# AN INTRODUCTION TO STATISTICS

Shirleen Luttrell

Sandy Lake Academy

2011

# ACKNOWLEDGMENTS

I share this book because as a teacher with limited resources and time, I found myself in a predicament during Spring 2010, and I realized there might be other teachers who find themselves in similar positions. I had a textbook that only contained one chapter of statistics and a curriculum that expected a semester's worth. I also taught at a school that had no budget for anything more than my classroom books. What to do? Hunt online for worksheets? Photocopy textbooks I owned? And then I thought about my mentor, Keith Calkins, who said it was easier to write what you envisioned than to hunt hours for something compatible because in the end your book would tie it all together for easy future reference. So… this is my collection compiled to teach the course as I envisioned it to be. Of course, nothing takes the place of classroom discussion and notes that supplement and further explain with pictures and such.

Like the Academy Awards, I want to give credit where credit is due. I owe everything to my Lord and Saviour Jesus Christ. He gave me the strength to continue when I was unexpectedly given a new course half way through the term in a field I had never taught. On top of that I was already preparing for six different classes, in charge of yearbook and assistant in our fundraising program. My stress levels were high trying to balance everything! So praise God who gives us what we need to carry on.

I want to thank my mother secondly because she keeps her eyes out for mathematical material. She was in a used book store one day and found a Statistics book. She called me that night and asked if I wanted her to go back and buy the book! Most of my material comes from her shopping sprees. She saved me a lot of time that semester.

Thirdly, I want to thank my mentor Keith Calkins. He first hired me in 1998 as an assistant whose primary job was to write out a web-book on number theory. From there I ended up helping to write two statistics books which he has since made many editions. You might still find those volumes at www.andrews.edu/~calkins. He also provided valuable input on my own book, pointing out some trouble spots. He understood that statistics wasn't my favourite field but that I wanted to inspire my students in spite of my own lack of enthusiasm.

I also want to thank Bruce Wentzell and my sister Nicole Luttrell, who gave me a dash of realism and yet the encouragement to continue. When I was holed up in my office writing, Nicole was holding up our house. She took on more of the fundraising program, care of household maintenance, evening dishes and entertaining our grandmother who lives with us. She helped me relieve my stress by taking me for walks with my dog in tow or taking our grandmother for Sunday drives so I could have peace to write.

Lastly, and without whom I could not do such a thing, are my resources. I owe them much. I read and reread their books until they poured out of me. So I cannot claim this book as my very own. Everything I did come up with on my own was still heavily influenced by those on my bookshelves. You will find my sources cited at the end of my document.

Shirleen Luttrell, High School Teacher
Sandy Lake Academy, Nova Scotia, Canada

# INTRODUCTION TO STATISTICS

# STATISTICS LESSON 1

## LEVELS OF DATA MEASUREMENT

The term statistics has two basic meanings.[1] First, statistics is a subject or field of study closely related to mathematics. Descriptive statistics generally describes a set of data by graphically displaying the information or describing its central tendencies and how it is distributed. Inferential statistics tries to predict information about a population based on information from a sample. The second definition of statistics is the collection of methods used in planning an experiment and analyzing data to draw accurate conclusions.

In the above paragraph, the words population and sample were used. Population is a term describing a complete set of data. In general mathematics, this term is equivalent to the universal set. The term varies on its application. The population could be as broad as humans, people in North America, male Canadians, or Nova Scotia 15-19 male students. Sample is a portion of the population; in general math terms this would be equivalent to subset. If the population is Nova Scotia students, then a sample could be HRM students or SLA students.

To better understand a sample or population, people gather data. A parameter is a characteristic (data) of a population; whereas statistic is a characteristic of a sample. Data can be classified as being either qualitative or quantitative. The roots of these words will help define the type of data. Qualitative has a root from quality, so adjectives that describe the sample like colour and size are examples of qualitative data. Quantity is the root of quantitative, so any numeric data is an example of quantitative data. Quantitative data can be distinguished further as discrete or continuous data. Noting the differences in data may seem a trivial task, but its importance will manifest when trying to draw pictorial representations of the data. Some graphs or pictures work better for discrete data.[2]

Discrete data have a finite number of possible values. Examples are {small, medium, large}, {red, white, blue}, and {poor, fair, excellent}. Continuous data have infinite possibilities. Gas bought at a service station is an example of continuous data: {1 L, 1.05L, 1.0005L...}.

Data is obtained by measuring some feature of a population or sample. Based on the type of measurement, certain mathematical operations can be performed. Identifying the type of measurement used before proceeding with any analyze of the data is important.

The levels of measurement spell the acronym NOIR: [3]
Nominal: data that has no order, thus names or label of categories.
Example: Nissan, Honda, Toyota.
Ordinal: data that has an order, but intervals are not meaningful.
Example: poor, fair, well.
Interval: data with order and meaningful intervals, but no reference point.
Example: Celsius, Fahrenheit
Ratio: data with order, meaningful intervals, and has a reference point.
Example: Kelvin, test scores

We can say that Suzy scored twice as high as Wilma on a test, but we can't say 100F is twice as hot as 50F. That is why it is good to know what type of measurement you are using, so you don't perform inappropriate mathematical operations!

## STATISTICS LESSON 1 HOMEWORK [4]

1. Identify whether the data is qualitative or quantitative. If quantitative, is it continuous or discrete?
   a. Height of basketball players
   b. Style of shoes worn by classmates
   c. Number of people in a household
2. Complete the comparison: Parameter is to ____________ as statistic is to __________.
3. If a numeric data is not discrete, then it must be ____________.
4. What is the difference between statistics and statistic?
5. What is the difference between descriptive and inferential statistics?
6. In the following, a) identify the population and sample, b) identify whether the data is quantitative or qualitative, and c) If the data is quantitative, whether it is continuous or discrete.
   a. Sobeys wishes to determine how many cartons of eggs are damaged in shipment. For every 10 shipments of 1000 cartons, Sobeys examines every 50th carton to see how many cartons contain cracked eggs.
   b. A survey of Canadian family to determine the average number of pets uses a computer to select 3 provinces, then 10 counties in each, then 50 families in each county. Each family is asked how many pets they have.
   c. A biologist tranquilizes 400 male deer to measure their antlers to determine their ages.
7. A student asked his classmates "what is your favourite sport?" What type of data would (s)he get? What level of measurement is used?
8. A survey of distances students travel to SLA is taken. What type of data would be expected? What level of measurement is used?
9. Students made cookies for each other in the fall. Some were asked how they liked the cookies. The responses varied from "oooh, it's gross" to "yummy, I want more!" How would you categorize the responses? What level of measurement and what type of data are you using?
10. Identify the following as discrete or continuous:
    a. Yesterday's record shows two students were absent.
    b. Volvo sold 84,000 cars in 1997.
    c. A 1999 VW Bug weighs 3,600 pounds.
    d. The radar clocked a baseball at 98.4 mph.
11. Determine the level of measurement:
    a. Color of M&Ms.
    b. Final course grades of A, B, C, D, F.
    c. Daily high and low temperature of Halifax, Nova Scotia.
    d. Time (in days) for a sunspot to be visible from the earth.

# STATISTICS LESSON 2

## DISPLAYING DATA WITH CHARTS

Sometimes inherent properties of raw data have to be teased out. The following scores are from the last math test: 21, 21, 24, 28, 29, 33, 35, 38, 39 (out of 40). If there were a hundred such scores, you couldn't tell much by looking at it. It would take a while for patterns, like averages, to appear. So statisticians usually make graphs of their data.

Stem-and-Leaf Plot: Useful for only quantitative data. Stem can be any unit, and the leaf is the collection of data having the same stem. The above data would look like:

| | |
|---|---|
| 2 | 11489 |
| 3 | 3589 |

By this approach we see most people got under 75% (30/40)on their test. We can identify the mode, those data that appear the most, as 21. Organizing will be useful in finding other properties. If the data set is extremely large, you might see a split stem-leaf plot.

| | |
|---|---|
| 2 | 114 |
| 2 | 89 |
| 3 | 3 |
| 3 | 589 |

Frequency charts: Useful for either qualitative or quantitative data. Create categories based on the data and for each category list the sum of data that falls in it. The above data would fill a frequency chart in either manner, with the category across the top or running down the side:

| 20-29 | 30-39 |
|---|---|
| 5 | 4 |

| | |
|---|---|
| 20-29 | 5 |
| 30-39 | 4 |

If the data were qualitative (or nominal) like colors of skittles, then the 5 red, 3 blue, 10 green, and 4 yellow skittles would have a table like the following.

| Red | Blue | Geen | Yellow |
|---|---|---|---|
| 5 | 3 | 10 | 4 |

Relative Frequency Charts present the percentages of each category. That saves people the trouble of calculating. For example, in the Skittle example there are twice as many green as red, but in terms of the total skittles in the hand, how much are the green? Almost half the bag, or 10 out of 22 or 45.5%.

The relative frequency chart for the skittle would look like:

| Red | Blue | Geen | Yellow |
|---|---|---|---|
| 22.7% | 13.6% | 45.5% | 18.2% |

When creating Frequency Charts for quantitative continuous data, the category may present a range of values. In the event that the category (class) is a range of values, great care should be taken that each class contain the same class width. If studying students at SLA, the categories could be done four different ways:

| Elementary | | Secondary | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Boehner's | | Walker's | | Luttrell's | | Wentzell's | | Scott's | | | |
| 1-2nd grades | | 3-4th grades | | 5-6th grades | | 7-8th grades | | 9-10th grades | | 11-12 grades | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |

To clarify more about our frequency charts, terms need to be understood. A class is the same as a category. Class width is what types of data fall into that category (class). Class boundary is the number that separates each class. Class limit is the largest/smallest number that falls within that category. Class mark, used in graphing data, is the midpoint of the class. It is expected that class widths would be constant for frequency and relative frequency charts.

Example 1:

| Test score | 0-19 | 20-39 | 40-59 | 60-79 | 80-99 | 100-119 |
|---|---|---|---|---|---|---|
| Frequency | 2 | 7 | 8 | 11 | 11 | 9 |

The class width in the example above is 20. The class boundaries for the category 40-59 are 39.5 and 59.5. The class limits are 40 and 59. The class mark is 50.

Sometimes when reading statistical papers, you'll run across a cumulative frequency table. This table contains cumulative frequencies. Cumulative frequency tables are very similar to percentiles.

Example 2:
Using the data from example one, first calculate relative frequencies: 4.2%, 14.6%, 16.7%, 22.9%, 22.9%, and 18.8%.

| Test score | score < 20 | Score < 40 | Score < 60 | Score < 80 | Score < 100 | Score < 120 |
|---|---|---|---|---|---|---|
| Frequency | 4.2% | 18.8% | 35.5% | 58.4% | 80.3% | 100% |

Example 2 found the percentage of scores below a certain number. That is what cumulative frequency charts do. How much fell below 40? Sum the frequencies of those scores that fell between 0 and 19 to those that fell between 20 and 39. Not all charts are suitable for what you might be analyzing. I wanted to know how many students passed my exams at certain levels. So I made something like a cumulative frequency exam but it was in reverse, it cumulated scores above a specified number. You will find that some statisticians rely on several charts about the same data set because each chart or graph has a unique way of highlighting relationships. So you as the student need to know all the variations.

Example 3:

| Test score | Score> 0 | Score > 20 | Score>40 | Score> 60 | Score> 80 | Score > 100 |
|---|---|---|---|---|---|---|
| Frequency | 100% | 95.8% | 81.2 | 64.5% | 41.6% | 19.7% |

Example 3 shows that 41.6% of my class got A or B on their final and 64.5% passed their exam.

A lot of magazines display statistics with pictographs, where a chart uses pictures of an object to show its relationship to another object either by size or quantity.

A common chart is the pie chart. The slices of the pie vary in size depending on the relative frequency of the data. Pie charts are great for qualitative data. To calculate the angle of the slice of pie, multiply the relative frequency to the number of degrees in a circle.

A word of caution: 1) not all tables involving percents sum to one (100%) which means there are some overlaps in relationships which can mislead your conclusions. You can always sum frequencies to find the percentages yourself. But if you only have percents, you can't find frequencies without knowing sample sizes. Sometimes it is necessary to know if your results are coming from a large enough sample.

## STATISTICS LESSON 2 HOMEWORK

1. You survey 20 shoppers to see what type of soft drink they like best, Pepsi or Coca-Cola. The results are: Pepsi, Pepsi, Cola, Cola, Pepsi, Cola, Cola, Cola, Pepsi, Pepsi, Cola, Cola, Pepsi, Cola, Pepsi, Pepsi, Cola, Cola, Cola, and Pepsi. Which brand is preferred? Make a frequency chart.[1]
2. A zoo asks 1000 people whether they have been to the zoo in the last year. Those responding yes were 592, those responding no were 198 and those who didn't respond were 210. Make a frequency table and explain why you need to include those who don't respond.[2]
3. Make a relative frequency table for question 2. Use the results to find the response rate of the survey.
4. The Survey of Study Habits and Attitudes is psychological test that evaluates college students' motivation, study habits and attitudes toward school. Create a stem-and-leaf plot of the following scores:[3]

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 154 | 109 | 137 | 115 | 152 | 140 | 154 | 178 | 101 |
| 103 | 126 | 126 | 137 | 165 | 165 | 129 | 200 | 148 |

5. The Modern Language Association provides listening tests that measure the understanding of spoken French. The range of scores is 0 to 36. Create a stemplot of these scores. Use split stems for bonus![4]

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 32 | 31 | 29 | 10 | 30 | 33 | 22 | 25 | 32 | 20 |
| 30 | 20 | 24 | 24 | 31 | 30 | 15 | 32 | 23 | 23 |

6. Create a pie chart for the data in the lesson about skittles:

| Red | Blue | Geen | Yellow |
|---|---|---|---|
| 5 | 3 | 10 | 4 |

7. Below is a chart of 2008 car sales. What is wrong with this chart?

| Manufacturer | Toyota | Chrysler | General M | Ford | Nissan | Honda | Hyundai |
|---|---|---|---|---|---|---|---|
| % of market share | 26% | 35% | 45% | 30% | 33% | 28% | 34% |

8. Create a pie chart of those attending SLA:

| Grades P-3 | Grades 4-6 | Grades 7-8 | Grades 9-10 | Grades 11-12 |
|---|---|---|---|---|
| 14 | 14 | 7 | 14 | 9 |

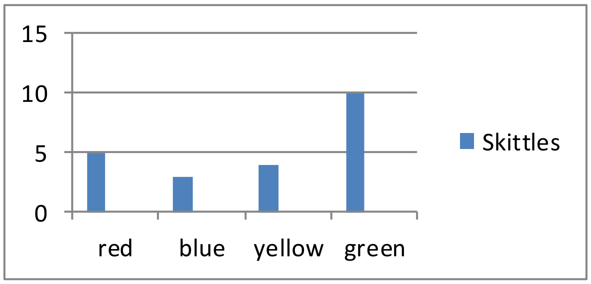
# STATISTICS LESSON 3

## DISPLAYING DATA WITH GRAPHS

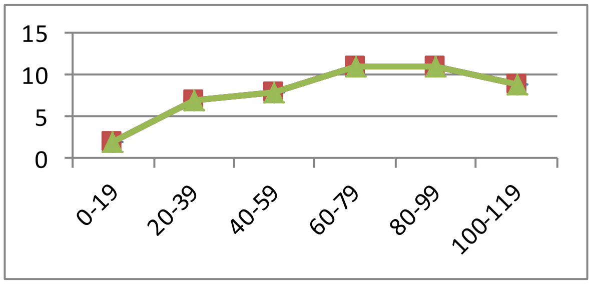
A histogram (aliases: bar graph, bar chart) is a graph composed of rectangles. The width of the rectangles is the category and the length of the rectangles is the frequency of that category. The difference with histograms and frequency polygons is that the frequency polygon plots the ordered pair of class mark versus frequency and then connects the dots.

Either histograms or frequency polygons graph the categories on the horizontal axis and frequency on the vertical axis. Sometimes you'll notice the rectangles formed on histograms have spaces between them and sometimes they don't, regardless of data not falling into a certain category. Depending on the categories, qualitative data tend to have rectangles separated by space – another indication of the disjointed categories. Quantitative data uses a number line as its horizontal axis, not allowing for spaces between the bars unless there are no data that falls within that category.

Example 1:
Skittles data shown by a histogram.

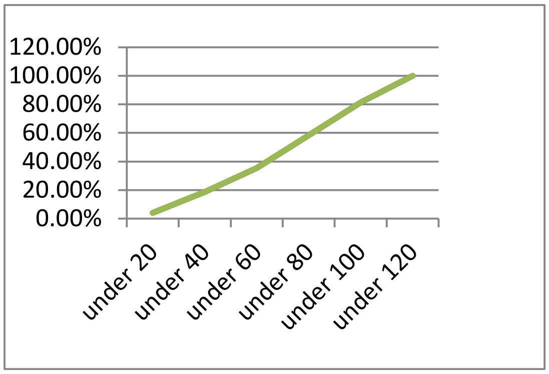


Example 2:
Exam scores in a frequency polygon.

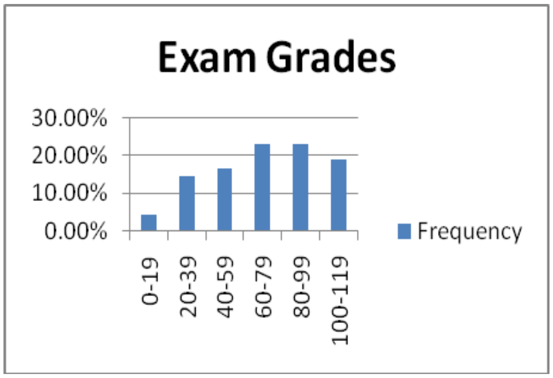


Note the plotting at the class marks (10, 30…).
Relative frequency polygons have the same horizontal scale as the frequency polgon, but the vertical is composed of percentages. Cumulative frequency polygons, also known as ogives, are commonly encountered. To create it, cumulate the percentages and plot them in relation to the category (horizontal axis). The graph should aways be increasing!

Example 3:
Using my exam scores in an ogive.



Example 4:
Exam scores in a histogram.



## STATISTICS LESSON 3 HOMEWORK

1. There are many ways to measure the reading ability of children. The results listed below are that of the Degree of Reading Power test. Make a histogram of the following 44 student scores.[1]

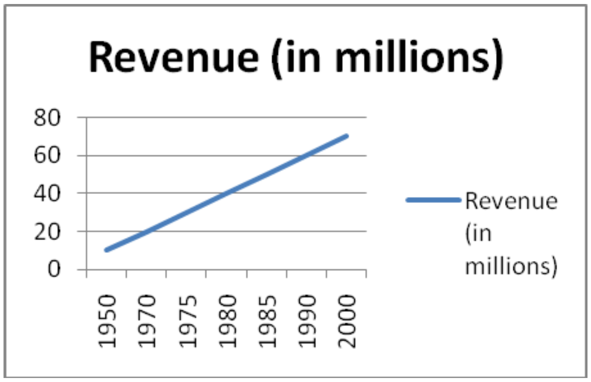
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 40 | 26 | 39 | 14 | 42 | 18 | 25 | 43 | 46 | 27 | 19 |
| 47 | 19 | 26 | 35 | 34 | 15 | 44 | 40 | 38 | 31 | 46 |
| 52 | 25 | 35 | 35 | 33 | 29 | 34 | 41 | 49 | 28 | 52 |
| 47 | 35 | 48 | 22 | 33 | 41 | 51 | 27 | 14 | 54 | 45 |

2. There were about 33,739,900 people living in Canada during 2009. Use the following data to make a frequency polygon showing the percentage falling within each age bracket.

| Age | 0-9 | 10-19 | 20-29 | 30-39 | 40-49 | 50-59 | 60-69 | 70-79 | 80-89 | 90+ |
|---|---|---|---|---|---|---|---|---|---|---|
| Percent | 10.7% | 12.6% | 13.9% | 13.5% | 15.7% | 14.2% | 9.8% | 5.9% | 3.2% | 0.6% |

3. Use the data in problem 2 with Canadian ages to create a pie chart showing the amount of Canadians who are youth (0-20), young adults (20-40), middle-aged adults (40-60), elderly (60+).

4. Explain why the following graph is misleading and how you would fix the problem.



5. Many describe their data based on the shape of a histogram. Here are some terminology you'll encounter and may even use:
   Bell-shaped: a big lump in the middle with tails on each side that taper about the same.
   Right Skewed: lump on left with a trail of data tapering off on the right.
   Left Skewed: lump on the right with a tail on the left.
   Uniform: All the bars have about the same height.
   Bimodal: Two peaks, or two modes.
   Symmetric: Draw a line down the middle of graph and each side is a reflection of the other side.

   Make a histogram from the following data set of test scores. Then describe the shape of the histogram using the terminology above.

