

## Introduction to Statistics

Welcome to AP Statistics. For those of you who thought that you were signing up for a math class, I feel that I should make you aware that I really do not consider statistics a math class. I consider statistics a class that uses math much like some of the science classes utilize math. In statistics, like science, we test hypotheses through the design and completion of observational studies and experiments.

Okay so you are asking yourself, if statistics is not really a math course what is it and why is it important?

First, I will address the question: **Why is statistics important?** It is important because I said it is. I am guessing that my answer didn't satisfy many or any of you. Okay, so how about this: How do you think Amazon grew to be so huge and powerful when it started out as online book store? The answer their use of statistics. How do you think that Japan became a major manufacturing power known for high quality automobiles when their country and industry was virtually destroyed by a world war? The answer is statistics. How do you think a baseball team with the lowest salary base became a divisional champion? The answer is statistics. How did a group of MIT nerds get banned from Las Vegas? The answer is statistics. How was an effective HIV screening test created from one that was completely unreliable? The answer is statistics. How are advertisements selected for your computer screen or smartphone? The answer is statistics. How did Target know that a teenage girl that they had never seen was pregnant before her father did? The answer is statistics.

Let's explore the story about Target and the teenage pregnancy in a little more detail.

**The background:** To make better use of their advertising budget, Target began utilizing statistics to better identify the purchasing patterns of their customers so that they could send advertisements that might be of greater interest to each customer. In other words, based on what a person purchased, Target was sending advertisements and coupons to them for things that Target was predicting that the customer would want to purchase.

To that end a teenage girl was receiving advertisements for baby items. As a result an angry father walked into a Target outside of Minneapolis demanding to talk to the manager: "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

The manager didn't have any idea what the man was talking about. He looked at the mailer. Sure enough, it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants. The manager apologized and then called a few days later to apologize again. On the phone, though, the father was somewhat abashed. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. She's due in August.

The Complete article can be found: <http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>

A closely related article and interview of the Target advertising Mastermind:

<https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=6&r=1&hp>

So now that you are a little curious, let's answer the question: **What is Statistics?**

**Statistics** is the study of how to collect, organize, display, describe, and analyze the data from a sample in order to make generalizations about a population.

I am certain that some or many of you are thinking, great another definition that my teacher believes explains everything, but in reality it leaves more questions than answers. To which I must respond, fair enough, but give me a few more minutes and I believe that I will be able to break down the definition and give you an overview of the course at the same time.

AP Statistics is divided into 4 sections:

- Collecting Data,
- Displaying and Describing the Collected data
- Analyzing the Collected Data
- Making inferences/decision based on the data collected.

The First Quarter will be devoted to Collecting Data and Displaying and Describing data. These two sections are the easiest sections of the course and it is imperative that you achieve really good grades in both of these sections for two major reasons. First, the collection of data is foundational to all other sections of the course and secondly the 2<sup>nd</sup> quarter which is comprised almost entirely of data analysis, **(the probability section)** is the most challenging for students. Yes, most students say probability is the hardest and their second quarter grades tend to be lower than that of their first quarter.

The final section of the course, making inferences, is where you actually get to make decisions based on data. Should we purchase the new machine, is the new drug more effective than the old one, can blank predict blank and with how much confidence.

I love the section on inferences and I feel it is empowering, because it is from this section that I can use all of the collected data to make an informed decision and not just a guess. Now I want to be perfectly clear about the word guess. **Nothing in statistics is certain**, however, with statistics I am able to determine the chance that I could be wrong. Typically I will choose to be correct 95% of the time. Which means that I accept the probability that I will be wrong about 5% of the time—but I am getting ahead of myself as these discussions will not occur in detail until the 3<sup>rd</sup> quarter.

As we begin this course, I should warn you that there are a significant number of definitions that must be learned in order to understand what we are trying to accomplish. So how important are the definitions? Very important—put it this way, I don't expect you to have to perform any mathematical calculations on your first test. That isn't to say that the first test is easy, it is just a recognition that statistics is different from all of the math classes you have taken thus far.

Incidentally, the first test is challenging for most students, not because the topics are difficult but because you are being asked to think and reason in a different manner. In previous math classes, you followed an explicit set of directions to find a solution to a problem. In statistics, you will be expected to not only know and apply the definitions and methods, but be able to relate them to the outside world. In fact, the more you know about the world and other subjects, the more connections you will be able to make and the more successful you will be in this course.

**So back to the definition and why we need statistics:**

**Statistics** is the study of how to collect, organize, display, describe, and analyze the data from a sample in order to make generalizations about a population.

To fully understand what this means, we need to understand the concept and importance of both a **population** and a **sample**. (we will define and discuss **sample** in detail next week)

**Population:** the entire group of subjects or individuals that is the subject of interest.

**Parameter**—a numerical measurement of a population (*rarely known*).

**Census**—the collection of data from each unit in the population.

*(Difficult if not impossible with a large population )*

In statistics, a population is the entire group of subjects or individuals that is the subject of interest while a sample is a subset of the population. For example: we may be interested in the percentage of students in your statistics class that have brown eyes. In that case, our population would be the students in your statistics class and it would be fairly easy to count and figure out the percentage of brown eyed students in the class. The percentage of brown-eyed students would be the parameter. Because we are checking the eye-color of the entire population, every student in the classroom, we are conducting a census.

The great thing about being able to conduct a census, is that we know the true population parameter. In this case, we would know the true proportion or percentage of students with brown eyes in the classroom. According to some of my students, the best part of being able to conduct a census knowing the true population parameter is that there is no need for statistical methods and no need for this class.

So why not perform a census all of the time? To that answer that question let's consider the percentage of students with brown-eye color scenario. What if my population of interest is not the students in my classroom but all of the students at Reagan? While counting that many students might be a challenge, and I might creep a few people out when I asked them for their eye color, it is probably still doable. However what if we wanted to know the percentage of people in San Antonio or the world who had brown eyes? Well then our population would be all of San Antonio or the world and it would be unrealistic to conduct a census count all of the brown eyed people in San Antonio or the world.

**Why don't we always collect data using a census?**

1. Sometimes it is impossible to conduct a census.
  - I cannot find the average size of bass in a particular lake without draining the lake at which point the bass is dead and there is no lake.
  - I cannot count the number of red blood cells in your body, without removing all of the blood from you and sucking your life away.
2. A census can be extremely expensive  
The 2010 U.S census cost **\$13 billion, approximately \$42** person—your tax dollars at work.
3. Due to the difficulty in execution, a census can very inaccurate
  - In the 2000 census, it is estimated that 6.4 million Americans were missed, however, there was an overcount of 36,000. Now how did that happen?
  - In 2010, it is estimated that in Hidalgo county in Texas 225,000 persons were not counted.

Date: \_\_\_\_\_

Compiled by: Loren L. Spencer

## Notes: Simulations

So while an accurate census is great, it is not always possible and may be cost prohibitive. In the event that it is not plausible to conduct a census, we are forced to settle for an estimate. So the question is how do we make an accurate estimate. Essentially we have two choices, we can take a sample or we can run a simulation. We will discuss in great detail how to utilize sample data, but we are going to begin with learning how to conduct a simulation.

**Simulation:** a way to model random events, such that simulated outcomes closely match real-world outcomes. By observing simulated outcomes, researchers gain insight on the real world.

- Chance/randomness must be employed
- Often Used because it is more economical than running a true experiment
- Typically a probability model can be used to generate the same information
- We Want things to be fair and without bias or prejudice in our method of selection

**Random:** An event in which we know what possible outcomes can occur but do not which outcome actually will take place.

**Component:** The most basic **situation** in which something happens at random—(Hint: the singular event that you are repeating)

**Outcome:** the result of a single component

**Trial:** The number of components necessary to occur to model a situation. A single run of sequence of events being simulated

**Response Variable:** variable that measures the outcome of each trial; **outputs (Y's)**

**Example:** How many heads would I get if I flipped a coin 100 times?

### Modeling a Simulation

**Step 1:** Identify the component to be repeated (a component is the most basic event of the simulation)

**Step 2:** Explain how to model the outcomes (Assign numbers to the possible outcomes)

**Step 3:** Explain how to simulate the trial-what are you going to do

**Step 4:** Clearly State what the Response Variable is—(this is the answer to the question)

**Hint:** How many random selections did it take to complete a trial?

**Step 5:** Run Several Trials (chart the results)

**Remember:** A trial is the desired outcome

**Step 6:** Analyze the Response Variable (take an average)

**Step 7:** State your conclusion in the context of the problem.

## Recipe for Success: Simulations

## Q 1.1

1. **Read the entire problem**      What is being asked?
2. **Identify the Question**      Explain what the question is asking in your own words
3. **Identify and define a success and component in context**
  - **Success:** What we want to happen
  - **Component:** What is being repeated
4. **Identify a Trial**      How many successes are required?
5. **Model the simulation**

**Assign 1 or 2 digit numbers** in proportion to the chance of success and failure.

  - 00 = 0 or 100; 01=1; 02 = 2; 03 = 3...09 = 9
  - 10 = 10; 11 = 11... 99 = 99
  - **Assign the numbers to be skipped or ignored**
6. **Address Duplications**

**Are repeats permitted:** Can something occur twice?

  - Percentages-usually can be duplicated
  - Specific items-usually cannot be duplicated  
(Occurs when the quantities of items are known)
7. **Explain how to run the simulation**
  1. **Explain how to run a trial**
    - Beginning from left to right I would select 1 or 2 digit numbers until there were \_\_\_\_ number of successes.
    - Count how many 1 or 2 digit values that were not skipped.
  2. **Tell how many trials are going to be run**
  3. **Find the average/mean of all the trials**
8. **Run the simulation & Make a Table**

Trial Number	Number of 2 Digit Values Counted (successes)
1	
2	
3	
	<b>Total</b>

  - Draw a line through the values that represent failures
  - Circle the values that represent successes  
(Do not forget about duplicates are they permitted or not)
  - Mark an X through Skips  
(These are numbers that are not possible- for instance when duplicates are not permitted)
  - Draw a vertical line at the end of a trial
  - Count the number of 2 digit numbers in the trial
  - Record the values in a table
  - Repeat for all necessary trials to complete the simulation
9. **Calculate the Expected number**

**Take the average/Mean**  
Sum the number of successes counted for each trial

  - Divide by the number of trials

### Conclusion:

Based on the simulation above, on average we would expect to have

\_\_\_\_\_ in order to find \_\_\_\_\_  
*Give the number & definition of component*      *Give the number & definition of successes*

### Notes: Simulations Scenarios

**Scenario 1** Every Monday a local radio station gives coupons away to 50 people who correctly answer a question about a news fact from the previous day's newspaper. The coupons given away are numbered from 1 to 50, with the first person receiving coupon 1, the second person receiving coupon 2, and so on, until all 50 coupons are given away. On the following Saturday, the radio station randomly draws numbers from 1 to 50 and awards cash prizes to the holders of the coupons with these numbers. Numbers continue to be drawn without replacement until the total amount awarded first equals or exceeds \$300. If selected, coupons 1 through 5 each have a cash value of \$200, coupons 6 through 20 each have a cash value of \$100, and coupons 21 through 50 each have a cash value of \$50.

- (a) Explain how you would conduct a simulation using the random number table provided below to estimate the distribution of the number of prizewinners each week.

I would assign 01-05 a \$200 valuation, and 06-20 a \$100 valuation and 21-50 a \$50 valuation. I would skip 00 and 51-99. Working from left to right I would select 2 digit numbers at a time, skipping any duplicates, until the sum of the valuation was greater than or equal to \$300. I would then count the number of selected 2 digit numbers between 01 and 50 which would be the number of prize winners for that trial. To increase accuracy, I would run the simulation 20 times and find the average number of winners and that would be the expected number of prize winners.

- (b) Perform your simulation 3 times. (That is run 3 trials of your simulation.) Start at the leftmost digit in the first row of the table and move across. Make your procedure clear so that someone can follow what you did. You must do this by marking directly on or above the table. Report the number of winners in each of your 3 trials.

<del>72749</del>	<del>13347</del>	<del>65030</del>	26128	49067	02904	49953	74674	94617	73317
	\$50 \$50	\$200							
<del>81638</del>	<del>36566</del>	<del>42709</del>	<del>33717</del>	<del>59943</del>	<del>12027</del>	46547	61303	46690	76423
		\$50	\$50 \$100	\$50 \$100					
<del>38149</del>	<del>46438</del>	<del>91579</del>	<del>01907</del>	72146	05764	22400	94490	49890	09258
	\$50 \$50		\$100 \$100						

Trial 1	3 winners	\$300
Trial 2	5 winners	\$350
Trial 3	4 winners	\$300
Average	4 winners	\$317

Based on the simulation above I would expect there to be 4 prize winners a week and I would expect my weekly payout to be about 317 dollars,

## Recipe for Success: Simulations

## Q 1.1

- |   |  |
|---|--|
| 1. Read the entire problem                                | What is being asked?   |
| 2. Identify the Question                                  | Explain what the question is asking in your own words  |
| 3. Identify and define a success and component in context | <ul style="list-style-type: none"> <li>• <b>Success:</b> What we want to happen</li> <li>• <b>Component:</b> What is being repeated</li> </ul>   |
| 4. Identify a Trial                                       | How many successes are required?   |
| 5. Model the simulation                                   | <p><b>Assign 1 or 2 digit numbers</b> in proportion to the chance of success and failure.</p> <ul style="list-style-type: none"> <li>• 00 = 0 or 100; 01=1; 02 = 2; 03 = 3...09 = 9</li> <li>• 10 = 10; 11 = 11... 99 = 99</li> <li>• <b>Assign the numbers to be skipped or ignored</b></li> </ul>  |
| 6. Address Duplications                                   | <p><b>Are repeats permitted:</b> Can something occur twice?</p> <ul style="list-style-type: none"> <li>• Percentages-usually can be duplicated</li> <li>• Specific items-usually cannot be duplicated<br/>(Occurs when the quantities of items are known)</li> </ul>   |
| 7. Explain how to run the simulation                      | <p>4. Explain how to run a trial</p> <ul style="list-style-type: none"> <li>• Beginning from left to right I would select 1 or 2 digit numbers until there were ____ number of successes.</li> <li>• Count how many 1 or 2 digit values that were not skipped.</li> </ul> <p>5. Tell how many trials are going to be run</p> <p>6. Find the average/mean of all the trials</p>   |
| 8. Run the simulation & Make a Table                      | <ul style="list-style-type: none"> <li>• Draw a line through the values that represent failures</li> <li>• Circle the values that represent successes<br/>(Do not forget about duplicates are they permitted or not)</li> <li>• Mark an X through Skips<br/>(These are numbers that are not possible-<br/>for instance when duplicates are not permitted)</li> <li>• Draw a vertical line at the end of a trial</li> <li>• Count the number of 2 digit numbers in the trial</li> <li>• Record the values in a table</li> <li>• Repeat for all necessary trials to complete the simulation</li> </ul> |
| 9. Calculate the Expected number                          | <p><b>Take the average/Mean</b></p> <p>Sum the number of successes counted for each trial</p> <ul style="list-style-type: none"> <li>• Divide by the number of trials</li> </ul>   |

Trial Number	Number of 2 Digit Values Counted (successes)
1	
2	
3	
	<b>Total</b>

**Conclusion:**

**Based on the simulation above, on average we would expect to have**

**in order to find**

*Give the number & definition of component*

*Give the number & definition of successes*



### Notes: Simulations Scenarios

**Scenario 2:** Suppose the probability that an exploratory oil well will strike oil is about 0.2. Conduct a simulation to answer the following questions. Assume that the outcome (oil or no oil) for any one exploratory well is independent of outcomes from other wells. Use the random number table below and conduct 20 trials. Clearly identify each trial on the table.

(a) Estimate the average number of wells that need to be drilled in order to strike oil.

I will work left to right choosing single digit numbers. 0 and 1 represent oil strikes & 2-9 represent wells that did not strike oil.

20 trials were run and 98 wells were drilled.

Based on the simulation below, we would expect to have to drill an average of 4.9 wells to strike oil.

(b) What is the probability that it will take fewer than 3 attempts to strike oil?

8 out of 20 times or 35% of the time oil was struck in less than 3 times.

(c) What is the probability that it will take exactly 6 wells to strike oil?

2 out of 20 times or 10% of the time oil was found on the 6<sup>th</sup> well.

47169	80410	03333	73856	85627	54351
36653	55390	20439	48605	45513	05458
76361	47409	14914	55280	70533	52960
20579	87054	59998	90071	67554	91237
96994	65965	73235	49260	45309	24660

Trial	# of Wells	Trial	# of Wells
1	3	11	2
2	4	12	9
3	2	13	4
4	1	14	2
5	1	15	3
6	19	16	6
7	10	17	2
8	2	18	8
9	7	19	2
10	5	20	6

**Notes: Simulations Scenarios**

**Scenario 3:** A certain professor has eight keys, but he never recalls which one fits his office door lock. He places all of the keys on a table and randomly selects and tries one key at a time. If the key does not fit he places the key back in the original pile and then randomly selects another key. (yes, he is senile and probably teaches statistics-don't laugh that would be rude) **Note: all the keys look the same and he is unable to tell which key he has already tried.** Conduct a simulation to answer the following questions. Use the random number table below and conduct 20 trials. Clearly identify each trial on the table.

(a) Explain how you would conduct a simulation using the random number table provided below to estimate the number of keys the professor would need to try in the lock.

I will work left to right choosing single digit numbers. 1 through 8 represents actual keys. Digits 9 & 0 are skips as they do not represent keys. Key number 8 is the correct key for the lock. Because keys are replaced, duplicates are permitted (a key could be tried multiple times). A trial will end when I find the correct key. I will run 20 trials.

(b) What is the expected number of tries needed for him to find the correct key?  
Based on the simulation below, we would expect the professor to have to try an average of 6.8 keys in order to find the one that fits the lock

(c) What is the probability it will take more than 4 tries to find the right key?

Based our simulation we would expect it to take more than 4 tries 55 % of the time. (11 out of 20 attempts)

Trial	# of Keys					
1	3					
2	4					
3	3	64831	78558	25961	07610	75464
4	13					
5	11					
6	11	34336	39840	24371	53548	01485
7	3					
8	4					
9	8	11792	38659	92620	48253	05370
10	7					
11	7					
12	6	65985	43392	21100	08763	37469
13	9					
14	11					
15	4	52822	48990	03648	34861	54680
16	4					
17	4					
18	3	31645	45552	78255	64794	21228
19	6					
20	16					
Total	137	38804	45687	85320	54654	76156
Avg.	6.85					01853

## Notes: Samples

**While economical, easy to run, and unbiased, simulations have 2 major limitations.**

- 1. To run a simulation we must know something about the population.** We must know the population parameter of interest. For instance, in the oil well simulation, we needed to know the percentage of time that we expected to strike oil. Unfortunately, much of the time we often don't know anything about the population. We don't know the percent of deformed blood cells in your body; we don't know what percent of the population supports a candidate.
- 2. Simulations do not allow us to test whether or not the claimed percentage of successes is correct.** When running a simulation we are forced to assume that the percentage of successes is correct. The oil well simulation problem claimed that we struck oil twenty percent of the time, but what if the percentage had been different than 20 percent and we the drilling company were unaware of the true percentage of successful wells? We assumed that the drilling company would strike oil twenty percent of the time and used that value in our simulation. However, simulations do not provide a mechanism to determine whether or not this was actually the case.

**So the question is:** what do we do when we don't know a population parameter and conducting a census is not appropriate? **The answer is:** take a sample. Well that sounds simple enough, but leads us to the next question: **What is a sample?** A sample is merely a subset of the population.

**Scenario:** If I wanted suggestions for a theme song for this year's Reagan students, instead of performing a census and asking all of the students at Reagan for their suggestions, I could take a sample or subset of Reagan students. Since a sample is a subset of the population there are several ways to go about finding a subset.

1. I could ask twenty-five students sitting together in the lunch room.
2. I could ask my freshmen advisory and ask for their idea of the theme song.
3. I could go into the Teacher's lounge and get suggestions for the song there.
4. I could attend the PTA meeting and ask for their suggestions.

So which of those methods would you suggest I use and how do you feel about them?

I am going to go out on a limb here and guess that you don't really like any of those methods. I am betting that you are concerned that your opinion for the Reagan Theme Song may not be represented.

While samples **can be** an excellent way to make determinations about a population, **samples must be representative of the entire population.** So the question is: how do we take a "good sample."

First, I want to stress again that **a good sample is one that is representative of the population.**

While that may seem obvious, it isn't as easy to do as one may think. If you are having trouble believing me, consider who I should ask about choosing a theme song for Reagan and give me a list of those people.

### Summary:

- A sample is a portion of a population which is studied to draw conclusions & about the characteristics of the whole population.
- A sample must be representative of the population to be useful.
- No conclusions should be drawn from poorly collected data or badly designed experiments or studies.

## Notes: Samples and Bias

We have touched on the need for "good samples," samples that are representative of the population of interest. We have also suggested that collecting a representative sample is not always easy to do. (*Did we ever get an agreement as to whom we should talk to about a theme song for the school*). So what are some of the pitfalls that prevent us from collecting samples that are representative of the population?

As preparation for those answers, I would like to ask you all to participate in a brief but very important activity. Without giving it much thought, please select number and circle it.

1 2 3 4

How many of you chose: 1. \_\_\_\_\_ 2. \_\_\_\_\_ 3. \_\_\_\_\_ 4. \_\_\_\_\_

Typically about 75% of the population chooses 3, 20% are divided between 2 and 4 and 5% choose 1. Because there are 4 numbers, each number should be chosen 25% of the time. So why didn't that happen. Psychologists have discussed a variety reasons why we overwhelmingly choose 3 and fail to choose 1. However regardless of the reason, this shows that in general we make biased selections and in statistics biased samples result in samples that are not representative of the population and are therefore worthless.

**So what is bias?** Bias is over or under-representing a component of the population in a sample.

The above example has some really practical applications to your life. You don't think so? Consider this: You go to the bathroom and there are three stalls. Assuming the bathroom is empty, which stall do you normally choose? For the vast majority of people, the answer is the second stall or middle stall which makes it the grossest and least clean stall of all. The next most chosen stall is the third stall which is the furthest one from the door. Assuming we are concerned about germs, hygiene etc. we should be choosing the first stall. Who knew that stats class would be so helpful with toilet training?

In the scenarios above we exhibited a bias. We over-represented the number 3 and the 2<sup>nd</sup> stall and we significantly under-represented the number 1 and the 1<sup>st</sup> stall. To have samples that are useful, we must eliminate bias through the design and execution of our studies and experiments.

**Summary:** The key point is that we must design studies in a manner that selects samples that are representative of the population. We must not over or under-represent segments of the population.

## Notes: Bias

Bias is the result of a bad study design and is sometimes referred to as **selection bias**, because the sample we **selected is biased**, that is to say that the selected sample does not represent the population.

As we stated earlier, biased data is absolutely useless and we have no methods to fix data that is not representative of the population of interest. If we happen to collect a biased sample, then we must discard it, start over and collect another sample.

In order to avoid collecting a biased sample, it is important to recognize the types of bias that exist so that we can create data collection methods that prevent the introduction of biased data into our sample.

Common types of statistical bias are:

**Under-coverage Bias**-Occurs when members of a population are not able to be selected or chosen for a sample because they are not able to be accounted for (*homeless, inmates, no land-line, college students*). It is desirable to have a listing of the population and then randomly select individuals from the listing for the sample. Obviously, the sample is only as good as the listing of the population and in this case the above mentioned groups are often left off of population lists.

**Non-response Bias**-Occurs when a member in the population is randomly identified and selected but refuses to participate in the poll or chooses not to return the survey or they were called but chose not to answer the questions in the poll. In the non-response bias situation, the sample group does mirror the population, but a person or persons in the sample group refused to participate.

**Voluntary Response Bias** occurs when anyone is permitted to choose to respond to a general invitation such is the case with radio call-in surveys; write in surveys; internet polls. The problem with this type of data collection is that it tends to over-represent people with strong opinions because they are the only ones who care enough to answer the poll. The person in the survey chooses to participate rather than be randomly selected from a listing of the entire population. There is no method to choosing the sample group. The respondents choose themselves.

**Response Bias** results from questions that are embarrassing/sensitive or are Non-neutral or poorly worded questions which due to their phrasing tend to lead people to a particular response. We minimize the chance of response bias by carefully wording and field testing questions. If a question is sensitive or embarrassing the respondent may choose to lie. Poorly worded questions tend to influence the responses of individuals.

**Example of a sensitive question where someone might be prone to lie:**

Coach Ham asks the team if anyone one of them had stayed out partying that weekend. Of course, none of the athletes were.

**Examples where the wording of the questions influenced the response:**

Are you in favor of providing school lunches for children that do not have enough to eat so that they have a chance to have a normal life?

Are you in favor of the government raising your taxes and taking your money away to pay for food for people who don't work? Obviously, the two questions will result in different responses.

## Notes: Bias Scenarios

**Scenario 1:** Interested in determining the percent of the population who believed in God, a surveyor stood outside a church on Sunday morning and asked all of the congregates a neutrally worded question about whether or not they believe in God. Will this survey produce biased results? Explain.

It is reasonable to assume that most people who attend church believe in God. Consequently, the percentage of people who believe in God is probably over-represented and the people who do not believe in God are under-represented. This case of SELECTION BIAS—selecting a really poor sample is subject to the following biases:

**Undercoverage Bias:** Those that do not believe in God are not likely to be at church.

**Response bias:** If I am outside of a church I may feel pressured to say that I believe in God because of all of those around me

**2008 Question 2:** A local school board plans to conduct a survey of parents' opinions about year-round schooling in elementary schools. The school board contacts 500 of the families by mail. The survey question is provided below.

*A proposal has been submitted that would require students in elementary schools to attend school on a year-round basis. Do you support this proposal? (Yes or No)*

The school board received responses from 98 of the families, with 76 of the responses indicating support for year-round schools. Based on this outcome, the local school board concludes that most of the families with at least one child in elementary school prefer year-round schooling.

- a. What type of bias is included in this survey?
- b. Could we just contact another 500 families and add their data to the original results?
- c. Name 2 ways to address the bias in this survey?
  - a. **Non-response bias:** Only 98 out of 500 families or 19.6% responded to the survey. It is likely that the opinions of parents who are most in favor of year round schools were the ones who responded and there are over-represented in the survey.
  - b. **NO, this would not be reasonable.** Even after increasing the sample size the non-response bias would still be present because there are still 402 of 500 families that did not respond from the first survey. Remember you cannot fix biased data.
  - c. **1<sup>st</sup>** The school board could contact the 402 families who did not respond and try to get their responses.  
**2<sup>nd</sup>** The school board could take a completely new survey using phone calls or person to person contact to try to eliminate the amount of non-response bias.



## Notes: Sampling Error

Recall that we defined **statistics** as the study of how to collect, organize, display, describe, and analyze the data from a sample in order to make generalizations about a population parameter of interest. Also remember that the parameter is a numerical measurement of a population which could be the mean, median, variance, range, percent...etc. From the sample, or subset of the population, we find estimates of population parameters. **The estimate of a population parameter that comes from a sample is known as a sample statistic and is often referred to as a statistic.** Thus, we talk about the sample mean, or sample median or sample variance...etc. being estimates or approximations for the corresponding population parameters.

For instance, if I wanted to know the true number of hours that Reagan students sleep on average per week, I could find an estimate by taking a sample 50 Reagan students. If I took a "good sample" one that represents the population at Reagan, my findings should be pretty close to the actual number of hours that are slept per night. However, the value I calculated from my sample is likely to differ somewhat from the actual number of hours slept. The difference between my sample statistic and the population parameter is known as sampling error or sampling variation.

More specifically, **Sampling Error or Sampling Variation** is variation due to random chance and can be described by a probability model and can be minimized by increasing the size of the sample. Sampling error is the recognition that every randomly selected sample from a population is likely to be different. Because there are so many variables or factors within a population, it is nearly impossible to design studies that account for every variable without drawing a census, which in itself is difficult, if not impossible. Because we cannot account for all of the variables, we involve chance in the selection of our sample to equalize the effects of variables that could not be accounted for. Even with a well-designed study or experiment, **the sample statistic is only an estimate of the population parameter.** The difference between the two is known as sampling error.

Recall that when we ran our simulations, there were multiple approaches that were all considered acceptable and led to different responses. However as we increased the sample size, the amount of variation in our responses decreased. It makes sense that if we increase our sample size that our result would be closer to the true population parameter. We can say that as we increase our sample size that the difference between the population parameter and the sample statistic will be less and that leads us to say that **the amount of variation diminishes as our sample size increases.** Obviously, our observations may be more varied as the sample size increases, but on average the sample statistics will vary less.

**Caution:** Please **do not confuse sampling error with Bias.** Sampling error is a term used to describe properly collected data that didn't actually match the population. In other words, given a proper method of data collection was employed, I would expect that my sample will vary or differ from the population the amount of variation or difference of a sample from the population is error. Error is normal and is predictable and therefore can be controlled. **The best way to minimize sampling error/sampling variation is by increasing the sampling size.**

Unlike sampling error, bias is problem in the design and/or method of collecting data. Increasing sample size will not fix biased data. Biased data cannot be fixed and must be discarded. To eliminate bias, we must carefully plan and execute how we collect our data.



## Notes: Observational Studies & Randomness

Obviously, it is important to collect samples that are representative of the population and by now you should be asking yourself: "how do we do that?" One primary method used to collect samples is the Observational Study. Observational studies are studies in which we gather data about subjects doing what they are already doing. The research of historical records or medical records and surveys are examples of observational studies. Most data is collected through observational studies. Observational studies do NOT assign treatments. **Because treatments are not assigned, causation cannot be proved.** However, observational studies can be statistically significant and can demonstrate a relationship, association or a correlation.

**Note: Only Experiments should be used to establish cause and effect relationships.**

**Most data is collected by observational study and not by controlled experiments.**

A common type of observational study is the survey of which there are several types including:

- The Census
- The Convenience Sample
- The Simple Random Sample
- The Systematic Sample
- The Cluster Sample
- The Stratified Sample

**The Census:** As we discussed earlier, a census involves the collection of data from each unit in the population. Unfortunately, a census is often impractical if not impossible to perform. On the positive side, if we perform an accurate census, we will have the population parameter of interest and inference procedures will not be necessary.

