

# Welcome to HE 220: Introduction to Epidemiology and Healthy Data Analysis

This course is an introduction to the statistics most commonly used in the public health field. We will begin with answering the question "why statistics".

## Why do I have to take statistics?

Numbers are all around us and often we organize and use them without much thought and certainly do not always call what we do "statistics". The numbers are collectively called "data" (by the way, data are plural- if I only had one number, it would be datum). The ability to understand the concepts behind statistics helps to organize and interpret the data in a way that is meaningful, useful, and - if done correctly- predictive and/or generalizable. When we talk about the field of public health, exercise sports science, or other health and human performance field, the ability to predict and generalize allows for data collected from sample studies to be inferred on broader populations. We will examine this further in Section 1.

As a student, you need to learn how to use and interpret statistics for more than just meeting a graduation requirement. Even if you never plan on gathering and analyzing data yourself, it is important that you are knowledgeable about symbols, vocabulary, statistical concepts, statistical tools, study design, and analysis procedures so you are able to interpret studies (especially journal articles) in your academic field. This way you will know if the data were used and interpreted correctly for inference and also allows for you to draw your own conclusions: the ability to understand the fundamentals help you make informed decisions outside the classroom (both consumer and citizen).

## How this resource works

This open education resource, or OER, has been designed as your learning material for HE 220. First and foremost it is free and easily available to all students. In addition, when developing the material, different learning styles were taken into account. At the start of each section you will find a printable topic preview that includes learning objectives, symbols, formulas, and key concepts. In addition, there are links to helpful tools and tutorials such as the use of a calculator, spreadsheets, or other calculating online statistical tool. Finally, you have access to a practice guide with step by step examples as well as practice problems and answers.

*It is important to note that there are many more statistical techniques than mentioned in this resource. The topics presented in this OER are only those necessary to meet the learning outcomes and objectives for the HE 220 course.*

## Section 1: Introducing Statistics

### Statistics Defined

Simply stated, *statistics* is the science of collecting, organizing, analyzing, and interpreting data. We use the resulting *data* to *describe* and *infer* on the population.

### Descriptive and Inferential Statistics

*Data* are measurements and/or observations sometimes called values, information, or even facts. For the purpose of this class, the data are considered information called *variables*: a trait or characteristic that can assume different values (i.e. age, height, blood pressure, level of education, number of courses taken).

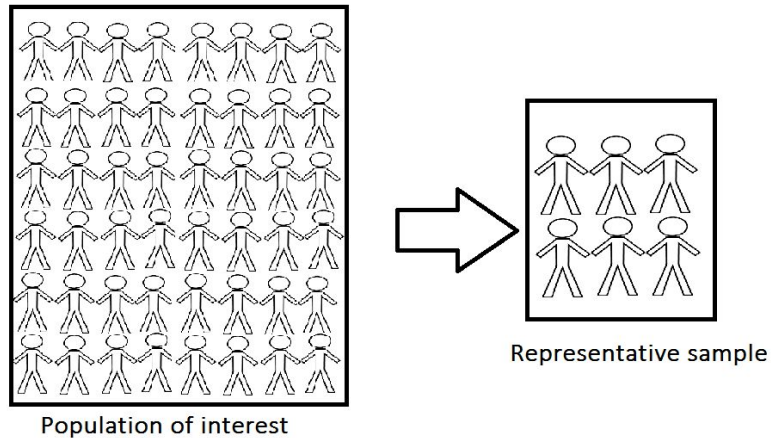
These data (remember, data are plural) can be used in two main areas: *descriptive statistics* and *inferential statistics*. Descriptive statistics do just as the name implies, they describe or summarize the data. This description is usually a way to present a large amount of data in a more precise manner and often happens numerically through tables, charts, and graphics. The key thing to know about descriptive data is that it cannot be used to draw conclusions about our population, but it is a good starting point to begin asking research questions and developing hypotheses. Conversely, through the use of *inferential statistics* you can make conclusions or generalizations about populations. This is where we would analyze our hypothesis.

### Samples and Population Data

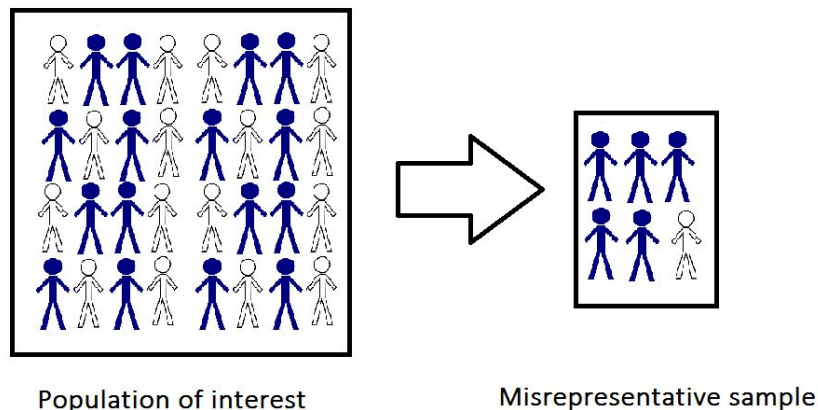
The goal of research is to find out certain characteristics of a *population*. A *population* consists of all subjects (human or non, but for health data and epidemiology, we will stick with human) that are being studied. The data collected, organized and used to describe a population are called *parameters*. However, when collecting data for a statistical study, often due to expense, time, location, size of the population, or other concerns, it is not practical if not impossible to gather information and for this reason, *samples* are selected. A sample is a subset of the population of interest. The data we collect, organize, and use to describe the sample are called *statistics*. It is important to note that there are many different ways to get a sample from your population: random sampling, systematic sampling, cluster sampling, and stratified sampling.

### Sampling

In reality, the only way to know the true measure or value of a population is to collect data (or observe) every subject within that specified population. Of course this might be possible if your goal was to collect the GPA of every student on campus and your campus was the final population to whom you are going to infer all of the data. When this full data collection happens, that is you have data from every subject in a population, this is called a *census*. Again, as mentioned earlier, for most populations, conducting a census is nearly impossible for financial, resource, or access reasons. This is why we sample. However, in order for the data we gather from our sample to be meaningful and give us the ability to make inferences about the population, we must seek a *representative sample* of the population. This means that the relevant characteristics of our sample subjects are generally the same as those of the population. If the subjects in our sample differ in a specific characteristic from the subjects of the population, *bias* may occur.



*Bias* is systematic in nature and can result in skewed - or distorted- data that then results in misleading data. Bias can come about in ways other than a misrepresentative population. Sometimes the data are collected in a manner that does not represent the population. For instance, only collecting data from people who responded to a newspaper ad or asking for volunteers to answer a survey - this is not random because there are certain characteristics unique to people who are willing to volunteer that might influence other ideas and/or behaviors . Other times a researcher lends bias if he or she has too strong of a personal stake in the research outcome and then may knowingly or unknowingly misrepresent the data results. There are also biases in the way data are presented. This will be covered in *Section 3: Summarizing and Describing the Data*. There are other ways bias may arise, but for now, we will move on to sampling methods.



### Sampling Methods

As stated previously, a good and meaningful statistical study must have a representative sample. Data may be collected through a variety of methods. In our field, this is generally a survey, an interview, or through the use of medical records. Each of these methods have drawbacks: low number or responses, question misunderstanding, response errors, or, in the case of medical records, difficult to access due to patient confidentiality laws. There are also issues with gathering samples that are unbiased- in other

words giving each person or *subject* in the population an equally likely chance of being selected. This idea of equal chance is also known as *random sampling*.

### Random Sampling

Random sampling is often the most preferred method to get a representative sample. It is often not very easy to start a random sampling so that the first pick is actually picked by chance (i.e. having everyone pick a number out of a hat then deciding to use only the subject who picked an even number). *Simple random sampling* holds the same idea but this time applied to a specific sample size. Often, calculators or other statistical programs have random number generators to help select the sample and, if the sample size is appropriate, allows for the use of statistical methods to analyze and accurately infer the results.

### Systematic Sampling

Another effective way to get a representative sample is through *systematic sampling*. This is when instead of chance for a starting point, a random starting point is selected (i.e. the tenth subject) and the selection continues by selecting from fixed intervals (every tenth subject thereafter). Systematic sampling is much easier to utilize than simple random sampling and allows for the researcher to add some set order, or system, to the selection process.

### Cluster Sampling

*Cluster sampling* begins with randomly selected *homogenous* groups and then collecting and using the data on all the subjects in that group. The term *homogenous* is used to describe subjects with the same, similar, or standardized characteristics. It is important to note that to utilize this method there is a higher risk of sampling error if the group-level information is not known or if the clusters are unequal in size. However, it is a very economical method (time, money, and resources) and often results in a larger sample size.

### Stratified Sample

Another method by which a population is divided into groups is called *stratified sampling*. This is done when a specific characteristic of the population is of importance to the study. One of the most common stratifications used for health outcomes is that of age. For instance, say the researcher is interested in a walking path used by persons aged 50 to 65. By breaking the population into specific age groups or *strata* then gathering a representative sample from that strata, the results can then be compared to different age strata. (Please note, this is only an example, strata can be other specific characteristics like geographic location, education level, political views, etc.)

---

Population → Parameter	Sample → Statistic
$\mu$ ("mu") = mean	$\bar{x}$ (x-bar) = mean
$\sigma$ (lowercase sigma) = standard deviation	s (lowercase s) = standard deviation
$\sigma^2$ = variation	$s^2$ = variation
n = total population	n = total population

---

## Section 2: Statistical Measures

### Data Types

Data variables come in two types: *qualitative data* and *quantitative data*. *Qualitative data* are data that can be used to characterize, categorize, and approximate (i.e. gender, eye color, zip code) but not measured on a standardized tool (i.e. a scale). The data are collected through observational methods such as surveys, interviews, photo and audio archiving, and other observational methods and then compiled into concepts or themes. The data are then used to describe and begin asking the question “why” - but not to make any claims. This is because observational data are almost always imprecise and often subjective to interpretation (for example - do we all have the same definition of the color sky blue?). To make claims, we must use *quantitative data*. These are data that have been generated through statistics and are measured or ranked (i.e. age, weight, time, cost, number of students). In short, the quantitative data are not only structured but more precise.

### Data Measures

In addition to determining if a variable is qualitative or quantitative, variables can be classified by their level of measurement. Understanding the types and difference between levels of measurement is very important because it determines what statistical techniques we can use to analyze our data. The four levels of measurement are *categorical*, *ordinal*, *interval*, and *ratio*.

*The categorical level of measurement* deals with qualitative data that can be put into categories that cannot be ranked or ordered. In fact, with categorical data (also called nominal data), we cannot use any type of mathematical operation – we can only name and separate the categories. For example, if we categorized based on eye color, we cannot rank those eye colors and say that green eyes come before blue eyes.

If we do want to rank or order our data – such as tallest to shortest or fastest to slowest, we use an *ordinal level of measurement*. Ordinal data have divisions like the categorical data but, in the case of ordinal data, we can now rank or order the data. It is important to know that there are no precise measurements or differences between the rankings. For example, we often see customer review of one to five stars. We can say five stars is higher than four stars and that four is higher than three, but we cannot measure that difference between stars in a meaningful way. We cannot say, in this example, that four stars are twice as good as two stars- we cannot assume precision to ordinal data.

As qualitative data, categorical and ordinal data do not have any set rules or guidelines for definition or measurement - as a reminder: what is blue (categorical) and how much blue is the bluest (ordinal)? This is why they are known as *nonparametric data* - no set parameters are used. And as a reminder, being qualitative and nonparametric means we can only describe what we have found and begin asking questions, even developing hypothesis.

*Interval* and *ratio* data are quantitative in nature and divided into two groups: *discrete variables* and *continuous variables*. Discrete variables are those variables with numbers that can be counted as whole numbers. For instance, the number of students in a classroom or the number of participants in a

seminar. Continuous variables can have an infinite number of values within a given range. A good example of this might be weight or height (not that anyone is an infinite number of inches tall). *Interval data* can be used to rank data but unlike ordinal data, in this case the differences between rankings are not only measurable and meaningful but precise. You might think of interval data as a number scale by which each position is a precise and equal distance from another. A good example of this is temperature. There is 1° F difference between each interval so that 60° F to 61° F is the same distance as 72° F and 73° F. One key feature of the interval data that will distinguish it from ratio data, is that there is no meaningful or true zero in the interval scale. Going back to our temperature example, a temperature of 0° F does not mean that there is an absence of temperature.

If we do have a true zero, then we use the *ratio level of measurement*. It has all the same characteristics of the interval scale but with a meaningful zero. A good example of this is height. We can measure height by feet and between each foot is 12 equally spaced inches but anything with a “0” height actually does indicate an absence of height. This true zero allows for a true ratio between values and may yield meaningful data between two different subjects in the population. For instance, if one subject weighs 120 pounds and another weighs 240 pounds then there is a 2 to 1 ratio (2:1). I can meaningfully say that the second person weighs twice as much as the first.

With the defined measurements of both interval and ratio data, we then call quantitative data parametric data. As you may guess, parametric data are those data with set parameters: an inch is always an inch, a foot always consists of 12 inches, etc. For this reason, parametric/quantitative data are regarded as both objective and reliable and therefore used to infer information on populations (analyze and address the hypothesis).

The table below is an example of data types and their properties.

Types of Data			
Level	Type of Data	Properties	Example
<b>Categorical</b>	Qualitative	Nonparametric	Gender, education level, eye color
<b>Ordinal</b>	Qualitative	Nonparametric	Rankings, Likert scales
<b>Interval</b>	Quantitative	Parametric	Temperature, year
<b>Ratio</b>	Quantitative	Parametric	Height, weight, blood pressure

## Section 3: Organizing and Displaying Data

As we learned in the introduction, data are everywhere. Sometimes data can become overwhelming and cluttered or difficult to interpret and find patterns when in raw form. We can help put some order to numbers by *organizing* our data. One way to organize simple data is with *frequency tables*. A basic frequency table is simply two columns of information: column one has the categories of data and column two has the frequency of each category. For example, what if the final grades for this course were as follows:

C, F, A, A, B, A, F, B, C, A, B, C, B, A, A, B, A, C, A, D

While all the grades are there, your mind might quickly focus on the F's or the C in the first space and skew your perception. However, when I put the data into letter categories in a frequency table, it makes the grade distribution easier to see.

Frequency Table for Final Grades	
Grade	Frequency
A	8
B	5
C	4
D	1
F	2

Before we go further, note that letter grades are often thought of as qualitative data because there is often no set parameter of what is an "A", what is a "B" were as I used the following chart, all the data are quantitative.

Frequency Table for Final Grades	
Grade	Frequency
99-90	8
89-80	5
79-70	4
69-60	1
59-50	2

Using a *frequency table* with quantitative data allows us to organize a wide spread of data (for instance, systolic blood pressures from 90 up to 222) into more manageable *class intervals*. Class intervals should be equal in *width* so that meaningful comparisons can be made between intervals. For example, look at the table above, the intervals are 99-90, 89-80, etc. Notice how there are 10 numbers in each interval (99,98,97,96,95,94,93,92,91,90) . The above example then has 5 intervals that are 10 numbers in width.

One challenge when first working with data for frequency tables it determining how many intervals to use as well as your *interval width*. For any data, the number of intervals can vary, but a good rule of thumb is no less than five and no more than twelve to fifteen (otherwise it might get overwhelming). To

determine the number and width of intervals you need, first you must find the data *range* (covered again in Section 4). Range is the difference between the highest and lowest values in the data set. In the systolic blood pressure numbers mentioned earlier, the highest value would be 222 and the lowest value would be 90 – a range of 132. If I divide 132 by 5, I would have class intervals of 26.4 or if I divide 132 by 15, I get a class interval of 8.8. Because I know that the categories of blood pressure have fairly small intervals, I might feel more comfortable using 15 intervals. Since 8.8 is not a whole number, and my systolic blood pressure readings are, I am going to stretch my class interval to 10. This means I will start my intervals at 90, but in this case end at 229 (it is okay that I do not have a value as high as 229 as long as I have a value in that interval).

