127 is,

$$1.0000 - 0.9641 = 0.0359 \text{ or } 3.59\%$$

Areas under the normal curve can also be associated with *probabilities*. In the above example we could say that the probability that some person would have an IQ score less than 127 is 0.9641 out of 1.0000 or about $\frac{96}{100}$ or $96\% = \frac{24}{25}$.

The probability that someone would have an IQ higher than 127 is about 4% or $\frac{1}{25}$.

Or, in a group of 25 people, chosen randomly, probably only one would have an IQ score of more than 127.

Example 3

The waiting-in-line time at a certain grocery store is normally distributed with a mean of 3.5 minutes and a standard deviation of 1.4 minutes.

a. What percent of the customers wait in line less than one minute?

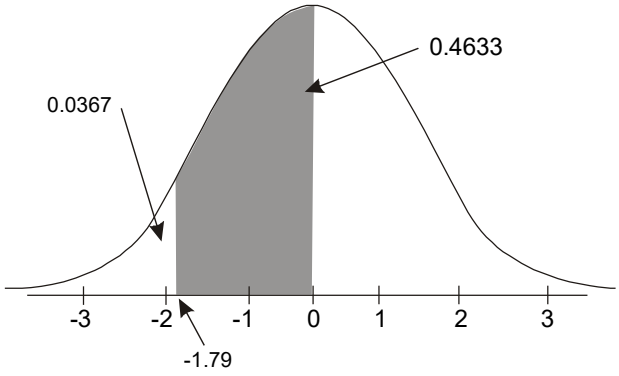
b. What percent of the customers wait in line more than 5 minutes?

c. What is the probability that a customer would have to wait in line for more than 7 minutes?

Solution

a. Convert 1 minute to a z score.

$$z = \frac{x - \bar{x}}{\sigma} = \frac{1 - 3.5}{1.4} = -1.79$$



From the table (Appendix A, see page 66), a z score of –1.79 yields the same area as 1.79. The area between 0 and –1.79 is 0.4633. The area to the left of –1.79 represents the proportion of the population that waits in line less than a minute,

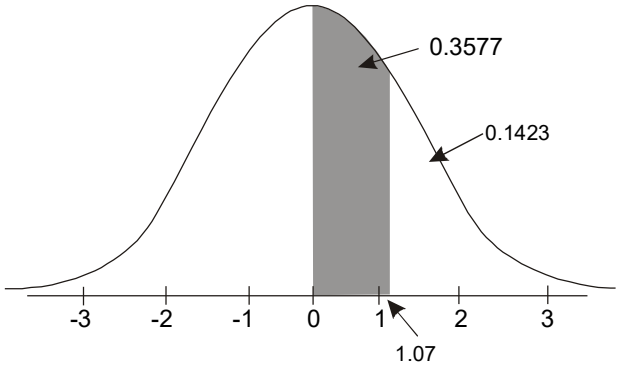
$$0.5000 - 0.4633 = 0.0367 \text{ or } 3.67\%$$

b. Convert 5 minutes to a z score.

$$z = \frac{5 - 3.5}{1.4} = 1.07$$

The area under the curve between 0.0 and 1.07 is 0.3577 and the area beyond 1.07 is

$0.5000 - 0.3577 = 0.1423$

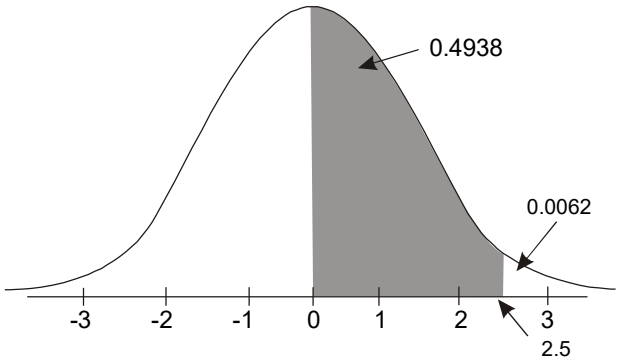


This means that 14.23% of the customers have to wait in line for more than 5 minutes.

c. Convert 7 minutes to a z score.

$$z = \frac{7 - 3.5}{1.4} = 2.5$$

The area under the curve between 0 and 2.5 is 0.4938. The area beyond 2.5 is $0.5000 - 0.4938 = 0.0062$.