140 122 119 99 92 90 90 88 85 82 82 81
80 80 77 74 74 73 72 71 70 70 69 69
69 68 68 68 67 66 64 64 62 60 59 59
58 58 56 56 56 56 55 54 53 53 50 47
35 32

# STATISTICS LESSON 4

## TYPES OF STATISTICAL SAMPLING

Before analyzing data, it is important to consider whether the sample size and method of collecting the sample are appropriate. Although a large sample is no guarantee of avoiding bias, too small a sample is a recipe for disaster.[1] Ten people from the province of Nova Scotia is most likely not a large enough sample to determine anything about what an average Nova Scotian is like! Additionally, some people tend to lie about personal information, so gathering data by asking them rather than measuring results would result in unreliable information. Collecting data has become a science in itself. Statisticians try to avoid anything that may skew their results; even asking the same question with different tones of voice can lead to unintentional results!! That's one reason why on standardized tests all the teachers read the same litany of directions.

There are about five methods of collecting data. The most common, and least biased, would be Random Sampling. In this sample, members of the population are chosen in such a way that all have equal chance to be measured. On a small scale, this would be the same as drawing a name out of a hat. Every member of the population has his or her name in the hat, and it's just a matter of which is drawn. Systematic Sampling measures every $k^{th}$ member of the population. This could be translated as enlisting every 10th person in the phone book. Stratified Sampling divides the population into subgroups and randomly samples each subgroup. To study Nova Scotians, a statistician could randomly select several households in each of the province's counties. This differs from Cluster Sampling, in which the population is divided into subgroups and one or two subgroups are exhaustively measured. The last sampling is the least used because of its tendency to be biased. It is Convenient Sampling, where the one doing the sampling conveniently chooses those being tested and often lets those being tested choose whether their results are used.

The type of questions used in collecting results can have a determining factor on the success of the research. Some questions are open-ended, as found in personal interviews which can elicit more information, and can be used by the questioner to guide the process to gather more information. Some questions are closed, like multiple-choice, which only gather information pertaining to that question. These closed questions can be coded easily and analyzed by computer programs. Each type of questioning has its pros and cons, so it's important for the statistician to reflect on his/her objective and which will be the easiest or most valuable in getting at the information they are seeking.

The main question to be answered when gathering data is, "Will this method ensure I have a good representation of my population?"

## STATISTICS LESSON 4 HOMEWORK

Determine which type of sampling is the following:[2]

1. Keith went through the telephone book and called every 89th person listed.
2. Four people divided the telephone book evenly and each called a random sample of their part.
3. All people with a 902 telephone exchange were called.
4. Every 5th block of 10 students arriving at Citadel High School on Feb 15th is exhaustively sampled about their love of hockey.
5. Read page 187-188 of your textbook and do problems (5.2) 1-5. (see endnote for textbook referred to)

# STATISTICS LESSON 5

## MEASURING THE CENTER

Describing the population from a representative sample is the basis of Descriptive Statistics. One way to describe the population is based on the shape of the sample's graphs. Another way is based on what's at the heart (center) of the data. There are four different numbers that measure the center of the data. Each number serves a different function, making each important. Thus, to get a better idea of the population, a statistician takes the time to identify each: Mean, Mode, Median, and Midrange.

<u>Mean</u>, commonly referred to as arithmetic mean, or average, is found by $\bar{x} = \sum \frac{x}{n}$. As my students like to say, add up all the numbers and divide by how many numbers there are.

<u>Mode</u> is the number occurring the most. Sometimes there is a tie for the most-occurring number. In such cases, the sample is bimodal (having two modes).

<u>Median</u>, is the number in the middle of the list. Make sure that the list is in order first. Sometimes there is no single number in the middle, rather two. Then an average (mean) is taken of the two middle numbers.

<u>Midrange</u> is the arithmetic mean of the lowest and highest data. Don't confuse this with range. Range is the difference between the high and low.

Example 1:
Joey had an ambitious teacher who gave him 14 quizzes in one quarter. [1] His grades were 86, 84, 91, 75, 78, 80, 74, 87, 76, 96, 82, 90, 98, 93. What was his 'average'?

Placing the scores in a list on the calculator and sorting them quickly gave the list as:
74, 75, 76, 78, 80, 82, 84, 86, 87, 90, 91, 93, 96, 98

You'll note that no number repeats, so there is no mode. The midrange is (98+74)/2 or 86. The mean is the sum of all those scores divided by 14. In this case the mean is 85. Having 14 scores, there is no single middle number in the list, so the median is the average of 84 and 86. The median ends up being 85 as well.

## STATISTICS LESSON 5 HOMEWORK

1. Does the mean, median, midrange, or mode have to be a number in the set?[2]
2. Why do you have to order data to find the median but not for mean?
3. Suppose the mean and median salary at a company is $50,000 and all the employees get a $1,000 raise. How would that affect the mean and the median? What about range?[3]

For questions #4-#7, find the mean, mode, median, and midrange.

4. Data of 1, 2, 3, 4, 5, 6, 7, 8, 9
5. Data of 10, 20, 30, 40, 50, 60, 70, 88
6. The age of the U.S. presidents upon initial inauguration: 57, 61, 57, 57, 58, 57, 61, 54, 68, 51, 49, 64, 50, 48, 65, 52, 56, 46, 54, 49, 51, 47, 55, 55, 54, 42, 51, 56, 55, 51, 60, 62, 43, 55, 56, 61, 52, 69, 64, 46, 54, 47.
7. Famous irrational numbers, truncated: 1.414, 1.618, 2.718, 3.141.
8. Do textbook problems p176(5.1): 1 & 2. (see endnote for textbook referred to)

# STATISTICS LESSON 6

## LAW OF AVERAGES: TYPES OF MEANS[1]

There are six different types of means, two of which are commonly used by students:

Arithmetic Mean, or average, is found by $\sum \frac{x}{n}$. Students use this when finding their average test scores.

Weighted Mean is commonly seen at exam time when students are trying to figure out their semester grades. Each quarter has an assigned weight of 40% and the semester exam has a weight of 20%. Weighted Mean takes into account that not all data are created equal. A formula could be written as $\sum wx$.

A third mean, the Quadratic Mean, will be used in the near future for finding standard deviation. This mean, also known as Root Mean Square (RMS), is typically used for data whose arithmetic mean is zero. Due to negative numbers and positive numbers cancelling out when adding, mathematicians circumvented the problem by squaring the data and then at the end of their calculation undoing the squaring by square rooting the value. The formula for RMS is $\sqrt{\sum \frac{x^2}{n}}$.

The Trimmed Mean pops up in statistics whenever there are outliers in the data. An outlier is a piece of data that is significantly different from the rest of the sample and can really make a difference in the mean. Take for instance a sample of common Nova Scotian wages. If everyone is making around $35,000 but you have one or two people making millions, the mean would be raised. Would having a mean between $60,000 and $100,000 be realistic to your sample? So before using a trimmed mean, a good analysis of the data using the shape of histogram, median, and midrange to help support your use is in order.

Geometric Mean is used in business for finding average rates of growth. It is the nth root of the product of the data. Its formula looks like $\sqrt[n]{\prod x}$.

The Harmonic Mean is found by dividing the number of data by the sum of the reciprocals of each data. This is common in calculating average speed. Its formula is $\frac{n}{\sum \frac{1}{x}}$.

The Frequency Mean is the same as obtaining the arithmetic mean from a frequency table. It's actually calculated like a weighted mean.

Example of frequency mean:

| Test scores | 55 | 60 | 70 | 75 | 80 | 90 | 95 |
|---|---|---|---|---|---|---|---|
| Frequency | 5 | 15 | 20 | 25 | 20 | 12 | 5 |

Frequency mean: (5*55 + 15*60+20*70+25*75+20*80+12*90+5*95)/102 = 74.6.

Notice how each score was multiplied by the number of people who score it. It's a quicker way of adding up 102 numbers where there are 5 55's or 15 60's. Remember, multiplication is another form of addition, i.e. 2 * 3 is the same as 3+3 or 2+2+2.

## STATISTICS LESSON 6 HOMEWORK[1]

1. The population of freshmen entering the MSc took a placement exam in which their scores were stratified into groups of 6 from the data set in descending order. Using a die, one from each strata was randomly selected to obtain the following same: 69, 68, 119, 59, 32, 56, 81, 77. Find the sample size, mean, mode, median, and midrange.
2. Calculate the average rate of growth for a portfolio with equal amounts earning the following annual interest rates: 5%, 10%, -5%, 20%, 15%.
3. Four students drive from Halifax to Fox Point at 100 kph and return at 80 kph. Find their average round trip speed, using the harmonic mean.
4. Zack measures the voltage in a standard outlet as 120 volts, -160 volts, 95 volts, and 10 volts at random intervals. Help him calculate the RMS voltage.
5. Using the inauguration ages from a previous homework (Lesson 5, #6), calculate the 10% trimmed mean and the 20% trimmed mean.
6. Calculate the GPA (weighted mean) for the following data: Biology, 5 credits, A- (use 3.667); Chemistry, 4 credits, B+ (use 3.333); College Algebra, 3 credits, A (use 4.000); and Health, 2 credits, C (use 2.000); Debate, 2 credits, B (use 3.000). Express your results to three decimal places.
7. A researcher finds the average teacher's salary for each state from the web. He then sums the salaries together and divides by 50 to obtain their arithmetic mean. Why is this wrong and what should he have done?
8. Given below are two sets of exam grades. Create a histogram for each and calculate the mean and the 10% trimmed mean for both. How does each class compare?

   10th: 103, 94, 89, 80, 76, 65, 64

   11/12th: 104, 88, 79, 75, 73, 65, 64, 60
9. A student got 85% for the first quarter, 92% on the second quarter, and 95% on the exam. If the quarters are weighed equally and the exam is 20% of the semester grade, calculate the semester grade.

# STATISTICS LESSON 7

## MEASURES OF DISPERSION

When describing your sample/population, one statistic/parameter is the dispersion of the data. Dispersion is a fancy term to say how spread out the data is. Variability is often used interchangeably with dispersion, but its definition includes a bit more. The more the spread, the higher the dispersion!

There are a couple of ways to measure how the data is distributed, or how far each element is from some measure of central tendency (average).

Range is the difference between the highest and the lowest data element. Range = $x_{max}$- $x_{min}$.

The most common measure of dispersion is standard deviation, the average distance an element is from the mean. Due to the fact some values are below the mean (being a negative value) and some above the mean (positive value), there is need for a quadratic mean of the distances. Hence the formula looks like: $\sqrt{\frac{\Sigma(\bar{x}-x)^2}{n-1}}$ = $s_x$. This is the formula for a sample standard deviation. Notice the difference in the formula for population standard deviation (μ = population mean): σ = $\sqrt{\frac{\Sigma(\bar{\mu}-x)^2}{N}}$.

It's interesting to note that statisticians use Greek letters when discussing population and Roman letters for samples. You'll also note that n-1 is smaller than N, thus causing the standard deviation for the sample to be bigger (if everything else is equal). That gives statisticians an added buffer in their work.

Another formula for standard deviation is common and slightly easier to derive: $\sqrt{\frac{n\sum x^2-(\sum x)^2}{n(n-1)}}$

The use of calculators such as TI-83 or TI-84 has made the calculation of $s_x$ a piece of cake. Try this example if you haven't found the mean, $\bar{x}$, or $s_x$ on your calculator before. Hit this key sequence on your TI-83/84: Stats, Edit, choose a list, type in 104, 88, 79, 75, 73, 65, 64, 60; Quit, Stats, Calc, 1-vars stats (followed by the name of your list).

It is not uncommon for an experiment to involve millions of events and associated data. It is the goal of many experiments to obtain very precise values, so great care is exercised to reduce systematic errors and also reduce the effect of random errors such as rounding. Since standard deviation is used frequently in other calculations, keep at least twice as many digits allowed for calculations. Otherwise, remember to keep "one more than the least number of significant digits in the data."

The last dispersion characteristic is variance. Variance = $\sigma^2$ or $s_x^2$ (square your standard deviation!). This measurement is the least common. It does play an applicable role later in statistics.

## STATISTICS LESSON 7 HOMEWORK

Find the range, sample standard deviation and sample variance for questions #1-#4.

1. Data of 1, 2, 3, 4, 5, 6, 7, 8, 9
2. Data of 10, 20, 30, 40, 50, 60, 70, 88
3. The age of the U.S. presidents upon initial inauguration: 57, 61, 57, 57, 58, 57, 61, 54, 68, 51, 49, 64, 50, 48, 65, 52, 56, 46, 54, 49, 51, 47, 55, 55, 54, 42, 51, 56, 55, 51, 60, 62, 43, 55, 56, 61, 52, 69, 64, 46, 54, 47.

4. Famous irrational numbers, truncated: 1.414, 1.618, 2.718, 3.141.

5. Make a histogram and determine the mean, standard deviation, variance, and range for the following 2010 Winter Math Exams:

   80, 87, 85, 47, 103, 94, 89, 80, 76, 65, 64, 104, 88, 79, 75, 73, 65, 64, 60.

6. Determine the mean, standard deviation, variance, and range for the following subgroups at SLA:

   a. 7/8th graders: 60, 65, 64, 35

   b. 10th graders: 74, 67, 67, 55, 59, 50, 52

   c. 11/12th graders: 80, 69, 58, 68, 61, 64, 55, 48

7. Calculate the mean and standard deviation from the following chart:

| Test scores | 60 | 65 | 70 | 75 | 80 | 90 | 95 |
|---|---|---|---|---|---|---|---|
| Frequency | 2 | 10 | 20 | 25 | 18 | 12 | 3 |

8. The following are 2010 Winter French Exams. Calculate the mean and standard deviation from the following chart:

| Test scores | 50-59 | 60-69 | 70-79 | 80-89 | 90-99 |
|---|---|---|---|---|---|
| Frequency | 5 | 15 | 20 | 25 | 20 |

9. There are different ways to 'curve' a test. For each method, find the new mean and standard deviation. Enter the data from #5 in to List 1 of your calculator and do the following calculations on your list to create your new data from which to calculate the new mean and $s_x$.

   a. Add 5 points to every score. (L1+5→L2)

   b. Take 75% of the scores and then add 20. (0.75*L1 + 20→L3)

   c. Compare the three lists. Which would you have chosen as a teacher? As a student? Would it influence your choice any if you had the higher or lower score?

10. Which has higher variability: a histogram with one rectangle or a histogram with several rectangles of the same height?

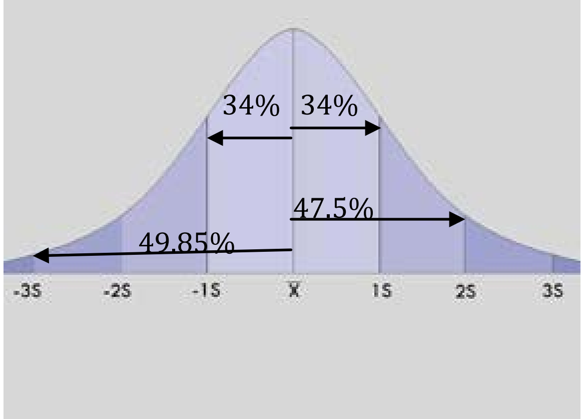
# STATISTICS LESSON 8

## NORMAL DISTRIBUTION

The distribution of independent random observations tends toward a Normal Distribution as the number of observations becomes large.   One of the first to characterize a normal distribution was Carl Friedrich Gauss.  For that reason, the distribution is usually named after him.  Normally distributed data graphed in a histogram or frequency polygon makes a bell shape; hence the other names for a graph of a normal distribution are:  Normal Curve, Bell Curve, or Gaussian curve.

The height of the curve represents the probability of the measurement being at that given distance away from the mean.  The total area under the curve being one represents the fact that we are 100% certain the measurement is somewhere around the mean.

The area under the curve represents the percentage of the population that falls within that range as indicated.  The Empirical Rule describes the Normal Curve as 68% of the population falls within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations of the mean.



So if 68% of the population is expected to be within one standard deviation of the mean, then 34% would be within one standard deviation above (or below) the mean.  The other percentages as stated by the Empirical Rule can be similarly divided.  You can even do a bit of algebra to figure out how much of the population lies between one and two standard deviations from the mean.

A standard normal curve will have the mean at 0 with $s_x$ =1.  Other applications of the normal curve do not have this restriction.  For example IQ tests have a mean of 100 and $s_x$ =15.  Other things which may take on a normal distribution include body temperature, shoe sizes, diameter of trees, etc.  It is also important to note the symmetry of the curve.  Some curves maybe slightly distorted or truncated, but still primarily conform to a heap or mound shape.

This is often an important consideration when analyzing data or samples taken from some unknown population.

A theorem much like the Empirical Rule is Chebyushev's Theorem. Whereas the Empirical Rule is only for a normal distribution, Chebyushev's is an approximation for any distribution. His theorem states that the percentage of data surrounding the mean can be approximated with $1-\frac{1}{k^2}$ where k is the standard deviation (k > 1). Since Chebyushev's Theorem approximates with any type of distribution, its values will be more on the conservative side: 2nd $s_x$ is about 75%, 3rd $s_x$ is about 89%, and 4th $s_x$ is 94%.

## STATISTICS LESSON 8 HOMEWORK

1. Find the mean and standard deviation for the data below:

| Profession | Teacher | Nurse Practitioner | Corporate Attorney | Computer Programmer | Accountant |
|---|---|---|---|---|---|
| Salary | 46,000 | 66,000 | 104,000 | 90,000 | 50,000 |
| Frequency | 130,000 | 200,000 | 50,000 | 50,000 | 75,000 |

2. Apply the symmetry of IQ distribution and the empirical rule to find the proportion of population which has an IQ between 85 and 130.
3. What does Chebyushev's Theorem say about the number of IQ between 85 and 115?
4. The Unibomber has been often cited to have an IQ of 170. Calculate how many standard deviations above the mean this corresponds to. [1]
5. Using the mean of 54.9 and standard deviation of 6.3, list the inauguration ages for any president beyond two standard deviations from the mean.
6. What percent of inauguration ages is within two standard deviations of the mean? Is this data a normal distribution?
7. Suppose you have a normal distribution with a mean of 110 and $s_x$ of 15. About what percentage of the values lie between 95 and 140?
8. Suppose you have a normal distribution with a mean of 110 and $s_x$ of 15. About what percentage of the values lie between 80 and 95?
9. The uniform distribution curve has height of one over the interval 0 to 1 and height of zero elsewhere. This means that data described by this distribution take values that are uniformly spread between 0 and 1. What percent of the observations lie above 0.8? Below 0.6? Between 0.25 and 0.75?[2]
10. The distribution of heights of adult American men is approximately normal with mean of 69 inches and a standard deviation 2.5 inches. Draw a normal curve on which this mean and standard deviation are correctly located (label horizontal axis). What percent of men are taller than 74 inches? What percent of men are shorter than 66.5 inches? Between what heights do the middle 95% of men fall?[3]
11. The army reports that the distribution of head circumference among male soldiers is approximately normal with mean 22.8 inches and standard deviation 1.1 inches. Use the empirical rule to determine what percent of soldiers have head circumference greater than 23.9.[4]
12. The length of human pregnancy is a normal distribution with mean 266 days and a $s_x$ of 16. How short are the shortest 2.5% of all pregnancies? How long are the longest 2.5%?[5]

# STATISTICS LESSON 9

## MEASUREMENTS OF POSITION

### Quartiles & Box plots

Just like data can be divided into two equal lists, those above the median and those below, data can be divided into four groups, quartiles. The three quartiles that split the data list into four equal groups are found the same way as the median. Take the group lower than the median and find the median of this newly formed group of data, call the median of this new group the lower quartile. From the group greater than the median, find the median of this new group. Call this the upper quartile. On the TI-83/84 calculator, you'll find these are called Q1, Q2 (median of the entire data set), and Q3.

Quartiles are a way to tell where a single data lies in relation with the entire set of data. A person with a test score falling in between Q1 and Q2 says that at least 25% of the tests are below his/her test. A person with a test score falling between Q3 and the maximum test scores, says he/she did better than at least 75% of the other scores.

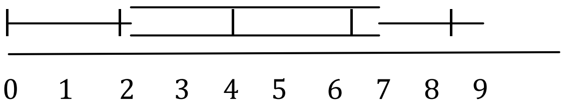
Another common unit of measure are deciles and percentiles. Deciles ($D_1$, $D_2$, $D_3$…$D_9$) are where the group of data is split into 10 equal sized groups. Percentiles ($P_1$, $P_2$,… $P_{99}$) are where the group is divided into 100 smaller groups. A person with an 85 percentile on a CTBS says that 85% of those tested had lower scores than he/she. A person with a $D_6$ has 6 out of 10 people with lower scores. Let's not confuse percentiles with percentages. If you received an 85% on a test that means you got 85% of the questions correct. But an 85th percentile has no bearing on how well you did. It could be that everyone failed the test, but 85% of the people did worse than you.