**The Convenience Sample:** The convenience sample is a sampling technique which selects subjects because of their convenient accessibility and proximity to the researcher. Convenience samples are rarely representative of the population and not every subject in the population has an equal chance of being selected. As a consequence, convenience sample findings are most often of no value, however, they are easy to perform if you are interested in collecting worthless data. In other words, **avoid them.**

**The purpose of Randomness.**

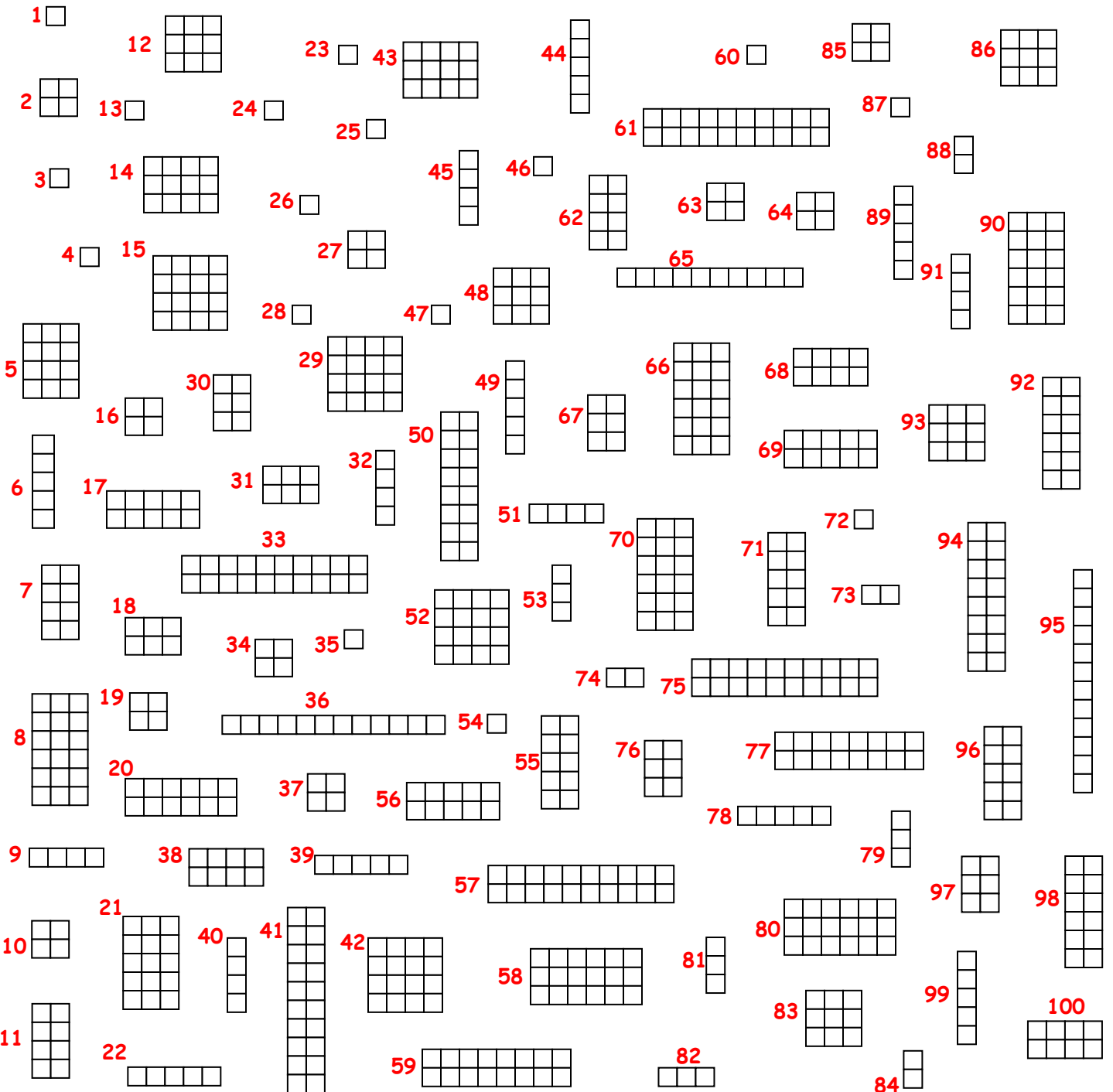
Recall that we want our samples to be representative of the population and that means we would desire that we would like every combination of traits to be proportionally represented in our sample. Unfortunately, if we really break it down, we are all unique individuals. Thus if we were to represent every combination of traits, we would need to perform a census and sample everyone. As we discussed previously, this is impractical if not impossible.

So how do we address the problem? The answer is the introduction of chance or randomness. Statisticians and mathematicians have found that randomness, tends to equalize the variations that we can't account for in our studies.

So you don't believe me. Try the exercise on the next page.

### Notes: Sample Surveys & Random Rectangles

So let's begin with your best guess as to the mean (average) number of squares in each figure? \_\_\_\_\_  
 Please use whole numbers.



**Judgment Sample** - Select a sample of 5 rectangles that, in your judgment, re representative of the rectangles. Record the size of your five rectangles and find their mean.

\_\_\_\_\_ mean = \_\_\_\_\_

Number each figure 1 to 100. The 100 figures make up our population of rectangles.

## Notes: Sample Surveys & Random Rectangles

So how do we determine the average number of rectangles per figure? Obviously, we could conduct a census and count the squares in each figure, and take an average. However, I am guessing that many of you would refuse to do the assignment and I am guessing that among those that did the assignment we would still have varying answers. So instead of conducting a census or just using our best guess or our judgment, let's utilize some sampling techniques that employ randomness. We are going to use samples of size 5 which are pretty small. We will begin with the simple random sample.

**The Simple Random Sample:** This is the basis for all statistical sampling and differs from census and the convenience sample in that randomness is utilized. In a simple random sample all subjects have an equal chance of being selected and every possible combination of subjects has an equal chance of being selected. That sentence is so important, that I am going to say it again: **In a simple random sample all subjects have an equal chance of being selected and every possible combination of subjects has an equal chance of being selected.**

So how do we conduct a Simple Random Sample? Typically a number is assigned to all subjects in the population and then a sample of numbers is randomly selected. Acceptable methods of selecting the numbers include: random number generators; mixed numbers in a hat; random number tables. If a **random number table** is used, it is necessary to describe the **selection process**, provide the **stopping point** and define how **duplicate values** are to be addressed. Unfortunately, it is often difficult to obtain a list of the entire population and there is a chance that our sample will not be representative of the population.

Use the random number generator on your calculator to select 5 numbers for our sample.

### Calculator Commands:

Press **MATH** → Highlight **PROB**  
 ↓ Highlight **8: randIntNoRep(**  
 Press **ENTER**  
**Lower: 1 Upper: 100 n: 5** Press **ENTER 2x's**

Locate the corresponding rectangles and record their size and find the mean.

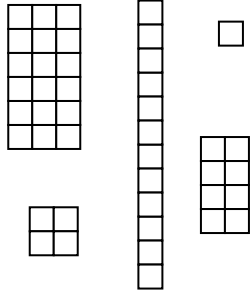
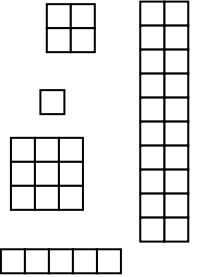
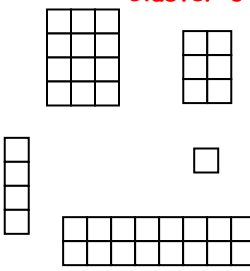
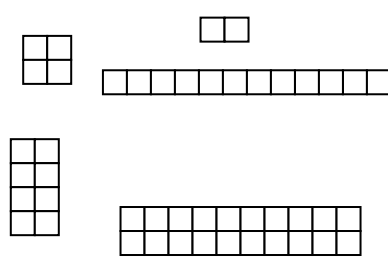
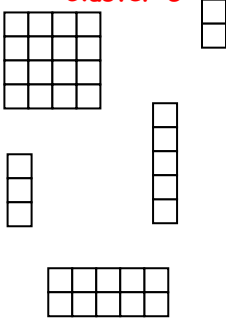
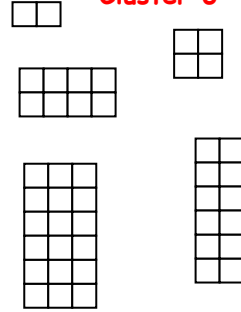
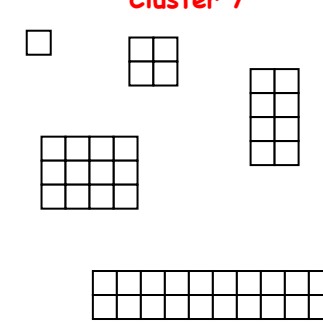
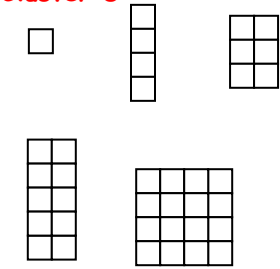
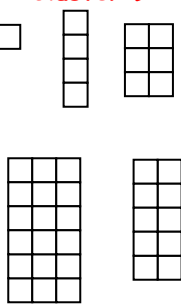
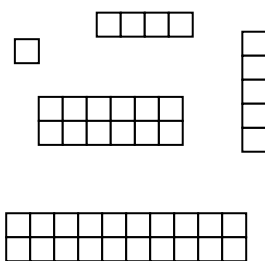
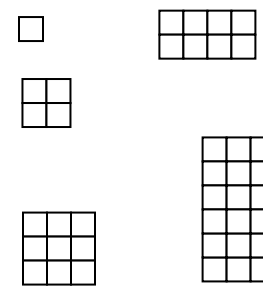
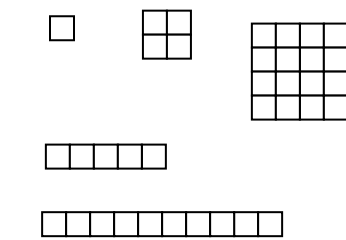
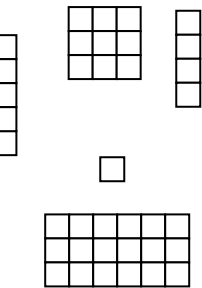
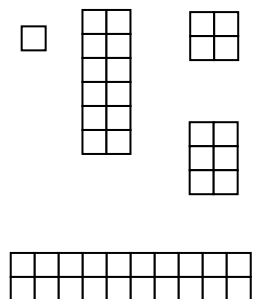
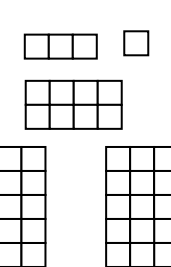
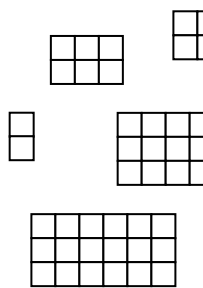
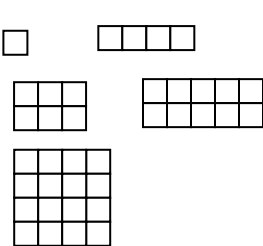
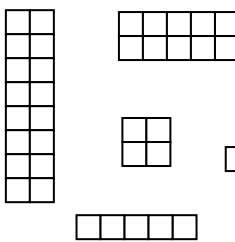
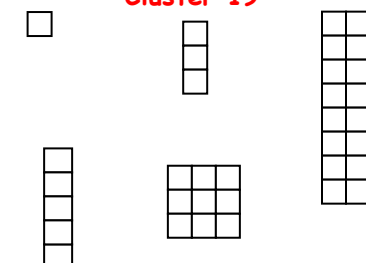
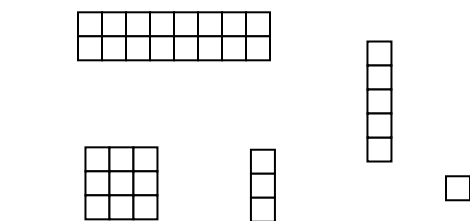
10	12	1	16	18		mean = <u>11.4</u>
69	95	26	50	58		

**The Systematic Random Sample:** The systematic sample also involves randomness and every individual has an equal opportunity of being selected, however every group is not possible. To perform a systematic random sample begin by dividing the population by your desired sample size which gives us  $n$ . We will be selecting every  $n$ th value. Systematic sampling is fast and easy and is a good method if there is no pattern in the population. We often use it when we want to survey people coming into a building (selecting every  $n$ th person) or every  $n$ th thing in a list.

In our case, we will divide 100 by 5 and get 20. Thus we will be selecting every 20<sup>th</sup> figure in the list. We will also use the  $n$  to find our starting point. We do this by using the random number generator and selecting one value 1 to 20. Please perform a systematic random sample.

4	5	16	3	5		mean = <u>6.6</u>
19	39	59	79	99		

### Notes: Sample Surveys & Random Rectangles

<p><b>Cluster 1</b></p> 	<p><b>Cluster 2</b></p> 	<p><b>Cluster 3</b></p> 	<p><b>Cluster 4</b></p> 	
<p><b>Cluster 5</b></p> 	<p><b>Cluster 6</b></p> 	<p><b>Cluster 7</b></p> 	<p><b>Cluster 8</b></p> 	
<p><b>Cluster 9</b></p> 	<p><b>Cluster 10</b></p> 	<p><b>Cluster 11</b></p> 	<p><b>Cluster 12</b></p> 	
<p><b>Cluster 13</b></p> 	<p><b>Cluster 14</b></p> 	<p><b>Cluster 15</b></p> 	<p><b>Cluster 16</b></p> 	<p><b>Cluster 17</b></p> 
<p><b>Cluster 18</b></p> 	<p><b>Cluster 19</b></p> 	<p><b>Cluster 20</b></p> 		

## Notes: Sample Surveys & Random Rectangles

**The Cluster Random Sample:** To perform a cluster sample, the population is broken into smaller groups or clusters that are representative of the population. In other words, each cluster is a microcosm of the population. Once the clusters are identified, we number them and then randomly select one or more clusters. Once a cluster is selected, we then perform a census within the cluster, which is to say, that within each selected cluster we survey or sample every subject.

Cluster sampling tends to be difficult to perform with humans. For example, how would you divide Reagan up if you wanted to perform a cluster sample? Consider how you might divide the state of Texas or San Antonio into clusters. Keep in mind that each cluster needs to mirror the population, which means that each cluster is very similar to the other clusters.

You might be thinking that cluster sampling isn't useful or important. However, there are several instances in which cluster sampling is important. Here are a three scenarios:

**Drawing Blood:** When a lab needs to test a person's blood for a disease. They draw a small sample and then place it under the microscope. That small sample of blood under the microscope is a cluster and is representative of all of the blood in someone's body.

**Baking:** When you make cookies or cakes or brownies, you mix all of the ingredients together thoroughly and then you take a spoonful and taste it and make adjustments as necessary. That spoonful is a cluster and is representative of all the batter that is being made. I certainly could use a chocolate chip cookie about now. Hint. Hint...

**Forrest Gump:** "Life is like a box of chocolates. You never know what you're going to get." However, my expectation is that every box of Russell Stover's Chocolate Truffles is the same. So if I wanted to perform a cluster sample, I would randomly select a box, (this would be my cluster). Now remember to perform a cluster sample, I must perform a census within the chosen group. To perform a census in this scenario would mean that I would need to take a bite out of every truffle. Oh how wonderful. Incidentally, I may need to increase my sample size which would meant that I would need more boxes of chocolate and would need to eat all of the chocolate. It's a tough job, but someone's got to do it.

The rectangles have been broken into 20 clusters of 5 which are representative of the population of rectangles. Number each cluster 1 to 20 and randomly select one of the clusters using the random number generator on your calculator

### Calculator Commands:

Press **MATH** → Highlight **PROB**

↓ Highlight **8: randIntNoRep(**

Press **ENTER**

**Lower: 1 Upper: 20 n: 1 Press ENTER 2x's**

Locate the corresponding cluster and record the number of rectangles in each figure and find the mean of your sample

4	6	2	12	18	
Cluster 16	Cluster 16	Cluster 16	Cluster 16	Cluster 16	mean = <u>8.4</u>

### Notes: Sample Surveys & Random Rectangles

Label each of the 5 sections of rectangles below based on the number of rectangles in each figure.

**Strata 1**

Strata 1 contains 20 numbered rectangles: 1 (1x1), 2 (1x1), 3 (1x1), 4 (1x1), 5 (1x1), 6 (1x1), 7 (1x1), 8 (1x1), 9 (1x2), 10 (1x2), 11 (1x2), 12 (1x2), 13 (1x2), 14 (1x2), 15 (2x1), 16 (2x1), 17 (2x1), 18 (1x1), 19 (1x1), 20 (1x1).

**Strata 2**

Strata 2 contains 20 numbered rectangles: 1 (2x2), 2 (2x2), 3 (2x2), 4 (1x4), 5 (2x2), 6 (2x2), 7 (2x2), 8 (1x4), 9 (1x4), 10 (2x2), 11 (2x2), 12 (1x4), 13 (1x4), 14 (1x4), 15 (2x2), 16 (1x4), 17 (1x4), 18 (1x4), 19 (1x4), 20 (2x2).

**Strata 3**

Strata 3 contains 20 numbered rectangles: 1 (1x4), 2 (1x4), 3 (2x2), 4 (2x2), 5 (2x2), 6 (1x4), 7 (2x2), 8 (1x4), 9 (2x2), 10 (1x4), 11 (1x4), 12 (2x2), 13 (2x2), 14 (1x4), 15 (2x2), 16 (2x2), 17 (1x4), 18 (1x4), 19 (2x2), 20 (1x4).

**Strata 4**

Strata 4 contains 20 numbered rectangles: 1 (1x8), 2 (2x2), 3 (2x2), 4 (2x2), 5 (2x2), 6 (1x4), 7 (1x4), 8 (1x4), 9 (2x2), 10 (1x8), 11 (1x4), 12 (2x2), 13 (1x4), 14 (2x2), 15 (2x2), 16 (1x8), 17 (2x2), 18 (2x2), 19 (2x2), 20 (1x4).

**Strata 5**

Strata 5 contains 20 numbered rectangles: 1 (2x2), 2 (2x2), 3 (1x4), 4 (1x8), 5 (1x4), 6 (2x2), 7 (1x4), 8 (2x2), 9 (2x2), 10 (1x4), 11 (2x2), 12 (2x2), 13 (2x2), 14 (2x2), 15 (1x4), 16 (1x4), 17 (2x2), 18 (1x8), 19 (1x4), 20 (1x8).

Number the figures in each section 1 to 20.

## Notes: Sample Surveys & Random Rectangles

**The Stratified Random Sample:** To perform a stratified sample the population is divided into groups called strata based on a common variable or characteristic. We then select a random sample from within each group. Although it is not required, typically, the size of the random sample within each group is proportional to the variables representation within the population. It is important to note that the strata are not determined randomly. Randomness is introduced through the selection within the strata. Please understand that although we take a simple random sample within each strata, a stratified random sample **IS NOT** a special case of the simple random sample.

We take stratified samples for two important reasons. First, a stratified random sample allows us to ensure that each of the subgroups of interest within a population are represented in the sample. Secondly, we are able to isolate the impact of the common characteristic that determined the groups. Please note that unlike cluster samples (where each cluster is a microcosm of the population), the individual strata or groups are not representative of the population.

Dividing groups by gender is one of the most common ways we stratify and we do this because we believe that in general there are differences in males and females. However, my go to example for stratified sample is our high school. For Instance, let's say that we were trying to choose a theme song for homecoming and instead of performing a census and surveying everyone, we decide to survey the opinions of 100 students. We could do a simple random sample, however we run a risk of having chosen all freshmen in the sample. I don't think the seniors would be very pleased to have freshmen making decisions for them.

However, if we divided Reagan by grade level (freshmen, sophomore, junior and senior) and then performed a simple random sample within each grade level. We would be assured of having every grade level represented. In addition, since we know that 28% of the school is comprised of freshmen, 26% are sophomores, 23% are juniors and 23% are seniors, we could proportionally represent each grade level and our sample would be representative of the Reagan population.

Use the random number generator on your calculator to select 5 numbers for our sample.

### Calculator Commands:

Press **MATH** → Highlight **PROB**  
 ↓ Highlight **8: randIntNoRep(**  
 Press **ENTER**  
**Lower: 1 Upper: 20 n: 1** Press **ENTER 2x's**

Locate the corresponding figure and record its size.

**Note:** Repeat the process 5 times (once for each strata)

Calculate the mean.

1	4	8	10	15	
<b>Strata 1</b> Element 3	<b>Strata 2</b> Element 11	<b>Strata 3</b> Element 5	<b>Strata 4</b> Element 20	<b>Strata 5</b> Element 6	mean = <u>7.6</u>

### Notes: Survey Scenarios

**2011 Question 3:** An apartment building has nine floors and each floor has four apartments. The building owner wants to install new carpeting in eight apartments to see how well it wears before she decides whether to replace the carpet in the entire building.

The figure below shows the floors of apartments in the building with their apartment numbers. Only the nine apartments indicated with an asterisk (\*) have children in the apartment.

<table border="1"> <tr><td>11*</td><td>12</td></tr> <tr><td colspan="2" style="text-align: center;">1st Floor</td></tr> <tr><td>14</td><td>13</td></tr> </table>	11*	12	1st Floor		14	13	<table border="1"> <tr><td>21</td><td>22*</td></tr> <tr><td colspan="2" style="text-align: center;">2nd Floor</td></tr> <tr><td>24</td><td>23*</td></tr> </table>	21	22*	2nd Floor		24	23*	<table border="1"> <tr><td>31</td><td>32</td></tr> <tr><td colspan="2" style="text-align: center;">3rd Floor</td></tr> <tr><td>34</td><td>33</td></tr> </table>	31	32	3rd Floor		34	33			
11*	12																						
1st Floor																							
14	13																						
21	22*																						
2nd Floor																							
24	23*																						
31	32																						
3rd Floor																							
34	33																						
<table border="1"> <tr><td>41</td><td>42</td></tr> <tr><td colspan="2" style="text-align: center;">4th Floor</td></tr> <tr><td>44</td><td>43</td></tr> </table>	41	42	4th Floor		44	43	<table border="1"> <tr><td>51*</td><td>52</td></tr> <tr><td colspan="2" style="text-align: center;">5th Floor</td></tr> <tr><td>54</td><td>53</td></tr> </table>	51*	52	5th Floor		54	53	<table border="1"> <tr><td>61</td><td>62</td></tr> <tr><td colspan="2" style="text-align: center;">6th Floor</td></tr> <tr><td>64</td><td>63</td></tr> </table>	61	62	6th Floor		64	63	<table border="1"> <tr><td colspan="2">* = Children in the Apartment</td></tr> </table>	* = Children in the Apartment	
41	42																						
4th Floor																							
44	43																						
51*	52																						
5th Floor																							
54	53																						
61	62																						
6th Floor																							
64	63																						
* = Children in the Apartment																							
<table border="1"> <tr><td>71</td><td>72</td></tr> <tr><td colspan="2" style="text-align: center;">7th Floor</td></tr> <tr><td>74*</td><td>73*</td></tr> </table>	71	72	7th Floor		74*	73*	<table border="1"> <tr><td>81</td><td>82</td></tr> <tr><td colspan="2" style="text-align: center;">8th Floor</td></tr> <tr><td>84*</td><td>83</td></tr> </table>	81	82	8th Floor		84*	83	<table border="1"> <tr><td>91</td><td>92*</td></tr> <tr><td colspan="2" style="text-align: center;">9th Floor</td></tr> <tr><td>94</td><td>93*</td></tr> </table>	91	92*	9th Floor		94	93*			
71	72																						
7th Floor																							
74*	73*																						
81	82																						
8th Floor																							
84*	83																						
91	92*																						
9th Floor																							
94	93*																						

- (a) For convenience, the apartment building owner wants to use a cluster sampling method, in which the floors are clusters, to select the eight apartments. Describe a process for randomly selecting eight different apartments using this method.

**Assign 1 to 1<sup>st</sup> floor, 2 to 2<sup>nd</sup> floor... and 9 to 9<sup>th</sup> floor. Place the numbers 1 through 9 in a hat and mix thoroughly. Draw 2 numbers without replacement. Survey all apartments on the floors represented by the two numbers.**

- (b) An alternative sampling method would be to select a stratified random sample of eight apartments, where the strata are apartments with children and apartments with no children. A stratified random sample of size eight might include two randomly selected apartments with children and six randomly selected apartments with no children. In the context of this situation, give one statistical advantage of selecting such a stratified sample as opposed to a cluster sample of eight apartments using the floors as clusters.

**With a cluster sample it would be possible to survey apartment clusters which have no children. The 3<sup>rd</sup>, 4<sup>th</sup> and 6<sup>th</sup> floor clusters do not have children. Using a cluster it is also possible to not select apartments that are adult only. Apartments with children are likely to have had more wear than apartments that have adults only. A stratified sample would ensure that apartments with children and those without were sampled.**

- (c) Why did we choose to stratify on apartments with and without children?

**We always stratify on the variable that we think will most impact our study. We were measuring carpet wear and we believe that having children impacts carpet wear differently than just having adults only**



## Notes: Survey Scenarios

**2010 Form B Question 2:** In response to nutrition concerns raised last year about food served in school cafeterias, the Smallville School District entered into a one-year contract with the Healthy Alternative Meals (HAM) company. Under this contract, the company plans and prepares meals for 2,500 elementary, middle, and high school students, with a focus on good nutrition. The school administration would like to survey the students in the district to estimate the proportion of students who are satisfied with the food under this contract.

Two sampling plans for selecting the students to be surveyed are under consideration by the administration. One plan is to take a simple random sample of students in the district and then survey those students. The other plan is to take a stratified random sample of students in the district and then survey those students.

- (a) Describe a simple random sampling procedure that the administrators could use to select 200 students from the 2,500 students in the district.

**Assign all students a number from 1 to 2500. Place the 2500 numbers in a bin and mix thoroughly. Draw 200 numbers without replacement and then survey the 200 students whose names correspond with the number drawn.**

- (b) If a stratified random sampling procedure is used, give one example of an effective variable on which to stratify in this survey. Explain your reasoning.

**Because of potential difference in food preferences as a student ages I would choose to stratify on school type. My three strata would be elementary, middle and high school. As people age their nutritional requirements and tastes often change, consequently food satisfaction is likely to differ by school type.**

- (c) Describe one statistical advantage of using a stratified random sample over a simple random sample in the context of this study.

**Stratifying based on school type, assures that satisfaction ratings are obtained from elementary, middle and high school other strategies could miss a type of school completely and that would be an issue if food satisfaction varied by age.**

### Notes: Survey Scenarios

**2013 Question 3:** An administrator at a large university wants to conduct a survey to estimate the proportion of students who are satisfied with the appearance of the university buildings and grounds. The administrator is considering three methods of obtaining a sample of 500 students from the 70,000 students at the university.

- (a) Because of financial constraints, the first method the administrator is considering consists of taking a convenience sample to keep the expenses low. A very large number of students will attend the first football game of the season, and the first 500 students who enter the football stadium could be used as a sample. Why might such a sampling method be biased in producing an estimate of the proportion of students who are satisfied with the appearance of the buildings and grounds?
- (b) Because of the large number of students at the university, the second method the administrator is considering consists of using a computer with a random number generator to select a simple random sample of 500 students from a list of 70,000 student names. Describe how to implement such a method.
- (c) Because stratification can often provide a more precise estimate than a simple random sample, the third method the administrator is considering consists of selecting a stratified random sample of 500 students. The university has two campuses with male and female students at each campus. Under what circumstance(s) would stratification by campus provide a more precise estimate of the proportion of students who are satisfied with the appearance of the university buildings and grounds than stratification by gender?

- (a) The sample may be biased because it is likely that the first 500 students arriving to the stadium are not representative of the entire population of students at the university. For example the 1<sup>st</sup> 500 might have substantially more school spirit and would likely give a greater percentage of favorable ratings for the buildings than the student body as a whole.
- (b) Assign each student a number 1-70,000. Use a random number generator to select 500 unique values. Duplicate values are to be ignored. The 500 students who correspond to those numbers will be interviewed.
- (c) We always stratify on the variable that we believe will have the greatest impact on the study. Because there are two separate campuses, it makes sense to stratify by campus rather than by gender if the buildings at each campus were built at a different time or are in different conditions or of different design.

## Notes: Experimental Design

Well-designed and conducted observational studies are a great way to collect data and are the most common type of data collection. A well-designed observational study can be statistically significant and demonstrate a strong relation or correlation. However, observational studies are limited and are unable to prove causation. Only well-conducted experiments have the power to prove causation.

So what is an experiment?

**An experiment:** is a controlled study that requires a treatment be assigned to one or more groups. The effects of the treatments on the response variable are then compared. Experiments must satisfy the principles of **control, random assignment, and replication.**

Given the simplicity of the definition: you might be inclined to wonder why more experiments are not conducted. You probably will guess two reasons right away. Yup, good old time and money. I never seem to have enough of either, although I did get a 1% pay raise this summer. Another reason experiments are not performed more often is ethics. Ethics? Remember the definition states that treatments must be assigned or imposed. Let's assume that I want to measure the impact of smoking tobacco cigarettes on lung function. I don't think it is difficult understand that it would be unethical for me to assign one group of people to smoke three packs of cigarettes while assigning another group to smoke three cigarettes a day and a third group to was assigned to no tobacco products at all.

Assuming that we have time, money and no ethical issues, what are the critical components to conducting a well-designed experiment?

The Critical Elements necessary for a well-designed experiment are: **random assignment, replication and control.**

**Random Assignment** is the principle of randomly assigning subjects or experimental groups to various treatment groups. We use random assignment to equal out or mitigate the differences in subjects or experimental units that we are unable to control in our design. In other words, humans have too many different traits to "control for" or account for in a study. To make up for this lack of control, we use random assignment to equalize those differences across the experimental groups which avoids any biased judgement in the placing of subjects into a group. **There must be at least 2 treatment groups.** One of the groups may be a control group. (discussed on the next page).

**Please note that random assignment is not random selection.** In random selection, we randomly selected subjects from the population of interest and observed the responses. In random assignment, we may actually have volunteers for the study, but we use randomization to assign the treatments to the volunteers.

**Replication** in science refers to the ability of duplicating the results. In statistics we refer to replication as the imposition of the treatment on multiple subjects in order to gain confidence that the result can be generalized across the population and is not a one person anomaly.