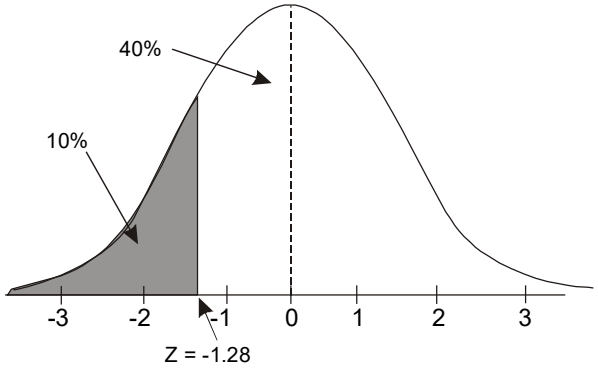
This means that 0.62%, or less than 1%, of the customers would have to stand in line for more than 7 minutes. The probability that someone would have to stand in line for more than 7 minutes is 62 in 10000 or $\frac{31}{5000}$.

Example 4

A certain tire company tested their new Treadmasters and found that the tires' tread life averaged 60000 km with a standard deviation of 7000 km. The company wants to sell the Treadmaster with a guarantee that they will last a certain number of kilometres. The company is willing to give a money back guarantee on 10% of its worst tires. At how many kilometres will 10% of the tires be worn out?

Solution

We need to find the z score that marks off an area under the curve of 10% or 40% from 0 to z. In the table, the closest value to 0.4000 is 0.3997, and this corresponds to a z score of 1.28.



To determine the kilometre value that is associated with a z score of –1.28, solve the z score formula for x.

$$z = \frac{x - \bar{x}}{\sigma}$$

$$-1.28 = \frac{x - 60000}{7000}$$

$$x = 60000 - 1.28\ (7000)$$

$$x = 51040 \text{ km}$$

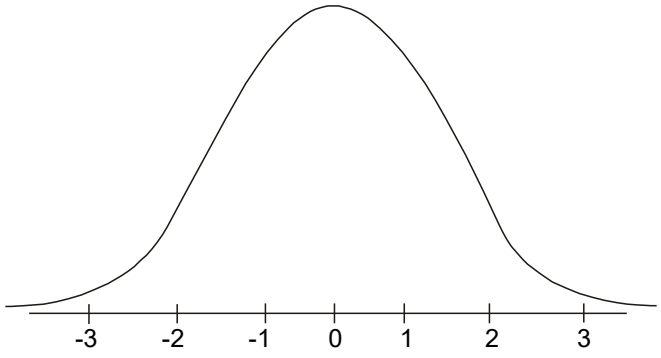
The company should guarantee tires that wear out before 51040 kilometres.

Now complete Exercise 6 and check your answers.

Exercise 6

Use Appendix A (see page 66) for the questions that follow.

1. Find the area under the normal curve between the following z scores.



a. z = 0 and z = 1.41 _____

b. z = -0.6 and z = 0 _____

c. z = -1.23 and z = 0.53 _____

d. z = 0.46 and z = 2.31 _____

e. z > 1.5 _____

2. The average resting heartrate for a normally distributed population of men was found to be 62 beats per minute with a standard deviation of 11 beats per minutes.

a. What percent of men have resting heartrates under 70 beats per minute?

_____

b. What percent of men have resting heartrates over 70 beats per minute?

_____

c. What percent of men have resting heartrates between 40 and 80 beats per minute?

_____

3. In a group of normally distributed women, the average height is 5 feet 4 inches (64 inches) with a standard deviation of 2.8 inches.

a. What percent of the women are between 5 feet and 6 feet ?

_____

b. What is the probability that a woman would be taller than 6 feet ?

_____

c. What is the probability that a woman would be shorter than 5 feet tall?

_____

4. A survey of college students enrolled in technology programs indicated that they spent an average of 29 hours a week outside of class time studying for their courses. The data was normally distributed with a standard deviation of 9 hours per week.

a. What percent of the students spent more than 40 hours per week studying?

_____

b. What percent spent fewer than 10 hours per week studying?

_____

c. What percent spent between 20 and 50 hours per week studying?

_____

5. Larry's lightbulb factory manufactures bulbs with an average life of 1000 hours and a standard deviation of 100 hours. To sell more light bulbs Larry wishes to give a

guarantee, but he is only willing to replace 5% of the lightbulbs sold. For how many hours should the lightbulbs be guaranteed?

6. Workers in a certain factory are given a bonus every time they assemble more than 300 toy cars in one eight hour day. The number of toy cars assembled each day by a worker is normally distributed with a mean of 270 cars and a standard deviation of 16 cars. What percent of the workers receive a bonus each day?

7. A radar unit measures the speed of passing cars on a highway. The speeds of the cars are normally distributed with a mean speed of 104 km/h.

   a. Find the standard deviation of the speeds if 3% of the cars are travelling faster than 115 km/h.

   b. Using the standard deviation found above, what percent of the cars are travelling at less than 90 km/h?

   c. What percent are travelling faster than 120 km/h?