Systolic Blood Pressure			
Class Interval	Frequency (count)	Relative Frequency	Cumulative Frequency
90-99	8	0.10	8
100-109	12	0.15	20
110-119	15	0.19	35
120-129	9	0.11	44
130-139	11	0.14	55
140-149	7	0.09	62
150-159	5	0.06	67
160-169	3	0.04	70
170-179	2	0.03	72
180-189	1	0.01	73
190-199	4	0.05	77
200-209	1	0.01	78
210-219	1	0.01	79
220-229	1	0.01	80
Total	80	1	80

Notice there are two other columns in my frequency table: *relative frequency* and *cumulative frequency*. Relative frequency of a category is simply the percentage (or fraction) of data that fell in that category. For example, in the 90-99 interval, I had a frequency of 8. When I divide 8 by 80 (the total number of subjects) I get .10 or 10%. I am saying that 8 out of 80 or 10% of the subjects had a systolic blood pressure between 90 and 99 mg/dl. As for the cumulative frequency – this is a “accumulation” of the data. In other words, the number of data in the interval plus all the categories before it. If I look at the cumulative frequency for the class interval 120-129 I have 44. This is the frequency for the categories 90-99, 100-109, 110-119, and 120-129 (8+12+15+9= 44). When I get to my final class interval, my cumulative frequency should add up to the total frequency count.

Another way the data are organized is through the use of a table or *spreadsheet* that contains multiple variables and/or responses of all the study subjects. Spreadsheets are electronic documents where our data are arranged in both columns and rows that can then be manipulated and used in statistical calculations. For this course we will use examples from Microsoft® Excel, however, there are many other electronic spreadsheets and statistical packages with spreadsheets that allow for similar calculations.



<b>Spreadsheet Example</b>								
Questions: See Key								
ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
1	1	2	0	72	20	3	0	1
2	2	3	0	67	30	0	1	1
3	1	1	1	66	19	3	0	2
4	2	2	0	75	20	0	0	2
5	1	2	0	70	23	1	0	1
6	1	1	0	64	18	4	0	2
7	2	3	1	66	21	1	0	1
8	1	3	1	65	21	3	0	1
9	1	2	1	61	20	0	0	2
10	1	2	1	64	20	0	0	1

KEY	
Question/Responses	
1.	Are you going into the medical field? (1= yes, 2= no, 3= undecided)
2.	How many years have you been in college?
3.	Are you male or female? (0= male, 1= female)
4.	What is your height (in inches)?
5.	What is your current age?
6.	What is your eye color (0= brown, 1= blue, 2= green, 3= hazel, 4 other)
7.	Have you smoked during the past week? (0=no 1=occasionally 2=daily)
8.	Do you have a pet? (1= yes, 2= no)

## Graphing Data

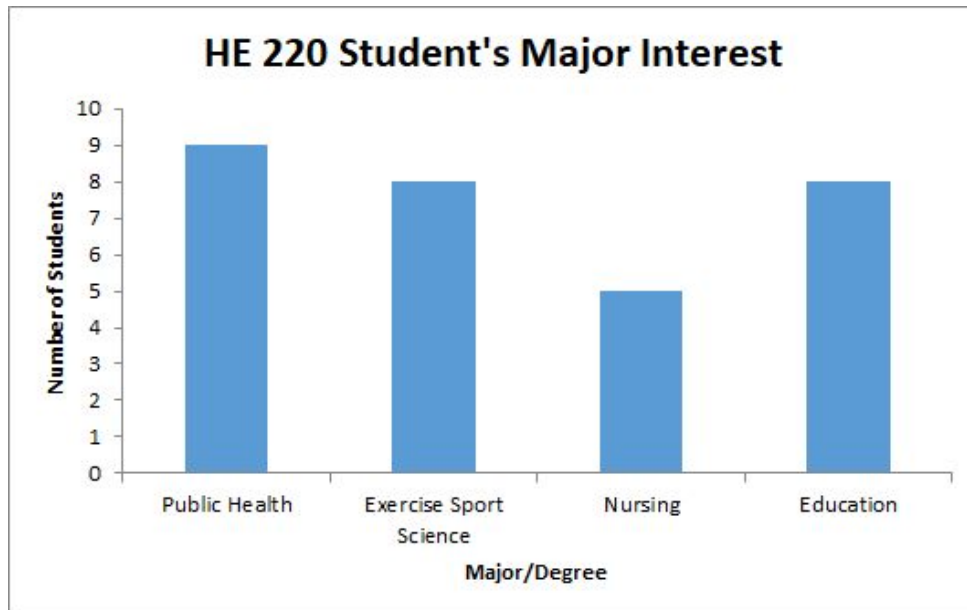
Once our data are organized, we can then begin using *visual graphics* to present our findings. When we use *visual graphics*, we appeal to our brains ability to take visuals and quickly find patterns such as groupings, gaps, and outliers. (Note: while graphics are easier to read, we are lacking some of the detailed information that you see in the table example).

It is very important that graphs are self-explanatory. This means that each graph has a descriptive title, the axes are labeled, and that the units of measure/observation are specified. However, graphs should still be simple – too much information can be confusing if not overwhelming.

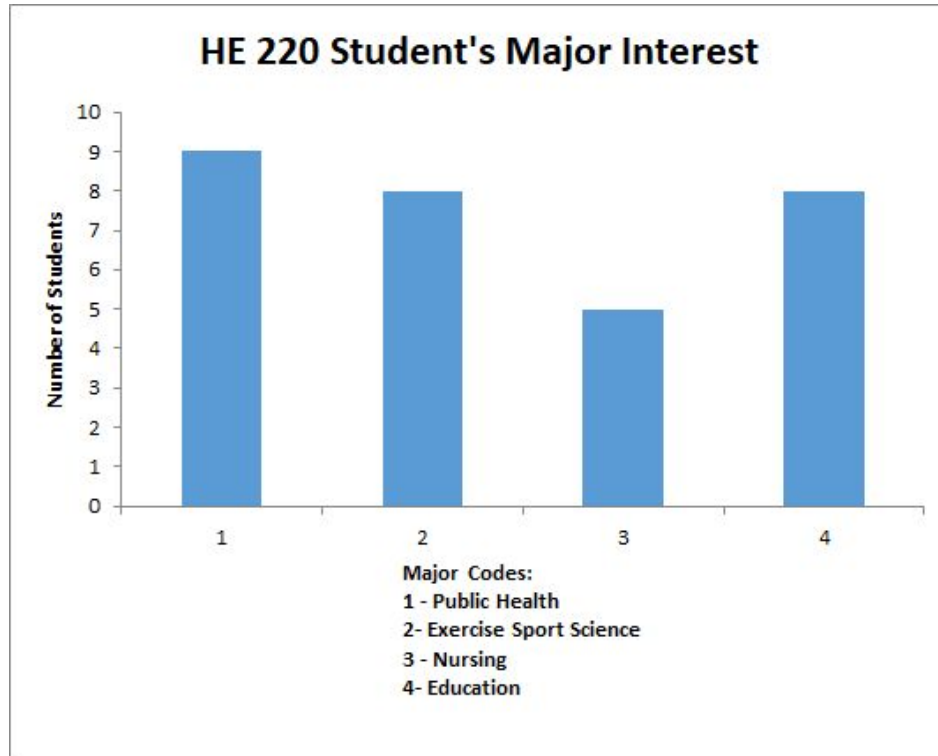
## Bar Charts

*Bar charts* are simple graphics that are used best when displaying categorical or ordinal data. Each bar stands vertically do not touch – this represents a single category for each bar. The widths of the bars are equal but the bar's height is what represents the frequency of the data- that is the taller the bar, the more variable values represented. Again, to make sure that the viewer knows the data are not continuous (interval or ratio) it is important to make sure the bars are equally spaced yet **do not touch**

one another. It is also important to the viewer that the vertical (up and down) axis begins at zero. However, if that is not possible or impractical, then broken bars should be used (please note, on some of your graphing programs, the setting to create the charts below are called "column charts").

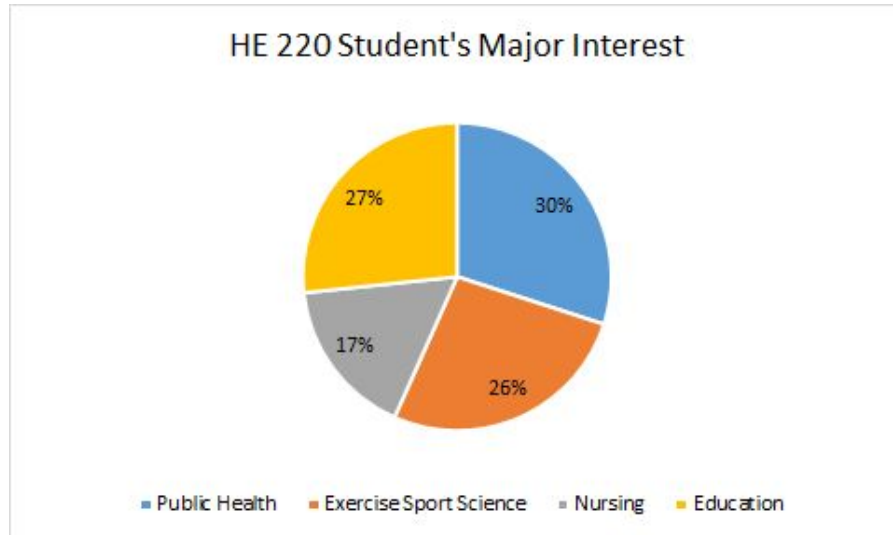


Notice how there are four categories of majors or degrees on the x-axis (horizontal) and that the bars do not touch. That is because these are named categories. However, even if I used a number code for each major, as you can see below, the bars still do not touch because they are separate categories.

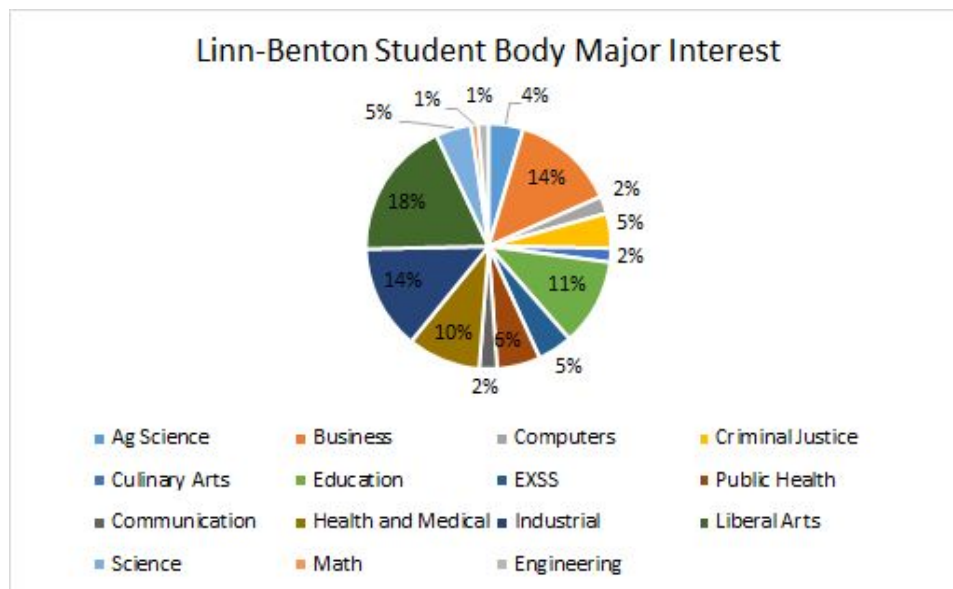


### Pie Charts

*Pie charts* are used to display categorical data where each "slice" or "wedge" of the pie represents a single category. Instead of height, like a bar chart, the width of the wedge corresponds with the frequency data for the specific category. Pie charts are only useful when there are a small number of categories. Otherwise they can become too crowded to determine the relative size of the categories. Notice in the chart below, the percentages are added on each wedge, this is not required however, if this were not the case, it might be hard to determine if there was a difference between Education (27%) and Exercise Sport Science (26%).

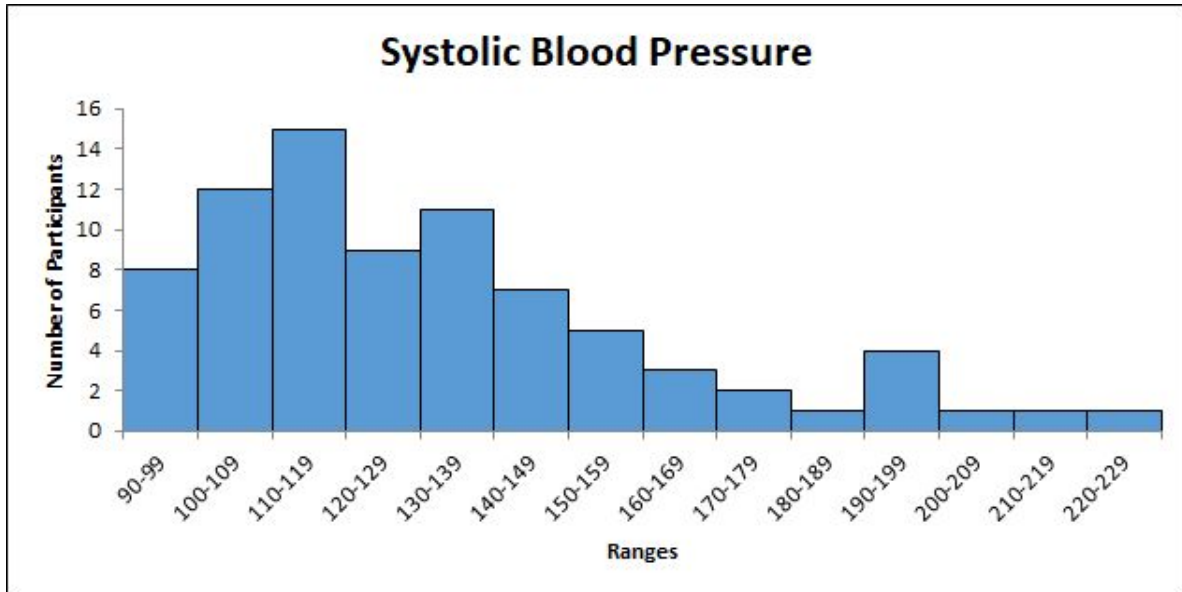


Now, if I were to ask all students at Linn-Benton Community College their area of interest, and even collapsed some of the majors as to not have so many categories, you can see where a pie chart is confusing and almost useless.

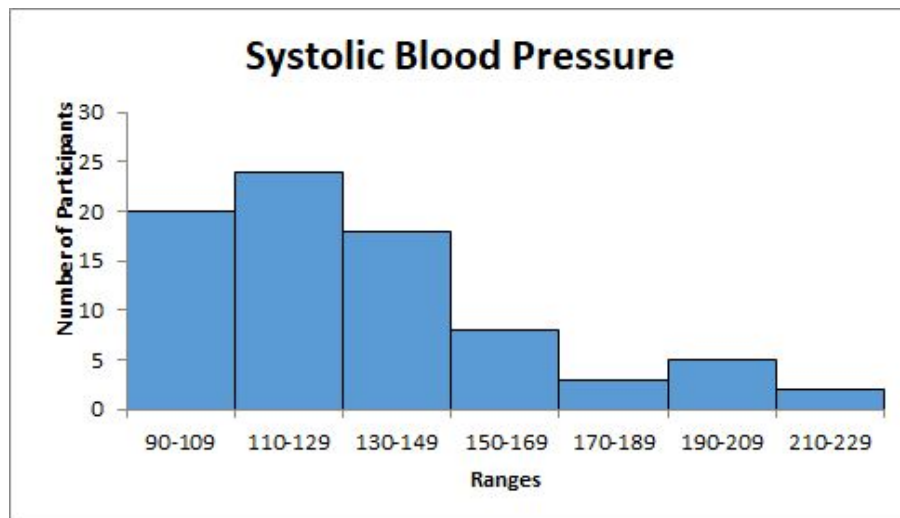


## Histograms

*Histograms* are likely the most common type of graph. It is simply a pictorial representation of the frequency tables presented earlier. In addition to frequency, it also shows us the distribution of the data. It is different than a bar graph in that **all the bars touch representing continuous data** (our quantitative data). Quite like the bar chart, the bars have equal widths but in the case of the histogram, these equal widths also represent equal interval categories. We created such intervals earlier with the systolic blood pressure data.