## Notes: Experimental Design

**Control** is the accounting for the impact of outside variables that are not of interest to us in our study. We design our experiments to remove or isolate the effects of external influences that we know will affect our study but are not of interest. We do this so that we can accurately measure the impact of the variables of interest. The following is a list of ways in which we exert control in our experiments: **Blocking, Blinding, and Control Groups.**

**Blocking:** Blocking is the separation of experimental units into groups with similar traits in order to control for the effect of outside variables that **we know** will have an impact or **confound** the results of the study. Confounding means that we are unable to determine which variable is causing the changes that we are observing. Blocking is not required, however if you block, always **block on the variable having the greatest impact** on the study.

**Note:** Blocking is to experiments as stratification is to surveys-**DO NOT** interchange the terms.

**Blinding:** In order to avoid a potential source(s) of bias, the subject or data collector or both are unaware as to the treatment that the subject is receiving. If both the subject and data collector are unaware of the treatment being assigned, the study is considered to be **double blind**.

**Control Group:** Typically a control group does not receive a treatment. The purpose of the control group is to provide a **baseline for comparison**. This allows us to determine whether the changes that are being observed can be attributed to uncontrolled for factors outside the parameters of our study. A **placebo** is often used as a method to establish a control group so that the subject and data collector believe that a treatment has been given.

The following scenario will be utilized to address the various methods utilized to control for the impact of outside variables that may impact the response, but are not the primary interest of the study.

**Scenario:** Researchers desire to test a new medication and treatment protocol for Alzheimer's disease. The new medication and treatment is thought to reverse the impact the disease has on memory function. Currently, there are no known side effects, but the drug has been only used on laboratory mice. As of this writing, no drug or treatment has been effective in reversing the effects of the disease.

**Blocking:** Because gender often impacts responses to medication, the researchers have chosen to block by gender.

**Single Blinding:** All of the patients in the study appear to be in the early stages of the Alzheimer's disease and are fully aware of what is going on. Consequently, the subjects are being told which treatment they are receiving as it would be depressing to learn that you were receiving a placebo or the old ineffective medication. The medication will be in containers marked A, B, and C and the patient will not be able to determine which medication they are receiving.

**Double Blinding:** The data collector will not be aware which medication the patient is receiving enabling the data collector to administer of the memory tests without bias.

**Control Group:** A Placebo is being administered to one of the patient groups. This will permit the researchers to be able to measure the progression of the disease and will let them know whether the changes being observed are due to the medication or are the result of some outside influence.

## Notes: Observational vs. Experiment

### Design an Observational Study and an experiment

**Example:** A study is to be designed to determine whether daily calcium supplements benefit women by increasing bone mass.

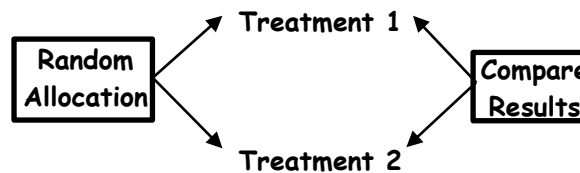
(a) How can an observational study be performed?

1. **Collect hospital records of women who have taken supplements and look at the changes in bone mass from the time they began taking supplements until some point in the future of the treatment.**
2. **Measure the bone mass of women currently taking the supplements and continue to observe them and take bone mass measurements**

**Note:** it would be wise to have a comparison group who were not taking the supplements

(b) How can an experiment be performed?

**Take a group of 100 women of the similar age and health status assign. Place all one hundred names into a bin and mix thoroughly the first 50 names drawn would receive the calcium supplements the remaining 50 would receive a placebo. A health professional would measure the bone mass of the women before taking the supplements and then three years after the supplements were taken. The bone mass of the women taken the supplements would then be compared to those who had not taken the supplements.**



### Which is More Appropriate a Study or Experiment?

**Example:** A study is to be designed to determine whether daily calcium supplements benefit women by increasing bone mass. Which is more appropriate an observational study or experiment and why?

**Either would be appropriate. However only the experiment could show causation. The experiment would take more time and cost more money.**

**Example:** A study is to be designed to examine the life expectancies of tall people versus those of short people. Which is more appropriate an observational study or experiment? Why?

**An observational study is more appropriate for ethical reasons in that we cannot assign people to be short or tall without placing them on the rack and stretching them or shortening them by cutting off body parts.**

**Example:** A study is to be designed to examine GPA's of students who use marijuana regularly and those who don't. Which is more appropriate an observational study or experiment? Why?

**An observational study is more appropriate for ethical reasons in that we cannot assign people to smoke marijuana.**

## Notes: Experimental Design Scenarios

**1999 Number 3:** The dentists in a dental clinic would like to determine if there is a difference between the number of new cavities in people who eat an apple a day and in people who eat less than one apple a week. They are going to conduct a study with 50 people in each group.

Fifty clinic patients who report that they routinely eat an apple a day and 50 clinic patients who report that they eat less than one apple a week will be identified. The dentists will examine the patients and their records to determine the number of new cavities the patients have had over the past two years. They will then compare the number of new cavities in the two groups.

- (a) Why is this an observational study and not an experiment?
- (b) Explain the concept of confounding in the context of this study. Include an example of a possible confounding variable.
- (c) If the mean number of new cavities for those who ate an apple a day was statistically significantly smaller than the mean number of new cavities for those who ate less than one apple a week, could one conclude that the lower number of new cavities can be attributed to eating an apple a day? Explain.

**(a) This is an observational study because:**

- No treatments were imposed
- The data used was existing data
- There was no random assignment of subjects to treatment groups

**(b) Confounding occurs when the effects of the variables cannot be distinguished from one another. In this case confounding would occur if it became impossible to measure the true effect on the number of cavities attributed to a group who eats more than one apple a week as compared to the number of cavities for a group that ate less than apple a week as the result of a related variable that is related to dental health. For instance, it is reasonable to assume that the group which consumes more than one apple a week has on an average a healthier overall lifestyle and that could be the cause for that groups having fewer cavities when compared to the group that eats less than one apple a week.**

**(c) Because this is not an experiment, it is not reasonable to draw a cause and effect relationship between eating apples and a lower number of new cavities.**

## Notes: Experimental Design Scenarios

**2006 B Number 5** When a tractor pulls a plow through an agricultural field, the energy needed to pull that plow is called the draft. The draft is affected by environmental conditions such as soil type, terrain, and moisture.

A study was conducted to determine whether a newly developed hitch would be able to reduce draft compared to the standard hitch. (A hitch is used to connect the plow to the tractor.) Two large plots of land were used in this study. It was randomly determined which plot was to be plowed using the standard hitch. As the tractor plowed that plot, a measurement device on the tractor automatically recorded the draft at 25 randomly selected points in the plot.

After the plot was plowed, the hitch was changed from the standard one to the new one, a process that takes a substantial amount of time. Then the second plot was plowed using the new hitch. Twenty-five measurements of draft were also recorded at randomly selected points in this plot.

- (a) What was the response variable in this study? Identify the treatments. What were the experimental units?
- (b) Given that the goal of the study is to determine whether a newly developed hitch reduces draft compare to the standard hitch, was randomization used properly in this study? Justify your answer.
- (c) Given that the goal of the study is to determine whether a newly developed hitch reduces draft compare to the standard hitch, was replication used properly in this study? Justify your answer.
- (d) Plot of land is a confounding variable in this experiment. Explain why.

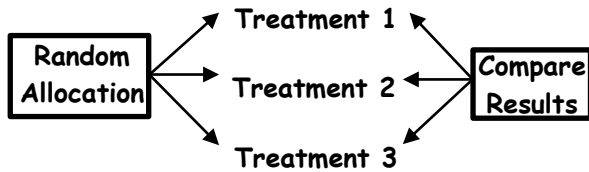
- (a) The response variable is the draft/energy needed to pull the plow. The treatments are the type of hitch that was used-the standard hitch and the new hitch. The experimental units are the 2 different fields.**
- (b) Randomization was properly applied because the two hitches, the treatments, were randomly assigned to the fields, the experimental units, which were plowed.**
- (c) Replication was not properly employed. Each treatment was applied to only one experimental unit. For replication to have occurred, the treatments/hitches would need to be used on multiple experimental units/plots of land.**
- (d) The plot of land is a confounding variable in this experiment because differing soil conditions between plots such as rockiness, soil compaction, root growth, and wetness will impact the draft. Thus, we are not able to determine whether the hitch or the plot is creating the differences in the draft. The plot is confounded with the hitch.**

# Recipe for Success: Completely Randomized Design

## Q 1.2

**Completely Randomized Design:** Subjects are assigned treatments by chance alone.

### 1. Draw the Diagram



- Draw the RA Box - Random Assignment
- Add Treatments - Label Accurately
- Draw the Compare Results Box

### 2. Explain the Random Allocation (Random Assignment)

- Place names in a hat, mix and draw
- Flip a Coin
- Assign a number and place in a hat, mix and draw
- Assign a number and use a random number generator (be certain to address repeats)

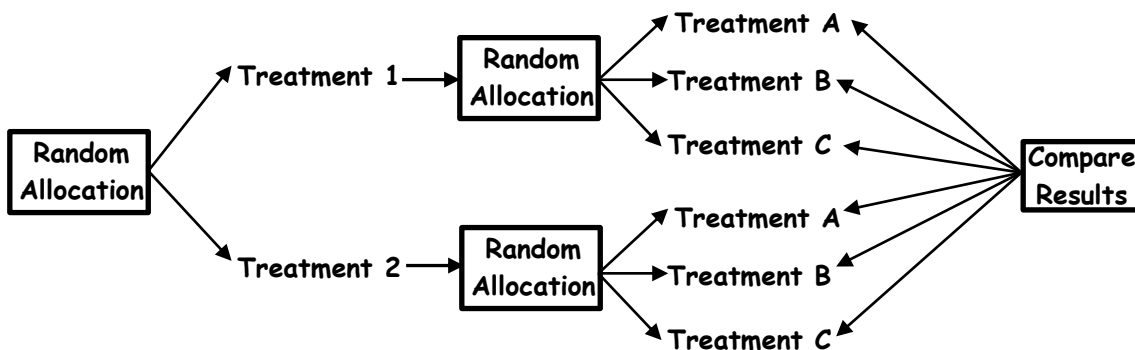
### 3. What is Being Assigned?

- Examine the Treatments
- Name Each Treatment Group (could be more than 1)  
**This is the Explanatory Variable(s) or Factor**

### 4. Identify the Results to Compare

- Specifically state what is being compared  
**This is the Response Variable**

## Multi-Variable Completely Randomized Design



**Explanatory Variables, Levels & Treatments**  
(Based on the Diagram)

- 1 Numerical Variable: 2 Levels 1 & 2
- 1 Letter Variable: 3 Levels A, B & C
- 6 Treatments:  $2 \times 3 = 6$  Treatments
  - Treatment 1 A
  - Treatment 1 B
  - Treatment 1 C
  - Treatment 2 A
  - Treatment 2 B
  - Treatment 2 C



## Notes: Experimental Design Scenarios

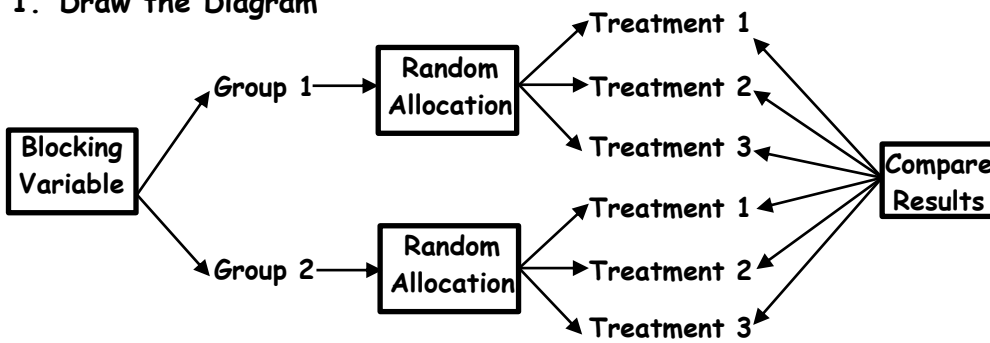
**2007 Question 2:** As dogs age, diminished joint and hip health may lead to joint pain and thus reduce a dog's activity level. Such a reduction in activity can lead to other health concerns such as weight gain and lethargy due to lack of exercise. A study is to be conducted to see which of two dietary supplements, glucosamine or chondroitin, is more effective in promoting joint and hip health and reducing the onset of canine osteoarthritis. Researchers will randomly select a total of 300 dogs from ten different large veterinary practices around the country. All of the dogs are more than 6 years old, and their owners have given consent to participate in the study. Changes in joint and hip health will be evaluated after 6 months of treatment.

- (a) What would be an advantage to adding a control group in the design of this study?
- (b) Assuming a control group is added to the other two groups in the study, explain how you would assign the 300 dogs to these three groups for a completely randomized design.
- (c) Rather than using a completely randomized design, one group of researchers proposes blocking on clinics, and another group of researchers proposes blocking on breed of dog. How would you decide which one of these two variables to use as a blocking variable?
- (a) Adding a control group to the study would allow researchers to determine whether the changes being observed were due to the treatments or due to variables outside of the study. For instance as dogs age, it is reasonable to assume that their joints will deteriorate. By having a control group we will know whether or not the treatments retarded the deterioration. In addition, a control group would allow us to determine whether joint pain might be attributed to weather. For instance, joints are typically less painful during warm months. So we wouldn't know if the dogs' joint pain responses were due to weather or the treatments if there were no control group.**
- (b) I would assign all of the dogs a unique number 1-300. I would place all of the numbers in a bin and mix thoroughly. The dogs associated with the first 100 numbers drawn, would receive glucosamine. The dogs that were associated with the next 100 numbers drawn would receive chondroitin. The remaining 100 dogs would be placed in the control group and would receive a placebo.**
- (c) The purpose of blocking is to create homogeneous experimental groups by controlling for the variable that would have the greatest impact on the study. I believe that clinics across the nation would have similar treatments and therefore I would NOT block by clinic. On the other hand, certain breeds of dogs are more susceptible to joint issues due to their size and genetics. Thus I would block by breed because I believe breed of dog would have a greater impact on the study than a clinic would**

# Recipe for Success: Randomized Block Design(s)

## Q1.2

### 1. Draw the Diagram



1. Draw the Blocking Variable Box
2. Label Groups Accurately
3. Draw Random Assignment Boxes
4. Label Treatments
5. Draw the Compare Results Box

### 2. Explain The Block

- Define the variable that is being blocked
- List the Groups in the Block

### 3. Explain the Random Allocation (Random Assignment)

- Place names in a hat, mix and draw
- Flip a Coin
- Assign a number and place in a hat, mix and draw
- Assign a number and use a random number generator (be certain to address repeats)

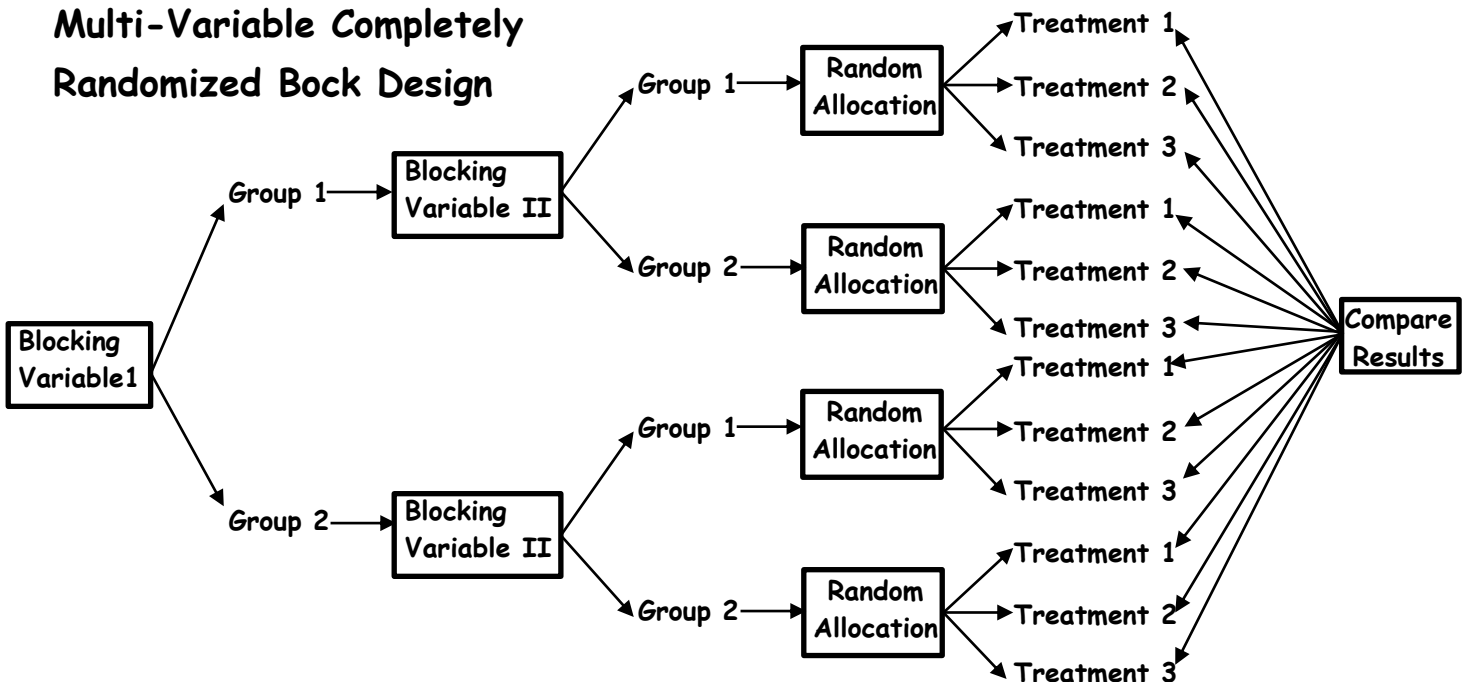
### 4. Identify the Results to Compare

- Specifically state what is being compared  
**This is the Response Variable**

### 5. What is Being Assigned?

- Examine the Treatments
- Name Each Treatment Group (could be more than 1)  
**This is the Explanatory Variable(s) or Factor**

## Multi-Variable Completely Randomized Block Design



Explanatory Variables, Levels  
Treatments  
(Based on the Diagram)

- Blocking Variable 1: 2 levels
- Blocking Variable II: 2 Levels
- Numerical Variable: 3 levels
- Treatments:  $2 \times 2 \times 3 = 12$  Treatments

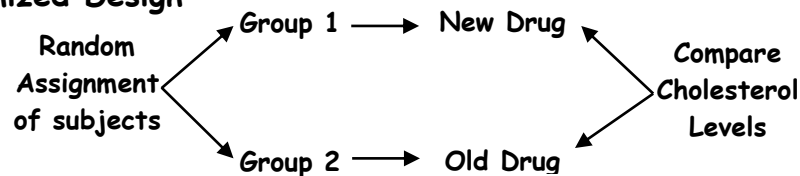
## Notes: Experimental Design Scenarios

**2000 Number 5** High cholesterol level in people can be reduced by exercise or by drug treatment. A pharmaceutical company has developed a new cholesterol-reducing drug. Researchers would like to compare its effects to the effects of the cholesterol-reducing drug that is currently available on the market. Volunteers who have a history of high cholesterol and who are currently not on medication will be recruited to participate in a study.

- Explain how you would carry out a completely randomized experiment for the study.
- Describe an experimental design that would improve the design in (a) by incorporating blocking.
- Can the experimental design in (b) be carried out in a double blind manner? Explain.

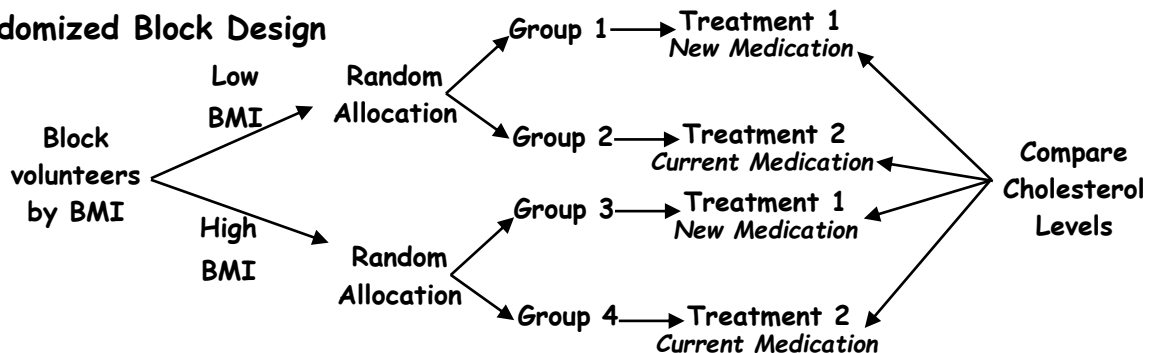
(a) Place the names of the volunteers in a bin and mix. Draw 50% of the names and assign to the old cholesterol drug protocol. The remaining 50% is assigned to the new cholesterol drug protocol. After an appropriate time compare the cholesterol levels between the 2 groups to determine which drug had been successful in lowering cholesterol the most

### Completely Randomized Design



(b) Block based on Body Mass Index (BMI) because those with a low BMI may respond differently than those with a high BMI. Randomly assign subjects within each block to the different drugs such that half of those in the low BMI group are assigned to the old drug and the remaining half are assigned to the new drug. This can be achieved by placing the names of those in the low BMI group in a bin and mixing. The first 50% would be assigned to the new drug and the remaining 50% would be assigned to current medication. Those in the high BMI group will also be randomly assigned so that half receive the new drug with the other half receiving the old drug. The random assignment can be achieved by placing the names of those in the high BMI group in a bin and mixing. The first 50% would be assigned to the new drug and the remaining 50% would be assigned to current medication. After an appropriate time cholesterol levels would be compared to determine which medication had been more effective in lowering cholesterol.

### Randomized Block Design



(c) Yes, this experiment can be carried out in a double blind manner. Label all the containers of the new drug A and the containers of the old drug B so that neither the subject nor the person administering the drug knows which drug had been administered. Obviously, a controller would know at the end of the test which drug had been given so that a conclusion can be drawn.

## Recipe for Success: Matched Pair Design (Each Group/Individual receives 2 Treatments)

### Q1.2

#### 1. Draw the Diagram



1. Draw the Random Assignment Box

2. Add Treatments ( I and II)  
Label Accurately

3. Add Treatments ( II and I)  
Label Accurately

4. Draw the Compare Results Box

#### 2. Explain the Reason for utilizing a Matched Pairs Design

- To control for the variation within an individual or subject
- Give a real life reason to control for variation by individual
  - I. **Lifestyle differences**  
(Water Proof Boots)
  - II. **Significant Variation within each person**  
  
(Susceptibility to Mosquito Bites)  
(Reaction to Medication)  
(Reaction time based on dominant and non-dominant hand)
  - III. **Control for unknown variables**  
(Each plot of land receives 2 plants)

#### 3. Explain the Random Allocation (Random Assignment)

- Place names in a hat, mix and draw
- Flip a Coin
- Assign a number and place in a hat, mix and draw
- Assign a number and use a random number generator  
(be certain to address repeats)

#### 4. Identify the Results to Compare

- Specifically state what is being compared  
**This is the Response Variable**

#### 5. What is Being Assigned?

- Examine the Treatments
- Name Each Treatment Group
- Give the order of the Treatments within each group  
**This is the Explanatory Variable(s) or Factor**

## Notes: Experimental Design Scenarios

**2005 B Number 3** In search of a mosquito repellent that is safer than the ones that are currently on the market, scientists have developed a new compound that is rated as less toxic than the current compound, thus making a repellent that contains this new compound safer for human use. Scientists also believe that a repellent containing the new compound will be more effective than the ones that contain the current compound. To test the effectiveness of the new compound versus that of the current compound, scientists have randomly selected 100 people from a state.

Up to 100 bins, with an equal number of mosquitoes in each bin, are available for use in the study. After a compound is applied to a participant's forearm, the participant will insert his or her forearm into a bin for 1 minute, and the number of mosquito bites on the arm at the end of that time will be determined.

(a) Suppose this study is to be conducted using a completely randomized design. Describe a randomization process **and identify an inference procedure for the study.**

**(b)** Suppose this study is to be conducted using a matched-pairs design. Describe a randomization process **and identify an inference procedure for the study.**

(c) Which of the designs, the one in part (a) or the one in part (b), is better for testing the effectiveness of the new compound versus that of the current compound? Justify your answer.

**(a) Place all 100 names of the participants in a hat and mix. The first group of 50 names will be assigned to the current compound. The remaining group of 50 names will have the new compound applied to their arm. The bins of mosquitoes will be numbered 1-100. Each participant will draw a number from a hat containing the numbers 1-100. The participant will then place their arm in the bin associated with that number for 1 minute. The number of bites for each treatment group will be compared to determine which compound worked the best. The average number of bites for each treatment group will be computed. The differences in the means of the two treatment groups will be compared**

**An appropriate test would be a two sample T-test with  $H_0: \mu_1 = \mu_2$  and  $H_A: \mu_1 > \mu_2$ ; where  $\mu_1$  is the mean number of bites for the group receiving the current repellent and  $\mu_2$  is the mean number of bites for the group receiving the new repellent.**

**(b) Each participant will be randomly assigned to a bin as above. First 50 names drawn from a hat will have the new repellent applied to their right arm and the current repellent applied to their left. The remaining 50 subjects will have the new repellent applied to their left arm and the current repellent applied to their right. Those who are assigned to an odd number will place their right arm in the bin first while those assigned to an even number bin will place their left arm in first. After one minute with the first arm, the second arm is placed in the bin. Then the difference in bites in each arm will be noted.**

**A paired t-test would be the most appropriate method of inference**

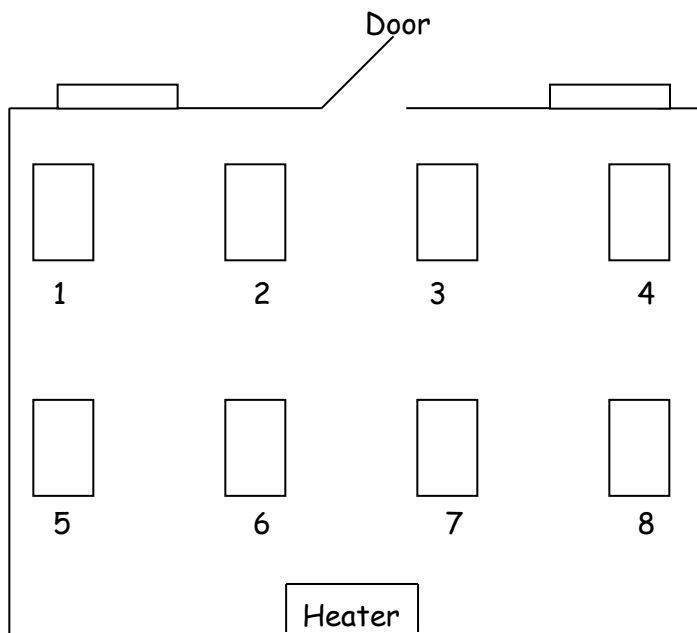
**$H_0: M_d = 0$  and  $H_A: M_d > 0$  where  $M_d$  is the difference between the old - new repellent.**

(c) The matched pairs design in part b is better than the completely randomized design because the matched pairs design helps control for potential sources of variation from person to person such as an individual's susceptibility to mosquito bites.

## Notes: Experimental Design Scenarios

**1997 Question 2:** A new type of fish food has become available for salmon raised on fish farms. Your task is to design an experiment to compare the weight gain of salmon raised over a six-month period on the new and old types of food. The salmon you will use for this experiment have already been randomly placed in eight large tanks in a room that has a considerable temperature gradient. Specifically, tanks on the north side of the room tend to be much colder than those on the south side. The arrangement of tanks is shown on the diagram below.

Describe a design for this experiment on weight gain that takes into account this temperature gradient.



The sources of heat gain or loss are the windows, the door, the heater and the north/south sides of the room. I would block the tanks so that each tank in the block would be equidistant from the source of heat gain or loss.

Block A Tanks (1,4)

Block B Tanks (2,3)

Block C Tanks (5,8)

Block D Tanks (6,7)

These blocks are the most homogenous blocks in regards to heat gain or loss.

For each Blocked pair, I would flip a coin to assign the fish food. If the coin showed heads, then the tank with the smaller number would receive the new fish food and the tank with the larger number would receive the older fish food. If the coin showed tails, then I would assign the older fish food to the smaller numbered tank in the block and the new fish food to the larger numbered tank.

After an appropriate amount of time, I would compare the average weight gain of the salmon in the block pairs to determine which food enhanced growth the most.

### Notes: Qualitative Data

Now that we have collected the data, how do we make sense of it? First, we need to determine the type of data that was collected. Is the data qualitative or is it quantitative?