_____

d. If there is a no tolerance rule in effect, and the posted speed is 100 km/h how many cars would be considered to be speeding?

_____

Answers are on page 69.

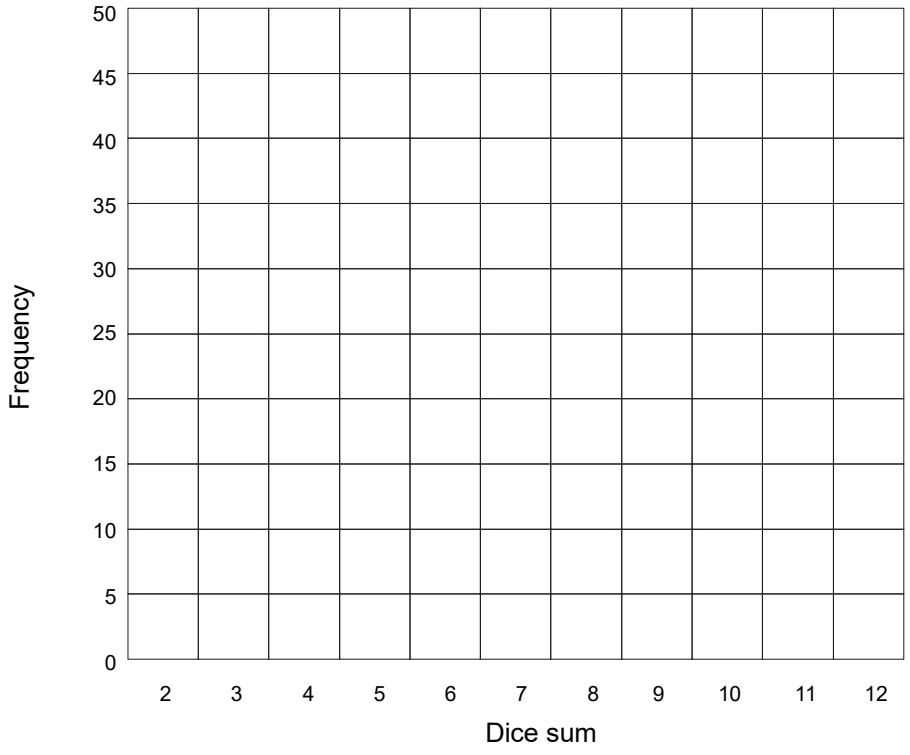
Activity 6: Rolling dice

With a partner, roll a pair of dice 150 times. Your partner should tally each roll of the dice, while you keep count of the number of rolls. Complete the tally sheet below and draw a histogram for this data.

a.

Dice sum	Tally	Frequency
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		

b.



c. Find the mean. $\bar{x}$ = _____

d. Find the standard deviation σ = _____

e. Find the interval $\bar{x} - \sigma$ to $\bar{x} + \sigma$ _____ to _____

f. What percent of the rolls lie within one standard deviation of the mean?

_____

g. What are $\bar{x} - 2\sigma$ and $\bar{x} + 2\sigma$ and what percent of the data lie within 2 standard deviations of the mean?

_____

h. What are $\bar{x} - 3\sigma$ and $\bar{x} + 3\sigma$ and what percent of the data lies within 3 standard deviations from the mean?

_____

_____

i. Does the data appear to be normally distributed? _____

j. What z score is associated with a roll of 9? _____

k. What percent of the rolls would be greater than 9? _____

Test this by rolling the dice ten times. How many times out of ten did a roll of 10, 11, or 12 occur?

_____

What percent is this? _____

Repeat: Roll ten more times and count how many times a 10, 11, or 12 was rolled.

Unit 7: Analysing survey data

Hardly a day goes by without the media reporting the results of some *survey*. Surveys are conducted to determine what people like or dislike, what their opinions are on various issues and what factors affect their lives.

Governments and businesses often use surveys in order to make decisions and to monitor the effectiveness of previous decisions.

We will restrict our analysis of survey data to YES-NO population surveys only. In a YES-NO survey, every member of the population answers a question with a YES or a NO. For example, "Do you smoke?" is a YES or NO type question. "How many cigarettes do you smoke every day?" is not a YES or NO question. If we were to ask every member of a population the YES or NO question we would be taking a *census* of the population. If 20% of the population answered YES to the question, this would be called a *20% yes population*.

It is often very expensive and very time consuming to take a population census. By using a smaller *sample* of the population, we can estimate the percentage of YES answers in the population.

For example, in a recent survey of 1000 Canadians, 55% responded YES to the question, "Is having a happy life the thing that matters most to you?" when compared to other things like health and freedom. Even though the survey only represents a small portion of the total population of Canada, its margin of error is calculated to be plus or minus 3.2% 19 times out of 20.