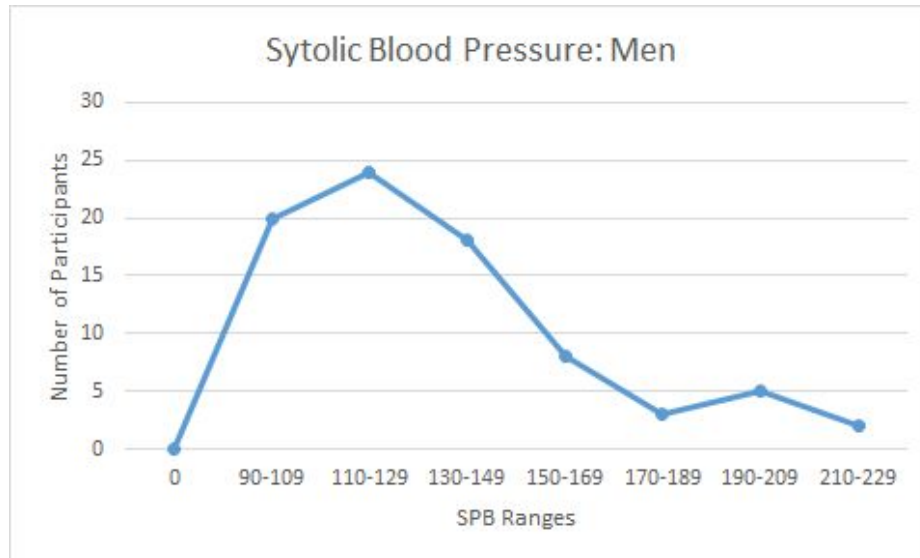


While the histogram represents the data from our frequency table, the fourteen intervals (with a 10 count width) could be reduced. In the example below, the same data are now presented in 7 intervals with a 20 count width. Both are correct and for this class, that will be your preference. But remember, the data must be meaningful, if I only used two intervals, I would not learn much about the systolic blood pressure of my participants.

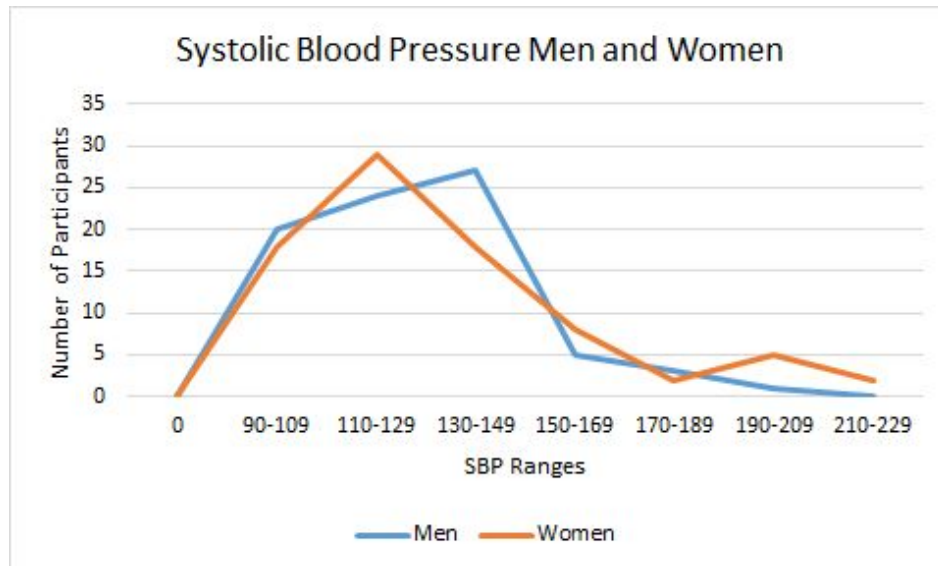


### Frequency Polygons

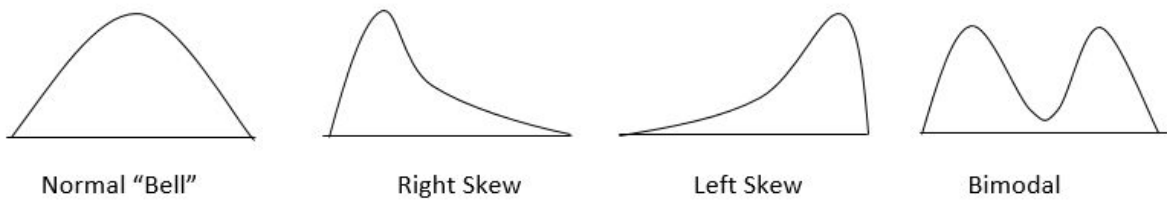
Following histograms, line graph – or *frequency polygon* – is also a widely used type of graph to show the distribution of quantitative data. Instead of using bars, the frequency polygon uses a series of **dots** (where the middle of the histogram bar would be) connected by lines. To “close” the data, you would want the line to touch the horizontal line to the left and to the right at the zero point.



As with histograms, *frequency polygons* should only be used for quantitative data – remember: data that are a continuous distribution. It might be that you want to compare two or more continuous data sets (with the same intervals) and in this case, frequency polygon is better than a histogram. As the example below shows, this allows for all the bars to be viewed (whereas if you stacked histograms, you may miss a lower bar).



Another key feature of the frequency polygon is the shape. The polygon may take on several different shapes and this gives us a quick “glance” at the data distribution. The examples below show a symmetrical or “bell-shaped” distribution,, a right skewed distribution, a left skewed distribution, and a bimodal distribution. (Further detail on each distribution will be presented in Section 4).



There are, of course, other ways to display data such as a stem-and-leaf display, a box-and-whisker plot, or an ogive. However, for the purpose of this class, we will focus on the frequency table and the graphics presented above.

## Section 4: Summarizing the Data

### Measures of Central Tendency

Previously we have learned how to organize and display the data. Now it is time to learn how to calculate and interpret descriptive statistics- namely *measures of central tendency* and *measures of variation*. The three most common measures of central tendency are the *mean*, *median*, and the *mode*. These three measures are used to help **describe** the entire data set, including the value at the middle of the data set, and the data distribution (or shape). The most common measures of variation - and those we use in this class are the range, variance, and standard deviation. Further discussion and the uses of measures of variation are found later in this section.

### Mean

The mean is the most commonly used measure of central tendency. Sometimes the term “average” is used interchangeably as they are solved using the same method. However, for this class we will continue to refer to the outcome as the arithmetic mean (or just mean). Before we present the formula to solve for the mean, first let’s differentiate between the two means we will be addressing in this course: population mean and sample mean (please refer back to Section 1 definitions of population and sample). The population mean is denoted with the Greek symbol  $\mu$  (mu) whereas the sample mean is denoted by the symbol  $\bar{x}$  (x-bar). The ability to distinguish between the two will become more important in future sections. In either case, the symbolic representation of the mean is represented as:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + \dots + x_n}{n}$$

In this example, we are using a sample mean ( $\bar{x}$ ) and the individual observations are represented by an  $x$  with the  $x_1$  being the first entry and the  $x_n$  being the last. The symbol  $n$  is the total number of observations. Notice how all the data points ( $n$ ) are included in the formula - this is a strength of the using the mean. Including all the data points means we have a more accurate and even sensitive representation of the data. However, there are two noticeable weaknesses: one is that we can only use the mean on quantitative data the other is that the mean is subject to skewing (distorting the bell shape) because it does include all data points. For example, if most all of your scores fell between 50 and 70 except for one outlier of 30, this moves your mean to the left (lower) and now the measure may not accurately reflect the raw data.

Putting this to practice, let’s reexamine our spreadsheet example from Section 2 and determine the average height (Q4) of our 10 participants using the formula above.

<b>Spreadsheet Example</b>			
Q4: Height			
ID	Q4	ID	
1	72	6	64
2	67	7	66
3	66	8	65
4	75	9	61
5	70	10	64



$$\bar{x} = \frac{72 + 67 + 66 + 75 + 70 + 64 + 66 + 65 + 61 + 64}{10} = \frac{670}{10} = 67$$

Our mean height was 67 inches. As previously noted, it is important to note that the mean can be affected by the values of the observations. For instance, let us say that our 75-inch observation was actually 50 inches, our mean would move down (or to the left) to 64.5 inches. This means values that appear very different (larger or smaller) or “outliers” then they can distort our mean and skew the data distribution.

$$\bar{x} = \frac{72 + 67 + 66 + 50 + 70 + 64 + 66 + 65 + 61 + 64}{10} = \frac{645}{10} = 64.5$$

### Median

The median is observed by sorting our observations from highest to lowest then taking the data point in the middle – the one that separates our data. For example, if our data set consisted of 60, 61, 62, 65, and 70 inches, the median would be 62. The median is a good measure of central tendency when you have outliers because it is only interested in the middle number. (Note: if we have an even number of data like our spreadsheet example below, then we need to take the average of the two numbers that are in the middle of our data) . However, it may not be a good representation of your data since only the middle number is considered. Follow along with the example below to find the median of the data set. You will notice that the numbers are the same but are now ranked from lowest to highest. Using this set, find the middle two numbers.

<i>Spreadsheet Example</i>	
Q4: Height	
Q4	
61	
64	
64	
65	
66	
66	
67	
70	
72	
75	

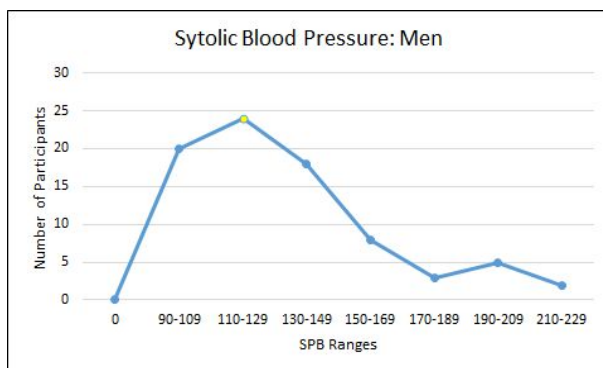
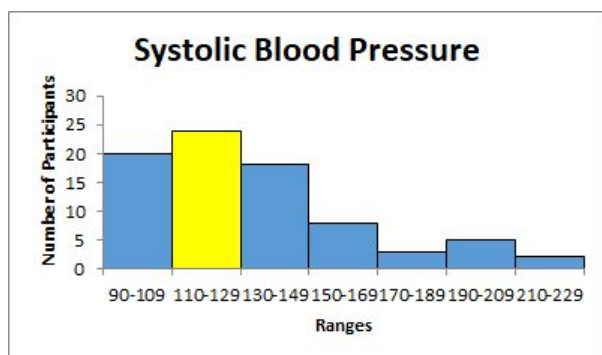
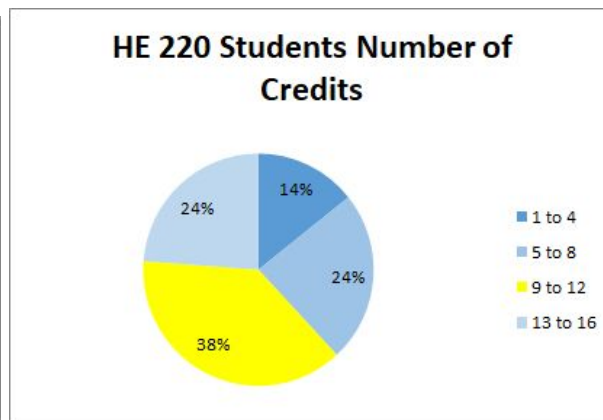
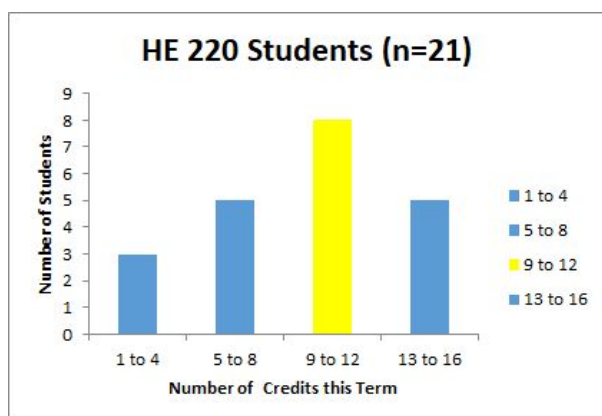
Did you notice there were two middle numbers and they are both 66? Mathematically speaking  $66+66 = 132 \div 2 = 66$  (but no need for math here). This means our median is 66. To show an example of different

numbers, had the numbers been 65 and 66, our median would then be  $65+66=132 \div 2 = 65.5$ . It is okay if our result is a number that is not in our original observation.

Again, what is important to note about median is that unlike a mean it is not influenced by outliers - it is very useful for observations that contain known outliers. One of the most common examples of median use is for house prices and/or personal income. Taking income as an example, if you included the CEO's salary down to an entry level salary then you would get a mean that is most likely distorted to the right (this means that the higher salary of the CEO pulled the mean up). However, looking at the median number might give you a better representation of the actual salaries.

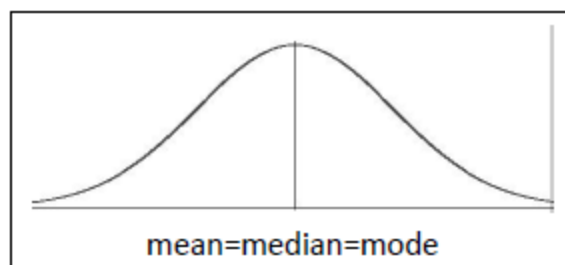
## Mode

The *mode* is the most frequently occurring observation in the data set. Because we can translate it both numerically and visually, it is used with both qualitative and quantitative data (categorical, ordinal, interval and ratio). Like median, it is unaffected by outliers and only takes the most frequently occurring number (s) into account. However, some data sets have more than one mode (look at the example above, both 64 and 66 occur twice) or perhaps have no mode at all. In a bar chart, histogram, or a frequency polygon, the mode (or modes) are the highest bar in our chart. In a pie chart, it is the largest "slice" of the pie.



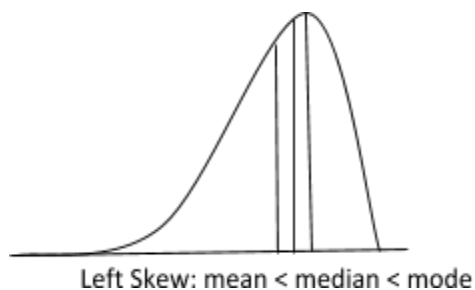
## Mean, Median, and Mode and Distribution Shape

In a symmetrical or "bell-shaped" distribution, the mean, median, and the mode occur at the same point in the data.



However, as mentioned earlier, data can be skewed by outliers and when this happens, the shape of the curve changes. While the mode is always the highest peak of the data, if we had a number that was an outlier to the left, or well below the mean, it would pull the curve to a negative or left skewed distribution by moving the mean lower than the median and mode (think back to the arithmetic mean description). Likewise, if there was an outlier to the right, or well above the mean, it would pull the curve to the right to a positive or right skewed distribution.

Note: Remember, this is a number line with the far left being the lowest number. Think of the left tail reaching way down to grab that low outlier to include it under the curve. Notice the mode still has the height as an outlier will not move the most frequently occurring number. In a right skewed distribution, there would be a mirror image of the one to the left with a tail reaching to the right and  $\text{mode} < \text{median} < \text{mean}$ .



### Measures of Variation

While it is important to know the central tendency measures from your data, it is also important to know if your observations are quite similar (little variation) or very different (large variation). There are three measures of variation we will examine in this class: *range*, *variance*, and *standard deviation*.

#### Range

The range of a set of data is simply the difference between the highest value (number) in the set and the lowest value in the set. While range is very easy to calculate it only takes into account two of the data values and tells us the spread or *width* of our data set. The data set below has already been ranked from lowest to highest and now it is simply taking the highest number ( $x_{\max} = 75$ ), and subtract from it the lowest number ( $x_{\min} = 61$ ) and you get  $75 - 61 = 14$  (or  $x_{\max} - x_{\min} = 14$ ).

<b>Spreadsheet Example</b>	
Q4: Height	
61	66
64	67
64	70
65	72
66	75

## Variance

Variance is used to determine how widely the values in our data vary. In other words, it identifies the spread of the data to determine how far our values vary from the mean. Variance is denoted as  $\sigma^2$  for a population and  $s^2$  for a sample (remember that  $\sigma$  is “sigma”, the Greek lowercase “s”). It is important to distinguish the difference as they are solved differently. Looking at the dialog box below. To determine the variance of our data, we must first determine the *mean deviation* of each value from our mean, then find the mean deviation squared, and finally the sum of squares. The steps to finding variance can be found in the textbox below.

<b>Spreadsheet Example for Height (Q4)</b>		
x (our data set value)	(x-x̄) (the mean deviation)	(x-x̄) <sup>2</sup> (the mean deviation squared)
72	5	25
67	0	0
66	-1	1
75	8	64
70	3	9
64	-3	9
66	-1	1
65	-2	4
61	-6	36
64	-3	9
<b>Sum of all Squares</b> $\Sigma (x-x̄)^2$		158

These are sample data and recall our mean of sample data is denoted by  $\bar{x}$ . Had this been population data, we would have used  $\sigma$ . When I solved the mean for these data earlier I found  $\bar{x} = 67$ . Now following the formula  $(x-\bar{x})$ , I will find the mean deviation for all 10 values. For example,  $x=72$ ,  $\bar{x} = 67$  would give me  $72-67=5$ . It is okay if you have a negative number, this simply means your value was below the average of 67.