**Qualitative Data/ Categorical Data:** Data that is used to describe an attribute (color, smell, how much pain you feel). **Can categorical data be numerical?**

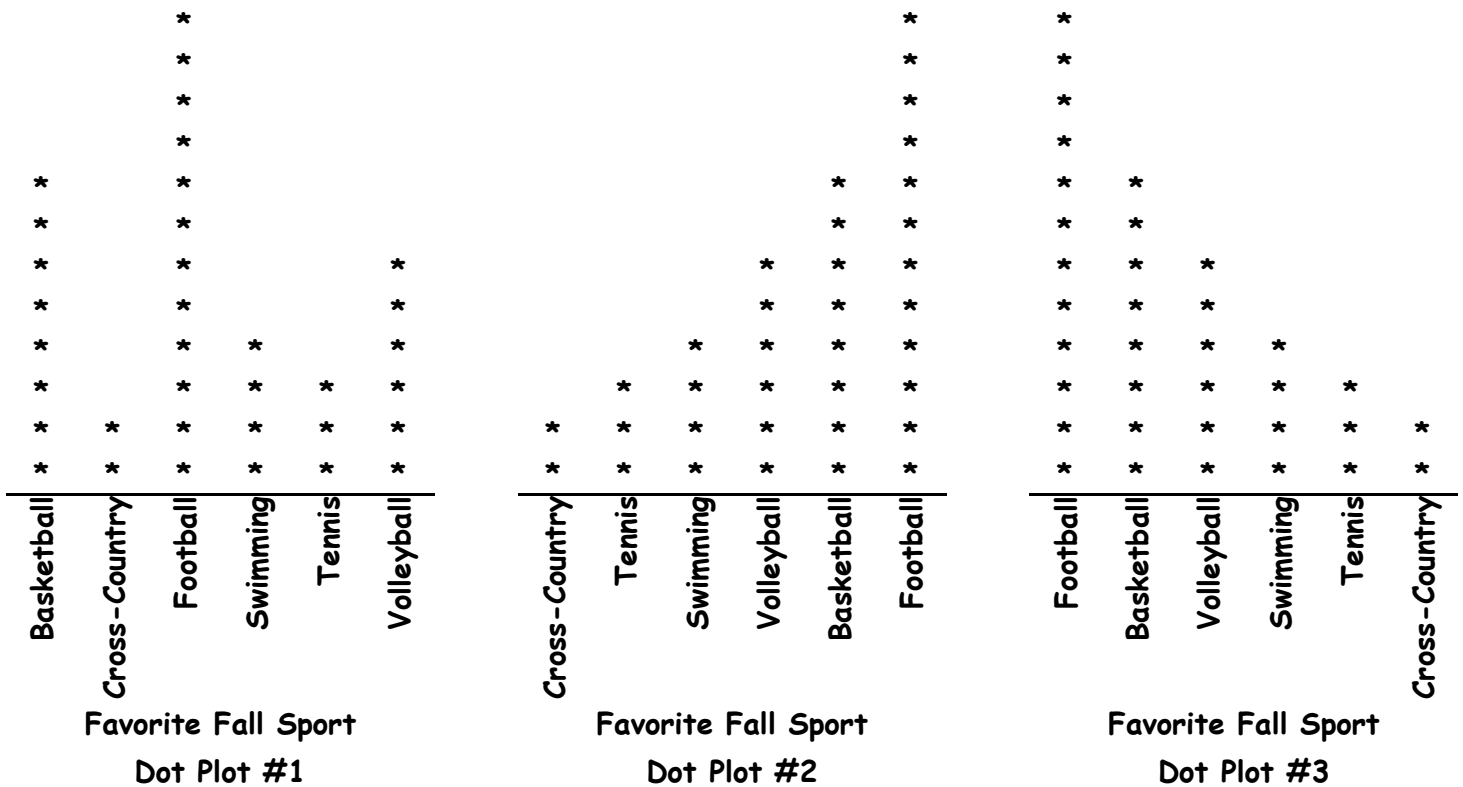
**Quantitative Data:** Data that can be measured and expressed numerically (example: height, weight, g.p.a, age. etc).

We will begin the section on displaying and describing data talking about qualitative data only because there it is less involved. Because qualitative data is not numerical in nature, we are unable to talk about the mean (average), median (the middle) or the range (largest-smallest). We may talk about the mode or modes which is the category(s) that occurs most often.

You have heard it said, that a picture is worth a thousand words and that is definitely the case with data. In the case of qualitative data the pictures or graphs that are primarily used are: dot-plots, bar-charts, pie-charts and segmented bar-charts.

**Scenario 1:** 35 students were asked to choose their favorite fall sport. The selections were as follows: swimming (4), basketball (8), football (12), tennis (3), volleyball (6) and cross-country (2).

Three dotplots have been created from the collected data. Which is correct and why?



Does the order matter for qualitative data? Why or why not?

**Note:** For categorical/qualitative data **DO NOT** describe shape, center, spread, or unusual features. **DO** talk about the mode-the group or category with the most.

Create three bar-charts for the scenario above. **Note:** the columns should not touch on bar charts.

## Notes: Types of Categorical Displays

Regardless as to whether or not a graph represents categorical or quantitative data, it is important to not only accurately display the data, but to provide a **title**, **label the axes** and **provide a key** or legend. Failure to label has resulted in zero points being awarded for an otherwise perfectly drawn display. In addition to the dotplot displayed on the prior page, the following is a list of some of the more common categorical displays.

**Bar Chart:** A graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. Because Bar charts represent categorical data, order does not matter and the bars should **NOT** touch one another. Bar lengths may convey actual data counts or they may represent relative frequencies/percents. If the bars represent actual counts or frequencies, then we know the exact number represented by the bar in the chart. However, if the bar chart is displaying relative frequencies (Percents), then we can look at the area of each bar and know its proportion of the sample relative to the other bars, but we will not have insight as to the actual number contained within the bar of the sample.

**Bar Chart vs. Histogram:** A histogram looks like a bar chart but is used to display quantitative data where order does matter and the bars must touch one another. We will spend more time discussing histograms in our discussion about quantitative data.

**Pie Charts:** A circle graph which is sub-divided into sectors based on their proportion to the whole. Please be aware that these are relative frequencies (proportions) and do not provide insight as to the number/size of the sample.

**Segmented bar chart:** A stacked bar chart in which each column is divided into segments which are proportional in size to that segment's representation within the population. Please be aware that these are relative frequencies (proportions) and do not provide insight as to the number/size of the sample.

**Independence:** Two events are considered to be independent, if the probability of one does not impact the probability of the other. If categories are independent then they are proportional. **Note:** We will continue develop the concept of independence throughout the remainder of this course.

**Scenario 2:** The following table provides the results which relates a person's IQ to a person's dominant hand. The table is illustrative of the proportional nature of independent categories.

**Actual Counts**

	Right Handed	Left Handed	Total
High IQ	190	10	200
Normal IQ	1,710	90	1800
Total	1900	100	2000

**Proportionality**

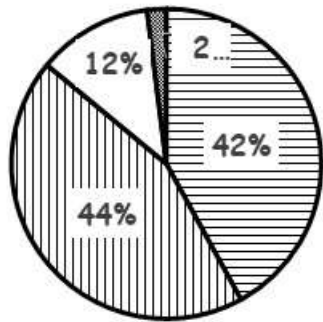
	Right Handed	Left Handed	Total
High IQ	.1	.1	.1
Normal IQ	.9	.9	.9
Total	1.0	1.0	1.0



**Notes: Pie Charts vs. Bar Charts**

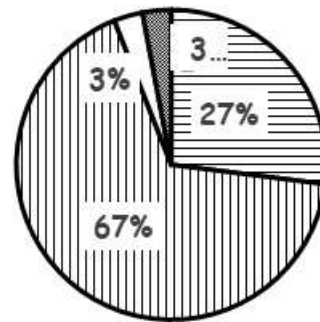
**Scenario 3:** The bar & pie charts represent a 2017 Pew Research poll of political affiliation by race.

**Party Affiliation Asian**



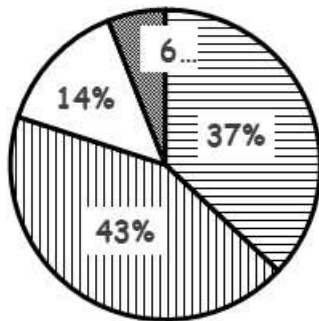
Independent   
  Democrat  
 Republican   
  Other

**Party Affiliation Black**



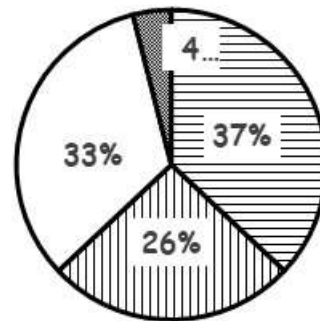
Independent   
  Democrat  
 Republican   
  Other

**Party Affiliation Hispanic**

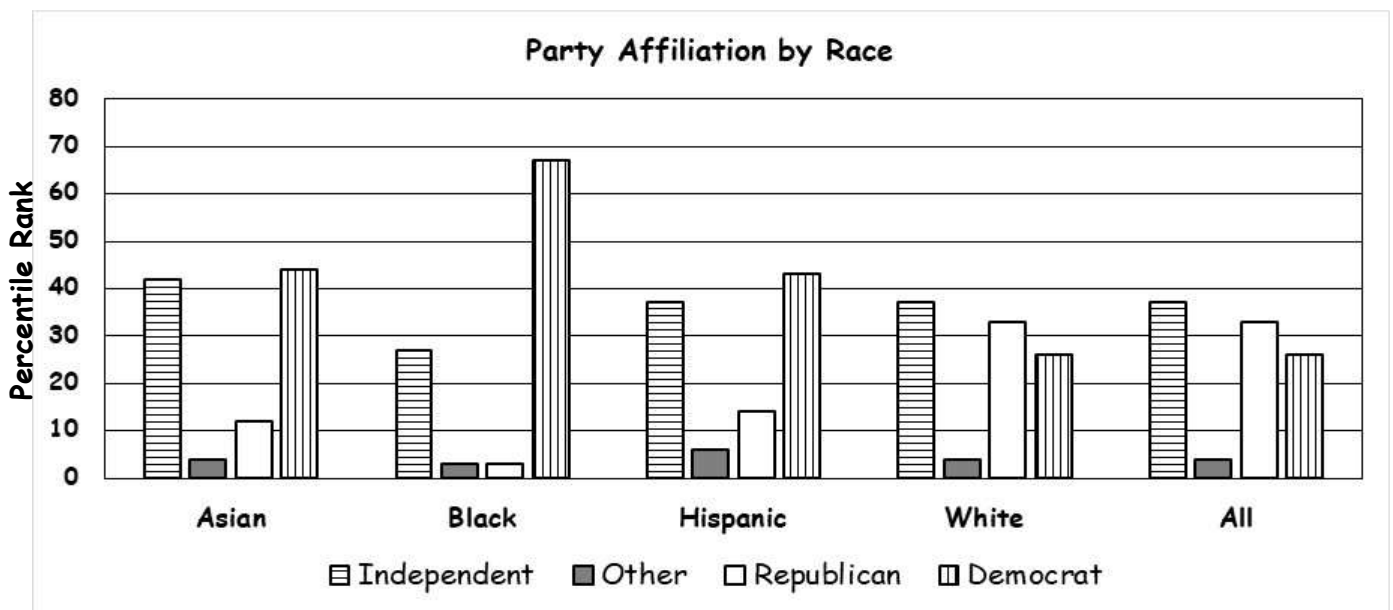


Independent   
  Democrat  
 Republican   
  Other

**Party Affiliation White**



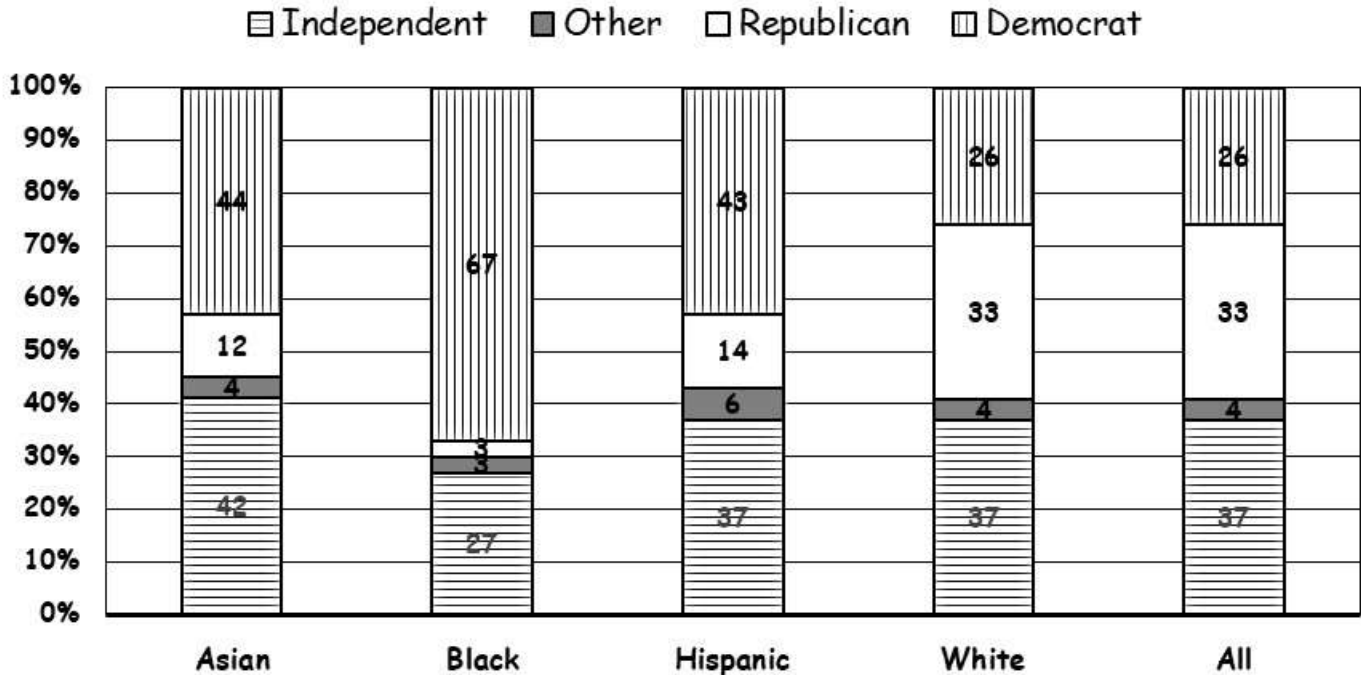
Independent   
  Democrat  
 Republican   
  Other



### Notes: Segmented Bar Charts

**Scenario 3 Continued:** The following segmented bar chart represents the same 2017 Pew Research poll of political party affiliation by ethnicity/race as the pie chart.

**PARTY AFFILIATION BY RACE**



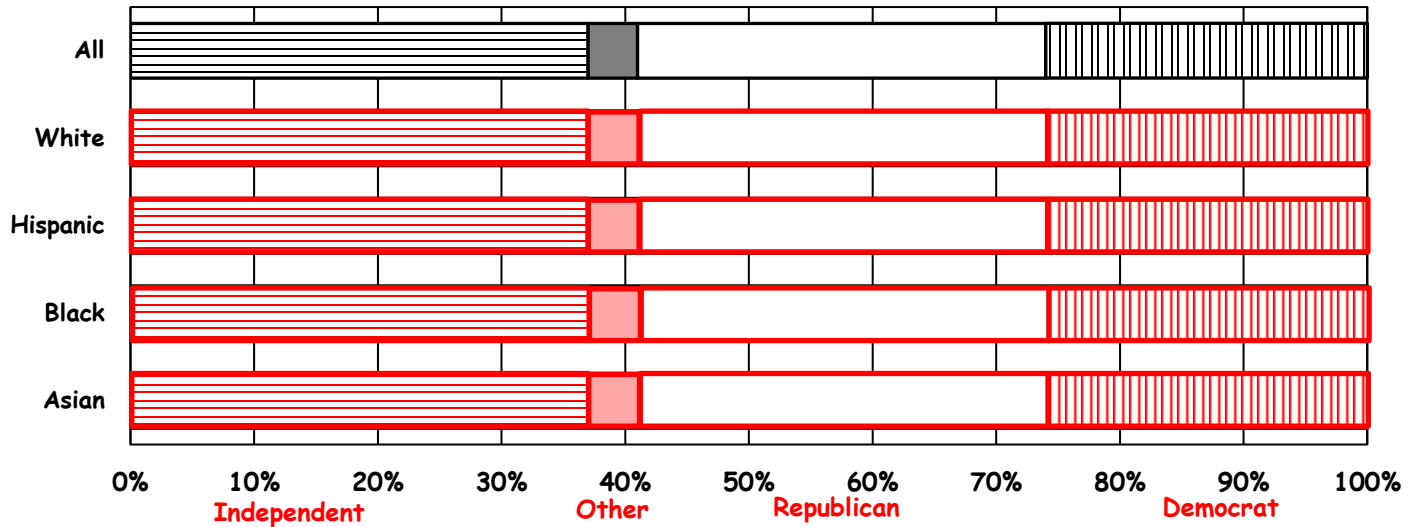
**Questions:** Using the segmented bar chart above to answer the following:

- Can we say that there are more than twice as many Black Democrats as there are Black Independents? Justify your response.  
**Yes we can. 67% of blacks are Democrat and only 27% of blacks are Independent.**
- Can we say that there are twice as many Black Democrats as there are White Republicans? Justify your response.  
**No we cannot. 26% of whites are Democrat and 67% of Blacks are Democrat, but we don't know how many White People vs. Black People there are.**
- Can we say that the number of Hispanic independents is about the same as the number of White Independents? Justify your response.  
**No we cannot. Although the percentages are similar, we do not know how many people are White nor do we know the number of Hispanics.**
- "Does political affiliation appear to be related to race?" Another way to ask the question is: "Are political party affiliation and race independent of each other?" More simply put, if the events are independent then the distribution of political affiliation would be the same for each race.  
**The distributions are different. The percentages are not the same. Party affiliation appears to be related to race. Race and party affiliation are not independent.**

**Notes: Independence**

**Scenario 3 Continued:** Based on the 2017 Pew Research data, party affiliation and race are not independent. However, for this exercise, assume that they are and the distribution of party affiliations follows that of the "All Race" distribution. Create a key and fill in the segments for each race.

**Party Affiliation By Race**



**Scenario 5:** Find the table value that results in perfect independence.

133	95
105	X

**Remember:** If the variables are independent they will have proportional values.

Hence:  $\frac{133}{105} = \frac{95}{x}$

$x = 75$

**Scenario 4:** Given that grades in statistics are independent of gender, complete the table below.

	Female	Male	Total
A's	30	S	54
B's	50	T	V
C's	20	U	W
Total	X	Y	Z

$X = 30 + 50 + 20 \rightarrow X = 100$

$S = 54 - 30 \rightarrow S = 24$

$\frac{30}{50} = \frac{S}{T} \rightarrow \frac{30}{50} = \frac{24}{Y} \rightarrow T=40$

$V=50+T \rightarrow V=50+40 \rightarrow V=90$

$\frac{30}{20} = \frac{S}{U} \rightarrow \frac{30}{20} = \frac{24}{U} \rightarrow U=16$

$W=20+U \rightarrow W=20+16 \rightarrow W=36$

$\frac{30}{X} = \frac{S}{Y} \rightarrow \frac{30}{100} = \frac{24}{Y} \rightarrow Y=80$

$Z=X+Y \rightarrow Z=100+80 \rightarrow Z=180$

## Notes: Two-Way Contingency Tables

**Two Way Contingency Tables** are useful in showing the relationships between 2 Categorical Variables and provide the frequency of occurrence (distribution) across each variable. This is only a brief introduction to 2-way tables. A thorough investigation of 2-way tables will be conducted during the probability section of this course.

**Marginal Frequencies:** Are the number of times a single variable resulted in a particular outcome. A Marginal frequency is the total of a single row or single column.

**Joint Frequencies:** Represent the intersection of a row and a column and provide the number of times that two separate variables yielded a particular outcome (this and that) both occurred. The key word for a joint frequency is **AND**.

**Scenario 6:** The table below provides the frequencies and relationship between Cholesterol level & Heart Attack Severity. Complete the table.

	Low Cholesterol	Medium Cholesterol	High Cholesterol	Total
Non-Fatal Heart Attack	29	17	18	64
Fatal Heart Attack	19	20	9	48
Total	48	37	27	112

**Questions:** Complete the table and answer the following:

How many variables are there? Name them

**2 Variables Heart Attack Severity and Cholesterol Level**

How many levels are there for each variable? Name them with their associated variable.

**Heart Attack has 2 -levels Non-Fatal and Fatal**

**Cholesterol has 3-levels Low, Medium, and High**

What is the sample size represented in the table?

**The sample size is 112**

List the joint frequencies.

**Non-Fatal, Low=29;      Non-Fatal, Medium=17;      Non-Fatal, High=18**

**Fatal, Low=19;              Fatal, Medium=20;              Fatal, High=9**

List the marginal frequencies.

**Low=48;                      Medium=37;                      High=27**

**Non-Fatal=64;              Fatal=48**

Are the variables independent? Justify your response?

**Cholesterol Level & Heart Attack severity are not independent there is a relation/association. The proportions are not the same:  $\frac{29}{48} \neq \frac{17}{37} \neq \frac{18}{27} \neq \frac{64}{112}$**

## Notes: Two-Way Contingency Tables

**Scenario 6 Continued:** Please use the table on page 42 to answer the following:

What percent of people in the sample had a fatal heart Attack? Is this a marginal or joint probability?

$$\frac{48}{112} \text{ or } 42.9\%. \text{ This is a marginal probability}$$

What percent of people in the sample have low Cholesterol? Is this a marginal or joint probability?

$$\frac{48}{112} \text{ or } 42.9\%. \text{ This is a marginal probability}$$

What percent of people in the study had a high cholesterol level and had a non-fatal heart attack? Is this a marginal or joint probability?

$$\frac{18}{112} \text{ or } 16.1\%. \text{ This is a joint probability}$$

What percent of people had a medium cholesterol level and had a fatal heart attack? Is this a marginal or joint probability?

$$\frac{20}{112} \text{ or } 17.9\%. \text{ This is a joint probability}$$

**Conditional probability** means that we have added a condition or restriction. As a consequence of the condition we are looking at a portion of the sample and are no longer looking at the entire sample. **Given** is the key word used to identify a conditional restriction and that word lets us know which marginal frequency we are being limited to. Because we are no longer using the total sample the denominator becomes the marginal frequency and not the sample total.

$$\frac{\text{Joint Frequency}}{\text{Marginal Frequency}} = \frac{\text{"and" "both"}}{\text{Given}} = \frac{\text{intersection}}{\text{Given}}$$

Given that a person had a fatal heart attack, what is the probability that they had high cholesterol?

$$\frac{\text{Fatal \& High}}{\text{Fatal}} = \frac{9}{48} \text{ or } 18.8\%.$$

Given that a person had a low cholesterol, what is the probability that they had a fatal heart attack?

$$\frac{\text{Fatal \& Low}}{\text{Low}} = \frac{19}{48} \text{ or } 39.6\%.$$

Given that a person had a moderate cholesterol level, what is the probability that they had a non-fatal heart attack?

$$\frac{\text{Non Fatal \& Moderate}}{\text{Moderate}} = \frac{17}{37} \text{ or } 45.9\%.$$

Please complete the table using percentages instead of frequencies

	Low Cholesterol	Medium Cholesterol	High Cholesterol	Total
Non-Fatal Heart Attack	25.9%	15.2%	16.1%	57.1%
Fatal Heart Attack	17%	17.9%	8%	42.9%
<b>Total</b>	<b>42.9%</b>	<b>33%</b>	<b>24.1%</b>	<b>1.0</b>

## Notes: Simpson's Paradox

**Simpson's Paradox:** when the results from combined grouping appears to contradict the results from the individual groupings. Simpson's Paradox arises when two or more sub-groups are combined to form a single group and there exists significant differences in the sizes of the sub-groups and the proportions in each group differ.

Test results for students from two teachers at a low performing school are summarized in the table below. Based on the table below make a determination as to which teacher is better at helping their students succeed. Justify your response.

	Teacher A	Teacher B	Total
Passed	130	125	<b>255</b>
Failed	15	20	<b>35</b>
Total	<b>145</b>	<b>145</b>	<b>290</b>

**Teacher A appears to be better than Teacher B.**

**Teacher A had  $\frac{130}{145}$  or 89.6% pass**

**Teacher B had  $\frac{125}{145}$  or 86.2% pass**

Further test score information became available in which the students were broken into two groups, the group that failed the prior year's test and those that had passed the prior year's test.

Passed the Prior Year's TAKS Test

	Teacher A	Teacher B	Total
Passed	110	30	<b>140</b>
Failed	5	0	<b>5</b>
Total	<b>115</b>	<b>30</b>	<b>145</b>

Failed the Prior Year's TAKS Test

	Teacher A	Teacher B	Total
Passed	20	95	<b>115</b>
Failed	10	20	<b>30</b>
Total	<b>30</b>	<b>115</b>	<b>145</b>

Based on the information contained in the subsequent tables which teacher did a better job? Justify your response and reconcile any differences with your prior response.

**Teacher A had  $\frac{110}{115}$  or 95.6% pass**

**Teacher A had  $\frac{20}{30}$  or 66.7% pass**

**Teacher B had  $\frac{30}{30}$  or 100% pass**

**Teacher B had  $\frac{95}{115}$  or 82.6% pass**

**Both of Teacher B's students outperformed the corresponding students of teacher A. Teacher B is better.**

What is the lurking variable in this situation?

**The number of failing students that each teacher had was substantially different**

**In this case, the differences in sub-groups can be attributed to a lurking variable that was masked when the sub-groups were combined.**

## Notes: C.U.S.S. & B.S.

We have been discussing non-numerical data or qualitative data which is used to describe attributes: gender, color, ethnicity, physical condition and many other things. We also recognized that while qualitative data may be non-numerical in that order doesn't really matter, we sometimes assign numbers for coding purposes, such as the UPC product code, or a social security number or girls are considered 1's and guys are considered to be zeroes. However, terms like mean or median and range made no sense, because there was no order. For instance, we could take the average of all the hair color in the room and come up with deep purple. We wouldn't average the girls and guys and come up with a "shim."

Now we are moving on to discuss Quantitative data which is truly numerical.

**Quantitative Data:** Data that can be measured and expressed numerically (example: height, weight, number of pets, g.p.a, age. etc).

In this case order does matter: Consider your class rank, or how much you can lift, or how fast you are, or how much money you make (I should say earn). In these cases, order absolutely matters. Of course, because you need some more definitions to contend with Quantitative data can be divided into two separate groups discrete and continuous. For now, we are going to only introduce the definitions, but we will explore both in detail.

**Discrete-**quantitative data that can be counted (example: the number of students)

**Continuous-** quantitative data that could take on any fractional value (volume, height, weight)

Prior to developing a common language to describe the distributions there are some very simple computations that can be made to help summarize the data. Let's begin by inputting the following data into the calculator:

9, 8, 7, 6, 5, 4, 3, 2, 1, -10

### Calculator:

Press **STAT**; highlight **EDIT**; & press **ENTER** (enter the data into  $L_1$ )

Press **2<sup>nd</sup> Mode/Quit**

Press **STAT** → highlight **CALC** highlight **1:1-Var Stats** & press **ENTER**

List: press **2<sup>nd</sup>** & press **1**; press **ENTER**

Frequency: Press **CLEAR**; press **ENTER**; press **ENTER**

- $\bar{x}$  = the mean (technically  $\bar{x}$  is the symbol for the sample mean) and is the sum of the values divided by the sample size
- $\sum x$  = the sum of all of the values
- $\sigma_x$  = the standard deviation of the population (we will discuss the calculations at a later date)
- $S_x$  = the standard deviation of the sample (we will discuss the calculations at a later date)
- $n$  = the sample size

### 5 Number Summary:

- minX = the minimum value of the data set-(it may or not be an outlier)
- $Q_1$  denotes the 1<sup>st</sup> quartile and 25% of the data set fall below this value
- Med = the middle most value when the data is listed least to greatest. 50% of the data set fall below this value
- $Q_3$  denotes the 3<sup>rd</sup> quartile and 75% of the data set fall below this value.
- maxX = the maximum value of the data set-(it may or not be an outlier)

## Notes: C.U.S.S. & B.S.

As we work through this section on quantitative data, we will learn to create data displays and we will develop a common language to describe the distributions of the data. We will begin with developing the common language which follows the acronym **C.U.S.S.**

**C: Measures of the Middle** (mean and median)

**U: Unusual Features** (gaps, clusters and outliers)

**S: Shape** (skewed left, skewed right, mound, symmetric, uniform, bi-modal and multi-modal)

**S: Spread** (variation, range, interquartile range, variance and standard deviation)

When describing or comparing distributions you will need to address all four of the parts of **C.U.S.S.** and you will need to B.S. those in your descriptions, that is to say you must **Be Specific**.

When it comes to being able to C.U.S.S. appropriately, all parts are important and should be addressed. However, I will give you a rank ordering of importance.

1. **Center and Spread:** These must always be addressed. Failure to do so will result in no points being awarded. (... and may God have mercy on your soul)— Billy Madison
2. **Shape:** If you want to receive full credit, you should be discussing the shape of the distribution
3. **Unusual Features:** If there exist unusual features they need to be addressed. If they are not addressed, you can expect point deductions.

### C: Measures of the Middle

- **Mean:** The average of all of the values (the mean is impacted by outliers and non-symmetric distributions). The mean is the balance point on a scale or see-saw.  
**Calculation:** Sum all of the values and divide by  $n$  the number of values summed.
- **Median:** the middle-most number (median is not impacted by outliers and non-symmetric distributions). 50% of the values are larger and 50% are smaller than the median.  
**Calculation:** Arrange the numbers smallest to greatest and find the middle number.

### U: Unusual Features

- **Gap:** a section in the distribution without a data point
- **Clusters:** Clusters are distinct groups of data as in two mounds. (**Note:** if clusters exist the centers spread and shape of the clusters must be given)
- **Outliers**-data points that are either too large or too small. In this course, we identify outliers utilizing 2 methods:
  1. More than 2-3 standard deviations above or below the mean. (We will learn about standard deviations, a measure of variation, in the near future.)
  2. More than 1.5 Interquartile ranges below the 1<sup>st</sup> quartile or more than 1.5 interquartile ranges above the 3<sup>rd</sup> quartile. (We will discuss the interquartile range and the 1<sup>st</sup> and 3<sup>rd</sup> quartile when we discuss box plots)

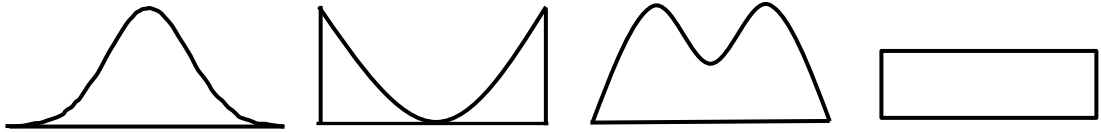


## Notes: C.U.S.S. &amp; B.S.

## S: Shape

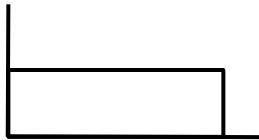
- **Symmetric:** The vertical line can divide the data into two matching mirror images. If the data is symmetric the mean and median will be equal.

Examples:



- **Uniform:** A graph that is approximately the same height.

Example:



- **Mound Shape:** The graph is symmetric with most of the data in the center of the graph. The data density diminishes as you move towards each tail. The mean, median and mode are all equal.

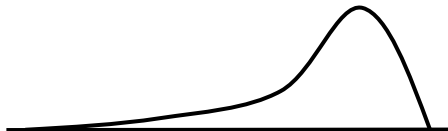
Example:



Because data is rarely if ever perfect, we need to qualify the above shapes as approximately symmetric or reasonably symmetric or reasonably mound shaped or approximately uniform.