In other words, if that survey was repeated 20 times, using a different 1000 Canadians each time, then 19 times out of 20 times, the number of YES responses to the above question would be between 51.8% (55% - 3.2%) and 58.2% (55% + 3.2%).

Sampling error for a 95% *confidence interval*.

If the sample size is n, then the sampling error of the percentage of YES answers in the population is approximately,

$$\frac{100}{\sqrt{n}}$$

When $n > 100$, then the accuracy of $\frac{100}{\sqrt{n}}$ is quite good.

Since 19 out of 20 is equivalent to 95%, we can be "confident" that the percentage of YES answers will be within the sampling error interval 95% of the time.

Example 1

The business association in Grissville surveyed 384 people and asked each if they had eaten dinner in a local restaurant at least once in the last week. 223 people responded YES to the survey. Find the 95% confidence interval (and sampling error) for this sample.

Solution

The proportion of YES answers in the population is $\frac{223}{384} = 0.581$ or 58%.

The sampling error, $\frac{100}{\sqrt{384}}$ is about 5.1%.

The 95% confidence interval is 58% - 5.1% and 58% + 5.1% or about 53% to 63%.

Now complete Exercise 7 and check your answers.

Exercise 7

1. Read the following newspaper clipping.

> The latest poll, conducted for Victoria radio station CBC shows Mr. Martin with 39.9 percent support, Ms. Wilson with 33.9 percent, and Mr. Yeung with 8.5 percent.

a. Why do the three percentages not add to 100%?

_____

b. Suppose this poll was accurate to plus or minus 3 percentage points 19 times out of 20.

i. What is the least support Mr. Martin could have 19 times out of 20?

_____

ii. What is the most support Ms. Wilson could have 19 times out of 20?

_____

iii. Considering the above and the percent of undecided voters, does Mr. Martin have a majority of the potential vote?

_____

2. A survey was conducted to determine whether bicycle riders should have to pay for a licence to ride on the city streets. 183 people were asked and 57 said yes.

a. What percent of the people responded "yes" to the survey? _____

b. What is the sampling error for this survey? _____

c. What is the 95% confidence interval for this survey? _____

3. An opinion poll reported that support for the Liberals stood at 58% with a sampling error of 2.5% 19 times out of 20. Use the sampling error formula to determine the sample size.

4. A certain poll found that 833 people out of 1240 thought that capital punishment should be reinstated for first degree murder. What is the 95% confidence interval for this sample?

5. If the poll in Question 4 was conducted in 1998, are the results still valid today?

6. Count the first one hundred letters in this sentence and then count how many times the letter "e" occurred and count every "e" as a YES response.

   a. What percent of the 100 letters were "e's"?

   b. What is the 95% confidence interval for the occurrence of the letter "e" in the English language?

   c. Repeat the above process with a different set of words and record the percent of "e's" found in the passage. Does this percent fall in the confidence interval range found in part b. above?

7. See if your calculator has random number function (RAN# button). It should produce three digit decimal numbers randomly. In other words, every number has an equal chance of showing up on your screen. Assume that every time a 3, 6 or 9 appears as the last digit of the random number, it is the same as receiving a YES response. Then,

theoretically a YES response should occur 30% of the time, since 3, 6 and 9 are three out of ten possible last digits in each random.

a. Find the sampling error for this 30% yes population if the sample size is 100 random numbers.

b. Generate 100 random numbers and tally the number of times a 3, 6 or 9 occurred as the last digit. What percent of the time did a 3, 6 or 9 occur?

c. What is the 95% confidence interval for this sample?

d. Check with the other students. Did their samples produce percentages within the 20 to 40 percent interval 19 times out of 20 times?

Answers are on page 69.

Activity 7: Smoking

When conducting a survey it is important to ask simple unambiguous questions. It is also important to select a sample that is representative of the population being surveyed. The following exercise should demonstrate the importance of sample size.

Ask various students the question, "Do you smoke?" (If there is any confusion about what you are asking, you could say, "Have you smoked a cigarette in the last 48 hours?") Record the number of YES responses and then calculate the percentage of YES responses.

Number of students asked	Number of YES responses	Percentage of YES responses
1		
2		
4		
8		
16		
25		
30		
35		

a. As the sample size increased, did the variation in percentages increase or decrease?

_____

b. According to Statistics Canada, 1996-1997, about 20% of the BC population are smokers. Are your results close to 20%?

_____

c. You sampled 35 college students. Are these students representative of the total college population? What problems might there be with your sample in terms of it being representative of the whole population?

_____

_____

Unit 8: A statistics project

Now it is your turn. You or your group will select a topic of interest, collect data regarding the topic, and statistically analyse the results. Choose a topic and then let your instructor know what it is before presenting your results in Exercise 8.