Now, I need to square all the mean deviations using the formula  $(x-\bar{x})^2$ . Here, my numbers may range from 0 to  $+\infty$  (infinity) but I cannot have a negative number. Next, I will sum all of the squared values I found together giving me my “Sum of Squares” (SS).

To find variance for our population ( $\sigma^2$ ), I would take the SS and divide it by N (total number of variable).

$$\sigma^2 = \frac{\Sigma (x - \bar{x})^2}{N}$$

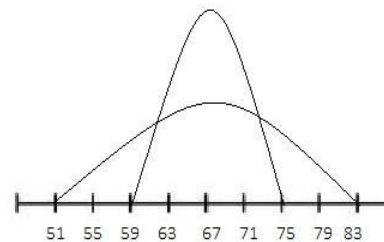
However, since the data to the left are sample data, I am solving for  $s^2$  and this means my denominator is now

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{(n - 1)} = \frac{158}{(10 - 1)} = 17.55$$

My variance for these sample data is 17.55.

## Standard Deviation

One of the most widely used measures of variation (and more useful measures in our class), *standard deviation* is also a measure of data spread but in this case we look more at how closely all the values are clustered around the mean of our data set. If we have a small standard deviation, the data are all bunched up around the mean (so we have little variation) then our curve peak will be tall. On the other hand, if our standard deviation is large, the data are spread out and our curve peak becomes shorter and wider.



As with other concepts in the class, standard deviation has different symbols for population data and for sample data. For population standard deviation we use the symbol  $\sigma$  (lowercase sigma) and for sample standard deviation we use  $s$ . We can find standard deviation by taking the square root of variance (remember that variance is denoted as  $\sigma^2$  and  $s^2$ ). So, looking at our variation example where  $s^2 = 17.55$ , then our standard deviation of this sample is standard deviation ( $s$ ) = the square root of variance ( $s^2$ ) then equals:  $s = \sqrt{s^2} = \sqrt{17.55}$ ,  $s = 4.19$ .

*NOTE: For now, the ability to solve for standard deviation is important. Once we get to Section 6, the normal distribution, we will examine the properties of standard deviation that make it so useful in health data analysis.*

When we are describing our data using the measures presented in this section, it is important to use a measure of central tendency that best represents our data. Along with the central tendency measure, a measure of variation is also important for descriptions. *For this class, you will find that our preferred measure of central tendency is the mean and measure of variation is the standard deviation.*

## Section 5: Probability and Counting Rules

In reality probability is often a hard concept to grasp because it has to do with future events. Probability is a measure of the likelihood or *chance* an event will happen. Think about weather forecast and your reaction when you hear on the news that there is a 20% chance of rain but instead it rains all day. What? That is because with predictions called forecasting, several factors are used in the formulation of that percentage. However, this is beyond the scope of this class (any of you going into the business or administration field will most likely encounter this type of probability in your upper-level courses). In this class we will look at basic probability and counting rules. In an effort to make the information flow in a more logical manner, we will examine different methods of probability in this section from the simplest to more complex. Whether simple or complex all probability methods have one thing in common: it is expressed in numbers ranging from 0 to 1. For reporting purposes this range of 0 to 1 can be converted to percentages ranging from 0% to 100%. What this means is that the likelihood of a future event happening is from never (0 or 0%) to absolute (1 or 100%).

Before we move to the probability methods, let us first look at a two terms used often in probability:

- An *experiment* is where we observe and measure an activity (or trials) for the purpose of collecting data.
- An *outcome* is a result of a single trial of an experiment.

### Classical Probability

One of the first types of probability studied (and perhaps the simplest) is *classical probability*. The simplicity comes because it does not rely upon an *experiment* but instead knowing our desired outcomes and total possible outcomes of any event. The use of classical probability means the assumption that all possible outcomes are equally likely to occur. For example, think of flipping a coin: a coin has two sides – a head and a tail - so we say that the likelihood of a head on one flip is 1 out of 2, or  $\frac{1}{2}$ , or 0.50, or 50%, and the same for the likelihood of a tail on one flip. As you can see, probabilities can be expressed as fractions, decimals, or even percentages. (For this class, you will be asked to present your findings as decimals or percentages).

Earlier you read that we need to know our desired outcomes or what we will call “favorable” outcomes and the total possible outcomes of any event. This translates to the formula:

$$P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

Looking again at the coin example, if I were to toss a coin and I wanted the outcome to be a tail, then my formula would be

$$P(\text{tail}) = \frac{1 \text{ tail}}{2 \text{ outcomes (1 head + 1 tail)}} = \frac{1}{2} = .50 = 50\%$$



## Basic Probability Rules

There are four basic probability rules that will help you decide if you are using probability tools correctly.

1.  $0 \leq P(E) \leq 1$ : The probability of any event ( $E$ ) ranges from 0 to 1 (a fraction or decimal).
2. If  $P(E) = 1$ , then it is certain that event ( $E$ ) will occur.
3. If  $P(E) = 0$ , then it is certain that event ( $E$ ) will not occur.
4. The sum of all sample spaces ( $S$ ) equal 1.



### Four Basic Rules Example with a Six-Sided Single Die

1. The probability of obtaining a 2 on a single role is  $1/6 = .1667$  ( $0 \leq .1667 \leq 1$ )
2. The probability of obtaining a number 1 through 6 on one role is  $6/6 = 1$
3. The probability of obtaining a 7 on a single role is  $0/6 = 0$
4. There are six sample spaces (numbers 1,2,3,4,5, and 6) =  $1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 6/6 = 1$

## Complementary Events

If we look at the example above, we see that the sample spaces ( $S$ ) for the outcomes of rolling a single die is 6. If I want my event  $E$  to be obtaining an even number on a single role, then my outcomes are 2,4, or 6. The *complementary event*, or  $\bar{E}$  ("not  $E$ ") would be obtaining an odd number of 1,3, or 5. The outcomes of an event and the outcomes of a complementary event add up to 1. The statements of *complementary events* can be seen below:

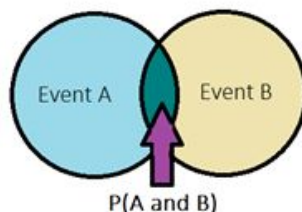
$$P(E) + P(\bar{E}) = 1 \quad \text{or} \quad P(\bar{E}) = 1 - P(E) \quad \text{or} \quad P(E) = 1 - P(\bar{E})$$

For this example,  $P(E) = 3$  even numbers/ 6 total outcomes or  $3/6 = 1/2 = .5 = 50\%$  and the complementary event of  $P(E) =$  "not an even number" =  $1 - .5 = .50$  or 50%. Right now, it may seem just as easy to logically say 50% because you know that there are six total outcomes, half are even numbers, the rest odd. The remembering that all outcomes are equal to 1 or 100% and the formula above can be helpful with large data sets to find the "not" ( $\bar{E}$ ) instead of the desired outcome ( $E$ ) and subtract from the total of 1 or 100%.

## Multiplication and Addition Rules

To examine questions about the probability of two or more things happening at once (called *compound events*), we must understand the *multiplication rule* and the *addition rule*. But first, let us review the terms "*independent*" and "*mutually exclusive*". Events are *independent* if the chance or probability of one has no effect on the chance or probability of another. For instance, just because I get a head on the first coin toss does not mean I am due a tail on the next coin toss. Each coin flip has equally likely chances ( 1 of 2 or 50%) of landing on a head. If events are *mutually exclusive*, then that means the events cannot occur at the same time: I cannot get both a head and a tail on a single coin toss. It is important to note that these two concepts are not the same.

With the *multiplication rule*, we can determine the probability of two or more independent events occurring. For instance, I can find the probability of tossing two coins and getting a head on both. Remember, these are independent because the outcome of the first coin does not affect the outcome of the second coin. To find the probability of the events occurring in sequence, I first find the probability of each and then multiply them together: P(head on coin one) and P(head on coin two) =  $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ . Notice the word “and” in the coin toss wording. For probability, I will assume that the word “and” means multiply (not intuitive, I know). Sometimes the word “and” is replaced with the symbol  $\cap$  - called an intersection.



**Multiplication Rule Formulas:  $P(A \text{ and } B) = P(A) \cdot P(B)$  or  $P(A \cap B) = P(A) \cdot P(B)$**

Within the multiplication rules we have *conditional probability*. Conditional probability is simply the probability of one event occurring “**given**” that another event has already occurred. For instance, maybe we want to see the probability that an HE 220 student passed the second quiz given they had passed the first quiz. The formula will be written as

$$P(A|B) = P\left(\frac{A \text{ and } B}{B}\right) \text{ or } P\left(\frac{A \cap B}{B}\right)$$

This formula is then interpreted as the probability of event A occurring given Event B has already occurred. The horizontal line between the A and B is the symbol for “given” in this formula. Please note that the event B in the numerator and denominator do not cancel one another out.

#### Multiplication Rule Example

Gender	Education Level		Totals
	High School	College	
Men	142	139	281
Women	154	165	319
Totals	296	304	600

Q: What is the probability that a person in this study is a man?

A: Use basic formula = 281 men/600 total = .468 or 46.8% are men.

Q: What is the probability that a person in this study is a man with a college education?

A:  $P(A \text{ and } B) = P(\text{man and college education}) = 139 \text{ men with college education} / 600 \text{ total} = .2317$  or 23.17%

Q: What is the probability that a person selected in this study is a man given they have a college education?

A:  $P(A|B) = P(\text{man given college education}) = 139 \text{ men with a college education} / \text{all } 304 \text{ participants with a college education} = 139/304 = .4572$  or 45.72%



The *addition rule* can also be used to determine the probability of two or more events. If the events are mutually exclusive, I might use the formula

$$P(A \text{ or } B) = P(A) + P(B)$$

Notice the word “or” in the equation. This is my key word for the use of the addition rule. For example, let us say we have 3 urgent care centers, 2 quick care centers, and 1 emergency room in a 15-mile radius. If a person is deciding which to visit for a deep cut, find the probability that it is either an urgent care center or an ER. Notice I have given you 6 total facilities (this is my “N”, my denominator) and I want to know if it is one of the 3 urgent care centers or the 1 emergency room,  $P(\text{urgent care or ER}) = P(\text{urgent care}) + P(\text{ER}) = 3/6 + 1/6 = 4/6$  (remember, for this class solve to a fraction or percentage = .667 or 66.7%). These events are mutually exclusive because in this example, a location cannot be an urgent care center and an ER at the same time.

Some events we choose to examine are not mutually exclusive. For instance, if within our urgent care center, we have 10 staff members made up of 3 physicians and 7 nurses: 1 physician and 6 of the nurses are female. This means that a staff member can be both a nurse and female and may lead to counting a probability twice giving us a  $P > 1$  (and remember our first rule of probability: it has to be between 0 and 1). Because the events in this example are not mutually exclusive, then we must use the addition rule formula

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

The end of this formula, “ $P(A \text{ and } B)$ ”, removes any overlap from our two events (look at the venn diagram on the previous page). Practice the following example on the selected four nursing characteristics.

To begin, and to make this easier, let’s put our data in a table:

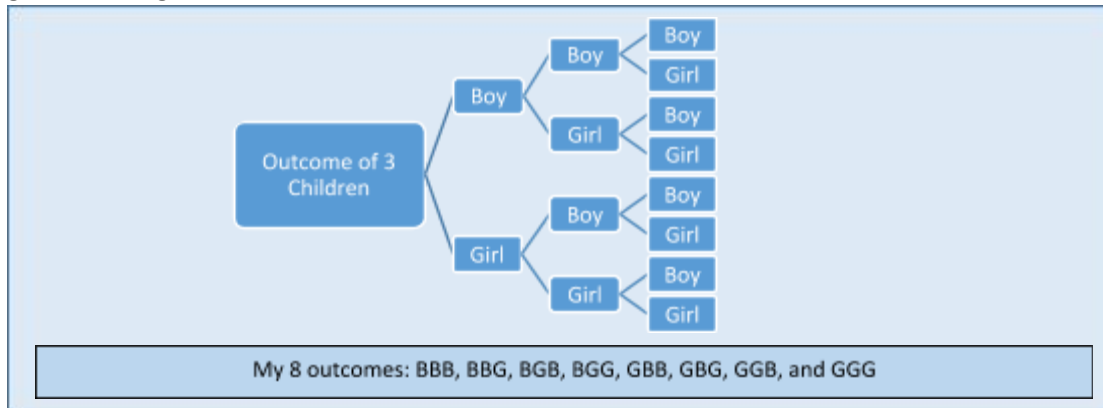
Staff	Female	Male	Total
Physician	1	2	3
Nurse	6	1	7
Total	7	3	10

Therefore, the  $P(\text{nurse or female}) = P(\text{nurse}) + P(\text{female}) - P(\text{nurse and female}) = 7/10 + 7/10 - 6/10 = 8/10$  (or 80%). Notice, if I had left off the overlap, I would have  $7/10 + 7/10 = 14/10$  and this gives a probability higher than 1 (1.4 or 140% - I cannot have more than everyone 100%).

### Counting Rules

The basic multiplication counting rule states that if one event can happen in  $n_1$  ways and a second event can occur in  $n_2$  ways then the number of ways of both occurring is  $n_1 n_2$ . For example, if you wanted to know in how many ways you could find the outcome of a toss of single die and the flip of a coin you know that there are six outcomes for the die (6) and two for the coin (2) then  $6 \times 2 = 12$ . There are twelve ways. Another example might be in how many ways could a couple has 3 children? Each child could be male or female (2 ways) so my answer is  $2 \times 2 \times 2 = 8$  (think of it as two outcomes “and” two

outcomes “and” two outcomes - remembering “and” means multiply). Another way to examine this is through a tree diagram.



However, tree diagrams can become very large and difficult to manipulate so remembering the multiplication rule is very helpful.

### Factorials

Denoted by  $n!$ , a factorial number is one solved by the multiplication of descending natural numbers. For example,  $4!$  is solved as  $4 \times 3 \times 2 \times 1 = 24$ . If I asked the question, “in how many ways can you line up 5 different cardio machines in your gym”, you would solve it as  $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$ . There are 120 different ways you could line up those machines. We will use factorials in *permutations* and *combinations* and it is important to note if you end up with a  $0!$ , that  $0! = 1$ .

### Permutations

*Permutations* are about the arrangement or ordering of specific objects. Here we examine the number of ways a selected group of objects ( $r$ ) from the total number of objects ( $n$ ) can be ordered and the order matters. For instance, let us say that a campus club was randomly selecting 4 students in our HE 200 class to receive a bookstore gift card. The card amounts were \$100, \$75, \$50, and \$25 with the first person drawn receiving the \$100, the second would get \$75, and so on. In this instance the **order matters**. If we had 18 students in the class and I am selecting 4, then  $n=18$  and  $r=4$  (think of it as 18 total  $n$  and I am pulling an “ $r$ ” sample of 4). To find the number of ways that the 4 can be selected, use the following:

$$P(n, r) = \frac{n!}{(n-r)!} = \frac{18!}{(18-4)!} = \frac{(18)(17)(16)\dots(1)}{(14)(13)(12)\dots(1)} = 73440$$

*There are 73,440 different ways we can select four people out of 18 when order matters.*

### Combinations

*Combinations* are similar to permutations except that **order does not matter**. Using our example above, if the 4 gift cards were all worth \$50, and all four selected from the 18 class members would get the same amount, order would not matter as long as you were one of the four selected. In this case you would use the following:

$$C(n, r) = \frac{n!}{r!(n-r)!} = \frac{18!}{4!(18-4)!} = \frac{(18)(17)(16)\dots(1)}{((4)(3)(2)(1)) \times ((14)(13)(12)\dots(1))} = 3060$$

There are 3060 different ways we can select any four people out of 18. Notice the new  $r!$  in the denominator – this takes away overlap and our resulting number is much lower.

## Sampling Probability

Two quick concepts to introduce before we get to the final topic in the are *sampling with* and *sampling without* replacement. They are quite simple: when sampling with replacement, the same number or subject may be selected several times and when sampling without replacement, once the number or subject has been selected, it may not be selected again.

### Sampling With and Without Replacement Example

Let us say you have purchased a raffle ticket at a local charity event. The drawing has several gifts available. If you held one of 60 tickets purchased, then when the first drawing occurred, you would have a  $1/60$  chance of winning. If you did not win and the first winning ticket was then removed, on the next drawing you now have a  $1/59$  chance of winning ( $60-1 = 59$ ) and so on - with each drawing you did not win, your chances increase because the denominator decreases. Let's change the scenario - what if it was a multiple draw ticket? In other words, after each drawing all 60 tickets went back into the pool. This means that every drawing, win or lose, you have a  $1/60$  chance of winning (or winning again).