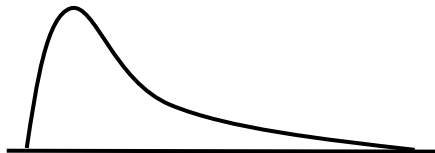
- **Skewed Left:** The tail is to the left and the mean is less than the median. The mean is to the left of the median. (imagine pulling on a sticky wad of gum towards the left)

Example:



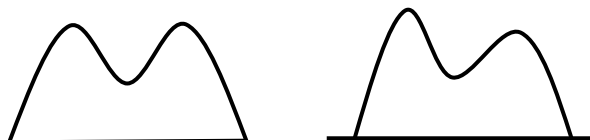
- **Skewed Right:** The tail is to the right and the mean is greater than the median. The mean is to the right of the median. (imagine pulling on a sticky wad of gum towards the right)

Example:



- **Bi-modal:** The distribution has two distinct peaks (modes). The peaks may not be equal in height.

Examples:



## Notes: C.U.S.S. & B.S.

### S: Spread

- **Range:** The range provides an idea as to how spread out the data is by subtracting the smallest value from the largest value. The range is a singular value and is never negative. Because the range uses the largest and smallest value it is greatly impacted by outliers.
- **Interquartile Range:** The interquartile range or IQR gives an idea as to how spread out the data is by focusing on the middle 50% of the data and is computed by subtracting the value of the 1<sup>st</sup> quartile from that of the 3<sup>rd</sup> quartile. We express it in this manner  $IQR = Q_3 - Q_1$ . Because the interquartile range focuses on the middle 50% of the data it is not impacted by outliers. IQR is very useful for skewed distributions.
- **Variance:** The variance takes an average of the data distances from the mean. Because some values are below the mean the distances are negative and obviously some of those distances are greater than the mean and are positive. As a consequence, the distances are squared and then summed.  $\sigma^2$  is the symbol for the true population variance.

**Remember:** we use means as a measure of center when the data tends to be mound shaped and symmetric. Because variance uses the mean in the calculation we tend to reserve the use of the variance for distributions that are symmetric and mound shaped.

The formula for variance of the entire population is  $\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$  where  $\mu$  is the true population mean.

Unfortunately we rarely know the true population mean and have to rely on the sample mean which has the symbol  $\bar{x}$ . The symbol for variance of the sample is  $s^2$  and the formula is as follows:

$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$  You may notice that the sample differences are divided by  $n-1$  this is done because we are using the sample mean,  $\bar{x}$  an estimate of the true population mean, in our calculation.

- **Standard Deviation:** The standard deviation is just the square root of the variance. Because of advances in technology we are able to easily take square roots and as consequence we tend to talk about standard deviations more than we do variances. The symbol for the standard deviation of

the population is  $\sigma$  and the formula is:  $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$ . As expected, the symbol for the

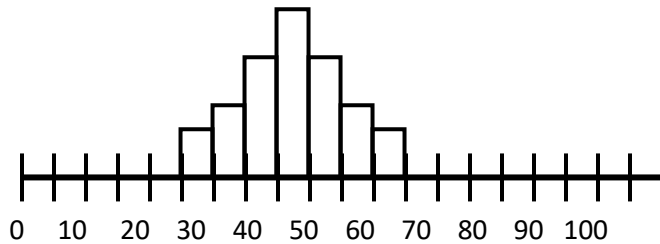
standard deviation of the sample is just  $s$  and the formula is:  $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$ . We will typically say something is so many standard deviations from the mean. Because we are relying on the mean, we usually reserve standard deviation for data that is mound shaped. The smaller the standard deviation the closer the data is to the mean.

- **Z-scores:** A Z-score is a ratio that provides a measure as to how far a value is from the mean and takes into account both the center and the dispersion of the data. Z-scores act as a ruler and can be used to compare different shaped distributions the basic z score formula is  $z = \frac{x - \mu}{\sigma}$  where **z is the number of standard deviations a value lies from the mean.**

**Practice: C.U.S.S. & B.S.**

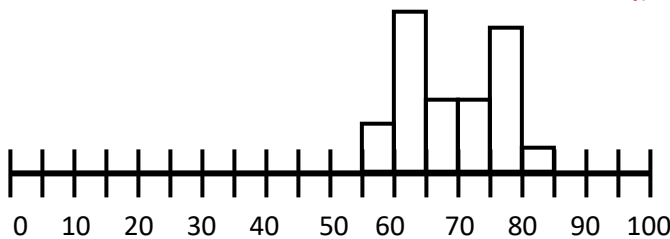
**Directions:** Describe the following graphs. Make certain that you C.U.S.S. We will worry about computations later.

1.



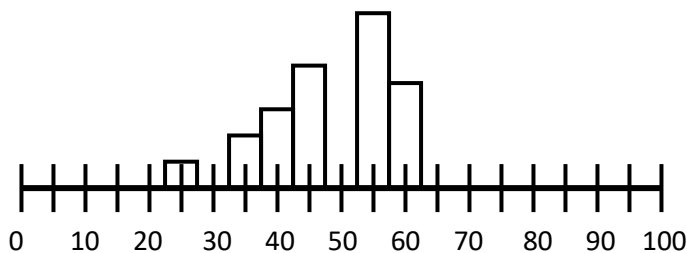
Centered at 47.5 with a range of 40. The cluster lies within the interval 30 to 70. The graph is reasonably symmetric and mound shaped with no unusual features

2.



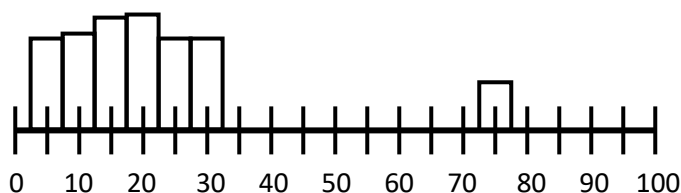
Centered at 70 with a range of 30. The cluster lies within the interval 55 to 85. The distribution is bimodal with peaks at 67 and 77. The graph has no gaps or outliers

3.



Centered at 50 with a range of 40. The cluster lies within the interval 22 to 62. The graph is skewed left. Its mode is at 57. There are gaps at 30 and 50

4.



The main cluster is reasonably uniform and centered at 17 and lies within the interval 3 to 33. The range of the distribution is 72 due to an outlier at 75

## Notes: Creating Stem and Leaf Plots

One of the most basic quantitative graphs is the stem and leaf plot which from this point forward I will refer to as a stemplot. Data values are placed in order from least to greatest and each data value is divided into both a stem and leaf component. The leaves are comprised of furthest digit to the right while the stem is comprised of the remaining digits. To visualize the data the leaves of data values are stacked onto a common stem. The advantage to this type of data display is that it maintains all of the original data values and it provides an idea of center shape, spread and any unusual features in a data set. Unfortunately, this type of graph becomes unwieldy with large data sets.

**Example:** The following table provides a list of weights in kilograms for the forwards and backs of a rugby team that toured parts of New Zealand.

Backs	120	85	90	83	90	81	110	84	113	77
	75	87	100	94	91	79	100	82	94	102

Forwards	128	100	89	105	105	101	115	108	114	99
	98	104	108	101	100	99	110	105	105	115

### Create a Stemplot for the Backs

**Step 1:** Input the data into  $L_1$  and follow the calculator commands below:

2<sup>nd</sup> STAT  
 → OPS  
 Highlight Sort A  
 Enter  
 2<sup>nd</sup> "column number"  
 Enter  
 STAT EDIT ENTER

**Step 2:** Create a legend

**Step 3:** Draw a vertical Line

**Step 4:** Input the Stems

**Step 5:** Input the leaves

Input the data for the forwards into  $L_2$ . Repeat the above Steps to create back to back Stemplots. Compare and contrast the 2 groups.

Backs	Forwards
975	7
754321	8 9
44100	9 889
200	10 00114555588
30	11 0455
0	12 8

10|3 = 103 kilograms

The weights for the backs is skewed to the right with a median around 90 Kilograms. The backs have a range of 45 kilograms and they lie within the interval 75 to 120 kg.

The weights for the forwards are reasonably symmetric and mound shaped with a median of 105 kilograms. The forwards weights have a range of 41 kilograms and lie within the interval 89 to 128 kilograms.

The weights of the backs are skewed to the right while that of the forwards is symmetric and mound shaped. The forwards median weight is 105 kilograms and is higher than the 94 median weight of the backs. The backs have a range of weights of 45 kilograms and are more spread out than the forwards range of weights which is 41 kilograms.

## Notes: Creating Histograms

A histogram is another type of frequency distribution whose class/bar widths have a height that is proportional to the frequency of the values in that class. Histograms are useful for large data sets. And they provide an idea of center, shape and spread and show unusual features of the data sets. However, individual data values are not included in histogram. **Note:** Bars should touch because the data is sequential/numerical in nature and both of the axes must be labeled.

**Example:** The following table provides a list of weights in kilograms for the forwards and backs of a rugby team that toured parts of New Zealand.

Bucks	120	85	90	83	90	81	110	84	113	77
	75	87	100	94	91	79	100	82	94	102

Forwards	128	100	89	105	105	101	115	108	114	99
	98	104	108	101	100	99	110	105	105	115

### Create a histogram for the Bucks

**Step 1:** Input the data into  $L_1$  and follow the calculator commands below:

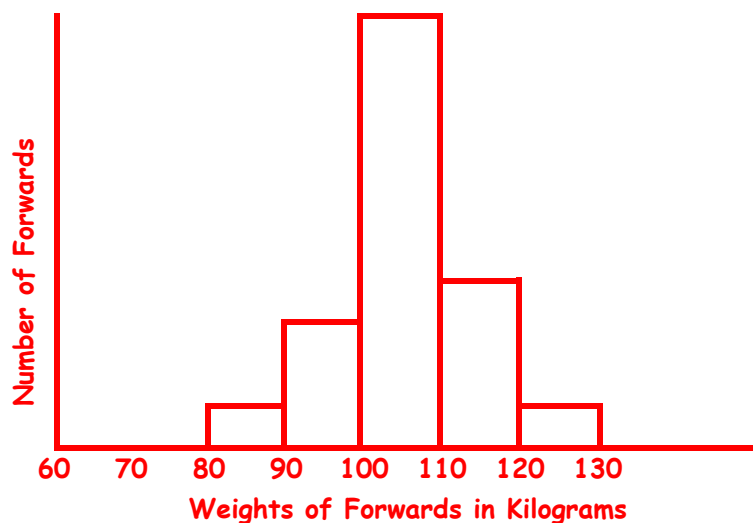
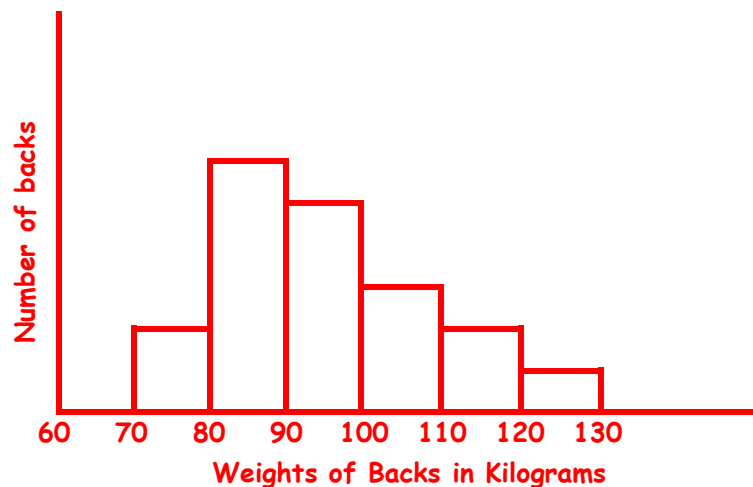
2<sup>nd</sup> STAT PLOT  
 ENTER  
 Highlight ON  
 ↓→highlight Histogram  
 ↓Xlist  
 2<sup>nd</sup> "column number"  
 Freq:1  
 Zoom 9  
 Window  
 Xmin 60  
 Xmax 140  
 Xscl=10  
 Graph & Trace

**Step 2:** Draw the graph

**Step 3:** Label the axis

**Step 4:** Title the graph

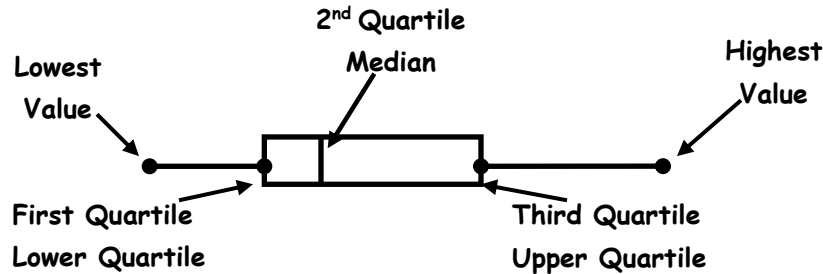
Input the data for the forwards into  $L_2$  & Repeat the steps above. Compare and contrast the 2 groups.



## Notes: Interpreting Boxplots

If I had to choose a data display that I thought was the most important for AP Statistics, I would without hesitation say that it was the box and whisker plot (saved the best for last). From this point forward we will refer to the box and whisker plot as the box plot.

**Box Plot:** a visual representation of the **5 number summary**. Each section of a boxplot contains 25% of the data. A boxplot can give some idea as to shape in particular symmetry and skewness. However exact shape of a distribution as well as, the mean and variance cannot be determined from a boxplot display.



**Median:** The data point that occurs in the **middle** when the numbers are placed in order from least to greatest. (not impacted by outliers)

**Mean:** The average of the data. (greatly affected by outliers)

**Range:** The difference between the highest and lowest value. (Always affected by outliers)

**First Quartile:** The median of the lower half of the data. 25% of the data are below the value.

**Third Quartile ( $Q_3$ ):** The median of the upper half of the data. 75% of the data are below the value.

**IQR:** the interquartile range  

$$IQR = Q_3 - Q_1$$

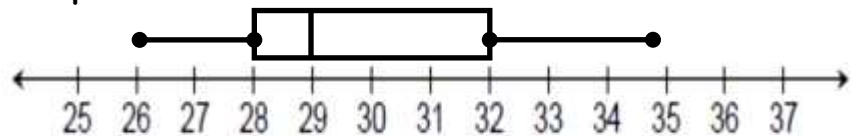
**Outlier:** A number that is too large or too small for compared to the rest of the data.

$$1^{\text{st}} \text{ quartile} - 1.5 \times IQR$$

$$3^{\text{rd}} \text{ quartile} + 1.5 \times IQR$$

**Note:** Box and whisker plots do not identify the mean.

Example:



- Identify the highest value  
**35**
- Identify the lowest value  
**26**
- Identify the median  
**29**
- Find the range  
 **$35 - 26 = 9$**
- Identify the first quartile  
 **$Q_1 = 28$**
- Identify the third quartile  
 **$Q_3 = 32$**
- Identify the interquartile range  
 **$IQR = Q_3 - Q_1 \rightarrow IQR = 32 - 28 \rightarrow IQR = 4$**
- 75% of the data are above what value?  
 **$Q_1 = 28$**
- 50% of the data are between which values? List all combinations.  
**26 to 29 and 28 to 32 and 29 to 35**
- Find the mean  
**Trick Question: You cannot determine the mean from a boxplot**

## Notes: Interpreting Boxplots

**Median:** The data point that occurs in **the middle** when the numbers are placed in order from least to greatest. **(not impacted by outliers)**

**Mean:** The average of the data. **(greatly affected by outliers)**

**Range:** The difference between the highest and lowest value. **(Always affected by outliers)**

**First Quartile:** The median of the lower half of the data. 25% of the data are below the value.

**Third Quartile ( $Q_3$ ):** The median of the upper half of the data. 75% of the data are below the value.

**IQR:** the interquartile range  

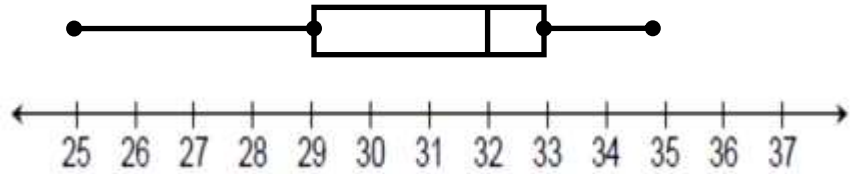
$$\text{IQR} = Q_3 - Q_1$$

**Outlier:** A number that is too large or too small for compared to the rest of the data.

$$1^{\text{st}} \text{ quartile} - 1.5 \times \text{IQR}$$

$$3^{\text{rd}} \text{ quartile} + 1.5 \times \text{IQR}$$

**Note:** Box and whisker plots do not identify the mean.



- Identify the highest value **35**
- Identify the lowest value **25**
- Identify the median **32**
- Find the range  

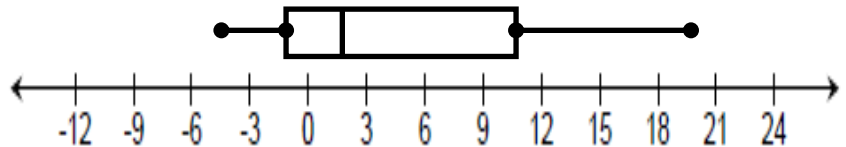
$$35 - 25 = 10$$
- Identify the IQR  

$$\text{IQR} = Q_3 - Q_1 \rightarrow \text{IQR} = 33 - 29 \rightarrow \text{IQR} = 4$$
- How small or large would a value be to be considered an outlier?  

$$X < Q_1 - 1.5(\text{IQR}) \rightarrow X < 29 - 1.5(4) \rightarrow X < 23$$
  

$$X > Q_3 + 1.5(\text{IQR}) \rightarrow X > 33 + 1.5(4) \rightarrow X > 39$$
- Is the data skewed?

**The boxplot is stretched to the left which means the data is skewed left and has a mean below the median**



- Identify the highest value **20**
- Identify the lowest value **-5**
- Identify the median **2**
- Find the range  

$$20 - (-5) = 25$$
- Identify the IQR  

$$\text{IQR} = Q_3 - Q_1 \rightarrow \text{IQR} = 11 - (-1) \rightarrow \text{IQR} = 12$$
- How small or large would a value be to be considered an outlier?  

$$X < Q_1 - 1.5(\text{IQR}) \rightarrow X < -1 - 1.5(12) \rightarrow X < -19$$
  

$$X > Q_3 + 1.5(\text{IQR}) \rightarrow X > 11 + 1.5(12) \rightarrow X > 39$$
- Is the data skewed?

**The boxplot is stretched to the right which means the data is skewed right and has a mean above the median**

## Notes: Interpreting Boxplots

**Median:** The data point that occurs in the **middle** when the numbers are placed in order from least to greatest. **(not impacted by outliers)**

**Mean:** The average of the data. **(greatly affected by outliers)**

**Range:** The difference between the highest and lowest value. **(Always affected by outliers)**

**First Quartile:** The median of the lower half of the data. 25% of the data are below the value.

**Third Quartile ( $Q_3$ ):** The median of the upper half of the data. 75% of the data are below the value.

**IQR:** the interquartile range  

$$\text{IQR} = Q_3 - Q_1$$

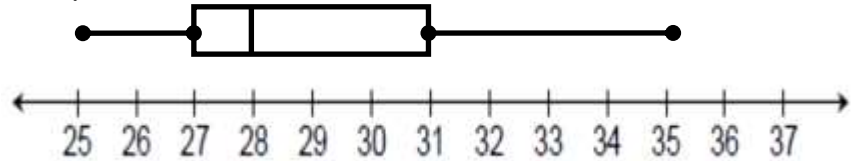
**Outlier:** A number that is too large or too small for compared to the rest of the data.

$$1^{\text{st}} \text{ quartile} - 1.5 \times \text{IQR}$$

$$3^{\text{rd}} \text{ quartile} + 1.5 \times \text{IQR}$$

**Note:** Box and whisker plots do not identify the mean.

**Example:**

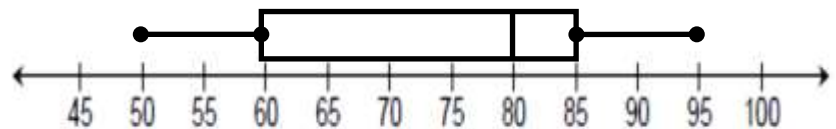


a) For which value of data is 50% of the values larger and 50% smaller? **The median which is 28**

b) What is the range of the data on the given graph?

$$35 - 25 = 10$$

**Example:** A box and whisker plot gives a picture of the test scores on a recent math test.



Using the box and whisker plot above answer the following questions:

a) True or false—We know that the class average was an 80.  
**False: Boxplots do not give the mean. The median is 80**

b) True or false—We know that half of the students made less than an 80.

**True: The median is 80 and half made higher & half lower**

c) True or false—We know that the lowest grade was a 45.

**False: The lowest score is a 50**

d) True or false—We know that at least half of the class passed the test. **True: The median is 80 which means**

**more than half scored higher than an 80**

e) True or false—We know that the difference between the highest and lowest score was 55.

**False: The difference between the highest and lowest score = 95 - 50 or 45**

f) True or false—50% of the class made a B or better.

**True: The median is 80, which is a B and half made higher**



## Notes: Creating Boxplots

### Box Plots

**Median:** The data point that occurs in **the middle** when the numbers are placed in order from least to greatest. (**not greatly affected by outliers**)

#### Creating Box Plots

**Step 1:** Press **STAT EDIT ENTER**

Input the data into column  $L_1$

**Step 2:** Press **2<sup>nd</sup> STAT PLOT** →

**Enter** → (turn stat Plot on)

Press **↓ → → Enter**

Press **↓ 2<sup>nd</sup> L<sub>1</sub>**

(Make certain Freq:1)

**Step 3:** Press **ZOOM 9**

#### 5 Number Summary

**Step 1:** Press **STAT EDIT ENTER**

Input the data into column  $L_1$

**Step 2:** Press **STAT → CALC ENTER**

(1-Var Stats)

**Note:** Make certain  $L_1$  is selected

**Step 3:** Press **ENTER 3 times**

**Step 4:** Arrow down

- Min = smallest value
- $Q_1$  = 1<sup>st</sup> quartile
- Med = Median/2<sup>nd</sup> quartile
- $Q_3$  = 3<sup>rd</sup> quartile
- Max = largest value

**Note:** Other information

- $\bar{x}$  = sample mean
- $n$  = sample size
- $s_x$  = sample standard deviation
- $\sigma_x$  = population standard deviation

**Box Plot:** A graphical display of data along a number line, divides the data into four parts called quartiles and identifying the lowest and highest data value along with the **median**

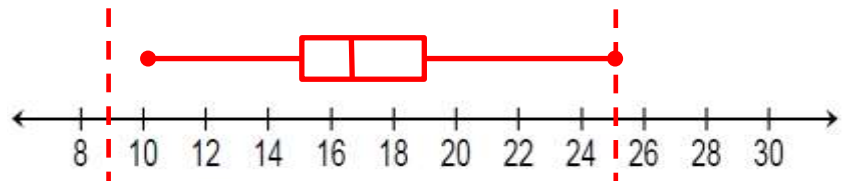
**Example:** Using the data below construct a box plot and show the construction of the fences for the outliers.

19, 17, 12, 24, 10, 19, 25, 15, 16, 15, 19, 16

$$Q_1 = 15 \text{ and } Q_3 = 19 \text{ and } IQR = Q_3 - Q_1 \quad IQR = 19 - 15 = 4$$

$$\text{Lower fence} = Q_1 - 1.5(IQR) \quad 15 - 1.5 \times 4 = 9$$

$$\text{Upper fence} = Q_3 + 1.5(IQR) \quad 19 + 1.5 \times 4 = 25$$



**Example:** Using the data below construct a box plot and show the construction of the fences for the outliers.

9, 8, 7, 6, 5, 4, 3, 2, 1, -10

$$Q_1 = 2 \text{ and } Q_3 = 7 \text{ and } IQR = Q_3 - Q_1 \quad IQR = 7 - 2 = 5$$

$$\text{Lower fence} = Q_1 - 1.5(IQR) \quad 2 - 1.5 \times 5 = -5.5$$

$$\text{Upper fence} = Q_3 + 1.5(IQR) \quad 7 + 1.5 \times 5 = 14.5$$



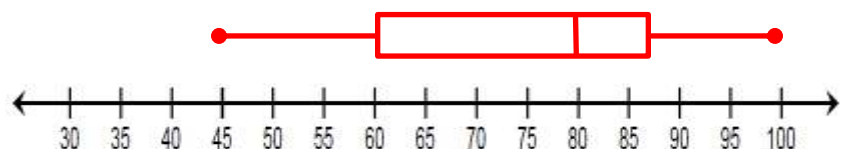
**Example:** Using the data below construct a box plot and show the construction of the fences for the outliers.

80, 90, 60, 90, 85, 65, 95, 55, 70, 80, 100, 45, 95, 80, 60, 75, 80, 60, 50, 85

$$Q_1 = 60 \text{ and } Q_3 = 87.5 \text{ and } IQR = Q_3 - Q_1 \quad IQR = 87.5 - 60 = 27.5$$

$$\text{Lower fence} = Q_1 - 1.5(IQR) \quad 60 - 1.5 \times 27.5 = 18.75$$

$$\text{Upper fence} = Q_3 + 1.5(IQR) \quad 87.5 + 1.5 \times 27.5 = 128.75$$

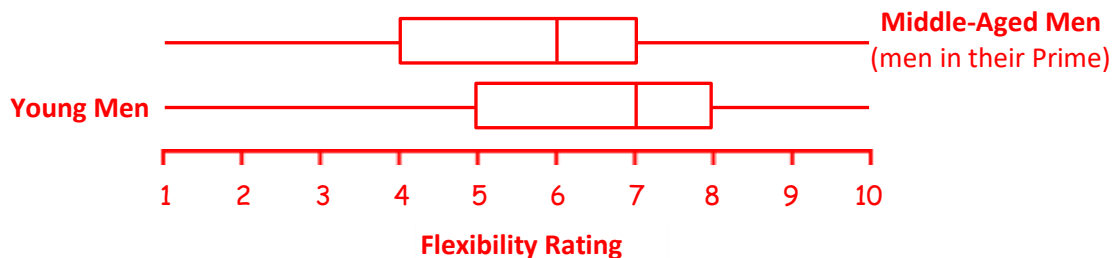


### Notes: Displaying and Describing Data Scenarios

**2000 Problem 3** Five hundred randomly selected middle-aged men and five hundred randomly selected young adult men were rated on a scale from 1 to 10 on their physical flexibility, with 10 being the most flexible. Their ratings appear in the frequency table below. For example, 17 middle-aged men had a flexibility rating of 1.

Physical Flexibility Rating	Frequency of Middle-Aged Men	Frequency of Young Adult Men
1	17	4
2	31	17
3	49	29
4	71	39
5	70	54
6	87	69
7	78	83
8	54	93
9	34	73
10	9	39

- (a) Display these data graphically so that the flexibility of middle-aged men and young adult men can be easily compared.
- (b) Based on an examination of your graphical display, write a few sentences comparing the flexibility of middle-aged men with the flexibility of young adult men.



The median flexibility scores differ by 1 with young men's scores being higher at 7 while the middle age men is at 6 which suggests that young men have more flexibility than middle aged men.

The distribution of middle age men is reasonably symmetric while that of young men is skewed to the left. The upper 50% of the young men have flexibility ratings higher than 75% of the middle-aged men.

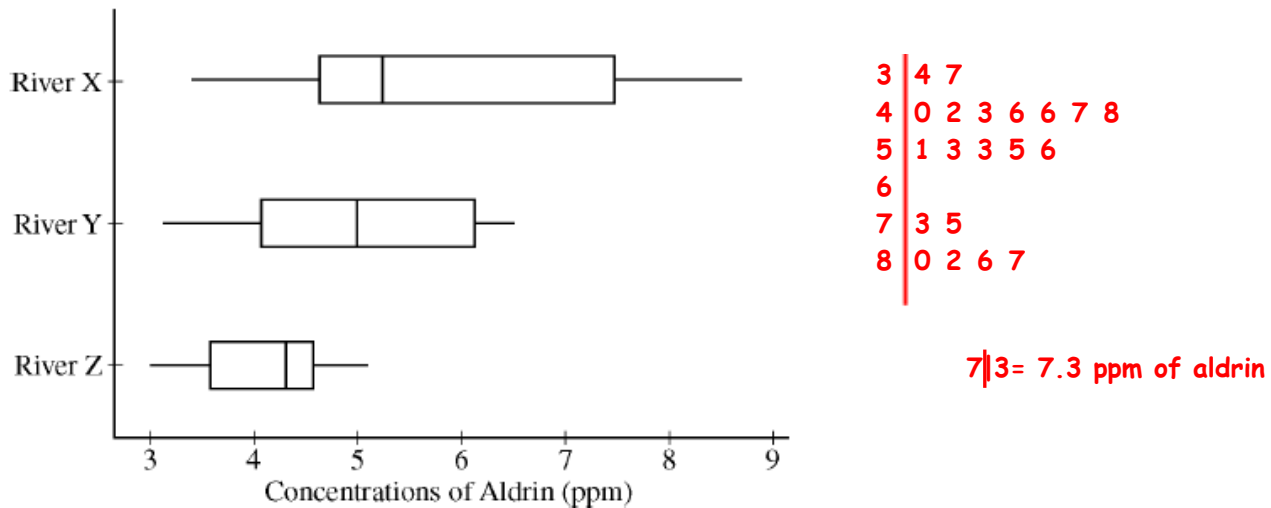
Middle aged men and young men flexibility scores both have a range of 10 and fall within the same values. Though the values are higher the interquartile range for both distributions is 3.

## Notes: Displaying and Describing Data Scenarios

**2010 Form B Problem 1** As a part of the United States Department of agriculture's Super Dump cleanup efforts in the early 1990s, various sites in the country were targeted for cleanup. Three of the targeted sites---River X,

River Y, and River Z---had become contaminated with pesticides because they were located near abandoned pesticide dump sites. Measurements of the concentration of aldrin (a commonly used pesticide) were taken at twenty randomly selected locations in each river near the dump sites.

the boxplots shown below display the five-number summaries for the concentrations, in parts per million (ppm) of aldrin, for the twenty locations that were sampled in each of the three rivers.



a) Compare the distributions of the concentration of aldrin among the three rivers.

b) The twenty concentrations of aldrin for River X are given below.

3.4	4.0	5.6	3.7	8.0	5.5	5.3	4.2	4.3	7.3
8.6	5.1	8.7	4.6	7.5	5.3	8.2	4.7	4.8	4.6

Construct a stemplot that displays the concentrations of aldrin for River X

c) Describe a characteristic of the distribution of aldrin concentrations in River X that can be seen in the stemplot but cannot be seen in the boxplot.

River X is skewed right and has the highest range, interquartile range and variability. River X also has the highest median with 50% of the readings being greater than 5.