Statisticians frequently use the Box-and-Whisker Plot (also known as Box Plot) to show the distribution of their data by graphing the quartiles. It graphs the 5-Number Summary: the minimum, the quartiles, and the maximum. With the minimum and maximum, you first determine an appropriate scale for a number line extending from the minimum and maximum. You place a tiny vertical line segment above the number line to indicate your extrema. Then you place a tiny vertical line segment above the number line for each of the quartiles. Then from these quartile markings, two adjacent rectangles are drawn where the sides are the quartile markings. Then connect the rectangles to your extrema to create the whiskers.

Actually the quartiles and box-and-whisker plots can all be done on your TI-83 or TI-84. First create a list under STATS, then exit out of the list, go to STATPLOT (above Y=) and chose a plot and type of graph. Make sure to change the list to the one you had entered your data. Then enter GRAPH. Of course you might have to go to WINDOW to change the scale.

Example 1: Digits are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
The minimum is 0, the maximum 9, and the median is the average of the two middle numbers, 4.5. The lower quartile is 2 and the upper quartile is 7.

The box-and-whisker for our example would look like:



The analysis would be that the data is uniformly spread, given the equal intervals between quartiles.

## STATISTICS LESSON 9 HOMEWORK

1. If Mrs Walker's baby's weight is at the median, what is her percentile?[1]
2. Matt takes a placement exam that has a mean of 40 minutes and a standard deviation of 6 minutes for those finishing the exam. Matt finishes at the 90th percentile. What percentage of students are still working on their exam?[2]
3. Rachel sleeps an average of 7 hours per night with a standard deviation of 15 minutes. What's the chance she will sleep less than 6.5 hours tonight?[3]
4. Find the 5-number summary of the following data. Make a box-and-whisker plot. Find the interquartile range, $Q_3$-$Q_1$.

   16 49 53 58 60 63 63 65 72 80 84 89 92 98

5. Find the 50th percentile, the 32nd percentile and the 8th decile of the following data:

   16 49 53 58 60 63 63 65 72 80 84 89 92 98

6. Is Q2 = (Q1 + Q3)/2 a true statement? If not, give a counterexample to show why it's false.[4]

7. Find the 5-number summary of the following Winter French exams. Make a box-and-whisker plot with 95, 68, 62, 50, 96, 92, 95, 85, 79, 78, 77, 75, 69, 64, 60, 60, 50, 98
8. Who would have fallen at the 86th percentile on the Winter French exams (see #7)? The 5th decile?
9. Given the data set {0, 2, 4, 5, 6, 7, 50}, find its quartiles. Are there any outliers?[5]
10. The definition of outliers has gotten precise over the years, being classified as mild or extreme outliers. If the data in question is 1.5 to 3 times the interquartile range above or below the upper or lower quartile, then it is a mild outlier. The outlier is extreme if it above 3 times the interquartile range plus/minus the upper/lower quartile. But some just use the rule of thumb that states an outlier is more than 2 standard deviations from the mean. In the case of question #9, is 50 a mild outlier or an extreme outlier?[6]

# STATISTICS LESSON 10

## MEASUREMENTS OF POSITION

### Z-Score

Often times a statistician needs to know how a certain data element relates to the overall set of data. One measure of comparing is to determine how many standard deviations away from the mean this data element is. The value calculated is called the z-score. One calculates it by dividing the difference between the mean and the data element by the standard deviation.

$$z = \frac{x_i - \bar{x}}{s_x}$$

A positive z-score indicates the data element is above the mean; a negative z-score indicates the data element is below the mean. Z-scores are always rounded to two decimal places.

Z-score is also called standard score. The calculation makes it possible to compare different data sets. It's similar to two students comparing test results from different teachers. They would convert their scores into percentages and talk about how they did in relation to their classmates. These data sets of tests have a mean greater than zero, and each have a different standard deviation. To compare, statisticians standardize the data, so that the mean is zero and the standard deviation is 1. A more precise way of saying it is to make the standard deviation the unit of measurement. Each score then is denoted by how many standard deviations above or below the mean it is.

Example 1:[1]

The EMT response time for an emergency in one city was found to be a normal distribution with a mean of 12 minutes and a standard deviation of 1.2 minutes. What is the probability that the response time is less than 10.8 minutes?

Solution:
The z-score for 10.8 minutes is -1. Since 68% of population is within one standard deviation of the mean, 34% is within 1 standard deviation below the mean. That makes 50%-34% or 16% of response times make it to the house before 10.8 minutes elapsed.

Example 2:
Using the same data as example 1, would an EMT response time of 16 minutes be unusual?

Solution:
The z-score of 10 minutes is $\frac{16-12}{1.2} = 3.33$. Since most data falls within 2 standard deviations, this would be unusual (extreme outlier) for an EMT response time to take 16 minutes.

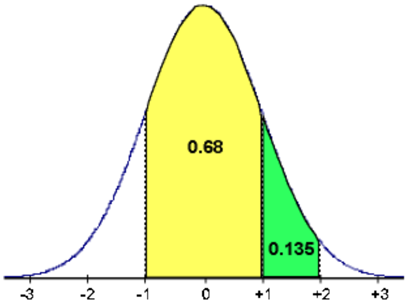
# STATISTICS LESSON 10 HOMEWORK

1. Graduating students from the MSc have a mean ACT score of 29. Calculate the z-score for their mean relative to the national mean of 21.0 and a standard deviation of 4.7. [2]
2. Graduating MSc students have a mean SAT score of 1279. Calculate the z-score for their mean relative to the national mean of 1016 and standard deviation of 157. [3]
3. In the data set of {0, 2, 4, 5, 6, 3, 6, 1, 1, 50}, calculate the z-score of the maximum. Is it an outlier? [4]
4. Three students take equivalent tests of neuroticism with the given results? Which had the highest standard score?[5]
   a. A score of 3.6 on a test for which the mean is 4.2 and the standard deviation is 1.2.
   b. A score of 72 on a test for which the mean is 84 and the standard deviation is 10.
   c. A score of 255 on a test for which the mean is 500 and the standard deviation is 15.
5. Two friends attend different universities and each take a math placement exam. Which of the following results indicates the higher relative level of math competence?
   a. A score of 60 on a test for which the mean is 70 and the standard deviation is 10.
   b. A score of 480 on a test for which the mean is 500 and the standard deviation is 15.
6. For men aged between 18 and 24 years, serum cholesterol levels (in mg/100 mL) have a mean of 178.1 and a standard deviation of 40.7. Find the z-score corresponding to a male, aged 18-24 years, who has a serum cholesterol level of 275.2 mg/100 mL. Is this level unusually high?[6]
7. The heights of six-year-old girls have a mean of 117.8 cm and a standard deviation of 5.52 cm. Find the z-score corresponding to a six-year-old girl who is 106.8 cm tall. Is this height unusual?[7]
8. A company conducts a survey of how long it took before customers hung up on its 3-minute phone message. Create a stem plot of the data below and discuss its shape: 2.4, 0.2, 3.0, 2.8, 1.5, 1.9, 0.7, 2.5, 1.3, 0.8, 2.1, 3.0, 0.4, 1.2, 3.0, 1.1, 0.3, 0.7, 1.8, 0.3, 1.0, 2.1, 3.0, 2.9, 0.5, 1.4, 3.0, 2.8, 1.2, 0.5, 0.5, 1.5, 0.9, 1.8, 0.6, 0.6, 0.7, 0.8, 0.8
9. Determine the quartiles and interquartile range for the data in problem #8. What is the z-score for a hang-up of 1.2 minutes?
10. Suppose the heights of adult women are normally distributed with a mean of 65 and a standard deviation of 2.5 inches. What percentage of women are between 57.5 inches and 67.5 inches tall?[8]

# STATISTICS LESSON 11

## WHAT DOES THE AREA UNDER THE GAUSSIAN CURVE REPRESENT?

The area under a density curve is a proportion of the observations in a distribution. The total area under the Gaussian curve is equal to 100% or 1 since the curve represents the entire data set. Any question about what proportion of observations lie in some range of values can be answered by finding the area under the curve. Up until now, we have only had the Empirical Rule to use. We could calculate the area under the curve using what we know about the z-scores of 1, 2, and 3. The following diagram shows how to use the Empirical Rule to find the percentage of data between z-scores of -1 and 2.



But what if a z-score was not 1, 2, or 3? Some past statistician recognized that if all normal distributions were the same when standardized, then the calculations of the area under curve that is less than a given z-score could be placed into a single table for easy reference. For years, people used tables to find the area to the left of the z-score and then used the properties of the bell curve to obtain by simple math the area between two z-scores. Nowadays, the area between two points on the curve can be even more readily obtained by calculator. On the TI-83/84, use normalcdf( lower z-score, upper z-score) to find the area of the curve that lies within a certain range. This can be found on the TI-83 under [Dist] or [2$^{nd}$][Vars].

Examples:
The level of cholesterol in the blood is important because high cholesterol levels may increase the risk of heart disease. The distribution of blood cholesterol levels for 14-year-old boys found the mean to be 170 mg/dl and a standard deviation of 30 mg/dl. What percent of the boys had cholesterol between 170 and 240? What percent had cholesterol levels above 240?

Solutions:
The z-scores for 170 and 240 are 0 and 2.33. To find how many boys have cholesterol between 170 and 240, you need to find the area of the curve between z-scores 0 and 2.33. Type into a TI-83/84 the following: normalcdf(0, 2.33) and hit enter. The answer to the question of how many boys have cholesterol between 170 and 240 is 49%. To answer the question about what percent of cholesterol levels are above 240, you need to find another z-score to represent the highest possible level within the data. Since most data falls within a range of 4 standard deviations (2 above, and 2 below the mean), choosing a large z-score of 9

should be safe enough to use in your calculations. So enter normalcdf(2.33, 9) to find the percentage of boys with cholesterol above 240 is 1%.

(Upon reading some manuals, you might not even have to standardize the data first. You may just have to enter the range of data in question, followed by the mean and standard deviation. Check it out with the first example question: normalcdf(170,240, 170, 30). It does give the answer of 0.4901.)

## STATISTICS LESSON 11 HOMEWORK

1. In each case, sketch a standard normal curve and shade the area under the curve that is the answer to the question: [1]
    a. $z< 2.85$ b. $z > 2.85$ c. $z > -1.66$ d. $-1.66< z < 2.85$
2. In each case, sketch a standard normal curve with your value of z marked on the axis.[2]
    a. The point z with 25% of the observations falling below it.
    b. The point z with 40% of the observations falling above it.
3. The distribution of heights of adult American men is approximately normal with mean 69 inches and standard deviation 2.5 inches.[3]
    a. What percent of men are least 6 feet tall (72 inches)?
    b. What percent of men are between 5 feet (60 inches) and 6 feet tall?
    c. How tall must a man be to be in the tallest 10% of all adult men?
4. Scores on the Wechsler Adult Intelligence Scale (a standard IQ test) for the 20 to 34 age group are approximately normally distributed with a mean of 110 and standard deviation of 25.[4]
    a. What percent of people aged 20 to 34 have IQ scores above 100?
    b. What percent have scores above 150?
    c. How high an IQ score is needed to be in the highest 25%?
5. Repeated careful measurements of the same physical quantity often have a distribution that is close to normal. Here are Henry Cavendish's 28 measurements of the density of the earth, made in 1798.[5] (The data give the density of the earth as a multiple of the density of water.)

    5.50, 5.61, 4.88, 5.07, 5.26, 5.55, 5.36, 5.29, 5.58, 5.65, 5.57, 5.53, 5.62, 5.29,
    5.44, 5.34, 5.79, 5.10, 5.27, 5.39, 5.42, 5.47, 5.34, 5.46, 5.30, 5.75, 5.68, 5.85

    a. Make a stemplot showing that the data is reasonably symmetric.
    b. Find the mean and standard deviation, then count the number of observations that fall within in the intervals $(\bar{x} - s_x, \bar{x} + s_x)$, $(\bar{x} - 2s_x, \bar{x} + 2s_x)$, and $(\bar{x} - 3s_x, \bar{x} + 3s_x)$.
    c. Compare the percents of the 29 observations in each of these intervals with the Empirical Rule.
    d. What percent of the measurements of the earth's density is greater than 5.85?
6. The quartiles of any density curve are the points with area 0.25 and 0.75 to their left under the curve. What are the quartiles of a standard normal distribution? Drawing a bell curve labelling the z-scores -3, -2, -1, 0, 1, 2, 3 and using the empirical rule will give you a good start on estimating these values.

    More practice can be found in "Statistics Workbook for Dummies", chapter 6 questions, p83.

## REVIEW FOR QUEST 1:

Answer the following questions with the data:
31, 32, 32, 34, 35,43, 24, 13, 19, 23, 23, 45, 13, 13, 54, 45, 12, 75, 23, 46, 54, 87, 12, 45, 78

1. Make a stem-and-leaf diagram.
2. Find the mean.
3. Find the mode.
4. Find the midrange.
5. Find the range.
6. Find the interquartile range.
7. Find the standard deviation.
8. Find the z-score for the value 87.
9. Draw a Boxplot.
10. Is 87 an outlier?
11. What is the 80th percentile?
12. The score of 54 is what percentile?
13. Make a relative histogram.
14. How much of the data falls between the data of 54 and 87?
15. Explain how the empirical rule applies to this data.
16. Explain how Chebyshev's Theorem with k=2 applies to this data.
17. Keith wants to determine the average jumping height of a 17-year-old male Nova Scotian. Match the following sampling methods: Systematic, Stratified, Convenience, Cluster, and Random.
    a. He measures only the 11/12th graders at SLA.
    b. He measures all 11/12th graders at three local high schools.
    c. He measures a random sample of 11/12th graders at 10 provincial high schools.
    d. He gets a list of 11/12th graders in Nova Scotia and measures every 60th student on the list.
    e. He assigns each 11/12th grader a number and lets his calculator pick out 100 random numbers.
    f. He measures those whose numbers have been selected.
18. Given 95, 85, 75, and 110. Calculate the mean should the values represent:
    a. speed in kilometres per hour.
    b. portfolio rates of growth.
    c. high school test scores.
    d. quarter grades with two exam grades of 80 and 88.

Name: ____________________________________ Math Class: ________________
Date: ___________________ Quest: **Statistics Lessons 1-11.**

SHOW WORK. A calculator and a 3” by 5” note card is allowed on this test. You must work alone. Do your work on the space provided.

Use the following data collected from a Geography 10 test given in February to answer the questions on this page:

38, 35, 8, 25, 28, 25, 6, 43, 0, 20, 45, 23, 16, 27

$\frac{}{10}$

1. Create a stem-leaf plot and a histogram for the data. Describe its shape.

$\frac{}{5}$

2. Fill in the answer based on the Geography 10 scores.

____ a. mode ____ f. mean
____ b. midrange ____ g. sample standard deviation
____ c. median ____ h. sample size
____ d. lower quartile ____ I. upper quartile
____ e. range ____ j. interquartile range

$\frac{}{5}$

3. Calculate the z-score for the top score on the Geography 10 tests. Is this person’s score *unusual*?

$\frac{}{5}$

4. According to the empirical rule, how much of my normal population should lie within 2 standard deviations? Are my Geography 10 scores normal? How much of my data actually lies within 2 standard deviations?

—
5

5. Scores on the Wechsler Adult Intelligence Scale for the 20 to 34 age group are approximately normally distributed with a mean of 110 and a standard deviation of 25. Use the empirical rule to answer the following:

a. About what percent of people have scores above 110?
b. About what percent have scores above 160?
c. If someone's score were reported as the 16th percentile, what would the score be?

—
5

6. Brianna gets homework scores of 100%, test scores of 85%, 80% on projects and 100% on citizenships.

| Grade Weight | Citizenship | Projects | Homework | Tests |
|---|---|---|---|---|
| Geography | 10% | 30% | 30% | 30% |
| Mathematics | 10% | 10% | 30% | 50% |

What are her grades for Geography and Mathematics?

—
5

7. Your parents speed at 120 kph to keep up with a caravan of students heading to Clements Park, but on the way home they go 100 kph. What is their average speed?

—
5

8. Match the following sampling types with the most appropriate example:

____ random A. You choose 7/8th and 9/10 homerooms and select few
____ cluster B. You choose the first five students you meet after class
____ convenience C. Every 5th student passing the water fountain is sampled
____ systematic D. Everyone in the P-3 and 11/12 grades are surveyed
____ stratified E. Use of a random number generator on a TI-83 to pick which students are surveyed.

—
5

9. Circle the appropriate word and cross out the inappropriate:
Nissan analyzes its sales. It wants to determine which models are most popular. Categorizing models is a type of (qualitative, quantitative) data. The level of measurement is (nominal, ratio, ordinal). Nissan decides against a chart that lists models under labels of unpopular, so-so popular, most popular because its level of measurement would be (nominal, ratio, ordinal) which isn't as high a level as (ratio, nominal). They choose a histogram that displays number of sales.

# STATISTICS LESSON 12

## BASIC PROBABILITY

An experiment is any process that generates one or more observable outcomes. The set of all possible outcomes is called the **sample space**. Tossing a coin has a sample space of {H, T} and a rolled dice has {1, 2, 3, 4, 5, 6}. An **event** is any outcome or set of outcomes in the sample space. The probability of an event is a number from 0 to 1 (0% to 100%), inclusive, that indicates how likely the event is to occur. A probability of zero is an event that cannot occur and a 100% is an event that must always occur. The probability of an event occurring is found by creating a fraction where the denominator is the size of the sample space (sum of all possible outcomes) and the numerator is how many ways that event occurs. The probability of rolling a six on one dice is 1:6 (1/6 or 16.7%); the 1 is from only one way of rolling a one and 6 from the size of the sample space. The probability of rolling a sum of 6 on a pair of die is 5:36 or 13.9%. The event could have occurred five different ways as seen in the set {1+5, 2+4, 3+3, 4+2, 5+1} taken from the sample space {1+1, 1+2, 1+3…} whose size is 36.

Two events are **mutually exclusive** if they have no outcomes in common. Two mutually exclusive events cannot both occur in the same trial of an experiment. If two events are mutually exclusive, then the probability of either event occurring is the sum of their individual probabilities. This is called the Addition Rule of Probabilities: P(A or B) = P(A) + P(B) where P(A) is the probability of event A happening. Note that in mathematics, the 'or' allows for one or other or both to occur. If your mother said you could have ice cream or cake for dessert, the mathematical interpretation would allow for both whereas the English grammar allows for only one.

The **complement** of an event is the set of all outcomes that are not contained in the event. The complement of event A occurring is the same as the event that A doesn't occur. If P(A) = p, then P(not A) = 1-p. The probability of a dice rolling one is 1:6 or 16.7%. The probability of not rolling a one is 1-0.167 or 83.3%. The "probability of at least one" is the same as the complement of the probability of zero.

Two events are **independent** if the occurrence or non-occurrence of one event has no effect on the probability of the other event. Some mistakenly confuse independence and mutually exclusive. The chart below helps to distinguish between the two:

| **Mutually Exclusive** | **Independent** |
|---|---|
| Refers to two possible results for a single trial | Refers to two or more trials |
| Uses the word "or" | Uses "and" |
| P(A or B) = P(A)+P(B) | P(A and B) = P(A)×P(B) |

Multiplication Rule of Probabilities states the probability of two events occurring is the product of their individual probabilities: P(A and B) = P(A)×P(B). An example of the multiplication rule is finding the probability of rolling a dice twice and getting a two followed by a one. The solution can be obtained by tediously making a **tree diagram** or making a **list** of all thirty-six possibilities. Only one outcome (1:36 is 2.78%) is possible. The multiplication rule simplifies the work to P(2 and 1) = 0.167×0.167 = 2.78%. Take note to either round to three significant digits or give the exact fraction.

If an event were not mutually exclusive, the Multiplication Rule would be nullified. The Addition Rule would need altering also for those not mutually exclusive, to eliminate any double counting. The Addition Rule for Non-Mutually Exclusive Events is P(A or B) = P(A) + P(B) – P(A and B).