## Binomial Distributions

When we see the prefix "bi" it means "two". Binomial probability is used when there are two outcomes to the probability distribution. Here, the outcomes are known as "success" and "failures". (Don't be confused, it does not mean one is right and the other wrong). For instance, if you wanted to know the probability of rolling a fair die and getting a 2 - rolling a 2 would be a success and rolling any of the other five numbers would be the failure.

To determine if binomial probability is an applicable tool, a few requirements are needed:

1. There is a fixed number of trials or observations (i.e. I will roll the die 20 times).
2. There are only two outcomes- success or failure
3. The outcomes are independent from one another (i.e. just because I flip a coin and get a head on the first trial does not mean I am "owed" a tail on the next flip).
4. The probability of success is consistent from trial to trial.

There are also set notations commonly used with binomial distributions (do note if you look up outside information, there may be different symbols utilized):

$n$  = number of trials

$x$  = total number of "successes"

$P$  = probability of a success on an individual trial

$Q$  = probability of a failure on an individual trial (also known as  $1-P$ )

This leads us to the formula  $\frac{n!}{x!(n-x)!}P^xQ^{n-x}$

### Binomial Distribution Example

Let us say a survey on campus found that one out of every four HE 220 students used the tutoring center during the term. If we randomly select 10 students at the end of our class, find the probability that exactly 3 have used the tutoring center that term.

$$n = 10, x = 3, P = \frac{1}{4}, Q = \frac{3}{4}$$

$$P(3) = \frac{10!}{3!(10-3)!} \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^{10-3} = .2503$$

So, I can say the probability of finding exactly 3 (among the 10) that used the tutoring center is approximately 25% (25.03%).

I can expand the question and ask, what is the probability that no more than 3 used the tutoring center. When I do this, I know repeat the formula above using again  $x=3$ , but also running the formula on  $x=0$ ,  $x=1$ , and  $x=2$  and adding all of the results together (go ahead and give it a try, you should end up with .776 or 77.6%).

Now, let us make life a little easier, remember the requirement "The probability of success is consistent from trial to trial"? This consistency means there is a table that we can use in place of the formula. You can use almost any online binomial distribution table but for this class (and class consistency) it is preferred you use the binomial table in your Moodle course or the online binomial calculator (a user-friendly version is found on your Moodle section for this topic). A sample of the table is explained below.

**Binomial Table Explained**

Probability		0.01	0.05	0.1	0.2	0.25	0.3
n	x						
10	0	0.904	0.599	0.349	0.107	0.056	0.028
	1	0.091	0.315	0.387	0.268	0.188	0.121
	2	0.004	0.075	0.194	0.302	0.282	0.233
	3	0.000	0.010	0.057	0.201	0.250	0.267
	4	0.000	0.001	0.011	0.088	0.146	0.200
	5	0.000	0.000	0.001	0.026	0.058	0.103
	6	0.000	0.000	0.000	0.006	0.016	0.037
	7	0.000	0.000	0.000	0.001	0.003	0.009
	8	0.000	0.000	0.000	0.000	0.000	0.001
	9	0.000	0.000	0.000	0.000	0.000	0.000
	10	0.000	0.000	0.000	0.000	0.000	0.0000

Look! Our example data of  $n = 10$ ,  $x = 3$ ,  $P = \frac{1}{4}$ ,  $Q = \frac{3}{4}$ , we find .250 on the chart.

This is by no means all of the probability we can use to describe our data, however, for the purpose of this class, this concludes our probability segment.

## Section 6: Normal Distribution

Up to this point, we have examined methods that are used to describe the data. Now we will begin to look at ways to draw inferences from the data beginning with *normal distribution*. The normal distribution is actually a group of *continuous probability distributions* described by the normal distribution equation:  $Y = \{ 1/[ \sigma * \text{sqrt}(2 \pi) ] \} * e^{-(x - \mu)^2/2\sigma^2}$ . (Do not worry... we are not going to manipulate that formula by hand).

Normal distribution is important in our fields of study because it gives us parameters of a normal interval (think blood pressure, cholesterol level, IQ scores) by answering the question “is this normal” and the probability we will have a value that is within the normal limits. Normal distribution is often used because limitless phenomena approximate or closely approximate the normal distribution. The mathematical properties of the normal distribution are also easy to manipulate, making it a popular choice as a statistical tool. No matter the reason for use, it is important to note that this tool is the basis for our use of inferential statistics.

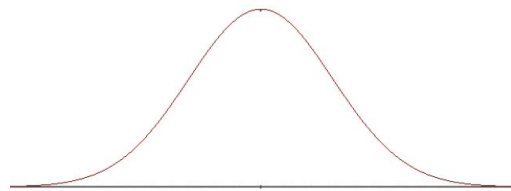
### Properties of the Normal Distribution

There are several properties of the normal distribution that are important to note:

1. It is unimodal and has a bell-shaped curve
2. It is symmetric
3. The mean, median, and mode are all directly in the middle
4. The total area under the curve is worth 1.0 or 100%
5. It is determined by its mean and standard deviation
6. There is a specific distribution for the area under the curve

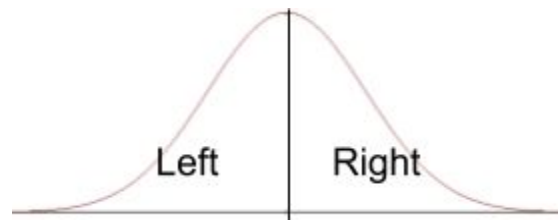
#### *Unimodal and Bell-Shaped Curves*

If you look back to the end of Section 3 you will see the shapes of data curves. Unimodal means one “hump” (so one mode) and bell-shaped means that along with that one “hump” there is a tail to each side.



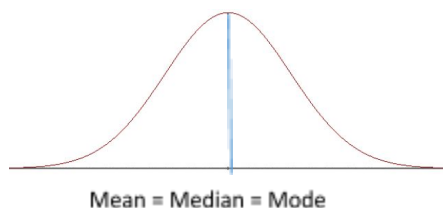
#### *Symmetric*

A symmetric curve means that the left side of the curve mirrors the right half of the curve.



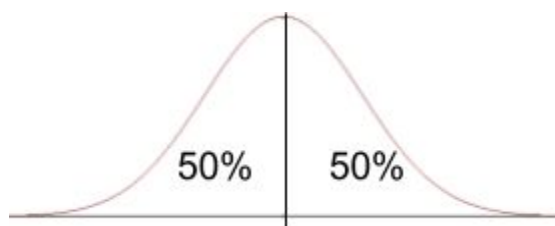
### Mean, Median, and Mode

Think back to our discussion on measures of central tendency. If the mean, median and mode are the same, and the mode determines the height, you can see where this would give us a bell-shaped and symmetrical curve.



### Curve Value

The normal curve surrounds all the data points from your data set. This means that it holds 100% of the data. Statistically speaking 100% = 1.00. Since the normal curve is symmetric and the mean=median=mode, then 50% (.5) of the data are on the left side and 50% (.5) are on the right side.



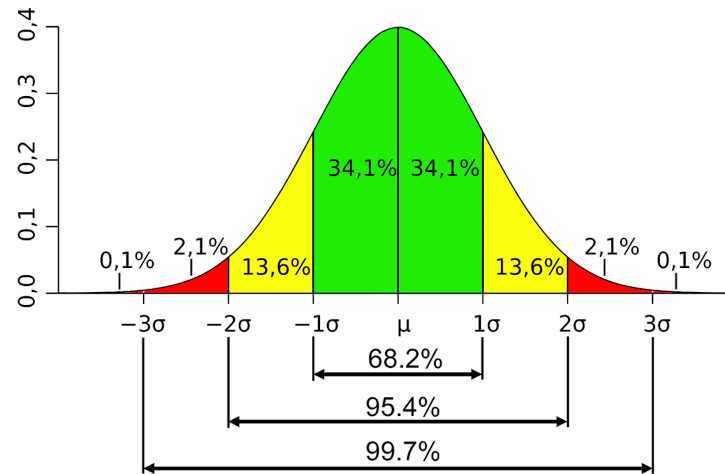
### Mean and Standard Deviation

We learned about both *mean* and *standard deviation* in Section 4. Since we are **standardizing** the curve we will say we have a mean of zero and a standard deviation of 1 unit. (No, do not go change every mean to 0). This statement,  $\mu = 0$  and  $\sigma = 1$ , allows us to use a standardized table to determine areas under the curve (not just the 50/50). You will find a copy of a Normal Distribution Table on Moodle or you may find many online. For this class, all examples will be given using the table from Moodle.

### The Distribution Under the Curve

Remember from Section 4 that a *standard deviation* is a measurement of the spread of the numbers. To take this further, there are some "standard" rules that apply in a normal distribution from the *empirical rule*. They are:

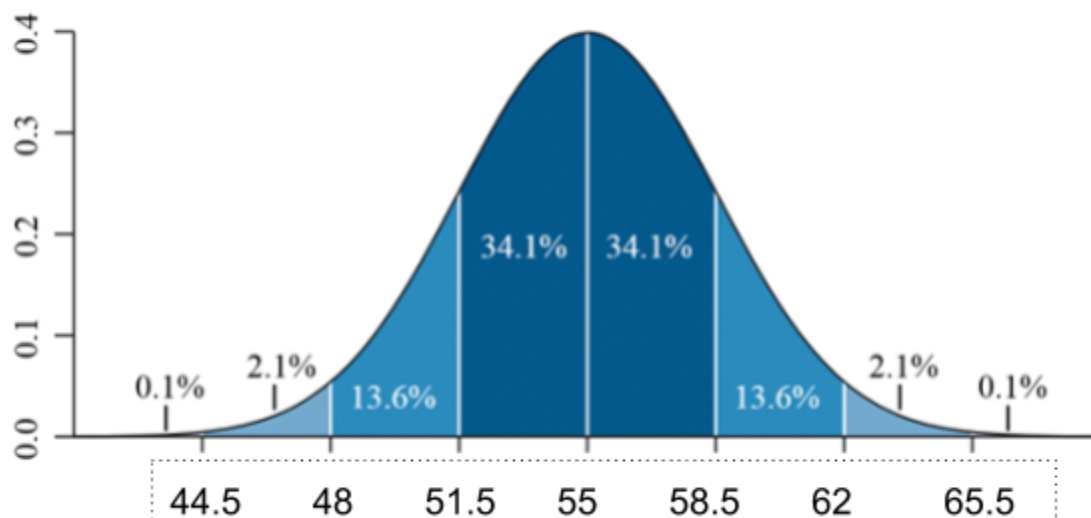
- About 68% (68.26%) of the data fall between  $\pm 1$  standard deviation
- About 95% (95.45%) of the data fall between  $\pm 2$  standard deviations
- About 99% (99.74%) of the data fall between  $\pm 3$  standard deviations



So how might this translate with real numbers/values? Follow along with the example below. (Please note that there are more examples in your Moodle help files as well as the practice packet for this class).

#### Normal Distribution and Standard Deviation Example

Let us say we had a mean ( $\mu$ ) of 55 and a standard deviation ( $\sigma$ ) of 3.5. This means that the data range for  $\pm 1$  standard deviation is  $55 \pm 3.5 = (51.5, 58.5)$ . For  $\pm 2$  standard deviations we now have  $55 \pm (2 \times 3.5) = 55 \pm 7 = (48, 62)$ . Finally, for  $\pm 3$  standard deviations we have  $55 \pm (3 \times 3.5) = 55 \pm 10.5 = (44.5, 65.5)$ . We can say that about 99% of data in this example fall between 44.5 and 65.5. Notice I am not mentioning about .01 or 1% of the data – it lies outside in the tails and is less than 44.5 on the left and greater than 65.5 on the right.





## Area Under the Normal Curve

The area under the normal curve is not a single curve but an infinite number of curves. Think of these infinite numbers of curves as a family and the family all have the same properties as listed earlier in this section. Since measuring every curve in the “family” is impossible, the area under a normal curve has been standardized. Remember above we had the statement  $\mu = 0$  and  $\sigma = 1$  where by the mean and standard deviation were transformed. This transformation allows us to use a standardized score or a **Z-score**.

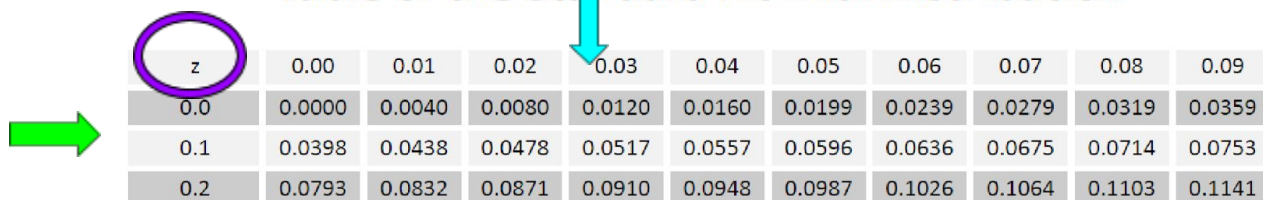
$$z = \frac{(x - \mu)}{\sigma} \text{ where } x = \text{the variable, } \mu = \text{the mean, and } \sigma = \text{the standard deviation}$$

With the use of the z-score and the Normal Distribution Table, we can now begin to find areas under the curve. Here is an example, let us say you were interested in what proportion of HE 220 students should score between the average of 185 and the score of 208 on the final exam (it is worth 245 points). From examining past data of the HE 220 final exam scores you find  $\mu = 185$  and  $\sigma = 10.33$ . My variable here is the score of the 220 exam, so  $x = 208$ . With this information, I can now work through the formula:

$$Z = \frac{(x - \mu)}{\sigma} = \frac{(208 - 185)}{10.33} = \frac{23}{10.33} = 2.23$$

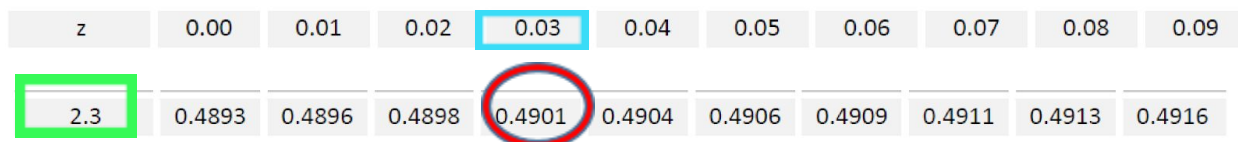
I find my  $Z = 2.23$ . But what does this mean? It means in these data my standard score is 2.23 but that still does not tell me much. Now you will need to take this z-score and look for it on your table to solve for the area under the curve. But first, let us examine how to read the chart. You start on the **vertical** for the first two digits, then the **horizontal** for the third digit.

### Table of the Standard Normal Distribution



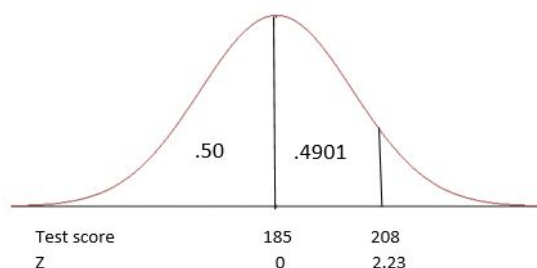
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141

When you look at your chart for  $Z = 2.23$ , you find the number **.4901**.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916

This tells me that the proportion scoring between 185 and 208 is about .4901 or 49.01% of the class. This might seem simple but determining areas can become confusing and for that reason is it best to draw out your curve and the areas of interest.



By looking at this graphic, I can see the area between 185 and 208 is .4901 and the area below the mean of 185 is .50 (Remember, half of my curve is worth .50 or 50% to the left and .50 or 50% to the right). Now look at the “z” below the test score, this is showing how my z= 2.23 falls on the number line. Notice the “0” at the mean, this is because if I deviated 0, then I am no different from the mean). Using simple math I can take the right half of my curve and subtract from it the area between the mean and 208 and find the proportion scoring above 208 points using this formula:  $.50 - .4901 = .0099$  or .99%. When I add the three areas together, they should equal 1.0 or 100% ( $.50 + .4901 + .0099 = 1$ ). Now I have accounted for the entire area under the curve.