A few possible topics are listed below. However, feel free to identify one of your own.

1. Number of cups of coffee (cans of pop) consumed by a student in one day.
2. Number of keys (or credit cards) carried by a person while at the college.
3. Number of cars passing a certain intersection every minute.
4. Number of cigarettes smoked per day by smokers.
5. Minutes spent studying last night.
6. Initial copyright dates on books in the library.
7. Resting heart rates of males (or females).
8. Total minutes per week spent on exercise by college students.
9. Age difference (in months) between oldest and youngest siblings (or between spouses).
10. Waiting time in line (grocery store or bank) during busy times.

Now complete Exercise 8 and check your answers.

Exercise 8

1. What topic have you chosen?

2. Before you collect the data, try to predict the mean and range of the data.

3. You must collect 30 or more data values. List your data below.

4. Present your data graphically below. Use either a line graph, bar graph or stem and leaf plot.

5. Describe the shape of your graph in 4. Is it normal, J-shaped, skewed or bimodal?

6. Calculate the following statistics for your data.

$\bar{x}$ = _____

median = _____

mode = _____

range = _____

σ = _____

Q_1 = _____

Q_3 = _____

7. What measure of central tendency best describes this set of data?

8. Which data value is higher than 90% of the rest of the data?

9. What percent of the data lies within,

a. one standard deviation of the mean?

_____

b. two standard deviations of the mean?

_____

c. three standard deviations of the mean?

_____

10. How do the actual mean and range compare with your predicted mean and range?

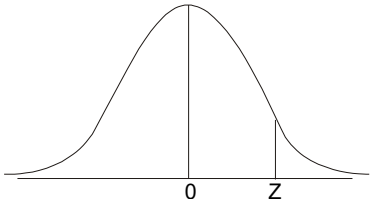
11. Was the sample you chose biased in any way? Were you biased in the way you collected the data?

12. If you were to do this project over again, how could you improve it?

13. Write a concluding statement about the data.

Appendix A

The entries in this table represent the area under the normal curve from 0 to z. Areas for negative values of z are obtained by symmetry.



$$z = \frac{x - \bar{x}}{\sigma}$$

Second decimal place in z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.454
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998									
4.0	0.49997									
4.5	0.499997									
5.0	0.4999997									

Appendix B

Activity 4: Quiz

Speed counts. Work as quickly as possible and do not use a calculator.

1. Add 1 + 3 + 5 + 7 + 9 + 11 + 13 + 15 + 17 + 19

_____

2. How many months have 28 days?

_____

3. Divide 8 by 0.

_____

4. Subtract 23.79 from 30.02.

_____

5. Write down the 5^{th}, 25^{th} and middle two letters of the alphabet.

_____

_____

_____

_____

6. Multiply 6.87 by 0.96.

_____

7. Name four countries in Africa.

_____

_____

_____

_____

8. Divide 1.3 by 0.52.

_____

9. Use the letters given to form three words across and three words down. (One letter per square.)

A B E E I R

T		N
		T

Answers are on page 75.

Answers

Exercise 1

1. The lengths of the bars suggest that men only have a life expectancy of one half or 50% that of women. This is a result of starting the years' scale at 75 years rather than at zero years. Actually, men have a life expectancy of $\frac{78}{81}$ or 96% that of women.

2. a. Only people with very strong opinions would bother to phone in.

   b. Most people do not respond to mail surveys, so the sample would be quite small. People who do not read books would most likely not respond. The survey might not be representative of all the geographical, cultural and economic sectors of Vancouver.

   c. Butler is probably at the very place where the smokers are hanging out.

Exercise 2

1. a. $\bar{x}$ = \$25 866.67, median = \$25 000, mode = \$25 000, range = \$50 000

   b. the mode c. the range

2. a.