## STATISTICS LESSON 12 HOMEWORK

1. Which of the following cannot be probabilities?[1] 4/3, 0, 0.9999, 1.000, 1.001, -0.2, 2, $\sqrt{2}$, $\sqrt{\frac{3}{4}}$
2. What is P(A) if event A is certain to occur?[2]
3. What is P(A) if event A is impossible?[3]
4. A sample space consists of 200 separate events that are equally likely. What is the probability of each?[4]
5. In a survey of 3630 college students, 1162 stated they have cheated on an exam. If one of these college students were selected, find the probability they were honest.[5]
6. Among 80 randomly selected blood donors, 36 were classified as group O. What is the probability that a person randomly selected will have group O blood?[6]
7. Among 400 randomly selected drivers in the 20-24-age bracket, 136 were in a car accident the previous year. If a driver in that age bracket is selected, what is the probability he/she will be in a car accident during the next year?[7]
8. A couple plans to have two children. List the different possible outcomes. Find the probability of having 2 girls. Find the probability of having one of each sex.[8]
9. On a quiz consisting of 3 true/false questions, a student guesses at each one. List the different possible outcomes. What is the probability of answering all three correctly?[9]
10. Both parents have brown/blue pair of eye-color genes, and each parent contributes one gene to a child. Assume that if the child has at least one brown gene, that color will dominate and they eyes will be brown. List different outcomes. What is the probability the child will have a brown/blue pair of genes? What is the probability the child will have brown eyes?[10]
11. Determine whether the two events are mutually exclusive for a single trial:[11]
    a. Selecting a student who attends statistics class and a student who has a computer
    b. Selecting a person with blond hair and a student with brown eyes.
    c. Selecting an unmarried TV viewer and a TV viewer who has an employed spouse
12. If P(A) = 0.45, then P($\bar{A}$) is ____.
13. If the probability of a baby being a boy is 0.513, then the probability a baby is a girl is ______.[12]
14. If P(A or B) = 1/3, P(B) = ¼, and P(A and B) = 1/5, find P(A)[13]
15. Among 200 seats available on a British Airways flight, 40 are reserved for smokers (including 16 aisle seats) and 160 are for non-smokers (including 64 aisle seats). If a late passenger is randomly assigned a seat, find the probability of getting an aisle seat or one in the smoking section.[14]

16. The Internal Revenue Service for the U.S. reports that 70% end up owing more money. One new auditor selects 8 tax returns, audits them, and boasts he collected additional taxes from all of them. What is the probability of doing what he boasts?[15]
17. A circuit requiring a 500-ohm resistance is designed with five 100-ohm resistors. There is a 0.992 probability that any individual resistor will work successfully. What is the probability that all five resistors will work successfully? What is the probability that none will work? What is the probability that at least four will work?[16]
18. A manager uses test equipment to detect defective computer disk drives. A sample of 4 different disk drives is to be randomly selected from a group consisting of 10 that are defective and 20 that are not. What is the probability that all 4 selected drives are defective?[17]

## MORE ON BASIC PROBABILITY

## LESSON 12 HOMEWORK CONTINUED[1]

1. To test the effectiveness of a new vaccine, researchers gave 100 volunteers the conventional treatment and 100 other volunteers the new vaccine. The results are shown in the table:

| Treatment | Disease Prevented | Disease Not Prevented |
|---|---|---|
| New Vaccine | 68 | 32 |
| Conventional Treatment | 62 | 38 |

   a. What is the probability that the disease is prevented in a volunteer chosen at random?
   b. What is the probability that the disease is prevented in a volunteer who was given the new vaccine?
   c. What is the probability that the disease is prevented in a volunteer who was not given the new vaccine?
2. A city council consists of six Democrats, two of whom are women, and six Republicans, four of whom are men. A member is chosen at random. If the member chosen is a man, what is the probability that he is a Democrat?
3. Mindy's chances of passing a precalculus exam are 80% if she studies and only 20% if she decides to take it easy. She knows that 2/3 of her class studied for and passed the exam. What is the probability that Mindy studied for it?
4. A circuit is used to control the temperature in a room. It performs correctly if at least 1 of 4 components does not fail. The probability of failure is 0.18. Find the probability that the critical function will be properly performed.
5. The Dover children, Eileen and Ben, are away at college. They visit home on random weekends, Eileen with a probability of 0.2 and Ben with a probability of 0.25. On any given weekend, what is the probability of both visiting? Neither visiting? Eileen visiting but not Ben?[2]
6. Terry Torrey has the following probabilities of passing the courses: Humanities, 90%; Speech, 80%, and French, 95%. What is his probability of passing all three? Failing all three? Passing at least one? Passing exactly one?[3]

# STATISTICS LESSON 13

## CONDITIONAL PROBABILITY[1]

Conditional probability is a term for dependent events. Conditional probability of event B occurring when event A has occurred is read as "the probability of B given A" and written as P(B|A). It is found by dividing the probability of events A and B both occurring by the probability of event A: P(B|A) = P(A and B)/P(A) where P(A)≠0.

Example 1: Talia tosses two coins. What is the probability that she tosses two heads, given she has tossed one head already?

Solution: P(A and B) = 0.25 which is the probability of two heads
P(A) = 0.75 which the probability of at least 1 head
P(B|A) = 0.25/0.75 = 1/3 or 33.3%

Example 2: Daniel Jones works in a laboratory where a drug promoting hair growth in balding men is being tested. He found that the number using the drugs with hair growth is 1600 and those with no hair growth is 800. His control group, those using a placebo, has 1200 men with hair growth and 400 with no hair growth. What is the probability that a test subject's hair grew, given he used the experimental drug?

Solution: P(H|D) = P(used drug and hair growth)/P(used drug)
P(drug use and hair growth)= 1600/4000 = 0.4
P(used drug) = 2400/4000 = 0.6
P(H|D) = 0.4/0.6 = 2/3 or 66.7%

It has occurred to statisticians and mathematicians that if P(A) and P(B) are independent, then the probability of event B occurring should be the same as if A occurred already, as in it doesn't matter if event A occurred or not. So they use the following as test for dependence: If P(B|A) ≠ P(B), then A and B are dependent.

## STATISTICS LESSON 13 HOMEWORK

1. Dakota, Keith, and Jenny compete in a series of daily 3-way races. For each race, the probability that Dakota wins is 1/2, Keith 1/5 , and Jenny 1/4. On a day that Dakota doesn't win, what is the probability Keith beats Jenny?

2. A manager of a store surveys 500 people exiting the store and found 250 bought something, 120 asked questions and bought something, and 30 people asked questions without buying. Based on the survey, determine whether a person who asked questions is more likely to buy than the average person.

3. In Ms Luttrell's class, 60% of the students have brown hair, 30% have brown eyes, and 10% have both. A student is excused early to go to a doctor's appointment.

   a. If the student has brown hair, what is probability that the student also has brown eyes?

b. If the student has brown eyes, what is the probability that the student does not have brown hair?

c. If the student does not have brown hair, what is the probability that the student does not have brown eyes?

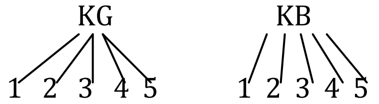
4. Two coins are tossed. What is the probability that one coin is heads if it is known that at least one coin is tails?

5. A Park Medical Center, in a sample group, there are 40 patients diagnosed with lung cancer, and 30 patients who are chronic smokers. Of these there are 25 patients who smoke and have lung cancer. Draw a Venn Diagram to represent this data. If the medical center has 200 patients, and one of them is randomly selected for a medical study, what is the probability that the patient has lung cancer, given that the patient smokes?

# STATISTICS LESSON 14

## PERMUTATIONS & COMBINATIONS

How many ways can Rachel wear khaki green or khaki beige pants with her five different blue tee-shirts? A tree diagram might be a nice approach to understand why 10 is the answer.

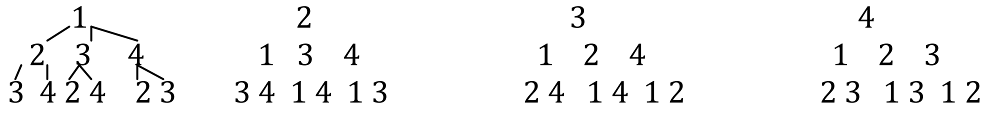


or N(A)*N(B) = 2 * 5 = 10

where N is the number of ways the event could take place.

**Permutation** is another name for an arrangement where the order is important. Examples of this occur in locker combinations, serving a 4-course meal, how one dresses, and how one spells a word. Permutation can be found using $_nP_r$ or $\frac{n!}{(n-r)!}$ where n is the number of elements and r is the number of items to be arranged.

Example: How many permutations are there with 1, 2, 3, and 4? $_4P_4$ or 4!/0! (where 0! =1) = 24. You can see this by filling in the blanks with the number of choices for each blank: $\underline{4}\cdot\underline{3}\cdot\underline{2}\cdot\underline{1}$ and then multiplying those choices! Or make a tree diagram



The last row shows that there are 24 branches to the tree, thus 24 different ways to arrange the numbers.

**Combination** is an arrangement where the order is not important. Thus having 4 numbers ordered 24 different ways is just being redundant, because a combination has only one way of arranging 1, 2, 3, and 4. If a permutation is concerned about how the food is carried to the table, combination is just concerned that the food is on the table! Combination is the permutation of an event divided by the redundant factors: $_nP_r$/ r! Or $\frac{n!}{r!(n-r)!}$ or $_nC_r$.

Example: At a local potluck, there are nine entrees, yet you only have room for six on your plate. How many combinations are possible for you?

Solution: 9C6 = 9!/3!/6! = 84 combinations possible.

**Permutations on a Circle**: If n elements are to be arranged on a circle, the permutations possible is (n-1)!.

**Permutations with Repeated Elements**: If seeking a permutation from n items but some items are duplicates, divide the permutation by the arrangements possible for each duplicate. An example would be to arrange the letters of MISSISSIPI: 10!/4!/4! = 37800 ways.

# STATISTICS LESSON 14 HOMEWORK

1. A snack shop makes subs using 5 meats, 3 types of bread, and 4 different seasoning choices. How many different types of subs can the employee? Make a tree diagram to represent the different variations of subs.

2. A six-question multiple-choice test is administered with each question having 4 possible answers. If all questions were answered, how many arrangements can the test be answered?[1]

3. How many different license plates can be formed by 3 letters followed by 3 digits? What's the probability of getting SDA-777?

4. How many ways can the letters in the set {S,T,U,D,Y} be arranged?
5. How many ways can the letters in the word TEACHER be arranged?

6. How many ways can a teacher of twelve students select five for the first row? What if order mattered? How would it change the results?

7. Sandy Lake Academy faculty and staff go to OPA's (Ela! Taverna) for a Christmas party. If there are 30 appetizers and 16 entrees, how many different special dinners are there?

8. There are 30 appetizers at Ela! Taverna with 8 cold appetizers, 16 hot appetizers and 6 soups and salads.
   a. How many arrangements can three different appetizers be brought to the table at once?
   b. How many arrangements of 5 different appetizers can be brought to the table individually?
   c. Which is a permutation or combination question? Why?
   d. How many ways can 3 diners order different appetizers?
   e. Suppose 7 diners arrive and each orders differently. How many ways can this be done?

9. What is the probability of rolling a sum of four on two dice?
10. Granma wants me to buy three different postal stamps for her collection. The post office has 15 types of stamps and only two of them are of the royal newlyweds. How many ways can I come home with at least one stamp of the royal couple?

11. Charlie Brown has 13 socks in his drawer, 7 blue and 6 green. He selects 5 socks to take on a trip. What is the probability of selecting the one sock that has a hole in it?[2]

12. Art is a good student. He figures that his probability of making an A is 0.92 for Algebra II and 0.88 for Chemistry. What is his probability of making at least one A?[3]

13. The ten digits (0-9) are arranged at random with no repeats. What is the probability that the numeral thus formed represents a number greater than 6 billion?

## STATISTICS LESSON 14 HOMEWORK CONTINUED[1]

1. Determine if the statement C(n,r) = P(n,r) is sometimes, always, or never true. Justify. C(n,r) is another way of saying $_nC_r$.

2. Determine if the following is a permutation or combination, then find the number of possibilities: a. choosing 2 different pizza toppings from a list of 6; b. seven shoppers in line at a checkout counter; c. arrangement of letters in the word *intercept*.

3. The principal of SLA wants to start a mentoring group. He needs to narrow his choice to five from a group of nine. How many ways can a group be selected?

4. Determine whether it's a permutation or a combination, then find the possibilities: the winner, second place and third place winners in a contest with 10 finalists.

5. Determine whether it's a permutation or a combination, then find the possibilities: selecting two of eight employees to attend a business seminar.

6. Determine whether it's a permutation or a combination, then find the possibilities: arranging the letters in the word *algebra*.

7. Determine whether it's a permutation or a combination, then find the possibilities: choosing 2 CDs to buy from 10 that are on sale.

8. Determine whether it's a permutation or a combination, then find the possibilities: selecting 3 of 15 flavours of ice cream at the grocery store.

9. How many different arrangements of the letters in the Hawaiian word *aloha* are possible?

10. Hanafuda is a Japanese game that uses a deck of cards made up of 12 suits, with each suit having four cards. How many 7-card hands can be formed so that 3 are from one suit and 4 are from another?

11. How many ways can 8 runners in an Olympic race finish in first, second, and third place?

12. How many possible ways of arranging 5 basketball players in a huddle (circle!)?

13. How many ways of arranging four different dishes on a revolving tray in the middle of a table?
14. How many ways of arranging six quarters with designs from six different provinces in a circle?
15. A jury list consists of 20 women and 20 men. Find the probability of selecting 12 and getting an all-male jury.

16. There are 6 women and 7 men on the committee for city park enhancement. A subcommittee of five members is being selected at random to study the feasibility of redoing the landscaping in one of the parks. What is the probability that the committee will have at least three women?[2]

# STATISTICS LESSON 15

## ODDS

“Odds against” is a ratio of the probability of A not occurring to the probability of A occurring. A typical method of calculating this would be P(-A)/P(A), where P(-A) is the probability of A not happening and P(A) is the probability of A happening. Notice that if these are expressed as fractions, the denominators of both the fractions in the denominator and the numerator could be the same. Thus the fraction could be simplified to the number of outcomes that are NOT event A over the number of outcomes of event A.

The “odds in favour” or “odds for” event A is the reciprocal of the “odds against”. This is otherwise stated as P(A)/P(-A). To say one has “long odds” would mean that it is unlikely to happen. Examples of such long odds are 10 to 1 and 100:1.

Odds are sometimes written as an unreduced fraction, but are typically reduced to a numerator of 1. They are more commonly expressed in the form a:b or a to b. When one says you have a 50:50 chance of getting heads on a flipped coin, this is a typical statement of odds, but can also be interpreted as saying you have a 50% chance of winning and a 50% chance of losing. Baseball and most such sports commonly use probability ratios. Batting average, or hits per at bat, would be a typical example. Given a batting average of 0.25 and a third of those hits being for extra bases (0.83), one could calculate either the odds against getting a hit (3 to 1); the odds of getting an extra base hit (11 to 1); or the odds of a hit being for extra bases (2 to 1).

Although the use of odds is easy to deal with, odds are awkward to use in calculations. That is why they are converted into probabilities.

Example: What are the odds against rolling a total of 4 using three regular six-sided dice.

Solution: There are three ways of rolling a 4 (1-1-2, 1-2-1, 2-1-1) out of 216 possibilities. The probability of not rolling a 4 is 213: 216, the odds against a 4 is 213:3 or 71:1.

# STATISTICS LESSON 15 HOMEWORK[1]

1. The probability of getting homework is 75% or 3:4. What are the odds of getting homework?
2. Odds against event A occurring is the ______ of P(-A)/P(A).
3. Odds are expressed in the form a:b where a and be are integers usually having no common ____.
4. Find the odds of rolling five dice with a sum of 30.
5. Suppose you read in the newspaper that the odds of Chester Area Middle School having a snow day before spring break are 2 to1, whereas Halifax Regional Schools have odds of 3 to 1. Convert the odds to probabilities of having a snow day.
6. What are the odds of Kelsey getting detention when she has a probability of 60% chance of ticking the teacher off to the point of being served detention?
7. If the class has the odds 3:2 of having a 'make-up' day, what is the probability of them having a make-up day?

# STATISTICS LESSON 16

## EXPECTED VALUE

The expected value or mean of a random variable is the average value of the outcomes. If the experiment is repeated a large number of times, the average approaches the expected value. To calculate the expected value of a random variable from a probability distribution, multiply each value by its probability and add the results. This is in actuality finding the weighted mean. Think about it: if all outcomes were the same chance, then you'd just add up the scores and divide by the number of scores. But some scores have a higher chance of being selected, so those need to weigh more in the formula.

Example: Find the expected value from the sum of two die.

| Sum | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Probability | 1:36 | 1:18 | 1:12 | 1:9 | 5:36 | 1:6 | 5:36 | 1:9 | 1:12 | 1:18 | 1:36 |
| Probability in decimal form | 0.028 | 0.056 | 0.083 | 0.111 | 0.139 | 0.167 | 0.139 | 0.111 | 0.083 | 0.056 | 0.028 |

Expected value is computed 2(0.028)+3(0.056)+4(0.083)+5(0.111)+6(0.139)+7(0.167)+ 8(0.139) + 9(0.111)+10(0.083)+11(0.056)+12(0.028) = 7.007. So theoretically, after an indefinite number of rolls, the sum of two die should, on average, be a 7.

Expected Value plays an important role in an area of application called decision theory. This touches every aspect of life: business, family decisions of pregnancies, and farming.

One reason Adventists do not support gambling is because of the wasted money that could be used for supporting the families. Some gamblers are so addicted to gambling that they waste all their earnings on a game leaving nothing aside for food and rent. Here is an example of why it doesn't pay to gamble.

Example: The probability distribution for a $1 instant-win lottery ticket is given below. Find the expected value and interpret the result.

| Win | $0 | $3 | $5 | $10 | $20 | $40 | $100 | $400 | $2500 |
|---|---|---|---|---|---|---|---|---|---|
| probability | 0.883 | 0.06 | 0.04 | 0.01 | 0.005 | 0.002 | 0.0002 | 0.00005 | 0.000004 |

Solution:
0(0.883)+3(0.06)+5(0.04)+10(0.01)+20(0.005)+40(0.002)+100(0.0002)+400(0.00005)+2500(0.000004) = 0.71. So on average, you expect to win $0.71 having paid $1 to enter the game. In reality, there is a net loss of $0.29 each time you play.

# STATISTICS LESSON 16 HOMEWORK[1]

1. Find the expected value of the random variable with the following probability distribution:

| Outcome | 0 | 1 | 5 | 10 | 100 |
|---|---|---|---|---|---|
| Probability | 0.43 | 0.32 | 0.24 | 0.10 | 0.01 |

2. Find the expected value of the random variable with the following probability distribution:

| Outcome | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Probability | 0.25 | 0.25 | 0.25 | 0.25 |

3. Find the expected value of the random variable with the following probability distribution:

| Outcome | 2 | 3 | 4 |
|---|---|---|---|
| Probability | ¼ | ½ | ¼ |

4. An office employs 5 people. A random variable is assigned to the number of people absent on a given day. The distribution is below. What is the probability that at least one person is absent? Find the expected value.

| Absent | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Probability | 0.59 | 0.33 | 0.07 | 0.01 | 0 | 0 |

5. An experiment consists of planting 4 seeds. A random variable assigns the number of seeds that sprout to each outcome. Complete the table and find the expected value for the experiment.

| Sprouted | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Probability | ? | 0.154 | 0.345 | 0.345 | 0.130 |

6. If you have a 1/10 probability of gaining $200, a 3/10 probability of losing $300, and a 6/10 chance of breaking even, what is your expected value?

7. The Halifax Regional Municipality seeks bids from several construction companies to build a new convention center in the city. One company bids on the job. If the bid is won, there is a 0.7 chance of making $1,500,000 profit and there is a probability of 0.3 that the contractor will break even. What is the expected value? Should the bid be submitted?

# STATISTICS LESSON 17

## SIMULATING EXPERIMENTS[1]

In order to estimate probabilities using the experimental approach, a large number of trials is needed. Because this approach is time-consuming, computer simulations are used to duplicate the conditions and thus be able to predict the results. Models of real-life are ubiquitous in the sciences and thus not uncommon in statistics. Another term for model is simulation, a process that behaves the same as the experiment so that similar results are produced.

Historically, random number tables were commonly used as a source of random numbers. Use of the digits of π is another source of random numbers. Most graphing calculators have random number generators that can be used to simulate simple probability experiments. To simulate an experiment with large numbers of trials, it is easiest to use a program that can keep track of the frequency of each outcome. A couple of programs exist on the TI-83/84 calculators: Int(_ + _ rand) or randInt(_,_,_) with the blanks the parameters of your experiment. These functions are found under Math/PRB. The first two blanks are the minimum and maximum number your outcome should be and the third blank is how many numbers you want displayed.

Simulations are commonly used for forecasting weather, analyzing nuclear power plants, and other applications where conducting experiments are challenging. But here are some simulations that can be done at home.

Example: Risk (a game of world dominance)
Within the game various battles are fought with the outcome determined by the roll of dice. The attacking army typically rolls three red dice to the defenders' two white dice. The two highest red dice are compared with the white dice. Each high pair in turn is compared and red wins only if it is bigger. The results by a computer program analyzed the probabilities as 0.372 red, 0.336 split, 0.293 white. Alternately one could toss dice repeatedly and tally the results found from int(1+6rand) or randINT(1,6,5).

Example: Family Size
If a couple plans to stop having children after having one child of each gender, assuming independence of events, then they can simulate family sizes with a program and determine what the average family size will be. They could do a toss of a coin where H(head) represents girls and T(tail) represents boys, or if they have a TI-83 then they could randINT(1,2,1) repeatedly where 1 is boy and 2 is girl and count family sizes. One example would be 1112 (family of 4), 112 (family of 3), 221 (family of 3), 12 (family of 2), 21 (family of 2). After a while they would have enough data to determine the average family size.