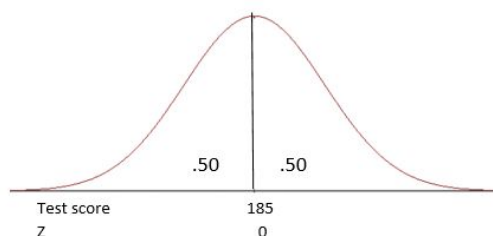
***Pull out your table and calculators and practice along with the following examples.***

#### **Normal Curve Examples:**

Keeping our data from above ( $\mu = 185$  and  $\sigma = 10.33$ ) the following examples show you have to work with the curve.

*Example 1:* What proportion of HE 220 students are expected to score 185 or above? ( $p > 185$ )

Here I can do two things, I could work the formula using the Z score, but I know that  $185 - 185 = 0$  so that means there is no deviation from the mean. I also know that the area to the right of the mean is .50 or 50%. So my answer is .50 or 50%.



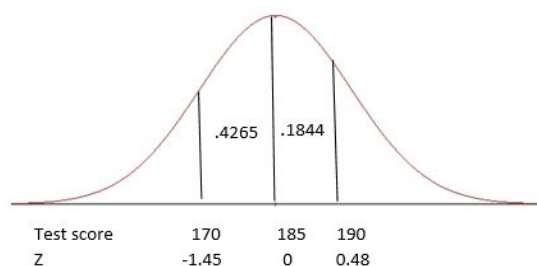
*Example 2:* What proportion of HE 220 students are expected to score between 170 and 190? ( $170 < p < 190$ ).

The corresponding Z scores are as follows:

$$Z = \frac{(x-\mu)}{\sigma} = \frac{(170-185)}{10.33} = \frac{-15}{10.33} = -1.45 \quad (\text{now to the table and you will find } .4265)$$

$$Z = \frac{(x-\mu)}{\sigma} = \frac{(190-185)}{10.33} = \frac{5}{10.33} = 0.48$$

Taking these Z scores, go to your table. Note that the negative does not matter, it simply denotes that that x value is to the left, or below, the mean.

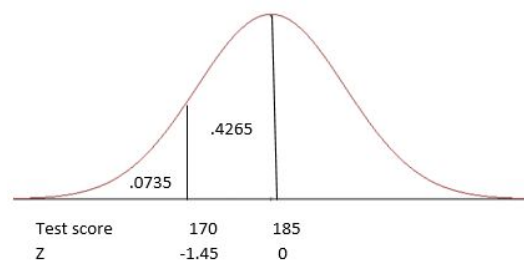


Now I am going to add the two areas:  $.4265 + .1844 = .6109$ . I can say that I expect approximately 61.09% of students to score between 170 and 190 on the final exam.

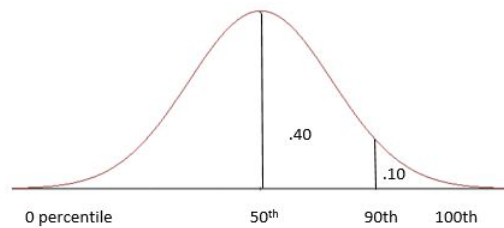
*Example 3:* What proportion of HE 220 students are expected to score below 170? ( $p < 170$ )

$$Z = \frac{(x-\mu)}{\sigma} = \frac{(170-185)}{10.33} = \frac{-15}{10.33} = -1.45 \quad (\text{again, from the table the area for a } z=-1.45 \text{ is } .4265)$$

I see my area between 170 and 185 is .4265 but the question is what is below 170. Since I always solve from the mean this means I need to do a little math. I know that half of the curve, and in this instance the area below our mean of 185, is worth .50 or 50%, I can then subtract the area I found on the chart ( $.5 - .4265$ ) and I find the area below 170 is .0735. I can now say that I can expect approximately 7.35% to score below 170 on the final exam.



*Example 4:* What if I decide to give the top 10% an “A” on the final. Now I need to find out what final exam score marks the top 10%. To do this, I need to remember that my curve is worth 100% and that 0% falls to the far left and 100% to the far right (of course with the 50<sup>th</sup> percentile falling in the middle).



In finding the top 10%, I know that  $100\% - 10\% = 90\%$  ( $1 - .1 = .9$ ). Remember we always solve from the mean so the area between the 90<sup>th</sup> percentile and the 50<sup>th</sup> percentile is 40% ( $.9 - .5 = .4$ ). We will still use our table, but this time we are going to look in the middle of the table for the number closest to .40. Once we find it, we will look to the left for the first two digits of the z-score and then up for the third digit. Please note – it doesn’t matter if it is higher or lower, just closest.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177

The number I find is .3997. When I look left I get 1.2 and when I look up I get 0.08. This tells me the Z-score for the 90<sup>th</sup> percentile is 1.28 ( $Z=1.28$ ). (NOTE: had I been looking for the lower 10%, my area between 10% and 50% is still 40% so my table would still give me  $Z=1.28$  but since 10% is lower than, or to the left of 50%, my Z is actual -1.28).

Recall the formula for Z is  $Z = \frac{(x-\mu)}{\sigma}$ , and you were given the  $\mu$  and the  $\sigma$  but not the Z or the x. We have found Z on the table and now we must solve for x. You can cross multiply to find the x or you might find this formula easier:  $x = (Z \times \sigma) + \mu$

Now let's fill it in:  $x = (1.28 \times 10.33) + 185 = 198.22$ . This means that anyone who scores a 198.22 or above on the final be in the top 10% and get an "A" on the final exam.

### **Final Thought**

As you can see, understanding the properties of the normal distribution and how to compute the areas under the curve, it is easier to compute expected proportions (probabilities) and begin to make inferences about the expected outcomes.

## Section 7: Comparing Means

In Section 6 we examined the area under the normal curve and learned how to determine if a specific variable was within “normal” limits of a population mean ( $\mu$ ). To do this we transformed the data using the z-score formula with the numerator being  $(x - \mu)$  and the denominator was the standard deviation,  $\sigma$ . Now we are going to move forward to compare a sample mean,  $\bar{x}$ , to the population mean,  $\mu$ .

As we have discussed in earlier sections, when we are looking to make inference on the population, it is quite often prohibitive to gather data on entire populations and that a representative sample would be most appropriate. Using the method presented in this section, we can now look at all the sample data and determine the sample mean giving us a distribution of sample means.

Before we move any further, it is important to introduce the *central limit theorem*. The central limit theorem states that if you have an adequate number of randomly selected, independent samples (generally with an  $n$  of 25 or greater) then the means of those samples will follow a normal distribution. This is true whether or not the population data is normally distributed. In addition, the mean of the distribution sample (this is denoted as  $\mu_{\bar{x}}$ ) is equal to the mean of the population ( $\mu$ ). Finally, the central limit theorem tells us that the standard deviation of the distribution sample ( $\sigma_{\bar{x}}$ ) is not equal to the population standard deviation ( $\sigma$ ) but is equal to the population standard deviation divided by the square root of the sample size:  $\sigma^2 = \frac{\sigma^2}{n}$ . The measure of variation presented by this formula  $\sigma^2 = \frac{\sigma^2}{n}$  is called the *standard error of the mean (SEM)*.

---

**Standard Error of the Mean (SEM): Population or Sample**

$$SE_{(\mu)} = \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad SE_{(\bar{x})} = \frac{s}{\sqrt{n}}$$


---

The standard error of the mean is like standard deviation in that it is a measure of variation. However, where we used the standard deviation  $\sigma$  for individual observations in Section 6, we will now use SEM, as our measure of variation for sample means.

---

**Comparison of Means Z-Score**

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} \quad \text{Where } \bar{x} = \text{the variable, } \mu = \text{the mean, } \sigma = \text{the standard deviation, and } n = \text{the sample size}$$


---

### Comparing Means Examples:

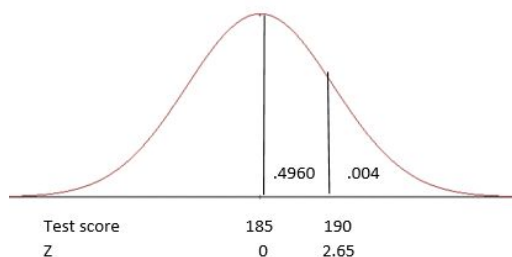
We will keep our data from Section 6 that  $\mu = 185$  and  $\sigma = 10.33$ . We will also continue to use the Table of Standard Normal Distribution as with Section 6.

*Example 1:* If final exam score of HE 220 is normally distributed with a mean of 185 points and  $\sigma=10.33$ , what is the probability that a sample size of  $n=30$  will provide a mean greater than 190?

The corresponding Z score is as follows:

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{(190 - 185)}{\frac{10.33}{\sqrt{30}}} = \frac{5}{9} = 2.65 \text{ (now, go to your table and you will find } z=2.65 = .4960)$$

Remember, this .4960 is the area from the population mean of 185 and the sample mean of 190. To finish answering the question, you need to find the “above” section:  $.5 - .4960 = .004$ . My answer is then the probability of finding a sample mean greater than 190 is .004 or 0.4%.



### Student's t Distribution

Many times the value of  $\sigma$ , the population standard deviation, is unknown. If we look at our formulas from Sections 6 and 7, we cannot solve the Z-score without it. When  $\sigma$  is unknown (or estimated) we then use sample standard deviation,  $s$ .

---


$$t = \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}} \text{ Where } \bar{x} = \text{the variable, } \mu = \text{the mean, } s = \text{the standard deviation, and } n = \text{the sample size}$$


---

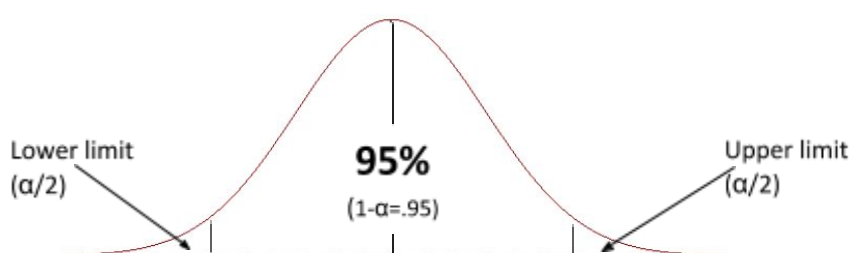
If you look at the t-distribution, it appears quite like the standard normal distribution and it has the same properties of being bell-shaped, unimodal, and symmetrical with a total value of 1.0 under the curve. However, for a t-distribution, the distribution is actually a group of curves that are based on the concept (a function of quantity and useful data) called *degrees of freedom*. In t-distribution, the variance is greater than 1 and as the sample size  $n$  increases, the distribution gets closer to the standard normal.

Degree of  
Freedom

$df = n-1$

## Section 8: Confidence Intervals and Sample Size

In Section 6 we moved into inferential statistics. One property of inferential statistics is the concept of *estimation*. This idea is that when we sample from a population, we will need to use the sample statistics to estimate our population parameters. While there are two ways to do this, a point estimate and a confidence interval, for the purpose of this class we will focus on *confidence intervals* (CI). A CI is an interval estimate that shows uncertainty associated with a sample statistic. In other words, if I sampled a population many times to get an interval estimate, some of our CI estimates would include the true or precise *population parameter* and some would not. For instance, if I set a 95% CI, then I would expect that among those samples 95% would contain the true population parameter. For our field of study, the most common CIs are 90%, 95%, and 99%.



### Confidence Intervals for the Mean – $\sigma$ Known

When  $\sigma$  is known, we can use the Z-score to help us estimate our CI. Remembering that when we first learned about Z-scores we learned that they were “transformed” so that they could be standardized. This means that when we use the CI formula for a known  $\sigma$ , we will use a standard number within our formula.

#### Confidence Interval CI Formula for Z

$$CI = \bar{x} \pm z \left( \frac{\sigma}{\sqrt{n}} \right)$$

In this case, the Z is dependent on the confidence level you select ( $1 - \alpha$ ). The Greek letter  $\alpha$  represents the total area of the two tails (see the upper and lower limits above). So, if I want to have an  $\alpha$  of .05 ( $\alpha = .05$ ), meaning that I want the probability of an incorrect analysis to be no more than 5%, then I take the full curve amount of 1.0 and subtract the alpha. The result is .95 or 95% interval that is correct. (NOTE: the end of the formula above-  $Z(\sigma/\sqrt{n})$  - is called the *margin of error*.) Once I decide upon  $\alpha$ , I can then find my Z. For this class, you will be given a t-Table where you can find Z. You can also find it in the table below.

Confidence Interval for Z±				
$\alpha$				
	.10	.05	.01	.001
Z	±1.645	±1.96	±2.58	±3.30
	90% CI	95% CI	99% CI	99.9% CI



Looking at the Z for  $\alpha = .05$  I see the number  $\pm 1.96$ . Using this Z and our Section 6 and 7 example data of  $\mu = 185$  and  $\sigma = 10.33$ , let us figure out the CI if our sample mean if  $\bar{x}$  was 183.5 with a sample size of 40.

$$\begin{aligned} \text{The 95\% CI of } \mu &= \bar{x} \pm Z \left( \frac{\sigma}{\sqrt{n}} \right) \\ &= 183.5 \pm 1.96 \left( \frac{10.33}{\sqrt{40}} \right) \\ &= 183.5 \pm 1.96 (1.63) \\ &= 183.5 \pm 3.19 \\ &= (180.31, 186.69) \end{aligned}$$

The upper and lower 95% confidence limits for these data are from 180.31 to 186.69. Since our *population parameter*  $\mu$  was 185, it is captured in the interval (180.31 < 185 < 186.69).

Now, let us run the data again using a 99% CI.

$$\begin{aligned} \text{The 99\% CI of } \mu &= \bar{x} \pm Z \left( \frac{\sigma}{\sqrt{n}} \right) \\ &= 183.5 \pm 2.58 \left( \frac{10.33}{\sqrt{40}} \right) \\ &= 183.5 \pm 2.58 (1.63) \\ &= 183.5 \pm 4.21 \\ &= (179.29, 187.71) \end{aligned}$$

The upper and lower 99% confidence limits for these data are from 179.29 to 187.71. Notice how the 99% confidence interval is wider than the 95% confidence interval. So we know for sure if our *population parameter*  $\mu$  is captured in the 95% limits, it will be in the 99% limits.

#### Margin of Error

Margin of error (ME) is our accepted level of sampling error – it quantifies our uncertainty that our results are based on chance, not a true finding. Here, Z is the same as you used in your CI.

$$ME = z \left( \frac{\sigma}{\sqrt{n}} \right)$$

#### Confidence Interval for the Mean – $\sigma$ Unknown

In Section 7, we learned that most of the time our value of  $\sigma$  is unknown and this means we must estimate it by using our sample statistic  $s$  (sample standard deviation). When estimating the population parameter  $\sigma$ , instead of a Z value when constructing a confidence interval, we will now use a  $t$  value.

As we learned in Section 7, the use of the  $t$  value also means use of *degrees of freedom* ( $df = n-1$ ). The  $t$ -value that you will use in the following equations always come from the  $t$  table taking  $df$  into account.

#### Confidence Interval CI Formula for $t$

$$CI = \bar{x} \pm t \left( \frac{s}{\sqrt{n}} \right)$$

To read the  $t$ -table, you will first find your  $\alpha$  from the bottom (think of it as  $100\% - \alpha = ?\%$ ):

z	0.000	0.674	0.978	1.000	1.282	1.645	1.960	2.328	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Then you will look to the left of the table and find your  $df$  vertically (the column):

df	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869

Finally, you will find where the two meet in the middle of the table. For instance, for a 95% CI with an  $n=25$  I would go to  $df = 24$  (remember, the formula for  $df$  is  $n-1$ ) and I find a  $t$  value of 2.064.

Using the data below let us assume our  $\sigma$  is unknown and we must use  $s$  to find the 95% CI.

$\mu = 185$ ,  $s = 9.65$ ,  $\bar{x} = 183.5$ ,  $n = 30$ ,  $df = 29$

$$\begin{aligned}
 \text{The 95\% CI of } \mu &= \bar{x} \pm t \left( \frac{s}{\sqrt{n}} \right) \\
 &= 183.5 \pm 2.045 \left( \frac{9.65}{\sqrt{30}} \right) \\
 &= 183.5 \pm 2.045 (1.76) \\
 &= 183.5 \pm 3.60 \\
 &= (179.9, 187.1)
 \end{aligned}$$

As we can see, our population parameter  $\mu$  is captured in the 95% confidence upper and lower limits of 179.9 to 187.1 (179.9 < **185** < 187.1).

There is a word of caution here. When interpreting CIs, we would say in the example above that we had a 95% chance of capturing the population parameter  $\mu=185$  within the interval. It is incorrect to say that 95% of the scores lie between the lower and upper confidence limits.