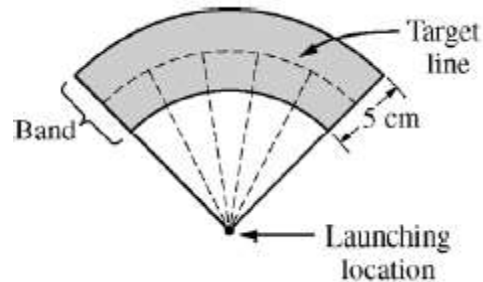
River Y is reasonably symmetric and has median of 5 and its range is the second largest.

River Z is skewed left and has the smallest range, IQR and Variance. It has a median of 4.3. All of its readings are below the upper 50% of river X

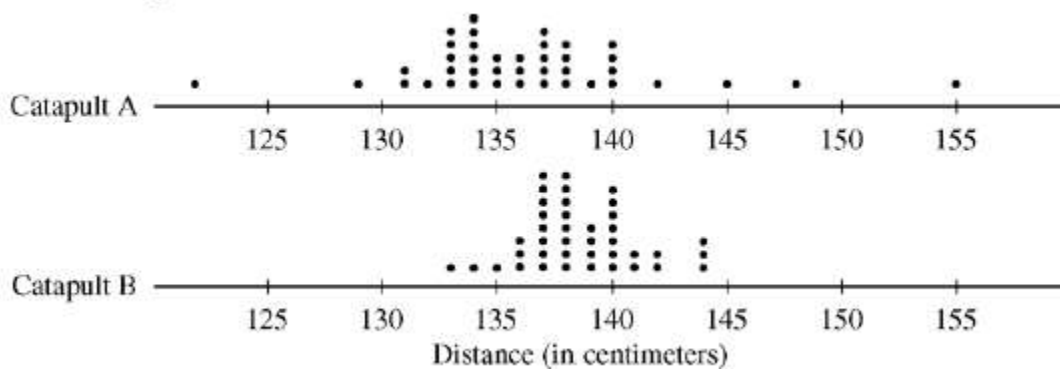
The gap at 6 cannot be seen in the box plot but is visible in the stemplot. The stemplot appears to show two clusters one centered at location 4 and one centered between locations 7 and 8.

## Notes: Displaying and Describing Data Scenarios

**2006 Problem 1:** Two parents have each built a toy catapult for use in a game at an elementary school fair. To play the game, students will attempt to launch Ping-Pong balls from the catapults so that the balls land within a 5-centimeter band. A target line will be drawn through the middle of the band, as shown in the figure below. All points on the target line are equidistant from the launching location.



If a ball lands within the shaded band, the student will win a prize. The parents have constructed the two catapults according to slightly different plans. They want to test these catapults before building additional ones. Under identical conditions, the parents launch 40 Ping-Pong balls from each catapult and measure the distance that the ball travels before landing. Distances to the nearest centimeter are graphed in the dotplots below.



- Comment on any similarities and any differences in the two distributions of distances traveled by balls launched from catapult A and catapult B.
- If the parents want to maximize the probability of having the Ping-Pong balls land within the band, which one of the two catapults, A or B, would be better to use than the other? Justify your choice.
- Using the catapult that you chose in part (b), how many centimeters from the target line should this catapult be placed? Explain why you chose this distance.

Both are reasonably symmetric however the Center of catapult b is higher at about 138cm versus catapult A at 135. The biggest difference is the variability in catapult A. Catapult A has both the longest and shortest shots. Catapult A has possible outliers on both sides of the data. Catapult B does not appear to have any outliers.

Because Catapult B has less variability I would choose catapult B. The smaller variability for catapult B would increase the probability of landing in a given band assuming the catapult is placed at the right distance from the target.

I would place the catapult at 138 cm from the target this happens to be the median value. At this distance 75% of the ping pong balls landed within a 5 cm band.

## Notes: Cumulative Frequency Graphs

As one moves from left to right cumulative frequency graphs increase from zero percent until they reach 100%. Cumulative frequency charts never decrease.

**Scenario:** The following graphs show the way 3 soccer clubs scored their points over a 10 month season.

1<sup>st</sup> Describe each of the shapes.

2<sup>nd</sup> Explain what each graph represents as far as to the distribution of goals throughout the season.

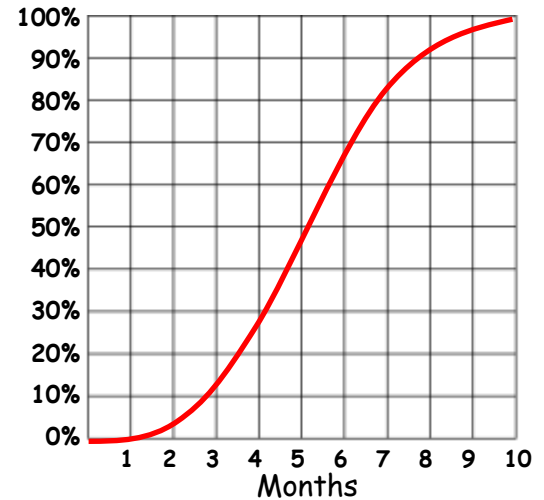
3<sup>rd</sup> Graph each as a cumulative frequency. Only focus on the shape of the graph

1<sup>st</sup> Graph

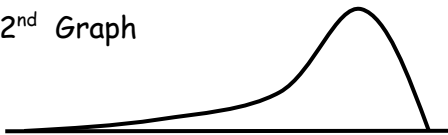


The distribution is reasonably symmetric and mound shaped with the mean, median and mode occurring at the midpoint.

The soccer team scored most of their goals in the middle of the season. They started a little slow and ended a little slow.

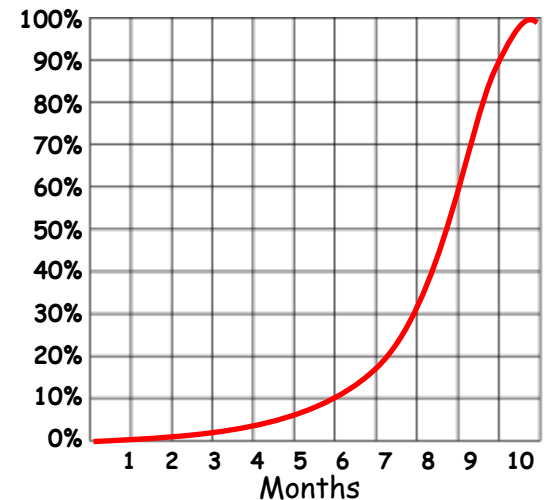


2<sup>nd</sup> Graph

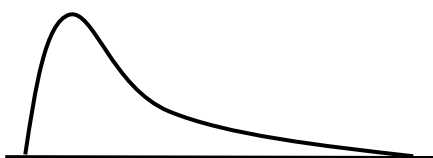


The distribution is skewed left with the mean to the left of the median.

The soccer team scored very few goals in the early part of the season but substantially increased their output towards the end of the season.

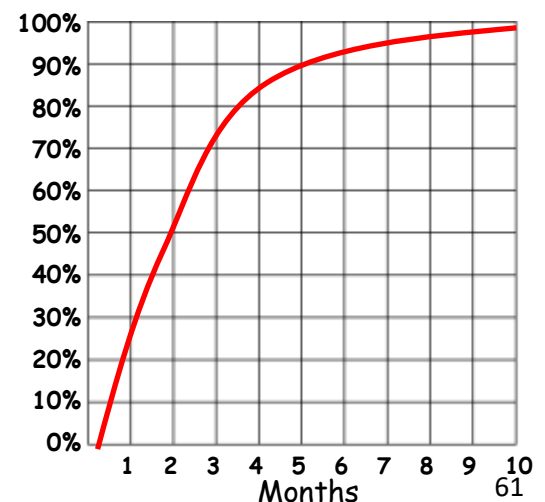


3<sup>rd</sup> Graph



The distribution is skewed right with the mean to the right of the median.

The soccer team scored most of their goals in the early part of the season. As the season progressed, they scored fewer and fewer.

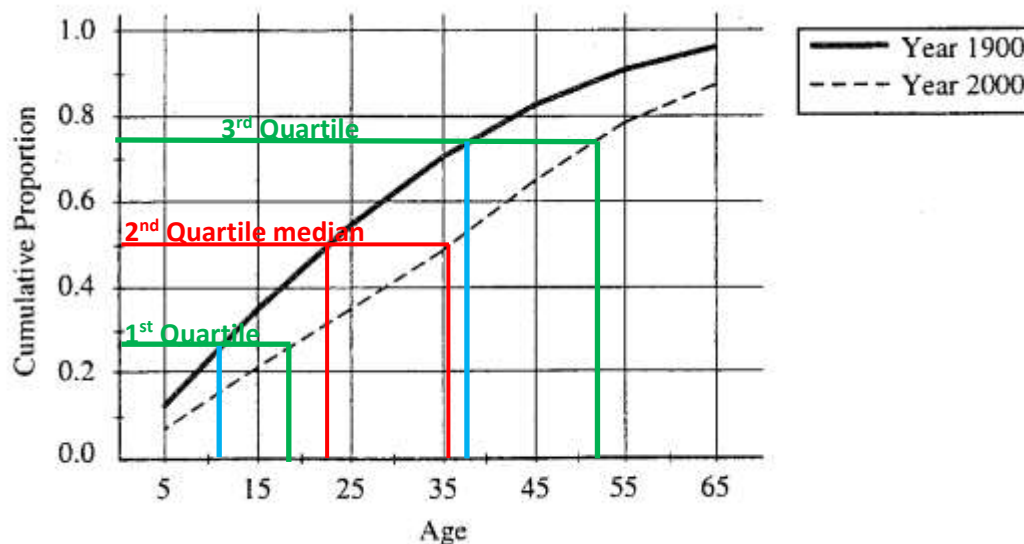


## Notes: Cumulative Frequency Graph Scenarios

AGE DATA

Age	1900	2000
5	0.121	0.066
15	0.344	0.209
25	0.540	0.344
35	0.700	0.480
45	0.822	0.643
55	0.906	0.781
65	0.959	0.870

**1999 Problem 1:** The table of data above provides the cumulative proportions for the United States population at selected ages for the years 1900 and 2000 (projected). For example, 0.344 or 34.4 percent of the population was at or under age 15 in 1900, while only 0.209 or 20.9 percent will be at or under age 15 in the year 2000. The graph below shows the cumulative proportions plotted against age for the years 1900 and 2000 (projected). The data and graph are to be used to compare the age distribution for the year 1900 with the projected age distribution for the year 2000.



- Approximate the median age for each distribution.
- Approximate the interquartile range for each distribution.
- Using the results from parts (a) and (b), write a sentence or two for a history textbook comparing the age distributions for the years 1900 and 2000.

(a) Median Age for 1900= 22.5 years Median Age for 2000 =36 years

(b) IQR=3rd quartile minus the 1st quartile. IQR for 1900 37-11=28 years  
IQR for 2000 53-18=35

(c) Based on the shift of the medians, the population as a whole is living longer in 2000 than they were in 1900. The middle 50% are living approximately from 18 to 53 years of age in 2000 versus living from 11 to 37 years of age in 1900. The change in IQR increased as well which means that there was a greater spread in the ages of people in the year 2000.



## Notes: Standardizing with Z-scores

**Z-scores:** A z-score is a ratio that can be used to determine how many standard deviations a value lies from its mean by taking into account a measure of spread (the standard deviation of the distribution) and a measure of center (the mean of the distribution). A negative value lies below the mean and a positive value falls above the mean. A value at the mean, has a z-score of zero.

For individual values we use:  $Z = \frac{x - \mu}{\sigma}$

For sample means we use:  $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$  or  $Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$  where  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Z-scores can be computed for any shaped distribution. Consequently, we can use z-scores as a ruler to compare scores from different distributions. We can say that this persons scored 1.2 standard deviations above the mean and this person scored 1.8 standard deviations above the mean and we can do that regardless of the underlying distribution and regardless of the sample size. Typically we will define values that are more than 3 standard deviations beyond the mean as outliers. In fact, we often say that a value that is more than 2 standard deviations beyond the mean is an outlier.

1. At a college the scores on the chemistry final exam are approximately normally distributed, with a mean of 75 and a standard deviation of 12. The scores on the calculus final are also approximately normally distributed, with a mean of 80 and a standard deviation of 8. A student scored 81 on the chemistry final and 84 on the calculus final. Relative to the students in each respective class, in which subject did this student do better?

**Chemistry**

$$Z = \frac{x - \mu}{\sigma}$$

$$\mu = 75$$

$$\sigma = 12$$

$$x = 81$$

$$Z = \frac{81 - 75}{12}$$

$$Z = .5$$

**Calculus**

$$Z = \frac{x - \mu}{\sigma}$$

$$\mu = 80$$

$$\sigma = 8$$

$$x = 84$$

$$Z = \frac{84 - 80}{8}$$

$$Z = .5$$

The student scored .5 standard deviations above the class mean for both exams which means that the student performed equally well on the 2 exams relative to the performance of the other students

What would a 60 on the Chemistry final equate to on the calculus final?

**Chemistry**

$$Z = \frac{x - \mu}{\sigma}$$

$$\mu = 75$$

$$\sigma = 12$$

$$x = 60$$

$$Z = \frac{60 - 75}{12}$$

$$Z = -1.25$$

**Calculus**

$$Z = \frac{x - \mu}{\sigma}$$

$$\mu = 80$$

$$\sigma = 8$$

$$x = ???$$

$$-1.25 = \frac{x - 80}{8}$$

$$x = 70$$

2. The table below describes the distribution of median household incomes in the 50 states.

n	Mean	SD	Min	Q <sub>1</sub>	Med	Q <sub>3</sub>	Max
50	51,742.44	8210.64	36,641	46,071	50,009	57,020	71,836

- a) Find and interpret the z-score for North Carolina with a median income of \$41,553.  $Z = \frac{x - \mu}{\sigma} \rightarrow Z = \frac{41,553 - 51,742.44}{8,210.64} \rightarrow Z = -1.24$

- b) New Jersey has a standardized score of 1.82.

Calculate New Jersey's median income.  $Z = \frac{x - \mu}{\sigma} \rightarrow 1.82 = \frac{x - 51,742.44}{8,210.64} \rightarrow x = 66,685.80$



### Notes: Standardizing with Z-scores

**2011 Question 1:** A professional sports team evaluates potential players for a certain position based on two main characteristics, speed and strength.

Speed is measured by the time required to run a distance of 40 yards, with smaller times indicating more desirable (faster) speeds. From previous speed data for all players in this position, the times to run 40 yards have a mean of 4.60 seconds and a standard deviation of 0.15 seconds, with a minimum time of 4.40 seconds, as shown in the table below.

	Mean	Standard deviation	Minimum
Time to run 40 yards	4.60 seconds	0.15 seconds	4.40 seconds

Strength is measured by the amount of weight lifted, with more weight indicating more desirable (greater) strength. From previous strength data for all players in this position, the amount of weight lifted has a mean of 310 pounds and a standard deviation of 25 pounds, as shown in the table below.

	Mean	Standard deviation
Amount of weight lifted	310 pounds	25 pounds

(b) Calculate and interpret the z-score for a player in this position who can lift a weight of 370 pounds.

$$Z = \frac{x - \mu}{\sigma} \rightarrow Z = \frac{370 - 310}{25} \rightarrow Z = 2.4$$

A player who can lift 370 pounds is 2.4 standard deviations above the mean for players at that position.

(c) The characteristics of speed and strength are considered to be of equal importance to the team in selecting a player for the position. Based on the information about the means and standard deviations of the speed and strength data for all players and the measurements listed in the table below for Players A and B, which player should the team select if the team can only select one of the two players? Justify your answer.

	Player A	Player B
Time to run 40 yards	4.42 seconds	4.57 seconds
Amount of weight lifted	370 pounds	375 pounds

**Note:** If something takes less time, it is faster. So a negative z-score would be associated with times that are faster than the average

**Player A**

$$Z = \frac{x - \mu}{\sigma}$$

**Player B**

$$Z = \frac{x - \mu}{\sigma}$$

$$\text{Running: } Z = \frac{4.42 - 4.60}{.15} \rightarrow -1.2$$

$$\text{Running: } Z = \frac{4.57 - 4.60}{.15} \rightarrow -0.2$$

$$\text{Lifting: } Z = \frac{370 - 310}{25} \rightarrow 2.4$$

$$\text{Lifting: } Z = \frac{375 - 310}{25} \rightarrow 2.6$$

Player B is slightly stronger than player A and is able to lift a weight that is 2.6 standard deviation above the mean vs. player A being able to lift 2.4 standard deviations above the mean. Player A is substantially faster and is 1.2 standard deviations faster than the mean vs Player B who is only .2 standard deviations faster. Even though Player B is .2 standard deviations stronger, **I would choose player A because he is a full standard deviation faster than Player B and only slightly less strong and both speed and strength are equally weighted**

## Notes: Calculating Means & Standard Deviations of Discrete Distributions

We previously calculated means, variance and standard deviations of data sets and used the following

$$\text{formulas: } \mu = \frac{\sum x_i}{n} \quad \sigma^2 = \sum \frac{(x_i - \mu)^2}{n} \quad \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

**Mean**
**Variance**
**Standard deviation**

What if instead of having a list of data, we knew how often certain results occurred either by simulation of experience? For instance, we know that if we roll a die we the chance of getting a 1 is 1/6 and 2 is 1/6 etc. So the question is how do we calculate the mean variance and standard deviation?

Well first of all we need to understand that when we roll a die, we can only get a 1, 2, 3, 4, 5 or 6—duh. However that is important to not because those values 1, 2, 3, 4, 5, and 6, represent the entire population. Further we know the chance of any one of those values is 1/6. I have represented this complex data in the table below.

Population	1	2	3	4	5	6
Chance	1/6	1/6	1/6	1/6	1/6	1/6

**Probability Distribution:** A table or an equation that links each outcome of a statistical experiment with its probability of occurrence. To be a valid probability distribution the following must be true:

1. The sum of the probabilities must be 1.
2. The probabilities must lie between 0 and 1 inclusive.

*There is no such thing as a negative probability.*

*There is no such thing as a probability greater than one.*

**Question:** Does the table represent a valid probability distribution? Justify your response.

**Yes, it does because the values lie within the interval 0 to 1 inclusive and sum to 1**

Now back to the original question: how do I calculate the mean, variance and standard deviation?

Fortunately, we have a simple formula to do this and if you are curious I can easily demonstrate where this formula came from. Just ask.

$$E(x) = \mu_x = \sum x_i p_i$$

$$(1/6)(1) + (1/6)(2) + (1/6)(3) + (1/6)(4) + (1/6)(5) + (1/6)(6) = 3.5$$

All we are doing is multiplying each outcome by its associated probability and then summing them.

**In the space above calculate the mean.**

Similarly the variance is found by squaring the difference between a mean and a value and then multiplying the result by the values probability the resulting products are then summed.

$$\sigma^2 = \sum (x_i - \mu)^2 p_i$$

$$(1 - 3.5)^2(1/6) + (2 - 3.5)^2(1/6) + (3 - 3.5)^2(1/6) + (4 - 3.5)^2(1/6) + (5 - 3.5)^2(1/6) + (6 - 3.5)^2(1/6) = 2.917$$

**In the space above calculate the variance.**

The standard deviation is the square root of the variance:  $\sigma = \sqrt{\sum (x_i - \mu)^2 p_i}$

$$= 1.7078$$

**Calculator Commands:** Enter the outcomes in column  $L_1$  the probabilities in column  $L_2$ .

Go to STAT CALC 1-VarStats. For List: enter  $L_1$  for Frequency enter  $L_2$ .

## Notes: Calculating Means & Standard Deviations of Discrete Distributions

Consider the following scenarios:

1. A commuter must pass through five traffic lights on her way to work. She estimates the probability model for the number of lights she hits as shown below. Calculate the expected value/mean and the standard deviation and variance showing all work.

X	0	1	2	3	4	5
Probability	0.10	0.25	0.30	0.20	0.10	0.05

$$E(x) = 2.1$$

$$\sigma = 1.3$$

$$\sigma^2 = 1.69$$

- a. Does the above scenario represent a valid probability distribution? Justify.

Yes, it does because the values lie within the interval 0 to 1 inclusive and sum to 1

- b. Calculate by hand the expected value and place in context of the problem:

$$E(x) = \mu_x = \sum x_i p_i \quad (0)(.10) + (1)(.25) + (2)(.30) + (3)(.20) + (4)(.10) + (5)(.05) = 2.1$$

On average a commuter can expect to have to stop at 2.1 of the 5 red lights

- c. Calculate by hand the variance:  $\sigma^2 = \sum (x_i - \mu_x)^2 p_i$

$$(0-2.1)^2(.10) + (1-2.1)^2(.25) + (2-2.1)^2(.30) + (3-2.1)^2(.20) + (4-2.1)^2(.10) + (5-2.1)^2(.05) = 1.69$$

2. Complete the probability distribution. Calculate the expected value, the standard deviation and the variance for the following:

X	2	4	6	8
Probability	0.15	0.3	0.35	

$$E(x) = 5.2$$

$$\sigma = 1.939$$

$$\sigma^2 = 3.76$$

$$1 - (.15 + .3 + .35) = .2$$

3. An insurance company charges \$1,600 annually for car insurance. The policy specifies that the company will pay \$3000 for a minor accident and \$10,000 for a major accident. The probability of a motorist having a minor accident during the year is .22, and of having a major accident, .03. Complete the probability table below for how much money the insurance company will make on a single policy. Hint: Decide what X is.

X	1,600	-1,400	-8,400
Probability	.75	.22	.03

$$E(x) = 640$$

$$\sigma = 4058400$$

$$\sigma^2 = 2014.55$$

## Notes: Transforming Random Variables

**Adding Constants:** Now that we know how to calculate the expected value and variance of a discrete random variable, what happens if we were to add a constant to every value in the distribution? Simply put nothing would happen to the measures of spread. The standard deviation, variance, range and IQR would remain unchanged. However the mean and median would change by the value of the constant.

**Consider the following scenario:** On the last test there were the following scores: 100, 73, 83, 100, 83, 92, 92, 28, 88, 68, 39, 92, 100, 92, 83, 78, 83, 92, 28, 96

Input the data using **STAT Edit** and calculate using **STAT Calc 1-var stats**

Mean = **79.5**

Median = **85.5**

Standard deviation = **21.86**

Variance = **477.86**

Range **100-28= 72**

IQR **92-75.5= 16.5**

As your teacher I decided to add 5 points to each grade. Recalculate the statistics:

**Hint: Highlight L<sub>2</sub> and press 2<sup>nd</sup> 1 + 5 and calculate using Stat Calc 1-var stats**

Mean = **84.5**

Median = **95.5**

Standard deviation = **21.86**

Variance = **477.86**

Range **105-33=72**

IQR **97-80.5=16.5**

**What happened to the measures of spread?**

**The measures of spread were not affected by the addition of the constant**

**What happened to the measures of center?**

**The measures of center changed by the value of the constant**

**Multiplying by Constants:** Multiplying every test score by a constant has the same effect as multiplying the measures of center by the constant. The same is true with the measures of spread with the exception of variance. Variance is the square of the standard deviation, thus variance is multiplied by the square of the constant.

**Consider the above scenario:** Adjust the original test scores by using the AP multiplier of 1.29

Mean = **102.55**

Median = **110.295**

Standard deviation = **28.2**

Variance = **795.24**

Range = **129-36.12= 92.88**

IQR **118.68-97.395= 21.285**

**What happened to measures of center, spread and variance?**

**The variance increased by the multiple  $(1.29)^2$  everything else increased by the multiple 1.29**

## Notes: Combining Random Variables

So what happens if we combine 2 random variables?

**Expected Values:** It turns out that the mean or expected values of random variables can be added or subtracted directly and no conditions apply.

(note: if we combine 2 normal distributions, the result is a normal distribution)

**Variations:** Unlike expected values, there are limitations with the combining of random variables when it comes to measures of spread. **First:** we can only combine variances. **Second:** we can only combine variance if the **random variables are independent**. **Third:** we can only add the variance. This means that we sum the variances when we are asked to sum them and we subtract them when we are asked to subtract them.

**Note:** If we wish to combine standard deviations, we must do the following: square the standard deviation; follow the rules for variances; take the square root of the resulting variance to find the standard deviation.

**Example:** **Given independent** random variables with means and standard deviations as shown, find the mean and standard deviation of each of these variables:

	Mean	Standard Deviation
X	10	2
Y	20	5

(a)  $3X$

$$3(10) = 30 \text{ \& } 3(2) = 6$$

(b)  $Y+6$

$$20+6 = 26 \text{ \& } 5$$

(c)  $X+Y$

$$10+20=30 \text{ \& } \sqrt{2^2 + 5^2} = \sqrt{29}$$

(d)  $X-Y$

$$10-20 = -10 \text{ \& } \sqrt{2^2 + 5^2} = \sqrt{29}$$

(e)  $X_1+X_2$

$$10+10 = 20 \text{ \& } \sqrt{2^2 + 2^2} = \sqrt{4}$$

(f)  $X \div 2 + Y$

$$10 \div 2 + 20 = 25 \text{ \& } \sqrt{\left(\frac{2}{2}\right)^2 + 5^2} = \sqrt{26}$$

(g)  $2Y + 20$

$$2(20)+20=60 \text{ \& } 2(5)=10$$

## Notes: Combining Random Variables Scenario

**Scenario:** A small business just leased a new computer and color laser printer for three years. The service contract for the computer offers unlimited repairs for a fee of \$100 a year plus a \$25 service charge for each repair needed. The company's research suggested that during a given year 86% of these computers needed no repairs, 9% needed to be repaired once, 4% twice, 1% three times, and none required more than three repairs.

1. Find the expected number of repairs this kind of computer is expected to need each year.

Show your work.  $(0)(.86) + (1)(.09) + (2)(.04) + (3)(.01) = .2$

**On average we expect each computer to need .2 repairs per year.**

2. Find the standard deviation of the number of repairs each year. Show your work.

$(0-.2)^2(.86) + (1-.2)^2(.09) + (2-.2)^2(.04) + (3-.2)^2(.01) = .5477$

3. What are the mean and standard deviation of the company's annual expense for the service contract?

$y = (25)(x) + 100$

**Mean =  $(25)(.2) + 100 = 105$  On average we expect to spend 105 dollars a year for repairs**

**Standard deviation  $(.25)(.5477) = 13.69$  dollars**

4. How many times should the company expect to have to get this computer repaired over the three-year term of the lease?

**$X_1 + X_2 + X_3 = .2 + .2 + .2 = .6$  on average we would expect a computer to need to be repaired .6 times over the three year time frame of the lease**

5. What is the standard deviation of the number of repairs that may be required during the three-year lease period? On what assumption does your calculation rest? Do you think this assumption is reasonable? Explain.

$X_1 + X_2 + X_3 = \sqrt{(.5477)^2 + (.5477)^2 + (.5477)^2} = .9486$  repairs

**The assumption rests on the fact that repairs are independent from year to year. I doubt that this is reasonable as things that break once tend to break more readily than things that haven't broken and the failures may be related to the care and skill of the computer operator**

6. The service contract for the printer estimates a mean annual cost of \$120 with standard deviation of \$30. What is the expected value and standard deviation of the total **annual** cost for the service contracts on computer and printer?

**Computer+Printer =  $105 + 120 = 225$  dollars  $\sigma = \sqrt{(13.69)^2 + (30)^2} = 32.98$  Dollars**

7. Which service contract should the company expect to cost more each year? How much more? With what standard deviation?

**Printer-Computer =  $120 - 105$  Printer costs 15 dollars more.  $\sigma = \sqrt{(13.69)^2 + (30)^2} = 32.98$  Dollars**

## Notes: Bivariate Data- A Very Brief Introduction

Bi-variate Data, as you may have surmised by the name, bi-variate data is two variable data. (Typically the variables are related or we believe them to be.) We have already spent a good deal of time dealing with bivariate data that is categorical in nature. (recall the two way tables). We have not addressed Bi-variate data that is quantitative in nature. **We will be discussing bi-variate data in great detail during the spring semester when we study regression.** I am only mentioning it here because you need to know that we are not skipping it. We are only postponing it.

**You will be accountable for the information below.**

**Examples of Quantitative bivariate data:** strength and age; electricity generation and hours of sunlight; blood pressure and weight. All of these have numerical measurements and are more than just counts.

In Algebra 1, you dealt with quantitative bi-variate data extensively when you dealt with linear equations. Consider the slope intercept form of the line  $y = mx + b$ . X and Y are variables and m provides the relationship between those variables.

As you may recall, a great deal of time was spent finding the equation of a line from a graph. This is very similar to what we will be doing in this class. There is one primary difference. In Algebra, the points all fell on the line. In statistics, we learned that data is not perfect and thus it is highly unlikely that all of the data points will lie on the line. Fitting a line to the data is known as regression and the line is called the least squares regression line because our goal is to minimize the distances the line is from the actual data. The closer the line to the actual data the better the fit.

You need to know that the graph of the line that contains the points of interest is known as a **scatterplot**. The y axis records the responses and the x axis provides the inputs. The plot demonstrates the relationship between the two variables.

## Notes: Introduction to Probability

**Probability:** The ratio comparing what you want to happen to the total number of possible outcomes. How likely something is to happen. **ALL** probabilities will range from 0 to 1 inclusive.

$$\text{Probability} = \frac{\text{Number of favorable outcomes}}{\text{Number of total possible outcomes}} \quad \text{or} \quad \frac{\text{What you want to happen}}{\text{Things that could happen}}$$

- The probability of the impossible is 0 or 0%
- The probability of something guaranteed to happen is 1 or 100%

**Random phenomenon**-possible outcomes are known, but the outcome for a particular event is unknown.

**Sample Space** is the list of all possible outcomes

**Trial**-a single event ex. 1 roll of a dice, 1 spin of a spinner, 1 flip of a coin, 1 guess on a test.

**Law of Large Numbers**-as the number of trials increases the percentage of successes moves closer to the expected number of successes-the theoretical number of successes. This occurs in the long run or repeated trials. **Does Not apply to individual trials or events.**

**Remember:** The past does not affect the future if the events are independent.

**Independent events:** Events in which the probability that one event occurs in no way affects the probability of the other event occurring.

**Multiplication Principal**-if two events are independent their probabilities can be multiplied to find the likelihood that both events will occur at the same time.

**Equation**  $P(A \cap B) = P(A) \times P(B)$  this reads the probability of the intersection of A and B, also A and B, is equal to the probability of A times the probability of B.