Example: Monty Hall Problem
An old television game show called "Let's Make a Deal", hosted by Monty Hall, had three doors with a prize behind one. The contestant selects one door. The host opens one of the remaining doors revealing that it is empty. He offers a choice of keeping the open door or

switching to the unopened door. The fact that one should switch because the probability of winning then becomes 2/3 is far from obvious. A scenario for simulating would be having digits 1-3 represent door 1, digits 4-6 representing door 2, and digits 7-9 representing door 3. Before each round you pick which door has the prize and which door you picked first. Doing randInt(1,9,2) repeatedly would display the two doors in question and over time you can determine how often you'd win if you kept that door.

## STATISTICS LESSON 17 HOMEWORK

1. A student guesses answers to each of the 5 true/false questions on a quiz. Use the decimal expansion of π and even/odd to estimate the mean number of correct responses for 10 such students. Use even for correct and odd for incorrect responses. π≈3. 14159 26535 89793 23846 26433 83279 50288 41971 69399 37510

2. Use the decimal expansion of π to estimate the average number of rolls of a single die necessary to get a 6. Skip outcomes that are not 1 through 6.

3. Use the decimal digits of π to run simulations of the Monty Hall Problem. Determine the probability of winning without switching doors. Find the complement, which would be the probability of switching doors. Is it better to switch or not based on your simulations?

4. Generate 15 families, using your TI-83/84, stopping once each family has both a boy and a girl. Show your work and calculate the average number of children. To ensure everyone is getting the same values so that the teacher can quickly check work, do 0→rand on the calculator. This resets the calculator to the same spot in its random number generator table.

5. Enter the expression 0→rand followed by either a) int(1+6rand) and hit enter five times or b) do randINT(1,6,5). Use the first three numbers to simulate the three red dice in RISK and the last two numbers as the white. Calculate who won based on the rules in the example. Repeat for a total of 12 battles. Tabulate your results.

6. Enter the expression 0→rand. Then do randINT(1,6,2) for 25 times. Keep track of how many entries summed to be greater than 9. Calculate the probability of getting a sum greater than 9. This would simulate an experiment of rolling 2 dice and getting a sum greater than 9.

# STATISTICS LESSON 18

## BINOMIAL DISTRIBUTIONS[1]

Probability distributions may be either discrete or continuous. The graph of the distribution would show the probabilities for each category (class). The normal distribution is a good example of a continuous distribution – the random variable can take on any value. Today the focus is on a discrete distribution, the Binomial Distribution. The Binomial Distribution meets four criteria: 1) a fixed number of trials, 2) independent trials, 3) only two outcomes possible, 4) and probability remains constant throughout experiment.

Some notation has become very standard for Binomial Distributions. S (success) and F (failure) denote possible categories for all outcomes; whereas, p and q = 1-p denote the probabilities of S and F, respectively. The term success may not necessarily be what you would call a desirable result. For example, you may want to find the probability of finding a defective chip. Here the term success might actually represent the process of selecting a defective chip. The important thing is to remember your terminology: P(S) = p, P(F) = q = 1-p, where n = number of trials, x indicates number of success, p is the probability of success in any one trial, and q is the probability of failure in any one trial.

Example: Find the probability of having five left-handed students in a class of nine given the probability of being left-handed is 10%.

Solution: $P(5) = {}_9C_5(0.1)^5(0.9)^4 = 0.000827$ or 0.08% chance. The ${}_9C_5$ is from the combinations possible for having 5 left-handed students in a classroom of 9, the $(0.1)^5(0.9)^4$ is from the Multiplication Rule of Probability that states if seeking P(A and B), then calculate P(A)P(B).

Take a moment to graph the following probabilities of having the number of left-handed students in a class of 9. You will note that instead of a continuous curve like the normal distribution, the binomial distribution is discrete.

| # left | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| P(left) | 0.3874 | 0.387 | 0.172 | 0.0446 | 0.0074 | 0.0008 | 0.00006 | $3*10^{-6}$ | $8*10^{-8}$ | $1*10^{-9}$ |

Calculating P(x), the probability of getting x successes in n trials, is similar to the binomial expansion of $(x+y)^n$ and hence the same name, Binomial Formula: $P(x) = {}_nC_x\,(p^x)(q^{\,n-x})$ where x = 0, 1, 2, 3… n.

Here ${}_nC_x$ has the usual definition as entries from Pascal's Triangle and was defined in a previous lesson. The mean, variance and standard deviation of a binomial distribution can be found using the formulae: mean = $\mu = np$ and variance = $\sigma^2 = npq$.

There are two functions on the TI-83/84 calculator; by looking under DISTR (2nd VARS), you will find BINOMPDF and BINOMCDF. The CDF gives the cumulative frequency and the PDF, when given the two arguments of n and p, in that order, will output a list of n+1 probabilities for each value of x, with the first one being for x = 0. To see the entire list, scroll to the right or left (the blue and white arrow buttons). In other words, use PDF if the experiment is seeking the probability of x = *a* successes or use CDF if determining the probability of x ≤ *a* or x≥ *a* successes.

Example: 20 coins are flipped and each coin has a 50% chance of being a tail. Find the mean and standard deviation. Solution: n =20, p =0.5, q = 0.5 so mean is 20(0.5) = 10 and variance is 20(0.5)(0.5) = 5. That makes the standard deviation approximately 2.236.

## STATISTICS LESSON 18 HOMEWORK

1. Calculate the binomial formula by hand of getting probabilities of 0, 1, 2, 3, or 4 left-handed students in a class of 25 with a probability of 10% being left-handed. Compare results to calculator binompdf(25, 0.1).
2. Using the data from problem 1#, find the probability of getting more than 5 left-handed students in a class of 25. Compare results with 1-binomcdf(25,0.1).
3. A farmer plants 4 seeds. The probability of success (sprouting) is 65%. Find P(0), P(1), P(2), P(3), and P(4). Find the expected value of the number of seeds that will sprout.
4. Find the probability of getting exactly 4 girls in 10 births, assuming P(male)=P(female).[2]
5. In 1995, 60% of all homes had cable TV. If 10 homes were randomly selected, find the probability that exactly 6 had cable TV.[3]

6. There is a 90% chance that Domino's Pizza will deliver in 30 minutes or less. An executive tests a franchise by ordering 10 pizzas at random times to different locations. If the 90% rate is correct, find the probability that 2 or more pizzas will be delivered in 30 minutes or less.[4]
7. Air America has a policy of booking 14 persons on a flight that only seats 12. Past studies show only a 90% passenger attendance rate. Find the probability that if they book 14 persons, not enough seats will be available.[5]
8. Among the 6665 films rates by Motion Picture Association of America, 2945 were rated R. Twenty rated films were randomly selected. Find the mean, variance, and standard deviation for the number of R-rated films in randomly selected groups of 20.[6]
9. A safety board found that 47% of injuries were caused by failure of the plane's seat. 200 different airline-passenger injuries are to be selected. Find the mean, variance, and standard deviation for the number of injuries caused by seat failure in such groups of 200.[7]

10. In New York, 61.2% of motor vehicles accidents involved injuries. When 80 accident reports for one road are examined, it is found that 72 of them involved injuries. Is the injury rate for this road unusually high?[8]
11. According to the U.S. Department of Justice, 5% of all U.S. households experienced burglary last year. For randomly selected groups of 150 households, find the mean and standard deviation for the number of households that experienced at least one burglary last year.[9]
12. Review: Three widely-separated traffic lights on Dunbrak Street operate independently of each other. The probability that you will be stopped at any one of them is 40%. Calculate the probability that you will be stopped at one, exactly two, and all three lights. Which is more probable, being stopped at more than one light or at one or less lights? Justify your answer.

# STATISTICS LESSON 19

## APPROXIMATING BINOMIAL WITH NORMAL[1]

For n>69, one finds 70! exceeds calculator capabilities. It is instructive to examine the binomial distribution for large n and note how it compares with the normal distribution, especially when p = ½. As n increases, the probability distribution for values of p and q looks approximately normal. The common rule is that you can approximate the binomial with the normal when np and nq both exceed some magic number. That magic number is variously stated as 5, 10, or 15, depending on the conservative nature of the statistician. In this book, we will use 10.

Since the normal distribution is continuous and the binomial distribution is discrete, we often must apply a <u>continuity correction</u>. That is to say, x is no longer represented by a single value, but takes on a range of values from x-0.5 to x+0.5. Note the use of the continuity correction in finding the following probabilities statements: at least 64 (area to the right of 63.5), more than 64 (area to the right of 64.5), at most 64 (area to the left of 64.5), fewer than 64 (area to the left of 63.5), exactly 64 (area between 63.5 and 64.5).

Example: Calculate the probability of getting 10 tails when 20 coins are flipped, but using the normal approximation.

Solution: Note np=nq=10. Using the continuity correction, x-0.5 to x+0.5 and values of the mean and standard deviation found in the previous lesson, we can calculate the area between the two z-scores of 9.5 and 10.5 which are -0.22 and 0.22. Using normalcdf(-0.22, 0.22) or normalcdf(9.5, 10.5, 10, 2.236) on the TI-83, the area given as 0.1741 or 0.1769, depending on the method. This normal approximation of the binomial experiment says that there is a 17.7% chance 10 tails will be flipped in 20 coin tosses.

Example: Based on the U.S. Census, 12% of men have a bachelor's degree. If 150 U.S. men are randomly selected, find the probability that at least 25 of them have a bachelor's degree.

Solution: np=18, nq>18 so approximating with normal is reasonable. Mean = np=18, standard deviation is $\sqrt{150 \cdot 0.12 \cdot 0.88}$=3.98. Using normalcdf(24.5, 1000, 18, 3.98), 5.12% is the probability that at least 25 men have bachelor's, from a group of 150.

## STATISTICS LESSON 19 HOMEWORK

1. Use the normal approximation for the binomial to calculate the probability of getting 12 heads in 40 coin tosses. Compare to the probability of getting 28 heads.
2. A certain true-false exam has 100 questions and P(true)=50%. What is the probability of 70 correct questions? Using the normal approximation, what the chance of getting 70 or more questions correct?

3. Find the binomial probability. Indicate if a normal approximation is suitable and if so, find the normal approximation for the probability.[2]
   a. A multiple test of 40 questions has 4 possible answers for each. Find the probability of getting at most 40% correct if all the answers were random guesses.
   b. A limousine service plans to send vans to meet 220 flights over the next few months to an airport that has 80% flights on-time. Find the probability that the number of on-time arrivals are between 180 and 185 inclusive.
   c. Only 25% of teenagers have their own cars. If a marketing team randomly selects 600 teenagers of driving age, find the probability that at least 210 of them have their own cars.

## REVIEW FOR QUEST 2

1. Events that do not affect one another are called ______________ events.
2. In probability, any outcome other than the desired outcome is called a ________(failure, success).
3. The sum of the probabilities of an event and its complement is always ____.
4. The __________(odds, probability) of an event is the ratio of the number of ways the event succeed to the sum of the number of ways it can succeed or fail.
5. How many different ways can three books be arranged on a shelf?
6. From a group of 3 men and 7 women, how many committees of 2 men and 2 women can be formed?
7. How many different ways can the letters of *level* be arranged?

A bag contains 7 pennies, 4 nickels, and 5 dimes.

8. What's the probability of pulling out 3 pennies?
9. What's the probability of pulling out 1 nickel and 2 dimes?
10. Find the odds of pulling out 3 pennies.
11. Find the odds of pulling out 1 nickel and 2 dimes.
12. What is the probability of randomly selecting 2 yellow markers from a box that contains 4 yellow and 6 pink markers.
13. A box contains slips of paper numbered from 1 to 14. One slip of paper is drawn at a time. Find P(3 or 4).
14. A coin is tossed 4 times. Find P(no heads).
15. A coin is tossed 4 times. Find P(2 heads and 2 tails).
16. During the Gulf War, SCUD missiles hit 20% of their targets. Find the probability that out of 6 missiles fired, that two would hit their target.
17. Ten percent of African-Americans are carriers of the genetic disease sickle-cell anemia. Find the probability that out of 30 sampled, 7 would be carriers.
18. Diana tosses two coins. What's the probability that she has tossed 2 tails, given that she has tossed at least one tail?
19. A box contains forty $5 bills, sixty $10, twenty $20 bills, three $50, and three $100 bills. A person is charged $20 to select one bill. Find the expected value of his/her profit.

20. Sandy Lake Academy students are surveyed about their favourite hot lunches. Using the chart, what's the probability that a student likes pizza given they are in high school?

| | Pizza | Hot dog | Spaghetti |
|---|---|---|---|
| Elementary | 10 | 10 | 8 |
| High School | 14 | 6 | 9 |

21. From the previous question, what are the odds that students like pizza over the other options?
22. In a large population of students, 25% experienced feelings of math anxiety. If you take a random sample of 50 students from this population, what is the probability that exactly 2 students have experienced math anxiety?
23. From the previous question, what is the mean and standard deviation of the number of students in the sample who have experienced feelings of math anxiety?
24. From question #22, what is the probability that between 10 and 15 students would have experienced math anxiety? If normal approximation of this answer can be done, explain why.

Name: ______________________________ Math Class: ____________________
Date: ________________ Quest: **Statistics Lessons 12-19**.

SHOW WORK. A calculator and a 3" by 5" note card is allowed on this test. You must work alone. Do as much of the work on space provided; raise your hand if scratch paper is needed.

$\frac{\quad}{10}$

1. Answer the following questions based on the following scenario. Jar A has Easter erasers: 4 eggs, 3 ducks, and 2 rabbits. Jar B has 10 tiny pencil sharpeners.
   a. What is the probability you pick an duck eraser from Jar A?

   b. What is the probability that you pick an egg eraser and then a duck eraser from Jar A, assuming no replacements?

   c. What is the probability of picking a tiny pencil sharpener from Jar B?

   d. What is the probability of picking a duck eraser from Jar B?

   e. What is the probability of picking a duck eraser and then a rabbit eraser from Jar A, assuming you replaced the erasers after each pick?

$\frac{\quad}{5}$

2. The probability of passing Math 11/12 is 85% and the probability of passing English 11/12 is 90%.
   a. What is the probability of not passing Math 11/12?

   b. What is the probability of not passing either class?

   c. What is the probability of passing at least one of the classes?

$\frac{\quad}{5}$

3. Dartmouth Seventh-day Adventist Church is hosting an evangelistic effort. They decide to document the number of people who come from a variety of sources: 10 from door-to-door invitations, 20 from mailings, 5 from both door-to-door invitation and mailings, and 10 came with a friend.
   a. What's the probability that a person showing up came because of a mailing?

   b. What's the probability that a person who has received a door-to-door invitation had already received a mailing?

$\frac{\quad}{5}$

4. Kevin is cleaning out his locker and decides to stack his 7 textbooks on the provided shelf. His Bible is always on top and is not considered a textbook. How many arrangements can he do? Bonus (3 pts): If he keeps his 4 morning textbooks in a separate pile, how many arrangements can he have between the two piles?

$\frac{\phantom{00}}{5}$

4. There are 3 females in the Math 11/12 class of 9 students. Two of the females are juniors. Answer the following questions:
    a. How many ways can the teacher arrange the students so that each group of three has a female?

    b. The teacher needs to pick three students to help serve pizza, how many ways can she do this? Is this a combination or permutation?

$\frac{\phantom{00}}{5}$

5. Answer the following questions about odds:
    a. A senior has a 90% chance of acceptance into CUC. Rewrite in terms of odds. Reduce ratio.
    b. The odds of rain on Thursday is 8:3 and on Friday 10:4. Which has the higher probability of rain?

$\frac{\phantom{00}}{5}$

7. A woman conceives. Find the expected value of how many babies she'll give birth to.

| # of babies in a single pregnancy | 0 | 1 | 2 (twins) | 3 (triplets) | N-tuplets |
|---|---|---|---|---|---|
| Probability | 15% | 84% | 0.935% | 0.011% | 0.054% |

$\frac{\phantom{00}}{5}$

8. Calculate the average number of rolls on a single die that it takes to get a six. Simulate the experiment on your TI-83/84 with 10 trials. Make sure to use the command "0→ rand" before starting your simulation. Show your calculator output for partial credit.

$\frac{\phantom{00}}{10}$

9. A farmer plants 3 seeds. The probability of sprouting is 70%. Find the P(0), P(1), P(2), and P(3). Bonus (4 pts): find the expected number of seeds to sprout.

$\frac{\phantom{00}}{10}$

10. The farmer plants a thousand such seeds (70% sprouting) in his field.
    a. Why is this a binomial distribution?
    b. What is the mean of this distribution?
    c. What is the standard deviation of this distribution?
    d. What's the probability of having 900 or more sprout?

# STATISTICS LESSON 20

## MARGIN OF ERROR

Scientists usually repeat an experiment many times to prove a point. Similarly, statisticians will take multiple random samples to determine what best describes their population. The best statistic to be used as an estimator for the population is the sample mean for the population mean. After several samples have been taken, the different sample means can be averaged to be used as an even better estimator of what the population mean is. Due to variation in the samples, the standard deviation of the sample means is known as the **standard error of the mean**. If the standard error of the mean is small, then the sample means are closer to the true mean. If the standard error of the mean is large, then the sample means are far from the true mean. In other words, a small standard error means you don't expect the sample results to change much with repeated sampling.

As the sample size of the sample means increases, the standard error will decrease. This is reflected in the probability that the increased sample should contain the true mean. The formula for Standard Error of the Mean is $\sigma_{\bar{x}}= \frac{\sigma}{\sqrt{n}}$ Note that the standard error of the mean is a fraction of the standard deviation of the sample means. The standard error of the mean is a measure of the variability of the sample means and behaves like the standard deviation of the sample means. Thus intervals of the probability of the true mean lying within can be created.

Such intervals require the use of the Margin of Error. Margin of Error is the standard error of the mean multiplied by the appropriate z-score. The formula for margin of error is $E = z\alpha/_2 \cdot \frac{\sigma}{\sqrt{n}}$. The margin of error is the number which you add or subtract to the sample mean ($\mu \approx \mu_{\bar{x}}$) to say that you have such-and-such confidence that your true mean lies within that range. My students got a laugh when I wrote the formula as *ME=SEz* and remembered it for the rest of the year! If *ME* is troubling you, look at its units. In the formula, the Z-score which has a division by standard deviation will cancel with the standard deviation in the standard error. So you are left with a distance from the mean. Or you can think of the Margin of Error as having 1.5 boxes of oranges. Each box has 80 oranges. How many oranges do you have? You'd multiply 1.5 by 80 and get 120 oranges. The same principle exists with *ME*.

The same idea is with the Margin of Error. You are given a z-score which is a fraction of how many standard deviations you have. You have a standard error of the mean which indicates how much the average distance is. You want to know the total distance from the mean you are working with, so you multiply the z-score by the standard error of the mean.

In the formula there is a little symbol that has not been introduced in this book yet. It is alpha, α, which is a common symbol to represent the area under the curve in the tails to the left/right of a z-score. Calculating on the TI-83, the z-scores with 95% of the sample within reach of the mean is ± 1.96. The same can be found for any percentage of the sample. In the case of margin of error, the z-score is positive and is usually kept to 95% confidence =1.96, 99% confidence = 2.58, or 90% confidence = 1.64.

Example: A survey of 1000 dental patients shows that the average cost of a regular six-month cleaning is $150 with a standard deviation of $80. What's the margin of error for this result? Assume 95% confidence.

Solution: $1.96 \times 80 \div \sqrt{1000} = 1.96 \times 2.53 = 4.96$

## STATISTICS LESSON 20 HOMEWORK

1. Find the standard error of the mean: $\sigma = 9.8$ and $n = 81$.
2. Find the standard error of the mean: $\sigma = 2.5$ and $n = 250$.
3. Find the standard error of the mean: $\sigma = 3.8$ and $n = 361$.