### Sample Size (n)

Sample size (n) selection is important to help us in determining if there are indeed mean difference among our data. It also helps us when determining significance (Section 9). The larger the sample size, the better the chance of 1) replicating the population, 2) a low probability of finding the difference in means by chance, and 3) there is an increased likelihood of finding significance.

During the design phase of a study, it is important to determine what the sample size should be using a sample size calculation. There are many sample size calculators available online or you could follow the formula  $n = \left(\frac{Z \times \sigma}{ME}\right)^2$ . For this class, you will not be determining sample size but you should know how to determine if the sample sizes given to you are large enough.

When you have an existing sample size and you want to determine if the sample size was large enough to approximate the normal distribution of  $\hat{p}$  (sample proportion) you can do so with the following rule:  $np \geq 10$  and  $n(1 - p) \geq 10$ . ( $p$ =population proportion).

### Sample Size Example

Using our HE 220 final exam data from previous examples, let us say we are looking for proportion of passing scores among students and we find 72% of respondents pass with a 70% or better. If we wanted to run a test on a sample of 22, is 22 sufficient to follow normal distribution?

Using the rule:  $np \geq 10$  and  $n(1 - p) \geq 10$  where  $n=22$  and  $p = .72$

$$np = 22 \cdot 0.72 = 15.84 > 10, \text{ but } n(1 - p) = 22(1 - .72) = 6.16 < 10$$

So, while  $np$  of 15.84 is greater than 10, the  $n(1-p)$  of 6.16 is less than 10 so I will not expect these samples to follow a normal distribution.

## Section 9: Hypothesis Testing

In the public health field, a clinician might want to know if one treatment option is better than another, a personal trainer might wish to determine if a strengthening technique decreases overall run pace, a dietician could look to see if a specific diet decreases cholesterol levels, or a teacher might want to know if a new learning method is better than another. These are all questions of comparison – the foundation for hypothesis testing. Hypothesis testing is a decision-making procedure to evaluate claims about a population. In Section 7 we were interested in comparing the means of samples, however, with hypothesis testing we are interested in making statistical inferences about our findings and determine if the findings (called an “observed effect”) are real or by chance.

For the purpose of this class we will focus on hypothesis concerning the population parameter  $\mu$ . This means we will be utilizing statistics we covered in earlier sections: The Z test and the  $t$  test. We will also examine the p-value method at the end of this section.

---

### Hypothesis Testing for the Population Parameter $\mu$

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} \text{ and } t = \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}}$$


---

### Steps in Hypothesis Testing

To begin, we start with a *statistical hypothesis*. This is an estimate or “best guess” about a population parameter that may or may not be true. The procedures we take to accept or reject statistical hypotheses are called *hypothesis testing*.

There are two types of *statistical hypothesis* for every situation: a null hypothesis and the alternative hypothesis. The *null hypothesis* ( $H_0$ ) is a statistical hypothesis that states that there is no difference between the population parameter  $\mu$  and the sample means. (NOTE – in your upper-level classes you will also examine the differences between two parameters so the null hypothesis would be “there are no differences between the two population parameters”). Conversely, the *alternative hypothesis* ( $H_1$ ) states that there is a difference between a population parameter and the sample means. When you present the null and alternative hypothesis – you do so using mathematical indicators. Below is an example for hypothesis with no direction - meaning they just say different, not greater than or less than. Please see the example on page 44 for hypothesis with “direction”.

---

**Null hypothesis:**  $H_0: \mu = \mu_0$  “there is no difference between  $\mu$  and sample means”

**Alternative Hypothesis:**  $H_0: \mu \neq \mu_0$  “there is a difference between  $\mu$  and sample means”

---

Your next step is to determine which statistical test you will use: a Z test or a  $t$  test. Remember from Section 7 if we know our population parameter  $\sigma$ , we use a Z test. If the population standard deviation is unknown or estimated from the sample data ( $s$ ), we use a  $t$  test.

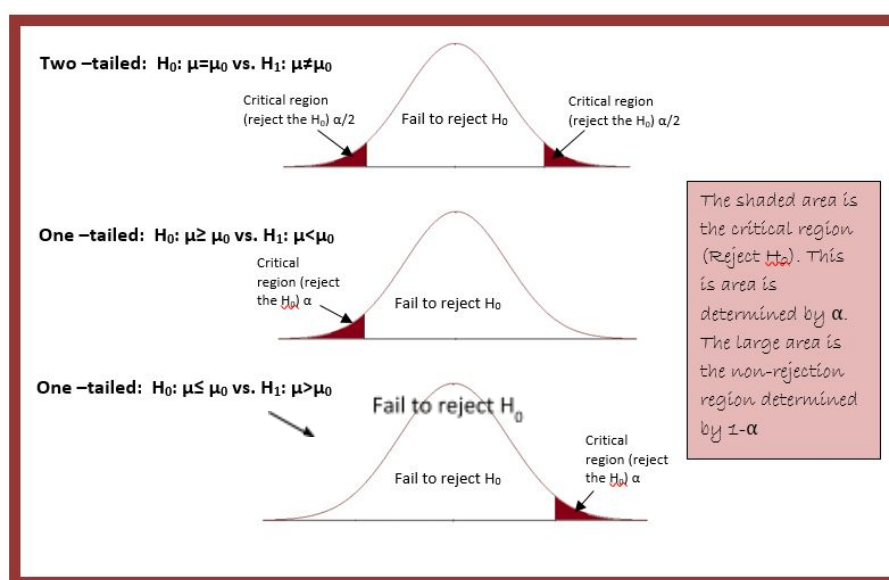
After you have selected the correct statistical test, you will then choose an appropriate *level of significance* ( $\alpha$ ). For this class we will practice using  $\alpha = .05$  and  $\alpha = .01$ . This is also the time to decide if you have a *one-tail* or a *two-tail* study so that you can determine the appropriate *critical value*.

The *critical value* (CV) is used to separate the critical region from the non-critical region of the distribution. You will use your significance level to select the CV from the table associated with your appropriate test. In most situations there is a separate Z table and  $t$ -table that you would consult to find the CV. For this purpose of this class the critical values for Z can be found on your  $t$ -table.

The CV can be either on the right side (+) or the left side (-) of the mean for a one-tailed test. Whether it is to the right or to the left is determined by the sign of your  $H_1$ . For example, if your hypothesis statement is  $H_0: \mu \leq \mu_0$  vs.  $H_1: \mu > \mu_0$ , the  $H_1$  of "greater than" (>) indicates a right side or positive CV (+CV). Alternatively, if your hypothesis is  $H_0: \mu \geq \mu_0$  vs.  $H_1: \mu < \mu_0$ , the  $H_1$  of "less than" (<) indicates a left side or negative CV (-CV).

The CV can also be *two-tailed* meaning that the value is on either side of the curve ( $\pm$ CV) and is also determined by the sign of your  $H_1$ . In this case, your hypothesis statement would be one of no difference:  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$ .

The area from the CV to the end of the tail is called the *critical region* or *rejection region*. This probability that your test statistic falls within that region is denoted by  $\alpha$ . If your study is *two-tailed* then both ends of the distribution contain the critical region. If your study is *one-tailed* then only one end of the distribution contains the critical region. Any way, if your test statistic falls within a *critical region*, then you reject the null hypothesis of no difference. If your test statistic does **not** fall within a critical region, but instead falls in the *non-rejection region*, then you would *fail to reject* the null. The *non-rejection region* is denoted as  $1 - \alpha$ .



Since we are only working with test of significance for one-sample Z or  $t$ 's, there are a series of steps that are necessary for you to learn. These include setting up your hypothesis, selecting the appropriate test statistic, deciding your significance level and resulting CV, calculating the test statistics and comparing it to the CV, and drawing the appropriate conclusions (reject, fail to reject, and a final statement).

### Six Steps in Hypothesis Testing

1. State the hypothesis using mathematical indicators  
 $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$   
 $H_0: \mu \leq \mu_0$  vs.  $H_1: \mu > \mu_0$   
 $H_0: \mu \geq \mu_0$  vs.  $H_1: \mu < \mu_0$
2. Choose a  $\alpha$  (significance level)
3. Select and calculate the appropriate test statistic (Z or  $t$ )
4. Determine the critical value and resulting critical region
5. Either reject or fail to reject the null hypothesis based on the region.
6. Summarize the results with a conclusion statement.

### DRAW IT OUT!

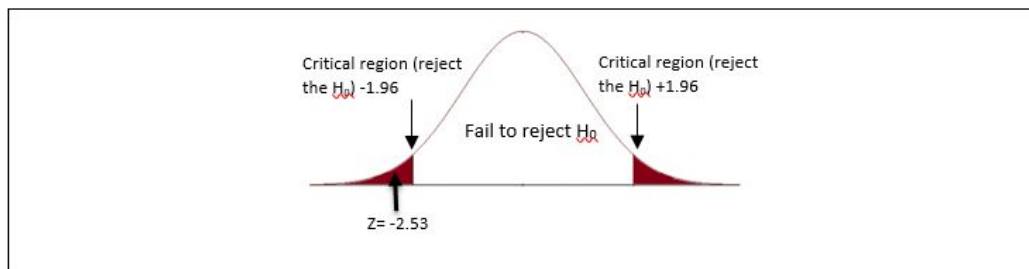
### Hypothesis Examples

#### Example 1:

To demonstrate the use of the six steps in hypothesis testing, let us reexamine the HE 200 Final data we have used in previous sections. Suppose that I tell you that throughout the history of the class, the final exam grades mean was 185. How could I support or reject that claim? First, I would draw a sample and in this case, I selected a sample size of 50 students. When I run the data on that sample of 50, I find a sample mean ( $\bar{x}$ ) of 181.3. Now I ask what is the likelihood of finding a sample mean of 181.3 in a sample of 50 from a distribution with a true mean,  $\mu = 185$  given that  $\sigma = 10.33$ ? Using the six steps presented above:

1.  $H_0: \mu = 185$  versus  $H_1: \mu \neq 185$ . I have selected the  $H_1$  of "different" because the narrative above did not offer any tail direction (i.e. it did not say "greater than 185" or "below 185"). Difference ( $\neq$ ) means that the  $\bar{x}$  could be higher or lower than  $\mu$ . This means my study is *two-tailed*.
  2. *Significance level:*  $\alpha = .05$
  3. *The test statistic* will be a Z because I know (and have been given) the population parameter  $\sigma$
- $$Z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{(181.3 - 185)}{\frac{10.33}{\sqrt{50}}} = \frac{-3.7}{1.46} = -2.53$$

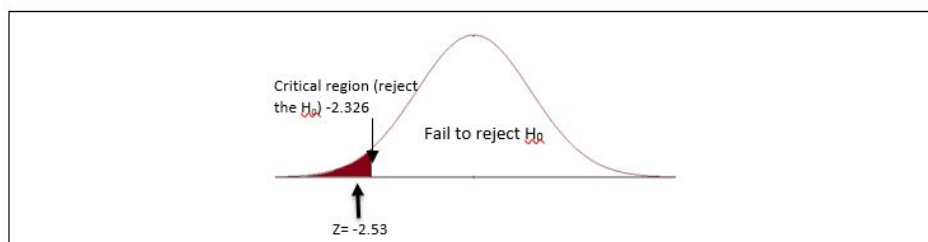
4. **Critical Value and Critical Region:** I know that I have a two-tailed test  $\alpha = .05$ . Looking at the top of the t-table I find the row titled “two-tails” then find  $\alpha = .05$ . Because this is a Z, I look down the .05 column to the Z row and find 1.96. Again, this is two tailed so my CV =  $\pm 1.96$
5. **Reject the null:** because the test statistic  $Z = -2.53$  falls within a critical region (in other words beyond the critical values of  $\pm 1.96$ ), I reject the null that the sample comes from a population not equal to 185. This is considered significant at the  $\alpha = .05$  level because the probability of it occurring by chance is less than .05
6. **Conclusion:** The mean test score of 181.3 is significantly different from the population mean of 185.



#### Example 2:

Let us run this data again but this time by asking the question “is the sample mean of 181.3 less than the population mean of 185” using an  $\alpha = .05$

1.  $H_0: \mu \geq 185$  versus  $H_1: \mu < 185$ . I have selected the  $H_1$  of “less” because the narrative above asked if  $\bar{x}$  was less than  $\mu$ . This means my study is *one-tailed* to the left (less than).
  2. **Significance level:**  $\alpha = .01$
  3. **The test statistic** will be a Z because I know (and have been given) the population parameter  $\sigma$
- $$Z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{(181.3 - 185)}{\frac{10.33}{\sqrt{50}}} = \frac{-3.7}{1.46} = -2.53$$
4. **Critical Value and Critical Region:** Now I have a one-tailed test  $\alpha = .01$ . Looking at the top of the t-table I find the row titled “one-tail” then find  $\alpha = .01$ . Because this is a Z, I look down the .01 column to the Z row and find 2.326. Being one-tailed to the left, my CV = -2.326
  5. **Reject the null:** because the test statistic  $Z = -2.53$  falls within the left critical region, I reject the null. This is considered significant at the  $\alpha = .01$  level because the probability of it occurring by chance is less than .01
  6. **Conclusion:** The mean test score of 181.3 is significantly less from the population mean of 185.



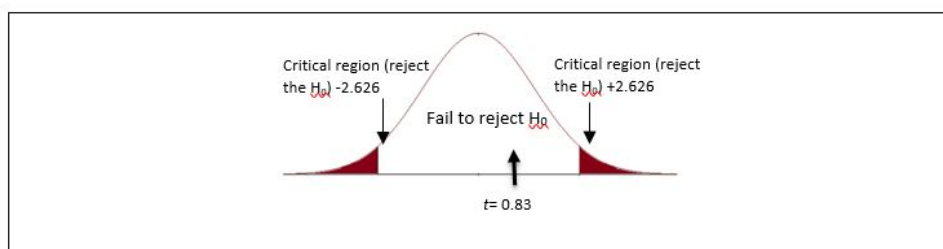
**Example 3:**

A survey of community colleges claims that the average number of students that skip a statistics class each week is 18.6. A random sample of 10 weeks during a term had a mean of 19.1. The sample standard deviation was 1.9. Is there enough evidence to reject the survey claim that there is a difference at  $\alpha = .05$ ?

1.  $H_0: \mu = 18.6$  versus  $H_1: \mu \neq 18.6$
2. Significance level:  $\alpha = .05$
3. The test statistic will be a  $t$  because we are using the sample standard deviation  $s$ .

$$t = \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}} = \frac{(19.1 - 18.6)}{\frac{1.9}{\sqrt{10}}} = \frac{0.5}{0.6} = .83$$

4. **Critical Value and Critical Region.** I know I have a two-tailed test because of the  $H_1$  of difference. As with before, look at the top of the  $t$ -table and find the row "two-tailed" then find  $\alpha = .05$ . This time I have a  $t$  so I will need to find and use degrees of freedom (df). Since  $n=10$ ,  $df=9$  and if I look at two-tailed,  $\alpha = .05$  for  $df=9$ , I find a CV of  $\pm 2.262$
5. **Fail to Reject the Null.** In this case the  $t$ -statistic of .83 does not fall in the critical region.
6. **Conclusion:** There is not enough difference to support the claim of a difference in the means.

**P-Values**

When making comparisons through hypothesis testing, the intention is often to report the findings as *statistically significant* (i.e.  $P < .05$ ). This  $P$  is the calculated probability of finding statistical result equal to or "more extreme" than the actual observation given the null hypothesis is true.

For this class all  $P$ -values will be calculated using an online tool found in your Moodle course shell. However, what you should understand is the interpretation of  $P$ -values (please note, the smaller the  $P$ -value, the greater the significance and the less change of results due to random sampling error).

P Value	Interpretation	Denotation
$P > 0.05$	The results are not significant, you would see the observed difference in up to 95% of studies due to random sampling error	none
$P < 0.05$	The results are significant, you would see the observed difference in up to 5% of studies due to random sampling error	*
$P < 0.01$	The results are significant, you would see the observed difference in up to 1% of studies due to random sampling error	**



P < 0.001	The results are significant, you would see the observed difference in up to 0.1% of studies due to random sampling error	***
-----------	--	-----

### Type I and Type II Errors

In hypothesis testing there is not 100% certainty because the test is based on probabilities. This means there is always a chance of obtaining an incorrect conclusion or *error*. The two types of errors that are possible in hypothesis testing are *type I* and *type II*.

#### Type I Error

If the null hypothesis of no difference is true, yet you reject the null and claim there is a difference, then you have committed a type I error. The probability of making such an error is  $\alpha$ , the significance level you set for your study. For example, in setting an  $\alpha = .05$  you are willing to accept that there is a 5% chance that you are wrong when you reject the null hypothesis. You can lower the risk of being wrong by lowering the value of your  $\alpha$ . However, this also means you might miss identifying a true difference if one really exists.