	daily mean	daily range
1	3	6
2	2.5	1
3	8	8
4	11.5	3
5	11	8
6	11.5	3
7	10	4

   b. 8.2º C

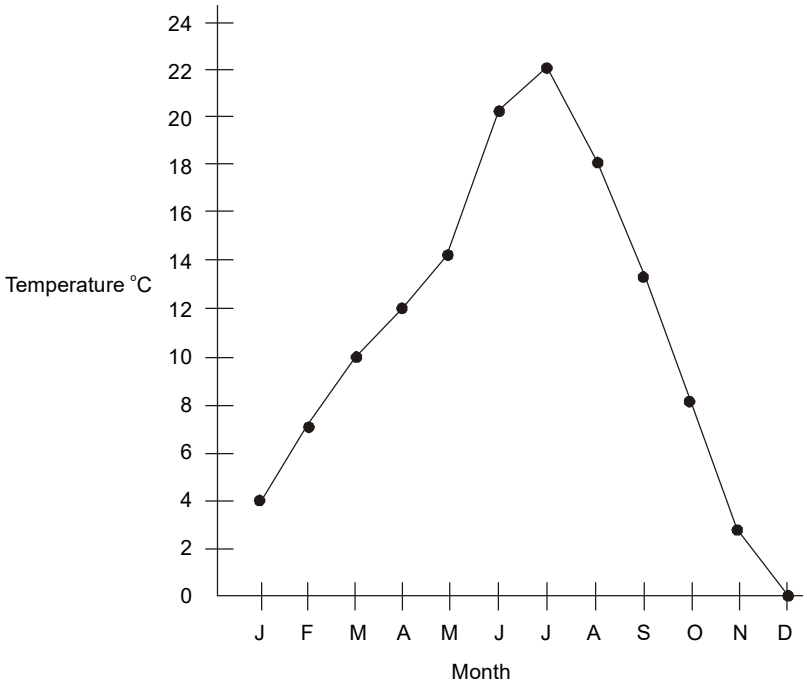
   c. 15º C

   d. 4.7º C

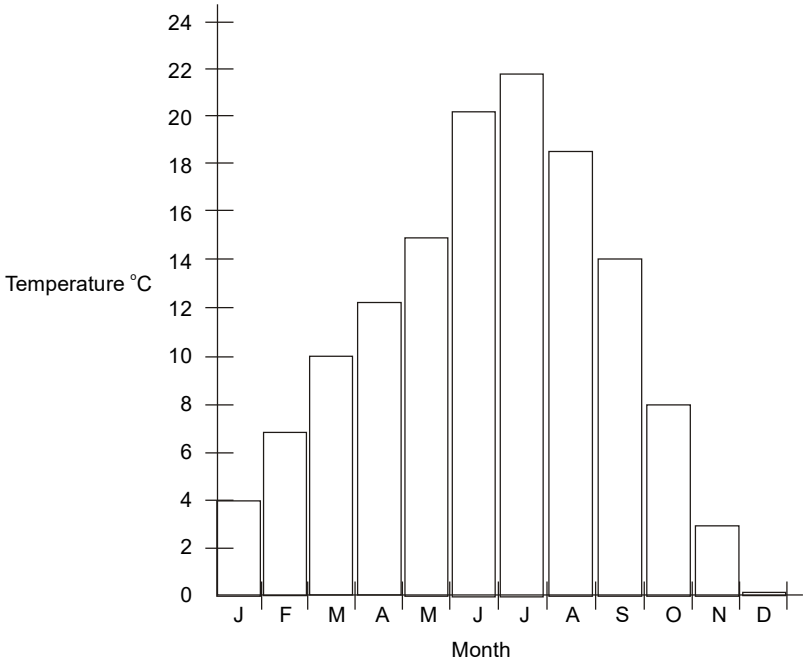
3. a. A b. A c. A d. B

4.

Average monthly temperatures for Vancouver over a one-year period (frequency graph)



Average monthly temperatures for Vancouver over a one-year period (histogram)



5. Computer course final grades.

4	0	2	2		
5	0	4	6	8	
6	6	6	8	9	
7	0	3			
8	0	4	5	8	9
9	3				

6. a. 25 students
   b. $\bar{x}$ = 63.5 seconds
   median = 63 seconds

7. $\dfrac{78\% + 78\% + 78\% + 78\% + 78\% + x}{6} = 80\%, \quad x = 90\%$

8. $\dfrac{182 + x}{260 + 100} = 0.75, \quad x = 88$ Neil's chances are poor.
   He has only averaged 70% on his tests thus far.

Exercise 3

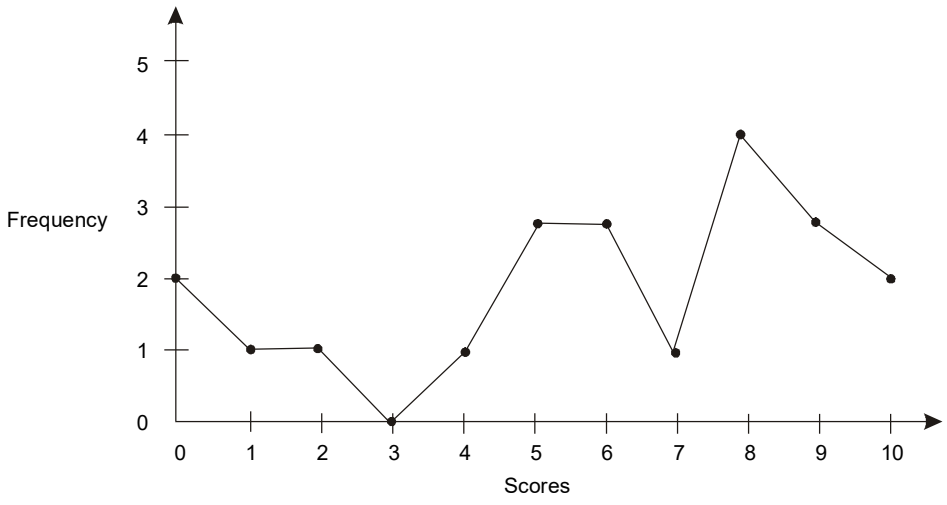
1. a. $P_{50} = 119$ and $P_{80} = 138$
   b. Jill scored at the 82nd percentile since $140 = P_{82}$. Jill can type faster than 82% of the other students.
   c. $90 = P_{12}$ or 90 is the 12th percentile. 12% or 15 out of 125 failed the test.
   d. $P_{20} = 98.5$