4. If the standard deviation of a sample set of data is 1.4 and the standard error of the mean is 0.014, how many values are in the sample set?
5. A potential dog owner checked out the statistics on the Nova Scotia Duck Tolling Retriever and found that with a sample size of 56 dogs, they have a life span of 13 years with a standard deviation of 2 years. What is the standard error of the mean?
6. Given a sample mean of 15.5 and a sample standard deviation of 2.42, with a sample size of 11, calculate the margin of error. Assume a 95% confidence interval.[1]

7. Given a sample size of 15.5 and a sample standard deviation of 2.42, with a sample size of 11, calculate a margin of error for a 99% confidence level.[2]
8. A **P-Value** is a way to express the confidence of our results. For a one-tailed test, it is the area under the curve to the right (or left) of our observed mean. Calculate the z-score using our observed mean of 15.5, the expected mean of 10.0 and the standard error of $\frac{2.42}{\sqrt{11}}$. Sketch this region on a normal curve. Calculate the area to the right of our z-score. Alpha, $\alpha$, is the term used to express the level of significance we will accept. For a 95% confidence, $\alpha=0.05$. If our P-Value is less than alpha, we can reject our hypothesis that the expected mean is 10. Should we reject our hypothesis?[3]
9. You sample 100 fish in Pond A at the fish hatchery and find that they average 5.5 inches with a standard deviation of 1 inch. Your sample of 100 fish from Pond B has the same sample mean, but the standard deviation is 2 inches. How do the margins of error compare?[4]

10. Suppose you conduct a study twice, and the second time you use four times as many people as you did the first time. How does the change affect your margin of error? (Assume the other components remain constant.)[5]
11. Suppose Sue and Bill each make a confidence interval out of the same data set, but Sue wants a confidence level of 80% compared to Bill's 90%. How do their margins of error compare?[6]
12. Suppose you find the margin of error for a sample mean. What unit of measurement is it in?[7]
13. How can you increase your confidence level and keep the margin of error small? Assume you can do anything you want with the components of the confidence interval. [8]

# STATISTICS LESSON 21

## CONFIDENCE INTERVALS

The value $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ is termed the standard error of the mean. It is used extensively to calculate the margin of error, which in turn is used to calculate confidence intervals. Remember, if we sample enough times, we will obtain a very reasonable estimate of both the population mean and population standard deviation. This is true whether or not the population is normally distributed. However, normally distributed populations are very common. Populations which are not normal are often ‘heap-shaped’ or ‘mound-shaped’. Some skewing might be involved or those dreaded outliers may be present. It is a good practice to check these concerns before trying to infer anything about your population from the sample.

Since 95% of a normally distributed population is within 1.96 standard deviations of the mean (99% confidence has a z-score of 2.58 and 90% confidence has a z-score of 1.64), we can often calculate an interval around the statistic of interest. The parameter in question is usually the mean.

A confidence interval is formed as: estimate ± margin of error. The estimate is the statistic being used to estimate the population parameter. Margin of error is $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, the length of the interval which is determined by the desired percent of confidence.

Example: Assume the population is the N.S. population with a mean IQ of 100 and standard deviation of 15. Assume further that we draw a sample of n=5 with the following values: 100, 100, 100, 100, 150. The sample mean is then 110 and the sample standard deviation is $\sqrt{500} \approx 22.4$. The standard error of the mean is then $\frac{\sqrt{500}}{\sqrt{5}}$=10. Then based on this sample, we can be 95% confident that the population mean lies between 110-1.96(10) and 110+1.96(10). In other words we expect 95% of the time our population mean will lie between 90.4 and 129.6.

Example: Given a sample mean of 15.5 and a sample standard deviation of 2.42, with a sample size of 11, calculate a 99% confidence interval.
Solution: The interval in question would be calculated by $15.5 \pm 2.58 \cdot \frac{2.42}{\sqrt{11}}$. That means there’s a 99% confidence that the population mean lies between 13.6 and 17.4.

However, if you are sampling without replacement (which affects probability used) and your sample size is more than, say, 5% of the finite population (N), you need to adjust (reduce) the standard error by multiplying it by the finite population correction factor, $\sqrt{\frac{N-n}{N-1}}$. If we can assume that the population is infinite or that our sample size does not exceed 5% of the population size (or we are sampling with replacement), then there is no need to apply this correction factor.

Note: Confidence intervals do not mean that the population mean falls within the given interval, it means we have confidence that our interval calculated will contain the population mean!

## STATISTICS LESSON 21 HOMEWORK

1. Find the interval about the sample mean that has a 95% level of confidence with $\sigma = 5.8$, $n = 81$, and $\bar{x} = 335$.
2. Find the interval about the sample mean that has a 90% level of confidence with $\sigma = 5.8$, $n = 81$, and $\bar{x} = 335$.
3. Find the interval about the sample mean that has a 99% level of confidence with $\sigma = 5.8$, $n = 81$, and $\bar{x} = 335$.
4. The mean height of a sample of 100 high school seniors is 68 inches with a standard deviation of 4 inches. Determine the interval in which there is a 90% chance that the mean height of the entire population occurs.
5. A botanist is studying the effects of a drought on the size of acorns produced by the oak trees. A random sample of 50 acorns reveals a mean diameter of 16.2 mm and a standard deviation of 1.4 mm. What is the interval about the sample mean that gives a 99% chance that the true mean lies within that interval?
6. There is a probability of 0.99 that the average life of a disposable hand-warming package is between 9.7936 and 10.2064 hours. The standard deviation of the sample is 0.8 hour. What is the size of the sample used to determine these values?
7. The Simply Crackers Company selects a random sample of 50 snack packages of their cheese crackers. The mean number of crackers per package is 42.7 and a $s_x$ of 3.2. If the true population mean should be 43 crackers per package, should the company be concerned about this sample? Explain.
8. The lifetimes of 1600 batteries used in radios are tested. With a 5% level of confidence, the true average life of the batteries is from 746.864 to 753.136 hours. What is the mean life of the battery in the sample and what is the $s_x$ of the sample?
9. Given a sample mean of 15.5 and a sample standard deviation of 2.42 with a sample size of 11, calculate the 90% confidence level.
10. The National Center for Education surveys 4400 college graduates about the time required to earn their bachelor's degrees. The mean is 5.15 years, and the standard deviation is 1.68 years. Construct the 99% confidence interval for the mean time required by all college graduates.[1]
11. In a study of the amounts of time required for room service delivery at a newly opened Radisson Hotel, 20 deliveries had a mean time of 24.2 min and a standard deviation of 8.7 min. Construct the 90% confidence interval for the mean of all deliveries.[2]
12. In a study of physical attractiveness and mental disorders, 231 subjects were rated for attractiveness, and the resulting mean and standard deviation are 3.94 and 0.75, respectively.[3]
    A. Use these sample data to construct the 95% confidence interval for the population mean.
    B. Use the sample $s_x$ as an estimate of the population $s_x$ and determine the sample size necessary to estimate the population mean, assuming you want 95% confidence and a margin of error of 0.05.
13. Find the 95% confidence interval for the mean of 100 IQ scores if a sample of 30 of those scores produces a mean and $s_x$ of 132 and 10, respectively. (Use population correction factor.)[4]

# STATISTICS LESSON 22

## OTHER DISCRETE DISTRIBUTIONS:[1]

## HYPERGEOMETRIC, POISSON, STUDENT T

The **Hypergeometric Distribution** is a cousin to the Binomial Distribution. Whereas the Binomial Distribution had constant probability throughout the experiment, the Hypergeometric Distribution's probability is variable. Often this occurs when the sampling is done without replacement from a small finite population. A classic example might be a lottery where 6 different numbers of 54 are selected. Due to the lack of replacement, we no longer have independence, thus our probabilities are not constant for each trial. However the other conditions for a Binomial Distribution are met.

The **Poisson Distribution** got its start with the Queuing Theory. Management of retail stores tried to reduce the frustration of customers by increasing the speed of the checkout lines. Although most grocery stores seem to have retained the multiple line system, many banks and fast food providers have created a single queue where customers wait for the next available cashier. Queuing theory leads one directly to the Poisson Distribution, named after the famous French mathematician Simeon Denis Poisson who first studied it in 1837. He applied it to such morbid results as the probability of death in the Prussian army resulting from the kick of a horse and suicides among women and children.

The Poisson distribution is the continuous limit of the discrete binomial distribution. It depends on the following four assumptions: 1) It is possible to divide the time interval of interest into many small subintervals; 2) the probability remains constant throughout the random time interval; 3) the probability of two or more occurrences in a subinterval is small enough to be ignored; and 4) occurrences are independent. Unlike the Binomial Distribution, Poisson Distribution has infinite population and is only affected by the mean, not the sample size nor probability as in Binomial.

The equation for the Poisson Distribution is: $P(x) = \frac{\mu^{x} \cdot e^{-\mu}}{x!}$

Example: On average there are three babies born a day with hairy backs. A. Find the probability that in one day two babies are born hairy. B. Find the probability that in one day no babies are born hairy.

Solution: a. $P(2) = 3^2e^{-3} /2 = .224$ b. $P(0) = 3^0e^{-3}/0! = .0498$

Example 2: Suppose a bank knows that on average 60 customers arrive between 10 and 11 a.m. daily. Thus 1 customer arrives per minute. Find the probability that exactly two customers arrive in a given one-minute time interval between 10 and 11.

Solution: Let $\mu$= 1 and x = 2. $P(2)=e^{-1}/2!=0.1839$.

The **Student t Distribution** gets its name from William Gosset who worked in an Irish brewery which did not allow publication of research. So he published under the pseudonym of Student. We know that large samples approach a normal distribution. Gosset showed small samples taken from an essentially normal population have a distribution characterized by the sample size. His work in the brewery required the use of small sample sizes. Anytime statisticians have a small sample, they most likely will do their calculations based on the t-distribution.

The following are properties of Student t Distribution:

1. The distribution varies for different size samples. Thus the need to know the degrees of freedom for the sample: n-1. The degree of freedom is the number of values that vary after restrictions have been imposed on all values.
2. Distribution is heap or mound shape, meaning it is less peaked and has fatter tails. As the sample size increases, n> 30, differences with normal distribution are negligible.
3. Mean is zero.
4. Distribution is symmetrical about the mean.
5. Variance is greater than 1. But variance approaches 1 as sample size increases.
6. Population standard deviation is unknown.
7. Population is essentially normal (unimodal and basically symmetric).

The Student t Distribution is similar to the normal distribution in that one calculates a t-score much like how one calculates a z-score: $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$. Then knowing the degrees of freedom, and the t-score(s), a statistician can estimate a confidence interval or percent of population with a certain range of values.

Example: what values of t would you use to find a 95% confidence interval for the mean of a population if n = 16?

Solution: If 95% is amount within the heap, 5% is in the tails, with 2.5% in each tail. Degree of freedom is 15. Using tdcf(lower t, upper t, degree of freedom), type into TI-83, tdcf(x, 20, 15) where *t* is a good starting point. Playing around with different *t*'s, the one closest to giving 2.5% is 2.13.

## STATISTICS LESSON 22 HOMEWORK[1]

1. Find the value of *t* where the area under the curve to the right of *t* is 0.05 when n=6.

2. Find *t* so that 99% of the area of the curve is to the right of *t* when t = 21.

3. Use your calculator to find t when n = 8 and the area to the right of t is 0.005.

4. A university researcher placed 12 randomly selected radon detectors in a chamber that exposed them to radon. The detector readings were: 91.9, 97.8, 111.4, 122.3, 105.4, 95.0, 103.8, 99.6, 119.3, 104.8, and 101.7. Calculate the t-score for the sample mean assuming the population mean is 105.

5. Using the previous radon detector question, calculate the area under the curve to the left of this t-score.

6. Suppose a bank knows that on average 60 customers arrive in a certain hour. Using a time interval of 1 minute, calculate the probability of exactly 1 customer arriving in a given 1-minute interval within that hour.

7. Suppose a bank knows that on average 60 customers arrive in a certain hour. Using a time interval of 1 minute, calculate the probability of NO customers arriving in a given 1-minute interval within that hour.

8. Suppose a bank knows that on average 60 customers arrive in a certain hour. Using a time interval of 1 minute, calculate the probability of more than 3 customers arriving in a given 1-minute interval within that hour.

9. What are the conditions a sample must meet to be classified as a hypergeometric distribution?

# STATISTICS LESSON 23

## OTHER CONTINUOUS DISTRIBUTIONS:[1]

## LORENTZIAN & VOIGT

Another commonly encountered distribution is the Lorentzian Distribution, also known as Cauchy Distribution. Augustin-Louis Cauchy (1789-1857) of France contributed rigor to mathematics and physics. His analysis clarified principles of calculus by developing the concepts of limits and continuity. Hendrik Lorentz (1853-1928) of Netherlands won the 1902 Nobel Prize for Physics for his theory on electromagnetic radiation. His work also refined reflection and refraction of light. The Lorentzian Transformation, time dilation and contraction superseded Galileo, Newton and Einstein's work on Relativity.

The **Lorentzian Distribution** is often used to describe the resonance behaviour of things like swings, a violin bow, goblets shattering to vocals, microphone feedback, measuring gene expression, edges in MRI brain images, muon spin relaxation theory, or rhythmic wind gusts that destroyed the Takoma Narrows Bridge. Like the Gaussian (normal) curve, the Lorentzian is symmetrical, unimodal, and continuous. The Lorentzian distribution tends to be lower with fatter tails. In fact the wings are so extended that the standard deviation is not defined!

**Voigt Profiles** are a convolution of Lorentzian and Gaussian distributions. Woldemar Voigt (1850-1919) of Germany created the Voigt Profiles for the use in spectroscopy, the study of radiation and matter as a function of wavelength. Voigt profiles allow us to measure the temperature and pressure of the emitting or absorbing layers in stellar atmospheres.

## STATISTICS LESSON 23 HOMEWORK

1. Match the distribution to its short description:

| | |
|---|---|
| ____ A. Normal | a. A combination of Lorentzian and Normal |
| ____ B. Binomial | b. 75% criteria is same as Binomial |
| ____ C. Student t | c. Fat tails, describes resonant behavior |
| ____ D. Poisson | d. Brewery worker used small samples |
| ____ E. Lorentzian | e. Two outcomes per trial, constant probability |
| ____ F. Hypergeometric | f. Bell curve follows empirical rule |
| ____ G. Voigt Profiles | g. Morbid studies lead to queuing and time studies |

# STATISTICS LESSON 24

## CENTRAL LIMIT THEOREM

Measuring the entire population may be impossible for researchers due to limitations on time, money, and other resources. Sampling is an important tool for determining the characteristics of a population. [1] The statistician needs to ensure randomization of the population has been done, because without randomization, generalizing the results from the sample is not possible. Usually we don't know the population's parameters (mean, standard deviation, etc), but we often want reliable estimates of them. There are essentially three things we want to know about any distribution: the center, its spread, and how it is distributed. The Central Limit Theorem helps approximate all three.

The Central Limit Theorem: As sample size increases, the sampling distribution of sample means approaches a normal distribution with the same mean as the population and a standard deviation equal to the standard error of the mean.

Stated another way, if you draw simple random samples of size n from any population whatsoever with mean μ and finite standard deviation σ, when n is large, the sampling distribution of the sample means $\bar{x}$ is close to a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. This is true regardless of the shape of the distribution of the original population.

Corollaries from the Central Limit Theorem:

1. For sample size n larger than 30, the distribution of the sample means can be approximated reasonably well by a normal distribution. The approximation gets better as the sample size n becomes larger.
2. If the original population is itself normally distributed, then the sample means will be normally distributed for any size n.

Example: Assume the population of the human body temperatures has a mean of 98.6°F. Also assume that the population standard deviation is 0.62°F. If a sample size of n=106 is randomly selected, find the probability of getting a mean of 98.2°F or lower.[1]

Solution: Because the sample exceeds a size of 30, we can conclude the distribution of sample means is normally distributed with a mean of 98.6 and $\sigma_{\bar{x}} = \frac{0.62}{\sqrt{106}} = 0.06$. Using normalcdf(0, 98.2, 98.6, 0.06), we can find the area under the curve to the left of 98.2, representing the possibility of the population mean being lower than 98.2, to be 0.0000000000132. The results show that if the mean of our body temperatures is really 98.6°F, then there is an extremely small probability of getting a sample mean of 98.2°F or lower when a106 subjects are randomly selected. University of Maryland researchers did obtain such a sample mean! There are two possible explanations: either the population mean is really 98.6°F and their sample represents a chance event that is extremely rare, or the population mean is actually lower than 98.6°F. Because the probability is so low, it seems more reasonable to conclude that the population mean is lower than 98.6°F. This type of reasoning is used in formal methods of hypothesis testing, which will be discussed later.

In applying the Central Limit Theorem, the use of $\sigma_x = \frac{\sigma}{\sqrt{n}}$ assumes that the population has infinite members. Applications are abundant of "without replacement," so if the sample size is greater than 0.05N, adjust the standard deviation by multiplying by the population correction factor.[2]

Example: In human engineering and product design it is often important to consider the weights of people so that airplanes or elevators aren't overloaded, chairs don't break, and other unpleasant things don't happen. Assume that the population of men has normally distributed weights, with a mean of 173 lbs and a standard deviation of 30 lb. If 36 different men were randomly selected from this population, find the probability that their mean weight is greater than 180 lbs.[3]

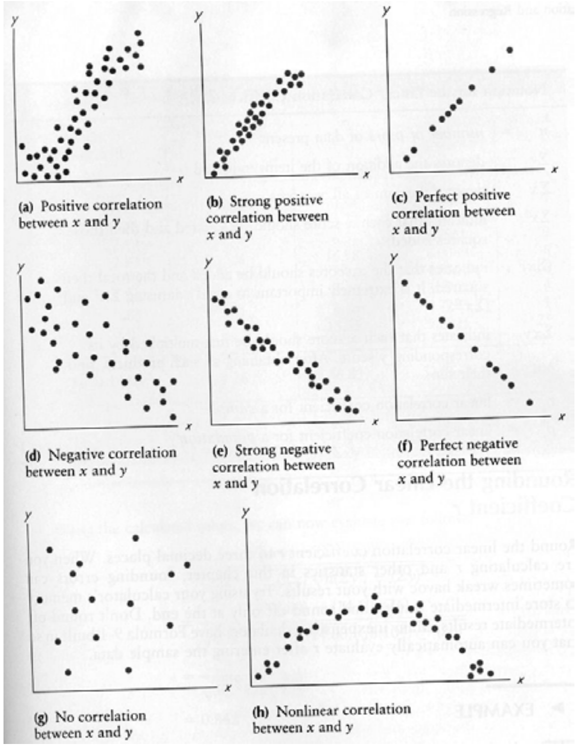
Solution: The standard error of the mean is 5, which estimates the standard deviation. Then typing in the TI-83 normalcdf(180, 4000, 173, 5), the solution is 8.08%.

## STATISTICS LESSON 24 HOMEWORK[4]

1. IQ scores are normally distributed with a mean of 100 and a $s_x$ of 15. If 25 persons were randomly selected, find the probability that the mean of their IQ scores is between 100 and 103.
2. For women aged 18-24, systolic blood pressures are normally distributed with a mean of 114.8 and a $s_x$ of 13.1. If 30 women in that age bracket are randomly selected, find the probability that their mean is above 120.
3. Assume the weights of paper discarded each week by households are normally distributed with a mean of 9.43 lb and a $s_x$ of 4.37 lb. Find the probability that 12 randomly selected households have a mean between 10.0 and 12.0 lb.
4. A study found that high school students spend 10.7 hours working each week on average with a $s_x$ of 11.2 hours. If 42 high school students are randomly selected, find the probability that their mean weekly work time is less than 12.0 hours.
5. The ages of commercial aircraft have a mean of 13 years and a $s_x$ of 7.9 years. If 35 commercial airlines are chosen for a stress test, find the probability that the mean age of this sample group is greater than 15 years.
6. In a study of job training, the times required to learn how to use a word processing system are found to have a mean and a standard deviation of 462.1 min and 76.3 min, respectively. If 32 subjects are randomly selected, find the probability that their mean is between 447.0 and 456.8 min.
7. A study of Reye's Syndrome found that children suffering from the disease had a mean age of 8.5 years and a standard deviation of 3.96 years; their ages approximated a normal distribution. If 36 of those children are randomly selected, what's the chance their mean age is between 7 and 10 years?
8. The lengths of pregnancies are normally distributed with a mean of 268 days and $s_x$ of 15 days. If 25 women were randomly selected for a special diet just before they become pregnant, find the probability that their lengths of pregnancy have a mean less than 260 days (assuming the diet had no effect).
9. More Central Limit: 5.3: 3-5, 7, 9, 12, 14. Read the investigations prior to #11.
10. More Confidence Intervals: 5.3: read p201-203, do #20, 21, 26, 27, 28, 29,
11. More Binomial Distribution: 5.4: 9, 11, 15-17, 26-28, 30, 32 (see notes for textbook)

# STATISTICS LESSON 25

## SCATTERPLOTS & CORRELATIONS



(a) Positive correlation between $x$ and $y$

(b) Strong positive correlation between $x$ and $y$

(c) Perfect positive correlation between $x$ and $y$

(d) Negative correlation between $x$ and $y$

(e) Strong negative correlation between $x$ and $y$

(f) Perfect negative correlation between $x$ and $y$

(g) No correlation between $x$ and $y$

(h) Nonlinear correlation between $x$ and $y$

A scatterplot[1] shows the relationship between two quantitative variables. The value of one variable, usually the explanatory variable, appears on the horizontal axis, and the value of the other variable, usually the response variable, appears on the vertical axis. Each individual data appears in the graph as a point. To interpret a scatterplot, first look for an overall pattern which reveals the direction, shape, and strength of the relationship.