#### Type II Error

If the null hypothesis is false (there is a difference), yet you fail to reject the null of no difference (accept that they are the same), you have made a *type II error*. The probability of making such an error is  $\beta$  (beta).  $\beta$  depends on the power of the test ( $1-\beta$ ). By making sure your sample size is large enough to detect a difference when one truly exists, you decrease your risk of committing a *type II error*. In addition, choosing a one-tailed test and increasing  $\alpha$  also help increase power.

	H <sub>0</sub> is true	H <sub>0</sub> is False (H <sub>1</sub> is true)
Accept H <sub>0</sub>	Correct Decision ( $1-\alpha$ )	Type II Error ( $\beta$ )
Reject H <sub>0</sub> (assume H <sub>1</sub> is true)	Type 1 error ( $\alpha$ )	Correct Decision ( $1-\beta$ )
TOTAL	1	1

## Section 10: Chi Square ( $\chi^2$ )

In Section 9 we used the Z test and *t* test to look for significance to make inferences about data. Recall that the data used in the Z and *t* are quantitative data and therefore not applicable to qualitative data. However, in the public health field, not all the variables we wish to examine are quantitative. We may need to categorize the population based on gender, diagnosis, or behavior and this type of data – if you recall from Section 2 – is categorical data. The categorical data can then be arranged in a table called a *contingency table*. The cells within the contingency table are the variables and the chi square test allows us to determine if there is an association between the variables.

### Contingency Table: GPA and Degree Major

Major	GPA			Totals
	4.0	3.0 to 3.9	2.0 to 2.9	
HDFS	15	22	11	48
HPM	12	27	6	45
EXSS	9	18	13	40
<b>Totals</b>	36	67	30	133

The Chi-square test is aimed at comparing *observed frequencies* to *expected frequencies* and determines how likely it is that the observed distribution is due to chance. (Please note, often a chi square test is often called a *goodness of fit* statistics because it determines how well the *observed* data distribution fits the *expected* data distribution).

The *observed frequencies* are exactly that, the data that we collected or “observed” ourselves. The *expected frequencies* are what we expected to see based on our hypothesis. When we examine the value of chi square compared to a Chi Square table to determine if the difference in the expected and the observed data are statistically significant.

---

#### Chi Square Statistic

$$X^2 = \sum \left( \frac{(O - E)^2}{E} \right)$$

Where “O” is our observed values and “E” is our expected values  
 Degrees of freedom (df) = (row-1)(column -1)

---

For example, let us say we wish to determine if a group of 100 Public Health majors prefers online courses to traditional lecture courses. Looking at data let us say we expect that 70% would prefer online while 30% prefer the traditional setting. However, after giving a survey, we find the following:

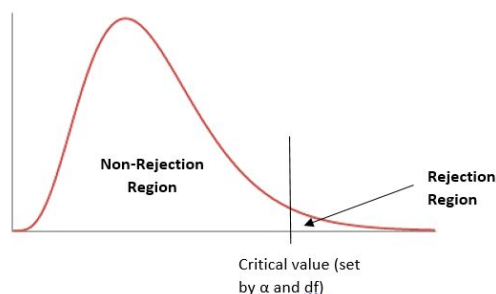
	Observed	Expected
Online	53	70
Traditional	47	30

Now, to work through the formula above, look at the following table:

	Observed	Expected	(O-E)	(O-E) <sup>2</sup>	(O-E) <sup>2</sup> /E
Online	53	70	-17	289	4.13
Traditional	47	30	17	289	9.63
<b>Totals</b>	100	100	0		<b>13.76</b>

This is the sum of  
 (Σ) all the (O-E)<sup>2</sup>/E.

My chi square ( $\chi^2$ ) is 13.76. To determine if this is significant, if it is unlikely to happen, we must first set  $\alpha$ , and then determine the degrees of freedom (df). For chi square, df is determined by the number of independent deviations (each  $O-E$ ) in the contingency table and solved with the formula degrees of freedom = (row - 1)(column - 1). In our example above, we have only a two-celled table so our degree of freedom = 1. Next, assuming an  $\alpha = .05$ , we go to the chi square table and find the corresponding critical value of df=1 at  $\alpha = .05$  and find 3.841. Chi square distributions are one-tailed to the right so if the resulting chi square value is greater than the critical value (to the right) then we would reject the null statement (this will be addressed later in the section and is based on the type of test).



In the previous example, your expected value “E” for each variable was known. However, this is not always the case. Often you have only the observed data and must estimate the expected value using probability. The best way to find the expected value for each observation is by putting your data into the contingency table as used earlier. To find the expected value for each observation, you use the row and column totals from that observation. For example, let us say we are observing if there is a difference between the proportion of AS degrees and selected major and AAOT degrees and selected major. The following observed data are presented:

	Health Major	Exercise Sport Science Major	Totals
AS Degree	62	92	154
AAOT Degree	71	55	126
Totals	133	147	280

To find the expected values of the observed data, you will work only from the total row and column (or perimeter). You can see that the observed value for AS degree seeking health majors was 62, to get the expected I will use the probability that they are a health major (133) and the probability that they are an AS degree seeker (154) and divide it by the total population:

$$\text{Health Major with AS expected} = \frac{(133 \text{ and } 154)}{280} = \frac{(133 \times 154)}{280} = 73.15$$

This step would happen for all four of the observed data presented above.

$$\text{Health Major with AAOT expected} = \frac{(133 \text{ and } 126)}{280} = \frac{(133 \times 126)}{280} = 59.85$$

$$\text{EXSS Major with AS expected} = \frac{(147 \text{ and } 154)}{280} = \frac{(147 \times 154)}{280} = 80.85$$

$$\text{EXSS with AAOT expected} = \frac{(147 \text{ and } 126)}{280} = \frac{(147 \times 126)}{280} = 66.15$$

Now using the chi square statistic, you can complete the following:

$$\chi^2 = \frac{(62-73.15)^2}{73.15} + \frac{(71-59.85)^2}{59.85} + \frac{(92-80.85)^2}{80.85} + \frac{(55-66.15)^2}{66.15} = 1.7 + 2.08 + 1.54 + 1.56 = 6.88$$

### Types of Chi Square Tests

Chi square tests can be used to determine the following:

1. If there is a difference between the observed frequency and the expected frequency (*goodness-of-fit test*)
2. If variables (or characteristics) are independent (*test of independence*)
3. If there is a significant difference between two proportions

Do note that the key assumptions to using a chi square test are that 1) your sample size is adequate and that no cell has less than a value of 5, and 2) if you are testing independence, that any one subject contributes data to only one cell. If, however, you are looking at non-independent data, like matched pairs in a before and after scenario, the appropriate choice is McNemar's *test*, discussed later in the section.

#### *Goodness-of-Fit*

As stated earlier, chi square is for qualitative data: specifically categorical data. If we have a categorical variable (i.e. chocolate candy color), and it has more than two sub-categories (i.e. green, blue, orange, red), and we want to test a hypothesis to determine if proportions or frequencies are similar across all the categories and “fits” the hypothesized set of proportions or frequencies, the goodness-of-fit test works best. Take for example the experiment we run during class to determine if the bag of candy we examined had the proportion of colors that the candy company claims it should have.

#### *Test of Independence*

The test of independence is a two-variable test to determine if there is an association between two categorical variables in one population. The null claim is that all the row and column variables are “independent” of each other or “not associated” with one another. This means our alternative hypothesis would be one of “dependence” (i.e. as one changes, so does another) or that of association.

#### *Test of Homogeneity*

The test of homogeneity is used to determine if the distribution of one variable is the same in multiple populations (two or more). The data are presented in a contingency table, just like the test of independence, but in this case, our null hypothesis is that the populations are homogenous (or equal) in respect to our characteristic or variable of interest. This means that our alternative hypothesis is that the populations are not homogenous (or not equal).

For all of the tests, the data must be random, mutually exclusive, and have large enough sample size. In addition, while their intents are different, they are calculated using the same formula, degrees of freedom, and chi square critical value chart.

### Six Steps in Hypothesis Testing: Chi Square

1. State the hypothesis of association, homogeneity, or change

$H_0$ : There is no significant difference between the observed and expected values

$H_1$ : There is a significant difference between the observed and expected values

-OR-

$H_0$ : There is no association between the variables (the categories are independent)

$H_1$ : There is an association between the variables (the categories are dependent)

-OR-

$H_0$ : The groups are homogenous

$H_1$ : The groups are not homogenous

-OR-

$H_0$ : There is no change pre and post intervention

$H_1$ : There is change from pre to post intervention

2. Choose a  $\alpha$  (significance level)
3. Select and calculate the appropriate test statistic (see below)
4. Determine the critical value and resulting critical region
5. Either reject or fail to reject the null hypothesis based on the region.
6. Summarize the results with a conclusion statement.

$$x^2 = \sum \frac{(O - E)^2}{E} \quad x^2 = \frac{n(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)} \quad x^2 = \frac{(b - c)^2}{b + c}$$

**DRAW IT OUT!**

### Example 1: Goodness of Fit

Therapeutic Bands, Inc. states that when you order a pre packaged mixed bulk order of their therapy bands, you will get 40% yellow (level 1), 30% red (level 2), 20% green (level 3), and 10% blue (level 4). You open one of your bags that has a 100 count and find that you have 45 yellow, 27 red, 18 green, and 10 blue. Is this finding consistent with Therapeutic Bands, Inc. claim?

	Observed	Expected	(O-E)	(O-E) <sup>2</sup> /E
Yellow	45	40	5	.625
Red	27	30	-3	.30
Green	18	20	-2	.20
Blue	10	10	0	0
Totals	100	100	0	1.125

1.  $H_0$ : there is no difference between the observed and expected proportions of colors of bands  
 $H_1$ : there is a difference between the observed and expected proportions of colors of bands
2. Assume  $\alpha = .05$

$$3. \text{ Using the goodness-of-fit } x^2 = \sum \frac{(O - E)^2}{E} = 1.125$$

4. Df = (4-1)(1-1) (I had four rows but only one column of variables) = 3 with a critical value of 7.81
5. Fail to reject the null.  $x^2 1.125 < 7.81$
6. There is no difference between observed and expected proportions of colors of bands

### Example 2: Test of Independence

An opinion poll surveyed a simple random sample of 200 students at LBCC about PE courses offered each term to determine if there was a gender difference by term. Respondents were categorized by

gender (male or female) and which term they preferred to take PE courses (fall, winter, spring). The following results were obtained:

	Term Preference			Row Totals
	Fall	Winter	Spring	
Male	32	25	40	97
Female	44	34	25	103
Column Totals	76	59	65	200

1.  $H_0$ : there is no association between gender and term selection (they are independent)  
 $H_1$ : there is an association between gender and term selection (they are dependent)
2. Assume  $\alpha = .05$

3. Using the test of independence 
$$x^2 = \sum \frac{(O - E)^2}{E} = 6.56$$
4. Df = (2-1)(3-1) = 2 with a critical value of 5.99
5. Reject the null.  $x^2 6.56 > 5.99$
6. There is an association between gender and term selection.

	Observed	Expected	$\frac{(O-E)^2}{E}$
Male Fall	32	$(97 \times 76)/200 = 36.86$	.64
Male Winter	25	$(97 \times 59)/200 = 28.615$	.46
Male Spring	40	$(97 \times 65)/200 = 31.525$	2.28
Female Fall	44	$(103 \times 76)/200 = 39.14$	.60
Female Winter	34	$(103 \times 59)/200 = 30.385$	.43
Female Spring	25	$(103 \times 65)/200 = 33.475$	2.15
	200	200	
			$x^2 = \sum \frac{(O-E)^2}{E} = 6.56$

### Example 3: Test of Homogeneity

In a study of the exercise habits of college freshmen, a researcher selects a random sample of 350 students -200 males and 150 females. Each student is asked which of the following types of exercise they like best: weight training, running (jogging), or yoga. The research question is "Do male preferences for these exercise types differ significantly from the female's exercise preferences?"

	Exercise Preference			Row Total
	Weight Training	Running/Jogging	Yoga	
Male	89	76	35	200
Female	38	62	50	150
Column Total	127	138	85	350

1.  $H_0$ : males who prefer weight training is the same as (homogenous) females who prefer weight training

$H_0$ : males who prefer running/jogging is the same as (homogenous) females who prefer running/jogging

$H_0$ : males who prefer yoga is the same as (homogenous) females who prefer weight yoga

$H_1$ : at least one of the null hypothesis statements are false (not homogenous).

2. Assume  $\alpha = .05$

3. Using the test of homogeneity 
$$x^2 = \sum \frac{(O - E)^2}{E} = 17.77$$

4. Df = (3-1)(2-1) = 2 with a critical value of 5.99

5. Reject the null .  $x^2 17.77 > 5.99$

6. At least one of the null hypotheses are not homogenous.

### Two-by-Two Contingency Tables

Quite often you will find public health research involves data that can be fit into a fourfold, or 2 x 2, table. This means there are two groups and two possible responses. In this case, as the table below will show, the frequencies (cells) are presented as  $a$ ,  $b$ ,  $c$ , and  $d$ . With the four cells, our degrees of freedom for any 2 x 2 table will be  $df=1$ .

	Intervention	Control	Totals
Yes	$a$	$b$	$a+b$
No	$c$	$d$	$c+d$
Totals	$a+c$	$b+d$	$n$

With this data, it is possible to compute chi square without needing to find the expected frequencies by using this formula:

$$X^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

Remember, these data are still mutually exclusive and the null hypothesis is still that of "no difference".

#### Example 4: Two-by-Two Contingency Table

A survey of students taking HE 220 found that 24 out of 40 students who had Math 95 as a prerequisite and 38 of 50 students with Math 111 as a prerequisite passed the course with a 90% or better. Is there evidence suggesting an association between prerequisite and scoring a 90% or better?

	$\geq 90\%$	$\leq 89\%$	Totals
Math 95	24	16	40
Math 111	38	12	50
Totals	62	28	90

1.  $H_0$ : there is no association between math perquisite and course completion grade

$H_1$ : there is an association between math prerequisite and course completion grade

2. Assume  $\alpha = .05$

Using the 2 x 2 formula 
$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{90[(24 \times 12) - (16 \times 38)]^2}{(62)(28)(40)(50)} = 2.65$$

3. Df = 1 with a critical value of 3.841  
 4. Fail to reject the null.  $\chi^2 2.65 < 3.841$   
 5. There is no association between math prerequisite and course completion grade

### McNemar's Test for Correlated Proportions

The chi square test can also be used to determine if two proportions have changed significantly on a dichotomous trait at two time points on the same population. This means the matched samples are not independent but the same participants measured twice (sometimes referred to as "within" samples). The chi square test for these type of data is called *McNemar's Chi Square*. This test is used most often with a before-and-after or pre-and-post study design. The idea is to determine if there is a change from before an intervention or event and after that event.

#### McNemar's Chi Square Test

$$\chi^2 = \frac{(b - c)^2}{(b + c)}$$

Degrees of Freedom (df) = 1

Since McNemar's chi square is for dichotomous data (two) with two points in time measured, it is displayed in a 2 x 2 contingency table. However, with McNemar's chi square, only cells *b* and *c* are utilized.

	Before (or first measure)	
After (or second measure)	Yes	No
Yes	<i>a</i>	<i>b</i>
No	<i>c</i>	<i>d</i>

#### Example 5: McNemar's Chi Square

Let us say you wish to measure the favorability of a bond measure on increasing spending on afterschool care for children before a town hall meeting and after the meeting to see if opinions have changed as a result of that meeting. The table below shows your findings:

	Wanted increased spending <b>before</b> the town hall	
Wanted increased spending <b>after</b> the town hall	Yes	No
Yes	29	37
No	17	49

1.  $H_0$ : the town hall had no impact on the favorability of the audience



$H_1$ : the town hall had an impact on the favorability of the audience

2. Assume  $\alpha = .05$

Using McNemar's Chi  $\chi^2 = \frac{(b-c)^2}{(b+c)} = \frac{(37-17)^2}{(37+17)} = 7.41$

3. Df = 1 with a critical value of 3.841  
4. Reject the null.  $\chi^2 7.41 > 3.841$   
5. The town hall had an impact on the favorability of the audience.
- 

#### **Important note on Chi Square**

While chi square does tell us if there is or is not a difference, when we reject the null of no difference and find significance, chi square does not tell us where that difference lies – just that it is there. For example, with our town hall meeting above, we can see there was a significant difference in the proportions before and after the meeting but we cannot say it moved people to be more or less favorable, just that the change occurred at a significant level.

---