**DO NOT USE THIS UNLESS THE EVENTS ARE INDEPENDENT!!!**

**Hint:** "Or" means **Add** & "And" means **Multiply** when doing probability.

**Example:** A fair number cube is rolled and a fair coin is tossed.

- a) What is the probability of rolling a 3 and getting a tails?

$$\frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

- b) What is the probability of rolling two 4's in a row followed by tossing 2 heads in a row

$$\frac{1}{6} \times \frac{1}{6} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{144}$$

- c) What is the probability of rolling a 5 or 6 and then tossing 3 heads in a row?

$$\left(\frac{1}{6} + \frac{1}{6}\right) \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{24}$$

**Example:** A bag contains 3 A's and 5 B's.

- a) What is the probability of drawing an A, rolling a 6 on the dice and tossing a tail on the coin?

$$\frac{3}{8} \times \frac{1}{6} \times \frac{1}{2} = \frac{1}{32}$$

- b) What is the probability of drawing a B and getting a heads on the coin and either a 2 or 3 on the dice?

$$\frac{5}{8} \times \frac{1}{2} \times \left(\frac{1}{6} + \frac{1}{6}\right) = \frac{5}{48}$$



## Notes: Basic Probability Scenario

**Scenario 1:** Five multiple choice questions, each with four possible answers, appear on your history exam. What is the probability that if you just guess, you

b. get none of the questions correct?

$$\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{243}{1024}$$

c. get all of the questions correct?

$$\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{1024} \quad \text{Yup, this proves it, I really am Cherokee}$$

d. get at least one of the questions wrong?

$$1 - \left( \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \right) = \frac{1023}{1024}$$

e. get your first incorrect answer on the fourth question?

$$\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{3}{256}$$

**Scenario 2:** The Masterfoods company manufactures bags of Peanut Butter M&M's. They report that they make 10% each brown and red candies, and 20% each yellow, blue, and orange candies. The rest of the candies are green.

a. If you pick a Peanut Butter M&M at random, what is the probability that

i. it is green?

$$1 - (.1 + .1 + .2 + .2 + .2) = .2$$

ii. it is a primary color (red, yellow, or blue)?

$$.1 + .2 + .2 = .5$$

iii. it is not orange?

$$1 - .2 = .8$$

b. If you pick four M&M's in a row, what is the probability that

i. they are all blue?

$$(.2) \times (.2) \times (.2) \times (.2) = .0016$$

ii. none are green?

$$(.8) \times (.8) \times (.8) \times (.8) = .4096$$

iii. at least one is red?

$$1 - (.9 \times .9 \times .9 \times .9) = .3439$$

iv. the fourth one is the first one that is brown?

$$.9 \times .9 \times .9 \times .1 = .0729$$

c. After picking 10 M&M's in a row, you still have not picked a red one. A friend says that you should have a better chance of getting a red candy on your next pick since you have yet to see one. Comment on your friend's statement.

**If you are selecting from a single bag of M&M's, your friend is correct your chance will increase because the events are dependent. However, if you are talking about a massive quantity of M&M's, your friend is incorrect in that the probabilities are essentially independent so what happened previously has little or no effect on the next pick**

## Notes: Introduction to Probability

**Probability:** The ratio comparing what you want to happen to the total number of possible outcomes. How likely something is to happen. **ALL** probabilities will range from 0 to 1 inclusive.

- The probability of the impossible is 0 or 0%
- The probability of something guaranteed to happen is 1 or 100%

**Dependent events:** A group of simple events where one outcome **does** affect the other.

**Example:** A bag contains 3 A's and 5 B's

- a) What is the probability of drawing two A's at the same time?

$$\frac{3}{8} \times \frac{2}{7} = \frac{6}{56} = \frac{3}{28}$$

- b) What is the probability of drawing a B on the first draw and then drawing an A without replacement?

$$\frac{5}{8} \times \frac{3}{7} = \frac{15}{56}$$

- c) What is the probability of drawing 3 A's in a row without replacement?

$$\frac{3}{8} \times \frac{2}{7} \times \frac{1}{6} = \frac{6}{336} = \frac{1}{56}$$

- d) What is the probability of drawing 2 A's in a row and then drawing a B without replacement?

$$\frac{3}{8} \times \frac{2}{7} \times \frac{5}{6} = \frac{30}{336} = \frac{5}{56}$$

- e) What is the probability of drawing an A replacing the letter and then drawing a B?

$$\frac{3}{8} \times \frac{5}{8} = \frac{15}{64}$$

- f) What is the probability of drawing a B replacing the letter and then drawing an A?

$$\frac{5}{8} \times \frac{3}{8} = \frac{15}{64}$$

**Complementary Event-** the probability that an event will not occur. The probability of the complement is equal to 1 minus the probability that the event will occur.

**Equation**  $P(A^c) = 1 - P(A)$  this reads the probability of the complement A occurring is equal to 1- Probability of "A" occurring.  $A^c$  is similar to NOT "A".

- a) What is the probability of not rolling a 3 on a six sided dice?

$$1 - \frac{1}{6} = \frac{5}{6}$$

- b) What is the probability not drawing an ace from a standard deck of cards?

$$1 - \frac{4}{52} = \frac{48}{52} = \frac{12}{13}$$

- c) If I draw 3 cards without replacement what is the probability that none are aces?

$$\frac{48}{52} \times \frac{47}{51} \times \frac{46}{50} = \frac{4324}{5525}$$

**At Least Questions:** If you have **at least** questions you must use the complement

**Example:** If a student guesses on a 5 question multiple choice quiz with 4 answer choices what is the probability that the student gets at least 1 wrong.

**The complement of getting it wrong is getting it right so  $1 - .75 = .25$**

**Thus  $1 - (.25 \times .25 \times .25 \times .25 \times .25) = .9990$**

**Example:** A machine has 4 components the probability that the components do not fail during the year are: .97, .99, .98 and .96. If any component fails the machine fails, what is the probability that it fails?

**The complement of fail is doesn't fail and they gave us those values as .97, .99, .98, .96**

**So  $1 - (.97 \times .99 \times .98 \times .96) = 1 - (.9035) = .0965$**

### Notes: Complement Scenarios

**Scenario: 2004 Question 4:** Two antibiotics are available as treatment for a common ear infection in children.

- Antibiotic A is known to effectively cure the infection 60 percent of the time. Treatment with antibiotic A costs \$50.
- Antibiotic B is known to effectively cure the infection 90 percent of the time. Treatment with antibiotic B costs \$80.

The antibiotics work independently of one another. Both antibiotics can be safely administered to children. A health insurance company intends to recommend one of the following plans for treatment for children with this ear infection.

- Plan I: Treat with antibiotic A first. If it is not effective, then treat with antibiotic B.
- Plan II: Treat with antibiotic B first. If it is not effective, then treat with antibiotic A.

(a) If a doctor treats a child with an ear infection using plan I, what is the probability that the child will be cured?

If a doctor treats a child with an ear infection using plan II, what is the probability that the child will be cured?

$P(A)$  = Probability of a Cure with antibiotic A or  $P(A)$  = Probability that A works

$P(B)$  = Probability of a cure with antibiotic B or  $P(B)$  = Probability that B works

$$\begin{aligned} \text{(a) } P(\text{Treatment I Cure}) &= P(A) + P(A)^c \times P(B) \text{ which is } P(A) + (1-P(A)) \times P(B) \\ &= (.60) + (1-.60) \times .90 = .96 \end{aligned}$$

$$\begin{aligned} P(\text{Treatment II Cure}) &= P(B) + P(B)^c \times P(A) \text{ which is } P(B) + (1-P(B)) \times P(A) \\ &= (.90) + (1-.90) \times .60 = .96 \end{aligned}$$

(b) Compute the expected cost per child when plan I is used for treatment.  
Compute the expected cost per child when plan II is used for treatment.

$$\begin{aligned} \text{(b) } E_{\text{cost}}(\text{Treatment I}) &= (50) + P(A)^c \times 80 \\ &= 50 + (.40)(80) = \$82.00 \end{aligned}$$

$$\begin{aligned} E_{\text{cost}}(\text{Treatment II}) &= (80) + P(B)^c \times 50 \\ &= 80 + (.10)(50) = \$85.00 \end{aligned}$$

(c) Based on the results in parts (a) and (b), which plan would you recommend? Explain your recommendation.

**(c) Because both treatment protocols have the same expected cure rate, the health insurance company will recommend Treatment I over treatment II in the attempt to minimize costs.**

### Notes: Sample Space

On any probability problem it is always a good idea to create a sample space. A sample space is a list of all possible outcomes along with their frequencies. This is especially true of games or events that are played twice. The classic example of this are dice questions

**Example:** What is the probability of rolling a fair six sided dice twice and getting sum less than 8?  
Create a sample space to answer the question.

		Die 1					
		1	2	3	4	5	6
Die 2	1	(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
	2	(1, 2)	(2, 2)	(3, 2)	(4, 2)	(5, 2)	(6, 2)
	3	(1, 3)	(2, 3)	(3, 3)	(4, 3)	(5, 3)	(6, 3)
	4	(1, 4)	(2, 4)	(3, 4)	(4, 4)	(5, 4)	(6, 4)
	5	(1, 5)	(2, 5)	(3, 5)	(4, 5)	(5, 5)	(6, 5)
	6	(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	(6, 6)

Given that the sum is less than 8, what is the probability that the sum was equal to 5?

$$\frac{4}{21}$$

Given that the sum is less than 8, what is the probability that the sum was greater than 4?

$$\frac{15}{21}$$

Given that the sum is less than 8, what is the probability that the sum is 3 or less?

$$\frac{3}{21}$$

**What if the die had been weighted and the probabilities were as follows:**

Number Shown	1	2	3	4	5	6
Probability	.10	.10	.10	.20	.20	.30

Calculate the probability of Doubles

Combination	(1, 1)	(2, 2)	(3, 3)	(4, 4)	(5, 5)	(6, 6)
Probability	$(.10)^2$	$(.10)^2$	$(.10)^2$	$(.20)^2$	$(.20)^2$	$(.30)^2$

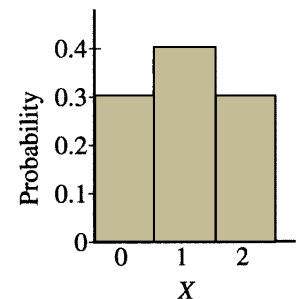
**Answer**  $.01+.01+.01+.04+.04+.09 = .20$

Calculate the Probability that the sum equals 6

Combination	(1, 5)	(2, 4)	(3, 3)	(4, 2)	(5, 1)
Probability	$(.1)(.2)$	$(.1)(.2)$	$(.10)^2$	$(.2)(.1)$	$(.2)(.1)$

**Answer**  $.02+.02+.01+.02+.02 = .09$

**Example:** A game of chance is played in which X, the number of points scored in each game, has the distribution shown at the right. What is the sampling distribution of the sum, Y, of the scores when the game is played twice?



Game 1 Score Distribution of (X)

Game 2 Score distribution of (X)	(0, 0)	(1, 0)	(2, 0)
	$(.3)(.3)=.09$	$(.4)(.3)=.12$	$(.3)(.3)=.09$
	(0, 1)	(1, 1)	(2, 1)
	$(.3)(.4)=.12$	$(.4)(.4)=.16$	$(.3)(.4)=.12$
	(0, 2)	(1, 2)	(2, 2)
	$(.3)(.3)=.09$	$(.4)(.3)=.12$	$(.3)(.3)=.09$

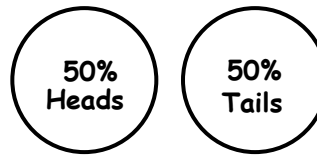
0	1	2	3	4
$.09$	$.12+.12$	$.09+.16+.09$	$.12+.12$	$.09$
$=.09$	$=.24$	$=.34$	$=.24$	$=.09$

## Notes: Classical Probability & Venn Diagrams

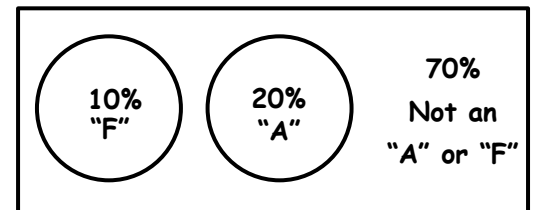
**Mutually Exclusive Events/Disjoint Events**-Events that cannot occur at the same time.

**Mutually Exclusive is the same as Disjoint**

**Example:** It is impossible to flip one coin and get both a heads and a tails on the same flip.



**Example:** You cannot make a 100% on the test and make an F on the test at the same time. Let's say the probability of an A is 20% the probability of an F is 10% and the probability of everything else sums to 70%.



**Complementary Event**- the probability that an event will not occur. The probability of the complement is equal to 1 minus the probability that the event will occur.

**Equation**  $P(A^c) = 1 - P(A)$  this reads the probability of the complement A occurring is equal to 1- Probability of "A" occurring.  $A^c$  is similar to NOT "A".

**Refer to the Examples Above.** Give the solutions and the associated Probabilities for each

The complement of heads is

**Not Heads or Tails**

The complement of tails is

**Not Tails or Heads**

The complement of "A" is

**Not A which is B, C, D, or F**

The complement of "F" is

**Not F which is A, B, C, or D**

**Independent Events**-Events that do not rely on one another. The outcome of one does not impact the outcome of the other. **If independent**  $P(A \cap B) = P(A) \times P(B)$

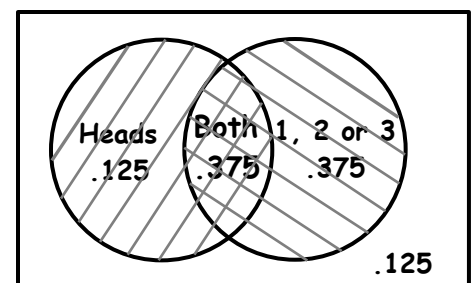
**Mutually Exclusive is not the same as Independent**

**Note:** Events **cannot** be mutually exclusive & Independent unless one event's probability is **ZERO**.

**Example:** A game requires that you roll a fair 4 sided die numbered 1 thru 4 and to flip a coin. The Venn Diagram illustrates the probability of getting a heads and rolling a 1, 2 or 3. Where did those probabilities come from?

**$P(1, 2, \text{ or } 3) = \frac{3}{4}$  and the probability of heads =  $\frac{1}{2}$**

**The probabilities are independent so they can be multiplied and  $\frac{3}{4} \times \frac{1}{2} = .375$**



**$P(\text{Heads}) = .5$  which equals  $.375 + .125$  and the  $P(1, 2, \text{ or } 3) = .75$  which equals  $.375 + .375$**

**General Addition Rule**-may be used anytime 2 probabilities are being summed but is necessary when 2 events that are not mutually exclusive are being summed. **Equation**  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$  **this reads (A union B), also A or B,** is equal to Probability of A plus the probability of B minus the probability of the intersection of (A and B)

**Note:**  $P(A \cap B)$  is the probability of both events occurring at the same time it is the overlap. There is not an overlap if the events are disjoint.

**Refer to the Game Example:** Use the formula calculate the probability of getting a 1, 2, or 3 or a head.

## Notes: Classical Probability & Venn Diagrams

**Example:** The Venn diagram illustrates the probability that a family has a cat or dog or both.

What is the probability of having both a Cat and a Dog?

$$P(C \cap D) = .12$$

What is the probability of having a Cat?

$$P(C) = .18 + .12 = .3$$

What is the probability of having a dog?

$$P(D) = .27 + .12 = .39$$

What is the Probability of not having a Cat?

$$P(C)^c = 1 - P(C) = 1 - .3 = .7$$

What is the Probability of not having a Dog?

$$P(D)^c = 1 - P(D) = 1 - .39 = .61$$

What is the Probability of not having both a Cat and a Dog?

$$1 - P(C \cap D) = 1 - .12 = .88$$

What is the Probability of having neither a cat nor a dog?

$$1 - (.18 + .12 + .27) = .43$$

Are having a dog or cat independent events? Justify your answer.

$$P(C) \times P(D) = P(C \cap D) \text{ and } (.3)(.39) \neq .12 \text{ therefore they are not independent}$$

What is the probability of having a cat or a dog? Use the formula:

$$P(C \cup D) = P(C) + P(D) - P(C \cap D) \text{ and } .30 + .39 - .12 = .57 \text{ so } P(C \cup D) = .57$$

What is the probability of having only a dog? What would probability of only a dog need to be for cat and dog to be independent events?

$$P(C \cap D) = P(C) \times P(D) \text{ and } .12 = (.30) \times P(D) \text{ so } P(D) = .40$$

Complete the Contingency table for the above situation.

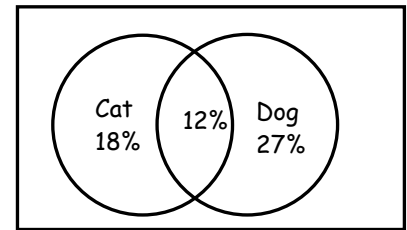
	Dog	No Dog	Total
Cat	.12	.18	.30
No Cat	.27	.53	.7
Total	.39	.61	1.0

Given someone has a cat, what is the probability that they have a dog?

$$\frac{P(\text{Intersection})}{P(\text{Given})} = \frac{P(\text{Cat} \cap \text{Dog})}{P(\text{Cat})} = \frac{.12}{.30}$$

Given that a person has a dog, what is the probability that they have a cat?

$$\frac{P(\text{Intersection})}{P(\text{Given})} = \frac{P(\text{Cat} \cap \text{Dog})}{P(\text{Dog})} = \frac{.12}{.39} \quad 78$$



## Notes: Venn Diagram Scenarios

Recall everyone's favorite antibiotic scenario.

**Scenario: 2004 Question 4:** Two antibiotics are available as treatment for a common ear infection in children.

- Antibiotic A is known to effectively cure the infection 60 percent of the time. Treatment with antibiotic A costs \$50.
- Antibiotic B is known to effectively cure the infection 90 percent of the time. Treatment with antibiotic B costs \$80.

The antibiotics work independently of one another. Both antibiotics can be safely administered to children. A health insurance company intends to recommend one of the following plans for treatment for children with this ear infection.

- Plan I: Treat with antibiotic A first. If it is not effective, then treat with antibiotic B.
- Plan II: Treat with antibiotic B first. If it is not effective, then treat with antibiotic A.

(a) If a doctor treats a child with an ear infection using plan I, what is the probability that the child will be cured?

If a doctor treats a child with an ear infection using plan II, what is the probability that the child will be cured?

(b) Compute the expected cost per child when plan I is used for treatment.  
Compute the expected cost per child when plan II is used for treatment.

(c) Based on the results in parts (a) and (b), which plan would you recommend? Explain your recommendation.

**Solve using a Venn Diagram:**

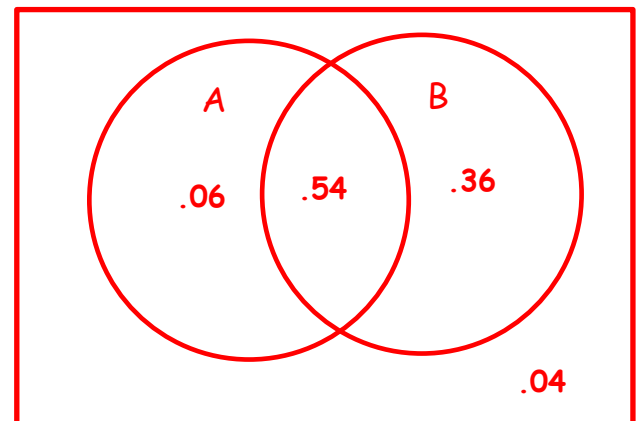
**Hint:** the events are independent, begin with the overlap

$$\begin{aligned} \text{Or } P(A) \cup P(B) &= P(A) + P(B) - P(A \cap B) \\ &= .60 + .90 - (.60)(.90) = .96 \end{aligned}$$

$$\begin{aligned} P(B) \cup P(A) &= P(B) + P(A) - P(A \cap B) \\ &= .90 + .60 - (.60)(.90) = .96 \end{aligned}$$

$$\text{Treatment I} = .60(50) + .40(130) = 82$$

$$\text{Treatment II} = .90(80) + .10(50) = 85$$



## Notes: 2-Way Tables & Conditional Probability

A 2-way table is way to organize the probabilities of two variables. The outside row and column totals are the **marginal** values and represent the probability of a variable's response out of the total population. The inside values of the table are **conditional** values and demonstrate the probability that the interaction of the 2 variables would result in a particular combination or intersection of responses.

**Consider the following Scenario:**

Suppose that 2% of a clinic's patients are known to have cancer. A blood test is developed that is positive in 98% of patients with cancer but is also positive in 3% of patients who do not have cancer.

**Step 1:** Identify the 2 variables

**Step 2:** Identify the one probability that is related to only one of the two variables and record its value at the bottom of the 1<sup>st</sup> column and fill in the remaining column.

		Cancer Status		Total
		Cancer	No Cancer	
Test Results	Positive	.0196 <small>(.02)(.98)</small>	.0294 <small>(.98)(.03)</small>	.049
	Negative	.0004 <small>(.02)(.02)</small>	.9506 <small>(.98)(.97)</small>	.951
	Total	.02	.98	1.0

**Step 3:** Label each variables potential value

**Step 4:** Using small writing, record the marginal probability of the columns into the conditional cells.

**Step 5:** Using small writing, record the remaining probabilities that are associated with the 2-variables.

**Step 6:** Multiply the values in the interior cells to find the conditional probabilities

**Step 7:** Sum each row's conditional probabilities to find the Marginal probabilities for each row.

a) What is the probability that the test is positive?

**.049**

b) What is the probability that the test is negative and the person has cancer?

**.0004**

c) Given the person has cancer, what is the probability that the test is negative?

$$\frac{.0004}{.02} = .02$$

d) If a person who is chosen at random from the clinic's patients is given the test and it comes out positive, what is the probability that the person actually has cancer?

$$\frac{.0196}{.049} = .40$$



## Notes: 2-Way Tables & Conditional Probability

**Example:** Leah is flying from Boston to Denver with a connection in Chicago. The probability her first flight leaves on time is 0.15. If the flight is on time, the probability that her luggage will make the connecting flight in Chicago is 0.95, but if the flight is delayed, the probability that the luggage will make it is only 0.65.

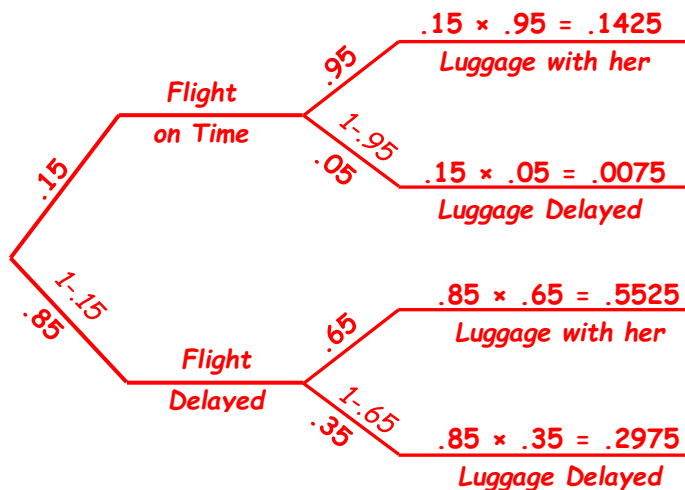
- (a) What is the probability that the luggage arrives in Denver with her?

**Make a 2-way table for the questions above.**

	Flight On time	Flight Delayed	Total
Luggage With her	<b>.1425</b> <i>.15 × .95</i>	<b>.5525</b> <i>.85 × .65</i>	<b>.695</b>
Luggage Delayed	<b>.0075</b> <i>.15 × (1 - .95)</i>	<b>.2975</b> <i>.85 × (1 - .65)</i>	<b>.305</b>
<b>Total</b>	<b>.15</b>	<b>.85</b>	<b>1.0</b>

**Draw a tree-diagram to answer the questions below**

- (b) Are the first flight leaving on time and the luggage making the connecting flight independent events? Explain.



## Notes: Conditional Probability & Contingency Tables

**Conditional Probability**- Gives the probability that event will happen given that another event has already happened. This is the same as a joint probability.

$$\text{Equation } P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ or } \frac{\text{conditional}}{\text{marginal}} \text{ or } \frac{\text{joint}}{\text{marginal}} \text{ or } \frac{\text{and/intersection}}{\text{given}}$$

- this means the probability of A given that B has occurred is equal to the probability of both events occurring divided by the probability of event B occurring.

**Example:** The table gives the results of a survey of the drinking and smoking habits of 1200 college students. Rows and columns have also been summed.

	Drinks	Doesn't Drink	Total
Smoke	315	165	480
Doesn't Smoke	585	135	720
Total	900	300	1200

A) What is the probability that a student smokes?

$$\frac{480}{1200} = .40$$

B) What is the probability that a student drinks?

$$\frac{900}{1200} = .75$$

C) What is the probability that a student smokes and drinks?

$$\frac{315}{1200} = .2625$$

D) What is the probability a student smokes given that the student drinks?

$$\frac{315}{900} = .35$$

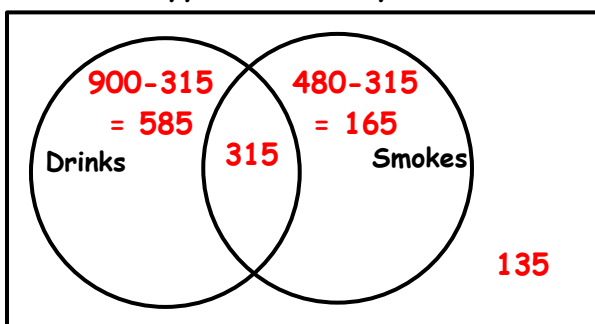
E) What is the probability that a student drinks given that the student smokes?

$$\frac{315}{480} = .65625$$

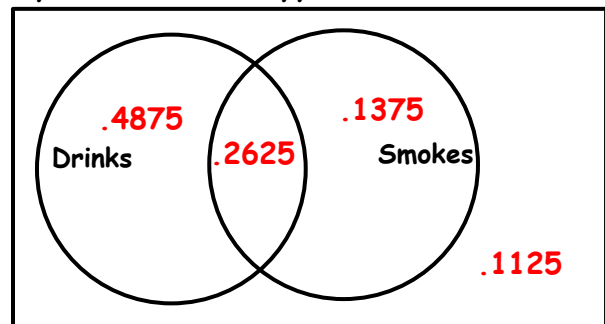
F) Why are the probabilities for D and E different?

Both D and E have the same overlap (drinkers and smokers intersection). They are different because neither is out of the entire sample of 1,200. They are out of two different subsets of the sample. One is taken from the subset of smokers while other is taken

Create a Venn Diagram using the actual values as opposed to the probabilities



Create a Venn Diagram using the probabilities as opposed to the values



## Notes: Conditional Probability & Independence

**Conditional Probability & Independence**—if  $P(A | B) = P(A)$  then the events are independent

Or if  $\frac{P(A \cap B)}{P(B)} = P(A)$  then independent. Also if  $\frac{P(A \cap B)}{P(A)} = P(B)$  then independent

**Example:** Recall the results of right-handedness and left handedness as relates to I.Q. scores.

	Right handed	Left Handed	Total
High IQ	190	10	200
Normal IQ	1,710	90	1800
Total	1900	100	2000

	Right handed	Left Handed	Total
High IQ	.095	.005	.10
Normal IQ	.855	.045	.90
Total	.95	.05	1.0

Previously we used proportionality to determine that hand dominance and IQ were independent.

**Demonstrate Independence using the multiplication rule:**

- Using Percentages (right table)

**If independent**  $P(A \cap B) = P(A) \times P(B)$

$$.045 = .05 \times .90$$

**.045 = .045 Therefore the events are independent**

- Using Actual Values (left table)

**If independent**  $P(A \cap B) = P(A) \times P(B)$

$$\frac{1710}{2000} = \frac{1900}{2000} \times \frac{1800}{2000}$$

**.855 = .855 Therefore the events are independent**

**Demonstrate Independence using Conditional Probability:**

- Using Percentages (right table)

**If independent**  $\frac{P(A \cap B)}{P(B)} = P(A)$

$$\frac{P(.045)}{P(.05)} = P(.05)$$

**.05 = .05 Therefore the events are independent**

- Using Actual Values (left table)

**If independent**  $\frac{P(A \cap B)}{P(B)} = P(A)$  then  $\frac{\frac{1710}{1800}}{\frac{1900}{2000}} = \frac{1710}{1800} = \frac{1900}{2000}$

**And .95 = .95**

**Therefore the events are independent**

### Notes: Conditional Probability & Contingency Tables Scenario

**2003 Form B Problem 2:** A simple random sample of adults living in a suburb of a large city was selected. The age and annual income of each adult in the sample were recorded. The resulting data are summarized in the table below.

Age Category	Annual Income			Total
	\$25,000-	\$35,001-\$50,000	Over \$50,000	
21-30	8	15	27	50
31-45	22	32	35	89
45-60	12	14	27	53
Over 60	5	3	7	15
Total	47	64	96	207

- (a) What is the probability that a person chosen at random from those in this sample will be in the 31-45 paid category?
- (b) What is the probability that a person chosen at random from those in this sample whose incomes are over \$50,000 will be in the 31-45 age category? Show your work.
- (c) Based on your answers to parts (a) and (b), is annual income independent of age category for those in this sample? Explain.

$$(a) P(31-45) = 89/207 = .43$$

$$(c) P(\text{over } 50,000 \cap 31-45) \mid P(\text{over } 50,000) = \frac{P(\text{over } 50,000 \cap 31-45)}{P(\text{over } 50,000)} = \frac{\frac{35}{207}}{\frac{96}{207}} = .365$$

(c) Let A be the age 31-45 category

Let B be the income category 50,000 and over

$$P(A) \mid P(B) = \frac{P(A \cap B)}{P(B)} \text{ and if independent, } P(A) \mid P(B) = \frac{P(A \cap B)}{P(B)} = P(A) \therefore P(A) \mid P(B) = P(A)$$

$$\text{If independent then } \frac{P(31-45 \cap \text{over } 50,000)}{P(\text{over } 50,000)} = P(31-45)$$

$$.365 \text{ does not equal } .43$$

However .365 does not equal .43 therefore the events are not independent. The age category 31-45 and the income category over 51,000 are not independent.