A common way to describe a scatterplot is to invoke its correlation. Correlation refers to a relationship between two or more variables. An example of correlation might be the number of cars in a city to the number of stoplights. The greater the number of cars, the greater the number of stoplights is an example of positive correlation. An example of negative correlation is the number of coyotes in an area to the number of new lambs reaching maturity. Negative correlation is where one increases, the other decreases and positive correlation is where they both increase or both decrease. It is important to note that correlation does not imply causation, only that a relationship exists. Further research would have to be done to prove causation.

One can tell if there is a correlation by how closely the scatterplot represents a known type of graph: linear, quadratic, sinusoidal, exponential, etc. Positive and negative correlations are used to describe a linear correlation and are directly related to the slope of the line.

The strength of the correlation is how close the points lie to a line with little scatter. The strength of a correlation is measured by the correlation coefficient ***r***.[2] Another name for the ***r*** is the Pearson Product Moment Correlation Coefficient in honour of Karl Pearson who developed it. The sample correlation coefficient is represented by ***r*** while the population correlation coefficient is denoted by the Greek letter ρ (pronounced rho). The closer ***r*** is to +1, the stronger the positive correlation. The closer it is to -1, the more negatively correlated. If |***r***| = 1, then the two variables are perfectly correlated. Yet the correlation coefficient of 0 does NOT mean there is no correlation, since there could be a nonlinear correlation. The following formula is to determine a linear correlation:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2} \cdot \sqrt{n\sum y^2 - (\sum y)^2}}.$$

Luckily students can find ***r*** when doing a regression curve on the TI-83/84. But that's not until the next lesson. Until then simplify the formula by using the TI-83, 1-VAR STATS to find the value of the summations. If a correlation coefficient is low, then the student should check for non-linear correlation.

## STATISTICS LESSON 25 HOMEWORK

1. Someone says, "There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. So sending lots of firefighters just causes more damage." What is wrong with this reasoning?[3]
2. A study of elementary school children, ages 5 to 11, finds a high positive correlation between shoe size and scores on a reading comprehension test. What explains this correlation?[4]

For the following questions, construct a scatterplot and describe its correlation.[5]

3. (1,1), (1,5), (2,4), (3,2).
4. (0,3), (1,3), (1,4), (2,5), (5,6).

5. Randomly selected subjects bicycle for one minute at 5.5 mph. Given are their weights with number of calories burned.

| Weight | 167 | 191 | 112 | 129 | 140 | 173 | 119 |
|---|---|---|---|---|---|---|---|
| Calories used | 4.23 | 4.69 | 3.21 | 3.47 | 3.72 | 4.45 | 3.36 |

6. In a study of the factors that affect success in a calculus course, data were collect from 10 different people. Scores on an algebra placement test are given along with their calculus achievement scores.

| Algebra | 17 | 21 | 11 | 16 | 15 | 11 | 24 | 27 | 19 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|
| Calculus | 73 | 66 | 64 | 61 | 70 | 71 | 90 | 68 | 84 | 52 |

7. The paired data below consist of the total weights (in pounds) of paper and plastic discarded by households in a week.

| Paper | 2.41 | 7.57 | 9.55 | 8.82 | 8.72 | 6.96 | 6.83 | 11.42 | 16.08 | 6.83 | 13.05 | 11.36 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Plastic | 0.27 | 1.41 | 2.19 | 2.83 | 2.19 | 1.81 | 0.85 | 3.05 | 3.42 | 2.10 | 2.93 | 2.44 |

More practice can be found at Statistics Workbook For Dummies[6] chapter 16: 1-22.

# STATISTICS LESSON 26

## LEAST SQUARES REGRESSION

It is one thing to be able to determine the correlation between two variables; it is another to be able to predict based on the perceived relationship. The power behind statistics is to be able to infer (predict). If the data is showing strong linear correlation, then a line that best fits the data can be found. This may start out with a random line drawn through as much data as possible. But to be as accurate as possible, you will want to minimize as much as possible the vertical distances between each data and the line.

Hence the technical name for the "best-fitting line" is the Least Squares Regression Line. The LSRL makes the sum of the squares of the vertical distances as small as possible. The LSRL has many other names: trend line, least-squares line, best-fitting line, regression line. As with any line, you will need the slope and y-intercept to write the equation. You can find the slope and y-intercept of the line (y = mx+b) by using the following formulae:

$$m = \frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2} \quad \text{and} \quad b = \frac{(\sum y)(\sum x^2)-(\sum x)(\sum xy)}{n(\sum x^2)-(\sum x)^2}.$$

If you place your independent (explanatory) variable into a list and the dependent (response) variable into another list on your TI-83/84 calculator, you can quickly determine the value of the summations to help in your calculations by doing 1-Var Stats. An even quicker method is to go to Stats/Calc and choose the option LinReg. Type in the lists you have chosen for your independent and dependent variables so that the command looks something like: LinReg(x,y) or LinReg(L1,L2). You will get the results where the calculator defines its slope and y-intercept and leaves it up to you to write the equation and use it for predictions.

The data might actually be a pattern other than linear. So you can test different non-linear regression curves on your calculator as well: QuadReg(x,y), ExpReg(x,y), SinReg(x,y). But how will you know which is the best regression curve? Look for the r or $r^2$ values on display with the equation. If it does not appear, then you need to go to DiagnosticsON, found under Catalogue. Enter the command, a DONE appears, and then re-enter your regression codes. The regression line or curve that best fits the data will be the one closest to ±1.

The coefficient of determination is the amount of variation in y that is explained by the regression line, $r^2$= explained variation/total variation. If r = 0.8, then $r^2$=0.64 which means 64% of the total variation in y can be explained by the regression line. So the closer $r^2$ is to 1, the closer to 100% of the variation is explained by the line.

Example: Sabra's parents are concerned she is short for her age. Their doctor has recorded the following in the chart below. Make a scatter plot. Find the LSRL. Predict her height at 40 and at 80 mos.

| Age (months) | 36 | 48 | 51 | 54 | 57 | 60 |
|---|---|---|---|---|---|---|
| Height (cm) | 86 | 90 | 91 | 93 | 94 | 95 |

Solution: The equation is y = 0.383x + 71.95 with an $r^2$ of 0.994. At 40 months, she is 87.3 cm and at 60 months, she is 102.59 cm.

## STATISTICS LESSON 26 HOMEWORK[1]

In the following questions, find the regression line. Find the predicted values where they are requested.

1.

| Paper | 2.41 | 7.57 | 9.55 | 8.82 | 8.72 | 6.96 | 6.83 | 11.42 |
|---|---|---|---|---|---|---|---|---|
| Household size | 2 | 3 | 3 | 6 | 4 | 2 | 1 | 5 |

2.

| Weight | 167 | 191 | 112 | 129 | 140 | 173 | 119 |
|---|---|---|---|---|---|---|---|
| Calories used | 4.23 | 4.69 | 3.21 | 3.47 | 3.72 | 4.45 | 3.36 |

3.

| HC | 0.65 | 0.55 | 0.72 | 0.83 | 0.57 | 0.51 | 0.43 | 0.37 |
|---|---|---|---|---|---|---|---|---|
| CO | 14.7 | 12.3 | 14.6 | 15.1 | 5.0 | 4.1 | 3.8 | 4.1 |

Find the best predicted value of CO (carbon monoxide) given that the HC (hydrocarbon) amount is 0.75 g/m.

4.

| Living area (hundreds of sq ft) | 15 | 38 | 23 | 16 | 16 | 13 | 20 | 24 |
|---|---|---|---|---|---|---|---|---|
| Taxes (in thousands) | 1.9 | 3.0 | 1.4 | 1.4 | 1.5 | 1.8 | 2.4 | 4.0 |

Find the best predicted value of taxes for a home with a living area of 1800 sq ft.

5.

| Temperature (Celsius) | -62 | -41 | -36 | -26 | -33 | -56 | -50 | -66 |
|---|---|---|---|---|---|---|---|---|
| Snow depth (in cm) | 21 | 13 | 12 | 3 | 6 | 22 | 14 | 19 |

Find the best predicted snow depth given a temperature of -60$^0$C.

6. Suppose you remember triangular numbers, but can't remember their formula. Enter the values of 1, 2, 3, 4 to represent the place in the series into L1. Now, enter 1, 3, 6, 10 into L2, these numbers representing the series. Now different types of regression to find the equation that best fits the data. Justify your answer. Based on your results, what would be the 10th triangular number?[2]

# STATISTICS LESSON 27

## HYPOTHESIS TESTING[1]

Inferential statistics is all about making predictions about the population from the sample. A prediction can be called hypothesis, if intending to make a true statement about the population. From science, hypothesis is a statement seeking verification before being called a theory. So how can we tell if the hypothesis is true?

The infamous Law of Contradiction is used in determining whether a hypothesis is true. The Law of Contradiction states that if two statements oppose each other (thus being mutually exclusive), then one statement is true and the other is false. So when determining whether your prediction is true about a population, you need to write out two statements involving your prediction so that you have two opposing views.

Such steps to proving the validity of the hypothesis can be summarized (thanks to Mario Triola):

1) Identify the hypothesis to be tested, making it the statement of equality.
2) Determine the alternate hypothesis, the statement that would be true if the hypothesis is false.
3) The hypothesis is $H_0$ (**null hypothesis**) and the **alternate hypothesis** is $H_a$ or $H_1$.
4) Select the **significance level,** α, based on the seriousness of a Type I Error. Values of 0.05 and 0.01 are common.
5) Identify the statistic that is relevant to the test and its sampling distribution.
6) Determine the test statistic, the critical values, and the critical region. Draw a graph and label parts.
7) Reject $H_0$ if the test statistic is in the critical region. Fail to reject if the test statistic is not in the region. (Failing to reject still hasn't proven $H_0$ is true.)

There are two ways to be in error when it comes to hypothesis testing. You either reject the null hypothesis when it is true (Type I Error) or you fail to reject the null hypothesis when it is false (Type II Error). The probability of rejecting the null hypothesis when it is true is called the **significance level**. The symbol for Type I Error is α (alpha) and the symbol for a Type II Error is β (beta). In other books, you might read the expressions "false positives," "false negatives," "true positives," "true negatives." A false negative would be rejecting the null hypothesis when it was true and a false positive is failing to reject the null hypothesis when it was false.

Here is a simple incident of a false positive. My sister told me a story about how my grandfather was drafted into the army to fight in WWII. Yet he got out because a tuberculosis test came back positive. The army didn't need men dying from diseases when they had enough casualties from the front lines, so they sent him home. Yet he never got sick with tuberculosis. That's an example of a false positive. It's also an example how stories can get twisted and exaggerated. My grandfather actually had surgery on his leg and it was slowly recovering, and because of it, he failed his physical.

The tails in a distribution are the extreme regions bounded by critical values. By examining $H_0$, we should be able to deduce whether a test is right-tailed, left-tailed, or two-tailed. The tail will correspond to the critical region containing the values that would conflict significantly with the null hypothesis. A useful check is to observe the inequality sign in $H_{1;}$ it should point in the same direction of the critical region.

Example 1: After analyzing 106 body temperatures, a medical researcher makes a claim that the mean body temperature is less than $98.6^0$F. Identify the Type I and II errors.

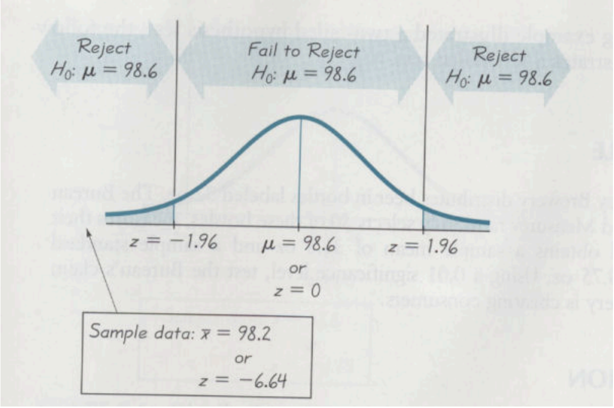
Solution: $H_a$ is $\mu < 98.6$. $H_0$ is $\mu \geq 98.6$. This test is left-tailed because the null hypothesis is rejected if the sample mean is significantly less than 98.6. The Type I error is to reject the null hypothesis when the population is really equal to or greater than 98.6. The Type II error is to fail to reject the null hypothesis when the population mean is really less than 98.6.

Example 2: A claim about the population mean weight of all aspirin tablets is tested with a significance level of $\alpha = 0.05$. The conditions are such that the standard normal distribution can be used. Find the critical values of the z-score if the test is (a) two-tailed, (b) left-tailed, (c) right-tailed.

Solution: In a two-tail test, the significance level is divided equally between the two tails. So there is an area of 0.025 in each tail. The z-scores that correspond to that are -1.96 and 1.96. If only one tail, then you are finding the z-score so that 95% falls above or below it. For 95% to fall above, the z-score would be found using normalcdf(z, 1000) and by playing around with the number you get -1.645. To get it so the area is below, the z-score is a +1.645.

Example 3: Test the claim that the mean body temperature is $98.6^0$F with a 0.05 level of significance.

Solution: $H_0$ is $\mu = 98.6$ and $H_1$ is $\mu \neq 98.6$. $\alpha = 0.05$. This is a two-tailed test. The z-score (using standard error in the calculations) is -6.64. We know that the z-scores that have 95% of the population within one standard deviation of the mean are -1.96 and 1.96. The z-score of -6.64 falls within the critical region, so we reject the null hypothesis. In other words, we conclude that there is sufficient evidence to warrant rejection of the claim that the mean body temperature of healthy adults is $98.6^0$F.


Reject $H_0$: $\mu = 98.6$
Fail to Reject $H_0$: $\mu = 98.6$
Reject $H_0$: $\mu = 98.6$
$z = -1.96$
$\mu = 98.6$ or $z = 0$
$z = 1.96$
Sample data: $\bar{x} = 98.2$ or $z = -6.64$

# STATISTICS LESSON 27 HOMEWORK

1. Express the null and alternate hypothesis: the mean IQ of statistic instructors is 120.

2. Express the null and alternate hypothesis: the mean weight of paper discarded by households in one week is less than 10 lb.

3. Express the null and alternate hypothesis: the mean time for undergraduates to earn a degree is greater than 5 years.

4. Express the null and alternate hypothesis: The mean annual age of US commercial aircraft is at least 10 years.

5. Find the critical z-scores for the given conditions: two-tailed test with $\alpha = 0.01$.

6. Find the critical z-scores for the given conditions: left-tailed test with $\alpha = 0.05$.

7. Find the critical z-scores for the given conditions: right-tailed test with $\alpha = 0.025$.

8. Identify the null hypothesis: There is sufficient evidence to support the claim that the mean age of college professors is greater than 30 years.

9. Identify the null hypothesis: There is sufficient evidence to warrant rejection of the claim that adult men have a mean height of 70 in.

10. In a hypothesis test, can $\alpha=0$? How?

11. We should arrange the null and alternative hypotheses so that the most serious error would be the rejection of a true null hypothesis. Because the probability of that error is the significance level, we should make $\alpha$ very small if the result of a Type I error is serious. Suppose we make a claim about the mean weight of all garbage disposed in a week. One study showed that the mean was 27.44 lb. Underestimating the true mean would be a very serious error because landfill and garbage pickup would be insufficient. On the other hand, overestimating will lead to waste in labour, equipment and land costs. Which error is more serious? Which null and alternative hypotheses should be used?

# STATISTICS LESSON 28

## MORE ON HYPOTHESIS TESTING[1]

Many professionals rely on a probability value to base their decision on whether or not to reject the hypothesis. This probability value is also known as a P-value, which is the probability of getting a value that is as extreme as the one found from the data, assuming the null hypothesis is true. Whereas traditional hypothesis tests results in a reject or fail to reject, the p-value approach measures how confident we are in rejecting a null hypothesis (similar to confidence level). In other words, the traditional methods would compare critical values (test statistic with critical values) and the P-value method compares significance levels.

If the P-value is less than or equal to the significance level, α, then rejecting the null hypothesis is a must, otherwise, fail to reject. The reason: if the P-value is less than 0.01, the data is highly statistically significant as strong evidence against the null hypothesis. If the P-value is between 0.01 and 0.05, it is statistically significant, and anything above 0.05 is insufficient evidence against the null hypothesis. Many statisticians consider first before they measure the results as to what significance level they are working with; otherwise they might be tempted to let the data influence the interpretation.

The P-value is the area under the curve bounded by the test statistic. And since the area under the curve is most likely determined by a z-score or t-score, which measures distance, the smaller the tail the more distance between the assumed value and known. The greater the distance, the more unlikely the null hypothesis is a true reflection of reality. Thus the smaller the P-value, the more likely it is to reject the null hypothesis.

Example: Using the same example of 106 body temperatures of a mean $98.2^0$F and $s_x$=0.62, use the P-value method to test the claim that the mean body temperature is $98.6^0$F.

Solution: $H_0$: μ=98.6, $H_1$: μ≠98.6. The z-score is -6.64. Using normalcdf(-200, -6.64) and multiplying by 2 because it is a two-tailed test, the P-value is 0.0002. Since the P-value is less 0.01, then we have significant evidence to reject the claim that the mean body temperature is $98.6^0$F.

Confidence intervals can also be used to determine the results of hypothesis testing. If we find a confidence interval which ends up estimating population parameter not contained in the interval, then we can reject the null hypothesis. For example, if a confidence interval is found for body temperatures to be 98.08 to 98.32 and our null hypothesis was a mean temperature of 98.6, then we can reject our null hypothesis. Just be careful when relating a confidence interval with a one-tailed test! For example, a one-tailed test with a 0.05 significance level corresponds to a 90% confidence level.

Example: University of Maryland had a sample mean of $98.2^0$F and $s_x$ of 0.62 for 106 body temperatures. Is there enough data to provide sufficient evidence to warrant rejection of the common belief that average temperature for a human body is $98.6^0$F?

Solution: We will use the probability of getting a sample mean of $98.2^0$F to determine whether there is significant enough difference to reject the commonly believed average. Assuming the mean is 98.6, we create a 95% confidence interval: $98.6 \pm 1.96(0.62/\sqrt{}106)$. Our sample mean of 98.2 falls outside of the interval (98.48, 98.82). Since the probability would be under 5% of falling outside that interval, we conclude that there is sufficient evidence to reject our assumption that the mean temperature is $98.6^0$F.

Example 2: A poll of 100 randomly selected car owners revealed that the mean length of time that they plan to keep their cars is 7.01 years and the $s_x$ is 3.74. The president of an auto company is trying to plan a sales campaign. Test the claim of the sales manager, who states that the mean length time for all car owners is less than 7.5 years. Use a 0.05 significance level.

Solution:
By the traditional approach, $H_0$ is $\mu=7.5$ and $H_1$ is $\mu\neq7.5$ and z-score = (7.5-7.01)/0.374 = 1.31. A significance level of 0.05 says we are working with a 90% area under the curve between the two z-scores -1.64 and 1.64. Since 1.31 does not fall in the critical region, we fail to reject the null hypothesis.

By the P-value method: Using the normalcdf(1.31, 99) and multiplying it by 2 for two tails, we get 0.19 as a result. This is much bigger than our significance level of 0.05. So we fail to reject the null hypothesis.

Using a Confidence Interval: Two tails would mean 90% confidence, 7.01 ±0.374(1.64) = (6.397, 7.623). Fail to reject since 7.5 is in the interval.

## STATISTICS LESSON 28 HOMEWORK

Do the following with the three approaches to hypothesis testing, if possible.

1. Test the claim that the population mean $\mu = 100$, given a sample of $n = 81$ for which the sample mean is 100.8 and $s_x = 5$. Test at the $\alpha = 0.01$ significance level.
2. Test the claim that $\mu \geq 20$, given a sample of $n = 100$ for which the sample mean is 18.7 and $s_x = 3$. Use a significance level of $\alpha = 0.05$.
3. Test the claim that a population mean equals 500. You have a sample of 300 items for which the sample mean is 510 and the sample standard deviation is 50. Test at the $\alpha = 0.10$ significance level.
4. Boston Bottling Company distributes cola in cans labelled 12 oz. The Bureau of Weights and Measures randomly selected 36 cans, measured their contents and obtained a sample mean of 11.82 oz and a $s_x = 0.38$ oz. Use a significance level of 0.01 to test the claim that the company is cheating consumers.
5. In a study of consumer habits, researchers designed a questionnaire to identify compulsive buyers. The scores obtained had a mean of 0.83 and $s_x$ of 0.24. Assume the subjects were randomly selected and the sample size is 32. At the 0.01 significance level, test the claim that the population mean is greater than 0.21.
6. In a study of distances traveled by buses before the first major engine failure, a sampling of 191 buses resulted in a mean of 96,700 miles and a $s_x$ of 37,500 miles. At the 0.05 significance level, test the manufacturer's claim that mean distance traveled before a major engine failure is more than 90,000 miles.

# STATISTICS LESSON 29

## INFERENCES FROM TWO SAMPLES[1]

**Inferences about Two Means:** If two samples from two separate populations are dependent, meaning they are related to one another, then calculate the differences between each statistic. Use the letter *d* in its abbreviation to refer to those differences in data, mean, and standard deviation. Let n represent the number of pairs of data. What is created now, assuming the samples were normal distributions, is a unique sample from which we can do a hypothesis test.