2. a. Q_3 or $P_{75} = 13$
   b. $13 = P_{70}$
   c. The discrepancy is a result of the relatively small (n = 70) sample size. As the sample size increases, these discrepancies become smaller.

Exercise 4

1. a. 10 b. 8 c. 6 d. 6 e. 3.1

f.



g. $\bar{x} - \sigma = 6 - 3.1 = 2.9$

$\bar{x} + \sigma = 6 + 3.1 = 9.1$

h. There are 15 scores between 2.9 and 9.1 or, 71% of all the scores lie within one standard deviation of the mean.

2. a. $\bar{x} + \sigma = 12.97$ and $\bar{x} - \sigma = 7.03$. $\frac{8}{12}$ or 66.7% lie within one standard deviation of the mean.

b. $\bar{x} + 2\sigma = 15.94$ and $\bar{x} - 2\sigma = 4.06$ All of the scores, or 100%, lie within two standard deviations of the mean.

3. All the data values are equal.

4. In the first class most of the scores are between

$\bar{x} \pm \sigma$ or $75 \pm 5 = 70$ to 80.

In the second class, $\bar{x} \pm \sigma = 70 \pm 15 = 55$ to 85.

It is more likely that the second class would have more scores of 85 or better.

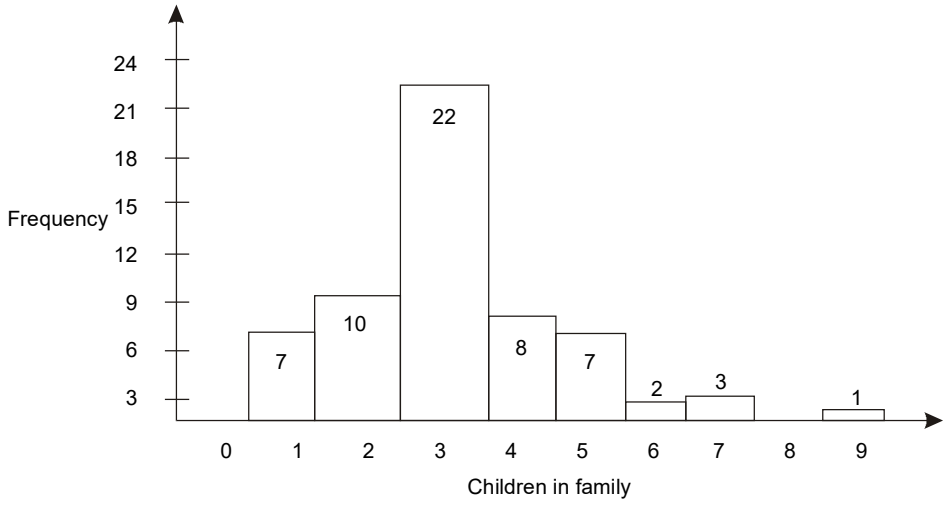
5. a. 55 kg b. 80.5 kg c. 13.6 kg d. 83.3%

6. a. 4.9 minutes b. 1.46 minutes

Exercise 5

1. a. 9-1 = 8

b.



c. $\bar{x} = 3.4$ d. $\sigma = 1.7$

e. 1.7 and 5.1 f. $\frac{47}{60} = 78.3\%$

g. 0.0 and 6.8 h. $\frac{56}{60} = 93.3\%$ i. -1.7 and 8.4

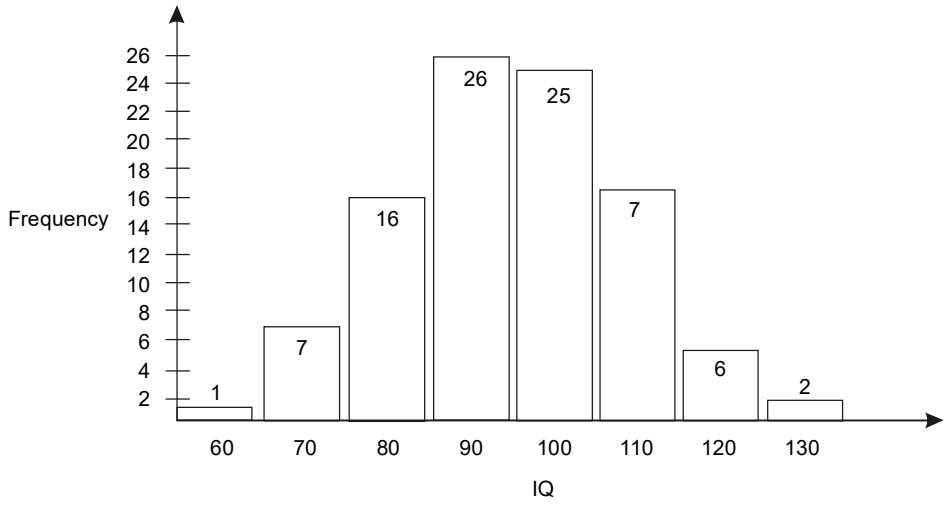
j. $\frac{59}{60} = 98.3\%$ k. yes

2. a.

Hours	Frequency
0.5	3
1.0	11
1.5	12
2.0	17
2.5	15
3.0	13
3.5	4

b. $\bar{x} = 2.07$
c. $\sigma = 0.78$
d. 58.7%
e. 96.0%
f. 100%

3. a.



b. 70 c. 99.5 d. 14.2 e. 67%

f. 95% g. 100% h. yes i. 142.1

Exercise 6

1. a. 0.4207 b. 0.2257 c. 0.5926 d. 0.3124 e. 0.0668

2. a. 76.73% b. 23.27% c. 92.67%

3. a. 92.15% b. 0.21% or about 1 in 500 c. 7.64% or about 3 in 40

4. a. 11.12% b. 1.74% c. 83.14%

5. $x = z\ \sigma + \bar{x}$ = -1.65 (100) + 1000 = 835 hours

6. About 3% of the workers.

7. a. $\sigma = \frac{x - \bar{x}}{z} = \frac{115 - 104}{1.88} = 5.85$ b. 0.84% c. 0.31%

d. 75.17%

Exercise 7

1. a. There are "undecided" voters.

b. i. 39.9% - 3% = 36.9% ii. 33.9% + 3% = 36.9%

iii. Yes. Even at Martin's worst and Wilson's best, Wilson cannot overtake Martin.

2. a. $\frac{57}{183} = 31.1\%$ b. $\frac{100}{\sqrt{183}} = 7.4\%$ c. between 23.7% and 38.5%

3. $2.5 = \frac{100}{\sqrt{n}}$ or $n = 1600$

4. $\frac{833}{1240} = 67.2\%$ and $\frac{100}{\sqrt{1240}} = 2.8\%$. 67.2% + 2.8% = 70% and 67.2% - 2.8% = 64.4%.
Rounding, the 95% confidence interval is 64% to 70%.

5. No. The results of polls are usually only valid for the day they are taken.

6. a. There are 18 "e's" or 18%.
b. 8% to 28%.
c. Check with your instructor.

7. a. $\frac{100}{\sqrt{100}} = 10\%$ c. 20% to 40% b. and d. Check with your instructor.

Appendix B - Quiz

1. 100
2. all the months have 28 days
3. meaningless – 8 can't be divided by 0
4. 6.23
5. E, Y, M and N
6. 6.5952
7. **Northern Africa**
   Algeria
   Egypt
   Libya
   Morocco
   Sudan
   Tunisia
   Western Africa
   Benin
   Burkina Faso
   Cape Verde
   Cote d'Ivoire
   Gambia
   Ghana
   Guinea
   Guinea-Bissau
   Liberia
   Mali
   Mauritania
   Niger
   Nigeria
   Senegal
   Sierra Leone
   Togo

   Eastern Africa
   Burundi
   Comoros
   Djibouti
   Ethiopia
   Kenya
   Madagascar
   Malawi
   Mauritius
   Mozambique
   Reunion
   Rwanda
   Somalia
   Tanzania
   Uganda
   Zambia
   Zimbabwe

   Middle Africa
   Angola
   Cameroon
   Central African Republic
   Chad
   Congo
   Equatorial Guinea
   Gabon
   Zaire
   Southern Africa
   Botswana
   Lesotho
   Namibia
   South Africa
   Swaziland
8. 2.5
9.

T	I	N
A	R	E
B	E	T