Example: Does it pay to take preparatory courses for standardized tests? Use the following results to determine the claim that the Allan Preparatory Course has no effect on SAT. Use a 0.05 level of significance. (Triola)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SAT before | 700 | 840 | 830 | 860 | 840 | 690 | 830 | 1180 | 930 | 1070 |
| SAT after | 720 | 840 | 820 | 900 | 870 | 700 | 800 | 1200 | 950 | 1080 |
| Difference | -20 | 0 | 10 | -40 | -30 | -10 | 30 | -20 | -20 | -10 |

Solution: If the course has no effect, the difference is to be 0, thus $\mu_d$=0. The alternate hypothesis is $\mu_d \neq 0$. The significance level is α=0.05. Because the sample is small and the mean is 0, we can use the student t distribution. The average difference between the data ($\bar{d}$) is -11.0 and the $s_d$= 20.2. Using these to calculate the t-score: (-11-0)/(20.2/√10) = -1.722. To have a significance level of 0.05, then the t-score should have been -2.262 (or smaller) or 2.262 (or bigger). Find that by tcdf(t,100, 9)= 0.025. Since the test statistic does not fall in the critical region, we fail to reject the null hypothesis. Thus there is insufficient evidence to reject the claim that the Allan Preparatory Course has no effect on SAT scores.

The example could have been done with the P-value approach. You would have found the P-value to have been greater than 0.05 by calculating 1-tcdf(-1.722,1.722). We would have rejected the claim only if the P-value was less than the level of significance. We could have also done a confidence interval for our $\mu_d$ by $\mu_d \pm z(s_d/\sqrt{n})$ and found our $\bar{d}$=-11 within our confidence interval of (-25.4,3.4).

If the two samples from two separate populations are independent and large (n>30), then we need to vary our approach to inferring. We will deal in the future with samples that are independent and small. For now, the larger samples tend to be normally distributed as the Central Limit Theorem states. Then the differences between the data when paired will also be normally distributed. We can calculate the mean differences the same way. But for the z-score, we use the formula $\frac{(\overline{x_1}-\overline{x_2})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}}}$, where the denominator is the formula for the standard deviation. Remember, since the population standard deviation is most often not known, we can use the sample standard deviation. The equation for the standard deviation can interpreted as the square root of the sum of variances (using standard error in place of the standard deviation). Why add variances? Because the variance of the differences between two independent variables equals the sum of each sample variance! Providing for this difference between independent and dependent samples of different populations, the work of inferring about our independent samples is the same.

Inferences about two variances and two proportions may be added in a future revision of this text, to further expand this topic.

## STATISTICS LESSON 29 HOMEWORK

1. Assume that you want to test the claim that the paired sample data come from a population for which the mean difference is 0. Assume a 0.05 level of significance, find $\bar{d}$, $s_d$, t, and the critical values.

| X | 3 | 4 | 4 | 6 | 8 | 10 | 12 | 11 | 14 | 17 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Y | 2 | 5 | 4 | 5 | 9 | 8 | 9 | 8 | 12 | 15 | 17 |

2. Using the sample paired in question #1, construct a 95% confidence level for the mean of all x-y values.

3. Malloy Advertising has prepared two different television commercials for women's jeans. One commercial is humorous, and the other is serious. A test screening involves 8 consumers who are asked to rate the commercials. At the 0.05 significance level, test the claim that the differences between the commercials have a mean of 0. Based on the result, does one commercial seem to be better?

| Consumer | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Humorous | 26.2 | 20.3 | 25.4 | 19.6 | 21.5 | 28.3 | 23.7 | 24.0 |
| serious | 24.1 | 21.3 | 23.7 | 18.0 | 20.1 | 25.8 | 22.4 | 21.4 |

4. A dose of a drug Captopril, designed to lower systolic blood pressure, is administered to 10 randomly selected volunteers. Construct the 95% confidence interval for $\mu_d$, the mean of the differences between the before and after scores.

| Subject | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Before | 120 | 136 | 160 | 98 | 115 | 110 | 180 | 190 | 138 | 128 |
| After | 118 | 122 | 143 | 105 | 98 | 98 | 180 | 175 | 105 | 112 |

5. Using the data in question #4, test the claim that the systolic blood pressure is not affected by the pill. Use a 0.05 significance level. Does it appear that the drug has an effect?

6. A teacher proposes a course designed to increase reading speed and comprehension. To evaluate the effectiveness of the course, the teacher tests students before and after the course. Construct a 95% confidence interval for the mean of the differences between the before and after scores.

| Student | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Before | 100 | 170 | 135 | 167 | 200 | 118 | 127 | 95 | 112 | 136 |
| After | 136 | 160 | 120 | 169 | 200 | 140 | 163 | 101 | 138 | 129 |

7. Use a 0.05 significance level to test the claim that the two samples come from populations with the same mean. In each case, the two samples are independent and have been randomly selected. The control group has n=40, mean = 79.6 and $s_x$ = 12.4 whereas the experimental group has size of 40, mean of 84.2 and $s_x$ of 12.2.

8. Use a 0.05 significance level to test the claim that the two samples come from populations with the same mean. In each case, the two samples are independent and have been randomly selected. The treated group has size of 32, mean of 8.49 and $s_x$ of 0.11. The untreated group has size of 60, mean of 8.41 and $s_x$ of 0.18.

9. Data were collected on the weights of men. For 804 men aged 25 to 34, the mean is 176 lbs and a $s_x$ of 35.0 lb. For 1657 men aged 65 to 74, the mean and $s_x$ are 164 lbs and 27.0 lbs, respectively. Construct a 99% confidence interval for the difference between the means of the men in the two age brackets.

Name: ______________________________ Math Class: ____________________

Date: ____________________ Quest: **Statistics Lessons 20-29.**

SHOW WORK for partial credit! A calculator and your notebook are allowed on this test. You must work alone. Do as much of the work on space provided.

5+2

1. Match the distribution to its short description:

| | |
|---|---|
| ____ A. Normal | a. A combination of Lorentzian and Normal |
| ____ B. Binomial | b. Morbid studies lead to queuing and time studies |
| ____ C. Student t | c. Fat tails, describes resonant behavior |
| ____ D. Poisson | d. Bell curve follows empirical rule |
| ____ E. Lorentzian | e. Two outcomes per trial, constant probability |
| ____ F. Hypergeometric | f. Brewery worker used small samples |
| ____ G. Voigt Profiles | g. 75% criteria is same as Binomial |

10

2. My niece Sabra is her parent's first child. They are extremely conscientious about her well-being. Having heard that Sabra was born smaller than most babies, they measure her every month to determine if she is showing any health problems.

| mos. | 0 | 1 | 3 | 6 | 9 | 12 | 15 | 18 | 21 |
|---|---|---|---|---|---|---|---|---|---|
| cm | 50 | 53 | 59 | 67 | 71 | 75 | 79 | 81 | 83 |

A. Find the best-fitting line for the given data.

B. Predict her size at her 3rd birthday using your equation.

C. Describe the correlation of these variables.

5

3. Identify the null and alternate hypothesis for each statement below:

a. The average score on the Algebra contest was 10.

b. The average family size is still under five persons.

10

4. High school students at SLA, numbering 29, took an Algebra Contest and scored on average 10. The standard deviation for the sample is 2.56.

a. (2 pts) Determine the standard error.

b. (2 pts) Find the margin of error.

C. (6 pts) Find the 95% confidence interval for the population mean.

5

5. Suppose a bank knows that on average 120 customers arrive in a certain hour. Using the time interval of one minute, calculate the probability of exactly one customer arriving in a given 1-minute interval within that hour.

____
5

6. A certain 7-8 grade class took an algebra contest and received the following scores: 2, 4, 6, 7, 8, 13. Calculate their class average and standard deviation. Calculate the t-score for their class average if the national average is 12.

____
5

7. Assume the amount of garbage discarded weekly in the Halifax Regional Municipality is normally distributed with a mean of 10 lbs per household and a standard deviation 4 lbs. Find the probability that 100 randomly selected households have a mean between 9 and 14 lbs.

____
5

8. *Compliments* distributes cola in cans labelled 12 ounces. Quality tests are done by randomly selecting 30 cans, measuring their contents and obtained a sample mean of 11.5 and a standard deviation of 0.26 oz. Use a significance level of 0.01 to test the claim that the company is cheating consumers.

____
5

9. Circle whether or not the statement is true or false:
True False A value of 0.865 indicates a strong negative correlation.
True False The probability of rejecting the null hypothesis is α.
True False Standard error of the mean measures variability and is used almost interchangeably with standard deviation.
True False Null hypotheses are rejected if the p-value is less than the significance level.
True False Regression curves are only calculated for linear patterns.

____
5

10. Suppose you are a lawyer who is trying to establish that a company has been unfair to minorities with regard to salary increases. Suppose the mean salary increase is 8% per year. Determine the null hypothesis and alternate hypothesis. What is the type I error and type II error? What would be the results of those errors?

____
0+10

Bonus - A teacher found that on average students scored 18 out of 25 on her true/false statements. Using a sample of 25 students with a $s_x$=2.3, determine the probability of students nation-wide getting 22 or more correct. Repeat the question by using each: Binomial, Student t, and Normal. Compare results. Explain which distribution is more appropriate for this question.

# NOTES AND REFERENCES

Please note that the assigned textbook for the class was Mathematical Modelling, book 2, published in Ontario by Nelson.

**Lesson 1**

1. Either Calkins, Keith. An Introduction to Statistics. Andrews University, 1999 version. www.andrews.edu/~calkins/math/webtexts/stattoc.htm lesson 1 or Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995, page 4.
2. Either Calkins, Keith. An Introduction to Statistics. Andrews University, 1999 version. www.andrews.edu/~calkins/math/webtexts/stattoc.htm lesson 1 or Triola, Mario. ElementaryStatistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995, page 5-7.
3. Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995, pg 10.
4. Questions for homework were taken mainly from Calkins, Keith. An Introduction to Statistics. Andrews University, 1999 version.www.andrews.edu/~calkins/math/webtexts/stattoc.htm, lessons 1 & 2.

**Lesson 2**

1. Rumsey, Deborah. Statistics Workbook For Dummies. New Jersey: Wiley Publishing, Inc; 2005, p12.
2. ibid, p 12.
3. Yates, Dan; David Moore; George McCabe. The Practice of Statistics: TI-83 Calculator Enhanced. W.H. Freeman, 2000, pg 20.
4. ibid, pg 20.

**Lesson 3**

1. Yates, Dan; David Moore; George McCabe. The Practice of Statistics: TI-83 Calculator Enhanced. W.H. Freeman, 2000, pg 15.

**Lesson 4**

1. Calkins, Keith. An Introduction to Statistics. Andrews University, 1999 version. www.andrews.edu/~calkins/math/webtexts/stattoc.htm, lesson 2.
2. Calkins, Keith. An Introduction to Statistics. Andrews University, 1999 version. www.andrews.edu/~calkins/math/webtexts/stattoc.htm, lesson 2 homework.

**Lesson 5**

1. Yates, Dan; David Moore; George McCabe. The Practice of Statistics: TI-83 Calculator Enhanced. W.H. Freeman, 2000, pg 34.
2. Rumsey, Deborah. Statistics Workbook For Dummies. New Jersey: Wiley Publishing, Inc; 2005, pg 55.
3. ibid, pg 57.

**Lesson 6**

1. Homework questions came mainly from Calkins, Keith. An Introduction to Statistics. Andrews University, 1999 version. www.andrews.edu/~calkins/math/webtexts/stattoc.htm, lesson 4 and lesson 4 homework

**Lesson 8**

1. Calkins, Keith. An Introduction to Statistics. Andrews University, 1999 version. www.andrews.edu/~calkins/math/webtexts/stattoc.htm, lesson 6 homework
2. Yates, Dan; David Moore; George McCabe. The Practice of Statistics: TI-83 Calculator Enhanced. W.H. Freeman, 2000, pg 71
3. ibid, pg 77
4. ibid, pg 79
5. ibid, pg 79

**Lesson 9**

1. Rumsey, Deborah. Statistics Workbook For Dummies. New Jersey: Wiley Publishing, Inc; 2005, pg 86.
2. ibid, pg 86.
3. ibid, pg 87.
4. Couldn't find this in any of my resources at home, so I'm assuming I got it from the textbook, now locked away at the school: Mathematical Modelling, book 2. Ontario: Nelson.
5. Calkins, Keith. An Introduction to Statistics. Andrews University, 1999 version. www.andrews.edu/~calkins/math/webtexts/stattoc.htm, lesson 7 homework.
6. Calkins, Keith. An Introduction to Statistics. Andrews University, 1999 version. www.andrews.edu/~calkins/math/webtexts/stattoc.htm, lesson 7.

**Lesson 10**

1. Rumsey, Deborah. Statistics Workbook For Dummies. New Jersey: Wiley Publishing, Inc; 2005, pg 92.
2. Calkins, Keith. An Introduction to Statistics. Andrews University, 1999 version. www.andrews.edu/~calkins/math/webtexts/stattoc.htm, lesson 7 homework.
3. Ibid
4. Ibid
5. Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995, pg 99.
6. ibid, pg 98.
7. ibid, pg 98.
8. ibid, pg 98.

**Lesson 11**

1. Yates, Dan; David Moore; George McCabe. The Practice of Statistics: TI-83 Calculator Enhanced. W.H. Freeman, 2000, pg 92.
2. ibid, pg 92.
3. ibid, pg 92.
4. ibid, pg 92.
5. ibid, pg 96.

**Lesson 12**

1. Yates, Dan; David Moore; George McCabe. The Practice of Statistics: TI-83 Calculator Enhanced. W.H. Freeman, 2000, pg 130.
2. ibid, pg 130.
3. ibid, pg 130.
4. ibid, pg 130.
5. ibid, pg 130.
6. ibid, pg 130.
7. ibid, pg 131.
8. ibid, pg 132.
9. ibid, pg 132.
10. ibid, pg 133.
11. ibid, pg 139.
12. ibid, pg 139.
13. ibid, pg 141.
14. ibid, pg 140.
15. ibid, pg 153.
16. ibid, pg 153.
17. ibid, pg 153.

**Lesson 12 Homework Continued**

1. Unless specified otherwise, this homework is a selection of questions found in Advanced Mathematical Concepts: Pre-Calculus with Applications. Glencoe, 2007, chapter 12.
2. Foerster, Paul. Algebra and Trigonometry: Functions and Applications, 2nd Edition. Addison-Wesley Publishing Co, 1990; pg 667.
3. ibid, pg 668.

**Lesson 13**

1. This chapter and homework is a summary of Conditional Probability as found in Advanced Mathematical Concepts: Pre-Calculus with Applications. Glencoe, 2007, chapter 12.

**Lesson 14**

1. Sullivan, Michael. PreCalculus, 7th Edition. New Jersey: Pearson- Prentice Hall, 2005; pg 843
2. Foerster, Paul. Algebra and Trigonometry: Functions and Applications, 2nd Edition. Addison-Wesley Publishing Co, 1990; pg 660.
3. ibid, pg 666.

**Lesson 14 Homework Continued**

1. Advanced Mathematical Concepts: Pre-Calculus with Applications. Glencoe, 2007
2. Sullivan, Michael. PreCalculus, 7th Edition. New Jersey: Pearson- Prentice Hall, 2005; pg 844.

**Lesson 15**

1. Calkins, Keith. Probability and Distributions. Andrews University, 2002 HTML version. www.andrews.edu/~calkins. Lesson 6 homework.

**Lesson 16**

1. This chapter is predominantly from Foerster, Paul. Algebra and Trigonometry: Functions and Applications, 2nd Edition. Addison-Wesley Publishing Co, 1990. However there are a few questions for which I cannot find the original source. These questions appear in none of my resources. I highly doubt that the source is the Mathematical Modelling, book 2 which is locked away at school. I am pretty confident that I made up the questions myself since I have done that a bit and because by this stage some kids were impatient for "drill and grill" questions. By writing the lessons, I could cater to *some* wishes☺

**Lesson 17**

1. Calkins, Keith. Probability and Distributions. Andrews University, 2002 HTML version. www.andrews.edu/~calkins. Lesson 7 and Lesson 7 Homework.

**Lesson 18**

1. Paraphrased lesson from Calkins, Keith. Probability and Distributions. Andrews University, 2002 HTML version. www.andrews.edu/~calkins, lesson 9 and lesson 9 homework unless specified otherwise.
2. Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995; pg 208.
3. ibid, pg 208.
4. ibid, pg 209.
5. ibid, pg 210.
6. ibid, pg 217.
7. ibid, pg 217.
8. ibid, pg 218.
9. ibid, pg 219.

**Lesson 19**

1. Paraphrased lesson from Calkins, Keith. Probability and Distributions. Andrews University, 2002 HTML version. www.andrews.edu/~calkins, lesson 9 and lesson 9 homework unless specified otherwise.
2. Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995; pg 273

**Test Review over Lessons 12-19**

1. The test review questions came mostly from pgs 883-885 of Holliday, Berchie. AdvancedMathematical Concepts: Pre-Calculus with Applications. Glencoe, 2007.

**Test over Lesson 12-19**

Questions come from Foerster, Paul. Algebra and Trigonometry: Functions and Applications, 2nd Edition. Addison-Wesley Publishing Co, 1990, my time working with Keith Calkins where we shared writing and distributing tests, and from my own interactions within the Sandy Lake Academy community.

**Lesson 20**

1. Calkins, Keith. Probability and Distributions. Andrews University, 2002 HTML version. www.andrews.edu/~calkins, lesson 12.
2. ibid
3. ibid
4. Rumsey, Deborah. Statistics Workbook For Dummies. New Jersey: Wiley Publishing, Inc; 2005, pg 122.

5. ibid, pg 121.
6. ibid, pg 121.
7. ibid, pg 121.
8. ibid, pg 123.

**Lesson 21**

I have no idea where the first nine homework questions come from. I checked all seven of the books I used and cannot find them. It's probably the textbook: Mathematical Modelling, book 2 that is sitting locked away in the school. There is only one chapter that really gets into statistics.

1. Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995; pg 303.
2. ibid, pg 304.
3. ibid, pg 305.
4. ibid, pg 307.

**Lesson 22**

This is a mixture of Keith Calkins' Probability and Distributions of 2002 (lesson 9), 2006 (lesson 10) versions and the latest I found while writing this book which was 2009-2010 version (lesson 8). You can try locating it too on www.andrews.edu/~calkins. Click on the link "Prob & Distr."

**Lesson 23**

This is a mixture of Keith Calkins' Probability and Distributions of 2002 (lesson 9), 2006 (lesson 10) versions and the latest I found while writing this book which was 2009-2010 version (lesson 8). You can try locating it too on www.andrews.edu/~calkins. Click on the link "Prob & Distr".

**Lesson 24**

1. Calkins, Keith. Probability and Distributions. Andrews University, 2002 HTML version. www.andrews.edu/~calkins , lesson 12.
2. Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995; Pg 257-258.
3. ibid, pg 260.
4. ibid, pg 261-263.

**Lesson 25**

1. Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995. Image on p477.
2. Calkins, Keith. Probability and Distributions. Andrews University, 2002 HTML version. www.andrews.edu/~calkins, lesson 15.
3. Yates, Dan; David Moore; George McCabe. The Practice of Statistics: TI-83 Calculator Enhanced. W.H. Freeman, 2000; pg 213.
4. ibid, pg 213.
5. Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995. My questions 3-7 were taken from pages 487-489: 3, 4, 8, 10, 11.
6. This question references chapter 16 of Rumsey, Deborah. Statistics Workbook For Dummies. New Jersey: Wiley Publishing, Inc; 2005. I found my students needed some extra help, so I cut and pasted questions 1-22 onto one two-sided sheet for them to try. I

would have deleted this question from my lesson but left it for future reference in case I need some extra examples.

**Lesson 26**

1. Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995. My questions 1-5 come from Triola's pages 504-505: 5, 10, 13, 14, 15.
2. Calkins, Keith. Probability and Distributions. Andrews University, 2002 HTML version. www.andrews.edu/~calkins . Lesson 15 homework.

**Lesson 27**

1. Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995. The majority of lesson 27 and its homework is a summarized version of pg 338-349 with the exception of example 3 and the image which are summaries of pg 352-353.

**Lesson 28**

1. Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995. The majority of lesson 28 and its homework is a summarized version of pg 355-365. Questions 1-3 are found on page 363, questions 4-5 on page 364, and question 6 on page 365.

**Lesson 29**

1. Triola, Mario. Elementary Statistics, 6th Edition. New York: Addison-Wesley Publishing Co, 1995. The majority of lesson 29 and its homework is a summarized version of pg 408-425. Questions 1-6 can be found on pages 416-417 and questions 7-9 can be found on page 425.

**Lesson 20-29 Quest**

The quest was built on ideas from several sources and an example from the Halifax Regional Municipality garbage with added actual local Canadian data.