## Notes: Combinations

**Factorial:** The product of an integer and all integers below it.

**Example:** 5 Factorial is written  $5! = 5 \times 4 \times 3 \times 2 \times 1$ ; 8 Factorial is written  $8! = 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$

**Note:**  $0!$  is equal to 1

**Combination:** a subset or selection of a group in which the order does not matter. It is often expressed as n choose k and written  $n(C)k$  which means n combination k.

We most often will use the notation:  $\binom{n}{k}$

**The Formula for**  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$  tells us k the number of different subsets that can be taken from a group of size n

- n is the size of the entire group;
- k is the size of the selection from the group

**How many different subsets of 3 can be chosen from a group of 5?**  $\binom{5}{3}$

$$\frac{n!}{k!(n-k)!} = \frac{5!}{3!(5-3)!} = \frac{4!}{3!(2)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1(2 \times 1)} = \frac{120}{6(2)} = \frac{120}{12} = 10$$

**Consider the following:** You have 4 people Aaron, Bob, Corina and Daniella that we call A, B, C, & D that we are going to make groups with.

**1<sup>st</sup> group is to be 4 people:**  $\binom{4}{4} = \frac{4!}{4!(4-4)!} = \frac{4!}{4!(0)!} = \frac{4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1(1)} = \frac{24}{24(1)} = \frac{24}{24} = 1$

**Remember order doesn't matter**, so there is only one way to have a group of 4 selected from a group of 4 and that is to choose all 4 Aaron, Bob, Corina and Daniella.

**2<sup>nd</sup> group is to be 3 people:**  $\binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4!}{3!(1)!} = \frac{4 \times 3 \times 2 \times 1}{3 \times 2 \times 1(1)} = \frac{24}{6(1)} = \frac{24}{6} = 4$

Using the letters we could have the following groups:

**A, B, C; B, C, D; C, D, A; or D, A, B**

**3<sup>rd</sup> group is to be 2 people:**  $\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4!}{2!(2)!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1(2 \times 1)} = \frac{24}{2(2)} = \frac{24}{4} = 6$

Using the letters we could have the following groups:

**A, B; A, C; A, D; B, C; B, D; or C, D;**

**4<sup>th</sup> group is to be 1 person:**  $\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4!}{1!(3)!} = \frac{4 \times 3 \times 2 \times 1}{1(3 \times 2 \times 1)} = \frac{24}{1(6)} = \frac{24}{6} = 4$

Using the letters we could have the following groups:

**A; B; C; or D**

**Notes: Combinations**

The Formula for  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$  tells us **k** the number of different subsets that can be taken from a group of size **n**

- **n** is the size of the entire group;
- **k** is the size of the selection from the group

**Example:** You are making a sandwich. How many different combinations of 2 ingredients can you make with cheese, mayo and ham? **3 choose 2**

**Answer:** {cheese, mayo}, {cheese, ham} or {mayo, ham}  $\binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3!}{2!(1)!} = 3$

How many combinations of all 3 ingredients are there? **3 choose 3**

$$\{cheese, mayo, ham\} \binom{3}{3} = \frac{3!}{3!(3-3)!} = \frac{3!}{3!(0)!} = 1$$

**Try the following:**

- An election ballot asks voters to select three city commissioners from a group of six candidates. In how many ways can this be done?

$$\binom{6}{3} = \frac{6!}{3!(6-3)!} = \frac{6!}{3!(3)!} = 20$$

- A four-person committee is to be elected from an organization's membership of 11 people. How many different committees are possible?

$$\binom{11}{4} = \frac{11!}{4!(11-4)!} = \frac{11!}{4!(7)!} = 330$$

- You are on your way to Hawaii (Aloha) and of 15 possible books your parents say you can only take 10. How many different collections of 10 books can you take?

$$\binom{15}{10} = \frac{15!}{10!(15-10)!} = \frac{15!}{10!(5)!} = 3003$$

- There are 12 standbys who hope to get on your flight to Hawaii, but only 6 seats are available on the plane. How many different ways can the 6 people be selected?

$$\binom{12}{6} = \frac{12!}{6!(12-6)!} = \frac{12!}{6!(6)!} = 924$$

- To win the small county lottery, one must correctly select 3 numbers from 30 numbers. The order in which the selection is made does not matter. How many different selections are possible?

$$\binom{30}{3} = \frac{30!}{3!(30-3)!} = \frac{30!}{3!(27)!} = 4060$$

## Notes: Geometric Distributions

**Bernoulli Trials**- a random experiment with exactly two possible outcomes, "success" and "failure", in which the probability of success is the same every time the experiment is conducted.

**Characteristics:**

1. two possible outcomes (success and failure).
2. the probability of success,  $p$ , is constant.
3. the trials are independent

**Note:** When we don't have an infinite population, the trials are not independent. However it is still okay to proceed as long as the sample is smaller than 10% of the population. **M&M example.**

**Geometric Distribution**-has only 2 possible outcomes {success ( $p$ ) and failure ( $q$ )  $q=1-p$ } Each trial is independent, the probabilities cannot change & the **number of trials not known.**

**Equation:**  $q^{k-1}p$   $k$  is the number of trials until the 1<sup>st</sup> success

$$\text{Mean or } E(X) \mu = \frac{1}{p} \quad \text{Variance or } \text{Var}(X) \sigma^2 = \frac{q}{p^2} \quad \text{Standard deviation } \sigma = \frac{\sqrt{q}}{p}$$

**Geometrics are usually phrased as:**

- What is the probability that the 1<sup>st</sup> success will occur on a given trial (Geometric PDF)
- What is the probability that the 1<sup>st</sup> success will occur no later than or by (Geometric CDF)

**Example:** Suppose only 17% of guys at some high school are honest. What is the probability that the probability that the 1<sup>st</sup> honest guy that a new girl at school meets will be the 5<sup>th</sup> guy to whom she is introduced?

$$\frac{(1-.17)}{NH} \times \frac{(1-.17)}{NH} \times \frac{(1-.17)}{NH} \times \frac{(1-.17)}{NH} \times \frac{(.17)}{H} = (.83^4)(.17) = .0807$$

**Calculator:** 2<sup>nd</sup> VARS ↓ E:geometpdf(

What is the probability that that the **first** honest guy she encounters will be no later than the 4<sup>th</sup> guy she meets?

$$1^{\text{st}}: \frac{(.17)}{H} = .17$$

$$2^{\text{nd}}: \frac{(1-.17)}{NH} \times \frac{(.17)}{H} = .1411$$

$$3^{\text{rd}}: \frac{(1-.17)}{NH} \times \frac{(1-.17)}{NH} \times \frac{(.17)}{H} = .117113$$

$$4^{\text{th}}: \frac{(1-.17)}{NH} \times \frac{(1-.17)}{NH} \times \frac{(1-.17)}{NH} \times \frac{(.17)}{H} = .09720379$$

$$P(1^{\text{st}}) + P(2^{\text{nd}}) + P(3^{\text{rd}}) + P(4^{\text{th}}) = .52541679$$

**Calculator:** 2<sup>nd</sup> VARS ↓ F:geometcdf(

How many guys does she expect to have to meet to find the first honest guy?

$E(X) = \frac{1}{p} = \frac{1}{.17} = 5.88$  **On average we would expect that she would have to meet 5.88 guys in order to meet the first honest one.**

**Notes: Geometric Distribution Scenarios**

**Example:** The probability that any terminal is ready to transmit is .95. How many terminals will need to be tested until we find one that is ready to transmit?

$$E(X) = \frac{1}{p} = \frac{1}{.95} = 1.0526 \quad \text{On average, we would expect to have to test 1.0526 terminals in order to find one that is ready to transmit.}$$

What is the probability that the first terminal ready to transmit is the 1<sup>st</sup> terminal tested?

$$P(\text{First}) = .95$$

What is the probability that the first terminal ready to transmit is the 3<sup>rd</sup> terminal tested?

$$P(\text{Third}) = \frac{(1-.95)}{F} \times \frac{(1-.95)}{F} \times \frac{(.95)}{S} \rightarrow (.05)^2(.95)^1 = .002375$$

What is the expected number of terminals that will need to be tested until we find one not ready to transmit?

$$E(X) = \frac{1}{p} = \frac{1}{.05} = 20 \quad \text{On average, we would expect to have to test 20 terminals in order to find one that is NOT ready to transmit.}$$

What is the probability that the first terminal not ready to transmit is the 10<sup>th</sup> terminal tested?

$$P(\text{Not Ready}) = .05; \quad P(\text{Tenth not Ready}) = (.95)^9(.05)^1 = .0315$$

What is the probability that the first terminal not ready to transmit is the 20<sup>th</sup> terminal tested?

$$P(\text{Not Ready}) = .05; \quad P(\text{20th not Ready}) = (.95)^{19}(.05)^1 = .0189$$

What is the probability that the first terminal not ready to transmit is the 30<sup>th</sup> terminal tested?

$$P(\text{Not Ready}) = .05; \quad P(\text{30th not Ready}) = (.95)^{29}(.05)^1 = .0113$$

What is the likelihood that the first terminal not ready to transmit is found before the 100<sup>th</sup> terminal is checked?

$$P(1^{\text{st}} \text{ Not Ready} < 100); (.95)^0(.05)^1 + (.95)^1(.05)^1 + (.95)^2(.05)^1 \dots + (.95)^{98}(.05)^1 = .9937$$

**Example:** 2% of cars have a defective seatbelt. How many cars would you expect to have to check to find one with a defective seatbelt?

$$E(X) = \frac{1}{p} = \frac{1}{.02} = 50 \quad \text{On average, we would expect to have to check 50 cars in order to find one with a defective seatbelt.}$$

What is the probability that the first defective seatbelt is the 50<sup>th</sup> seatbelt checked?

$$P(50^{\text{th}}) = (.98)^{49}(.02)^1 = .0074$$

What is the probability that the first defective seatbelt is found prior to the 51<sup>st</sup> seatbelt being checked?

$$P(1^{\text{st}} \text{ Defective} < 51); (.98)^0(.02)^1 + (.98)^1(.02)^1 + (.98)^2(.02)^1 \dots + (.98)^{49}(.02)^1 = \dots .6358$$



## Notes: Binomial Distribution

**Bernoulli Trials**- a random experiment with exactly two possible outcomes, "success" and "failure", in which the probability of success is the same every time the experiment is conducted.

### Characteristics:

1. two possible outcomes (success and failure).
2. the probability of success,  $p$ , is constant.
3. the trials are independent

**Note:** When we don't have an infinite population, the trials are not independent. However it is still okay to proceed as long as the sample is smaller than 10% of the population. **M&M example.**

**Binomial Distribution**-has only 2 possible outcomes {**success ( $p$ ) and failure ( $q$ )  $q=1-p$** } Each trial is independent, the probabilities cannot change & the number of trials is pre-determined/fixed.

1. two possible outcomes (success and failure).
2. the probability of success,  $p$ , is constant as is the probability of failure  $q$  which is  $1-p$
3. the trials are independent
4. **a specified number of trials will be completed whether there is one success, no successes or multiple successes**
5. Number of trials, successes and failures are discrete

**Equation:** 
$$\sum_{k=0}^n \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

- $n$  is the number of trials
- $k$  is the number of successes

Mean or  $E(X)=np$

Variance or  $\text{Var}(X) \sigma^2 = npq$

Standard deviation  $\sigma = \sqrt{npq}$

### Binomials are usually phrased as:

- What is the probability of some number of successes in a given number of trials?  
(Binomial PDF)
- What is the probability of at least some number of successes in a given number of trials? (Binomial CDF) usually  $1$ -Binomial CDF
- What is the probability of no more than some number of successes in a given number of trials?  
(Binomial CDF)
- What is the probability that the number of successes in a given number of trials are between 2 values? (Binomial CDF)--(Binomial CDF of larger value) - (Binomial CDF smaller value)

**Notes: Binomial Distribution Scenarios**

$$\text{Equation: } \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

Mean or  $E(X)=np$ Variance or  $\text{Var}(X) \sigma^2 = npq$ Standard deviation  $\sigma = \sqrt{npq}$ **Binomials are usually phrased as:**

- **(Binomial PDF):** What is the probability of an exact number of successes for a pre-determined number of trials?

**Calculator Commands**Press 2<sup>nd</sup> VARS

↓A: binompdf(

Press ENTER

trials: Enter the number of trials. Press ENTER

p: Enter the probability of success. Press ENTER

x value: Enter the exact number of successes. Press ENTER 3 times

**Examples:**

Suppose that the probability of having a defective item is .3. Given a box of 5 light bulbs:

- What is the probability of having 0 defective items?

$$\binom{5}{0} (.3)^0 (1-.3)^{5-0} \rightarrow \binom{5}{0} (.3)^0 (.7)^5 = .16807$$

- What is the probability of having 1 defective items?

$$\binom{5}{1} (.3)^1 (1-.3)^{5-1} \rightarrow \binom{5}{1} (.3)^1 (.7)^4 = .36015$$

- What is the probability of having 2 defective items?

$$\binom{5}{2} (.3)^2 (1-.3)^{5-2} \rightarrow \binom{5}{2} (.3)^2 (.7)^3 = .3087$$

- What is the probability of having 3 defective items?

$$\binom{5}{3} (.3)^3 (1-.3)^{5-3} \rightarrow \binom{5}{3} (.3)^3 (.7)^2 = .1323$$

- What is the probability of having 4 defective items?

$$\binom{5}{4} (.3)^4 (1-.3)^{5-4} \rightarrow \binom{5}{4} (.3)^4 (.7)^1 = .02835$$

- What is the probability of having 5 defective items?

$$\binom{5}{5} (.3)^5 (1-.3)^{5-5} \rightarrow \binom{5}{5} (.3)^5 (.7)^0 = .00243$$

- What is the standard deviation and expected number of defects in a box of 5 items?

$E(X) = np \rightarrow 5(.3) = 1.5$  on average we would expect to have 1.5 defective light bulbs in a box.

$$\sigma = \sqrt{npq} \rightarrow \sqrt{5(.3)(1-.3)} \rightarrow \sqrt{5(.3)(.7)} = \sqrt{1.05} = 1.0247 \text{ defective bulbs}$$

## Notes: Binomial Distribution Scenarios

$$\text{Equation: } \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

Mean or  $E(X)=np$

Variance or  $\text{Var}(X) \sigma^2 = npq$

Standard deviation  $\sigma = \sqrt{npq}$

- **(Binomial CDF):** What is the probability of **no more** than some number of successes for a pre-determined number of trials?—**no more, no greater than, less than or equal to, and the special case less than.**

### Calculator Commands

Press 2<sup>nd</sup> VARS

↓B: binomcdf(

Press ENTER

trials: Enter the number of trials. Press ENTER

p: Enter the probability of success. Press ENTER

x value: Enter the **maximum** number of successes. Press ENTER 3 times

### Examples:

- What is the probability of having no more than 3 defective items in a box of 5?

Let  $D =$  the number of defects in a box; 0 1 2 3 4 5

$P(D \leq 3) = P(D = 0) + P(D = 1) + P(D = 2) + P(D = 3)$

$$\begin{aligned} & \binom{5}{0}(.3)^0(.7)^5 + \binom{5}{1}(.3)^1(.7)^4 + \binom{5}{2}(.3)^2(.7)^3 + \binom{5}{3}(.3)^3(.7)^2 \\ & = .16807 + .36015 + .3087 + .1323 = .96922 \end{aligned}$$

$$\binom{5}{0}(.3)^0(.7)^5 + \dots + \binom{5}{3}(.3)^3(.7)^2 = .96922$$

- What is the probability of having at least 3 defective items in a box of 5?

Let  $D =$  the number of defects in a box; 0 1 2 3 4 5

$P(D \geq 3) = P(D = 3) + P(D = 4) + P(D = 5)$

$$\begin{aligned} & \binom{5}{3}(.3)^3(.7)^2 + \binom{5}{4}(.3)^4(.7)^1 + \binom{5}{5}(.3)^5(.7)^0 \\ & = .1323 + .02835 + .00243 = .16308 \end{aligned}$$

$$\binom{5}{3}(.3)^3(.7)^2 + \dots + \binom{5}{5}(.3)^5(.7)^0 = .16308$$

*Note: Calculator 1-BinomCDF of 2*

- What is the probability of having **2 to 4** defective items in a box of 5?

Let  $D =$  the number of defects in a box; 0 1 2 3 4 5

$P(2 \leq D \leq 4) = P(D = 2) + P(D = 3) + P(D = 4)$

$$\begin{aligned} & \binom{5}{2}(.3)^2(.7)^3 + \binom{5}{3}(.3)^3(.7)^2 + \binom{5}{4}(.3)^4(.7)^1 \\ & = .3087 + .1323 + .02835 = .46935 \end{aligned}$$

$$\binom{5}{2}(.3)^2(.7)^3 + \dots + \binom{5}{5}(.3)^5(.7)^0 = .46935$$

*Note: Calculator BinomCDF of 4 - Binom CDF of 2*

## Notes: Binomial Distribution Scenarios

**Binomial Equation:**  $\sum_{k=0}^n \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$

$n$  is the number of trials,  $k$  is the number of successes and  $q$  is the number of failures

### Binomials are usually phrased as:

- **(Binomial PDF):** What is the probability of an exact number of successes for a pre-determined number of trials?

#### Calculator Commands

Press 2<sup>nd</sup> VARS

↓A: binompdf(

Press ENTER

trials: Enter the number of trials. Press ENTER

p: Enter the probability of success. Press ENTER

x value: Enter the **exact** number of successes. Press ENTER 3 times

- **(Binomial CDF):** What is the probability of **no more** than some number of successes for a pre-determined number of trials?—**no more, no greater than, less than or equal to, and the special case less than.**

#### Calculator Commands

Press 2<sup>nd</sup> VARS

↓B: binomcdf(

Press ENTER

trials: Enter the number of trials. Press ENTER

p: Enter the probability of success. Press ENTER

x value: Enter the **maximum** number of successes. Press ENTER 3 times

### Example: Johnson's statistics exam has 20 true false questions

- What is the probability that he gets no more than 14 correct?

Let  $C$  = the number of correct answers;

0 1 2 3 . . . 13 14 15 . . . 19 20

$$P(C \leq 14) = P(C = 0) + P(C = 1) + \dots + P(C = 14)$$

$$\binom{20}{0}(.5)^0(.5)^{20} + \binom{20}{1}(.5)^1(.5)^{19} + \dots + \binom{20}{14}(.5)^{14}(.5)^6 = .9793$$

*Note: Calculator BinomCDF of 14*

- What is the probability that he gets exactly 14 correct?

$$P(C = 14)$$

$$\binom{20}{14}(.5)^{14}(.5)^6 = .0369$$

*Note: Calculator BinomPDF of 14*

- What is the probability that he gets less than 10 correct?

$$P(C < 10) \rightarrow P(C \leq 9) = P(C = 0) + P(C = 1) + \dots + P(C = 9)$$

0 1 2 3 . . . 8 9 10 . . . 19 20

$$\binom{20}{0}(.5)^0(.5)^{20} + \binom{20}{1}(.5)^1(.5)^{19} + \dots + \binom{20}{9}(.5)^9(.5)^{11} = .4119$$

*Note: Calculator BinomCDF of 9*

- What is the standard deviation and mean for the number of correct answers?

$$E(X) = np \rightarrow 20(.5) = 10 \text{ on average we would expect to get 10 correct answers}$$

$$\sigma = \sqrt{npq} \rightarrow \sqrt{20(.5)(1-.5)} \rightarrow \sqrt{20(.5)(.5)} = \sqrt{5} = 2.236 \text{ correct answers}$$

## Notes: Binomial Distribution Scenarios

### More Complex Cases for the Binomial CDF

- **(Binomial CDF):** What is the probability of **at least or more than** some number of successes for a pre-determined number of trials? **1-Binomial CDF**
- **(Binomial CDF):** What is the probability that the number of successes for a pre-determined number of trials are between 2 values? **(Binomial CDF of larger value) - (Binomial CDF smaller value)**

**Example:** Johnson's statistics exam has 20 true false questions

- What is the probability that Johnson gets at least 14 correct?

Let  $C$  = the number of correct answers; 0 1 2 3 ... 12 13 14 15 ... 19 20

$$P(D \geq 14) = P(D = 14) + P(D = 15) \dots P(D = 19) + P(D = 20)$$

$$\binom{20}{14}(.5)^{14}(.5)^5 + \dots + \binom{20}{20}(.5)^{20}(.5)^0$$

**Calculator:** At least is the complement so the  $P(D \geq 14) = 1 - P(D < 14) \rightarrow 1 - P(D \leq 13)$

$$1 - (2^{\text{nd}} \text{ vars binomcdf trials} = 20 \text{ p} = .5 \times \text{value} = 13)$$

- What is the probability that Johnson gets at least 10 correct?

Let  $C$  = the number of correct answers; 0 1 2 3 ... 8 9 10 ... 19 20

$$P(D \geq 10) = P(D = 10) + P(D = 11) \dots P(D = 19) + P(D = 20)$$

$$\binom{20}{10}(.5)^{10}(.5)^{10} + \dots + \binom{20}{20}(.5)^{20}(.5)^0$$

**Calculator:** At least is the complement so the  $P(D \geq 10) = 1 - P(D < 10) \rightarrow 1 - P(D \leq 9)$

$$1 - (2^{\text{nd}} \text{ vars binomcdf trials} = 20 \text{ p} = .5 \times \text{value} = 9)$$

- What is the probability that Johnson gets more than 10 correct?

Let  $C$  = the number of correct answers; 0 1 2 3 ... 9 10 11 12 ... 19 20

$$P(D > 10) \rightarrow P(D \geq 11) = P(D = 11) + P(D = 12) + \dots + P(D = 19) + P(D = 20)$$

$$\binom{20}{11}(.5)^{11}(.5)^9 + \dots + \binom{20}{20}(.5)^{20}(.5)^0$$

**Calculator:** More means you must subtract what you do not want so  $1 - P(D \leq 10)$

$$1 - (2^{\text{nd}} \text{ vars binomcdf trials} = 20 \text{ p} = .5 \times \text{value} = 10)$$

- What is the probability that Johnson gets 9 to 13 questions correct?

Let  $C$  = the number of correct answers; 0 1 ... 8 9 10 11 12 13 14 ... 19 20

$$P(9 \leq D \leq 13) \rightarrow P(D = 9) + P(D = 10) + \dots + P(D = 12) + P(D = 13)$$

$$\binom{20}{11}(.5)^9(.5)^{11} + \dots + \binom{20}{20}(.5)^{20}(.5)^0$$

**Note:** Calculator BinomCDF of 13 - BinomCDF of 8

- What is the probability that Johnson gets between 9 and 13 questions correct?

Let  $C$  = the number of correct answers; 0 1 ... 8 9 10 11 12 13 14 ... 19 20

$$P(9 < D < 13) \rightarrow P(D = 10) + P(D = 11) + P(D = 12)$$

$$\binom{20}{10}(.5)^{10}(.5)^{10} + \dots + \binom{20}{12}(.5)^{12}(.5)^8$$

**Note:** Calculator BinomCDF of 12 - BinomCDF of 9

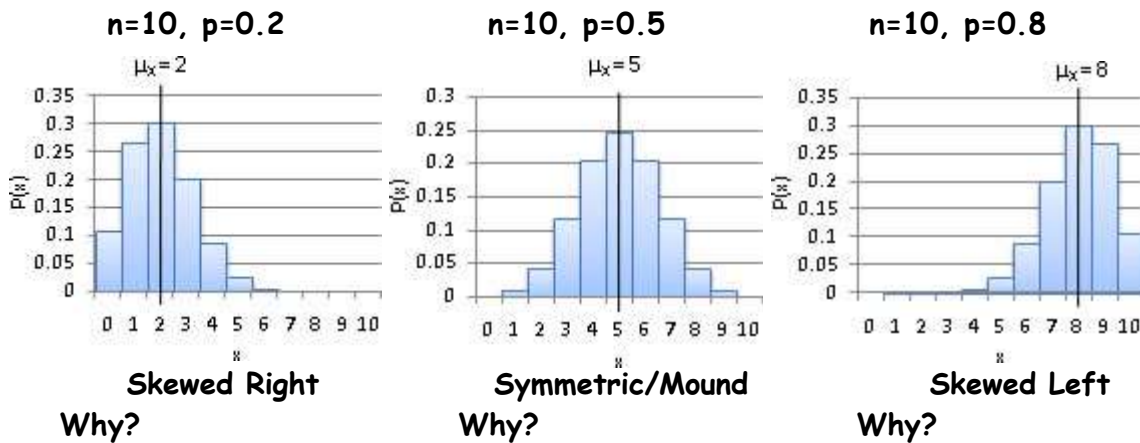
## Notes: Normal Approximation of the Binomial

If we have a very large number, the binomial becomes difficult to work with and can exceed the limits of our calculator. When this is the case we can use the normal distribution to approximate the binomial distribution.

### Requirements:

1. Two possible outcomes (success and failure).
2.  $np \geq 10$  number of successes is greater than 10
3.  $nq \geq 10$  number of failures is greater than 10

The best way to understand the effect of  $n$  and  $p$  on the shape of a binomial probability distribution is to look at some histograms, so let's look at some possibilities.

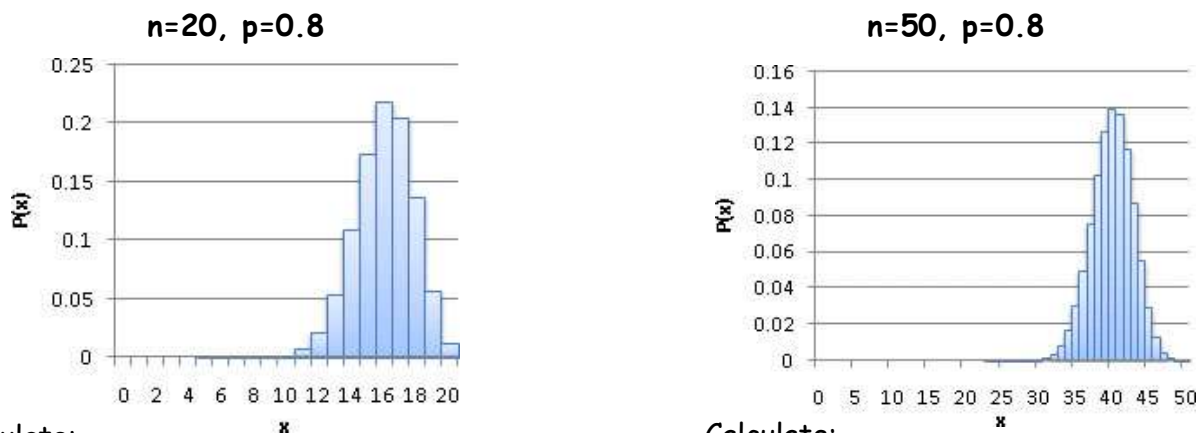


The tail is to the right.  
The mean is to the right  
of the median

The distribution is symmetric  
& mound shape.  
The mean & median are equal.

The tail is to the left.  
The mean is to the left of  
the median

Based on these, it would appear that the distribution is symmetric only if  $p=0.5$ , but this isn't actually true. Watch what happens as the number of trials,  $n$ , increases:



Calculate:

$$np = 20(.8) = 16$$

$$nq = 20(.2) = 4$$

Is the normal approximation appropriate?

No, because  $np$  is  $< 10$

Calculate:

$$np = 50(.8) = 40$$

$$nq = 50(.2) = 10$$

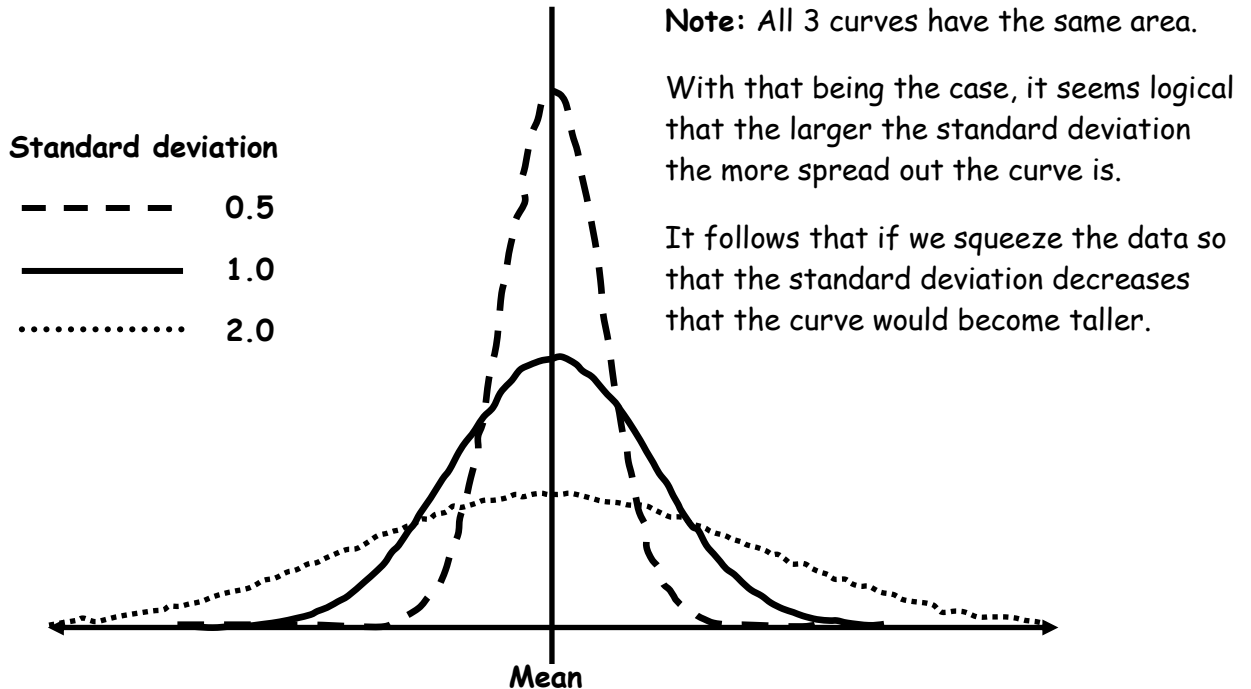
Is the normal approximation appropriate?

Yes, because both  $np$  &  $nq$  are  $\geq 10$

## Notes: Introduction to the Normal Curve

A **normal distribution** is an important theoretical idea used in statistics to model data applicable to large populations. The graph of a normal distribution is a symmetrical bell-shaped curve based on the mean and standard deviation of a sample, where the mean, median, and mode are the same. Below are the graphs of normal distributions with the same mean, but different standard deviations.

The mean is located at the center of the graph. Since this is also the median, half of the values lie above the mean and the other half lie below the mean. The shape of the curve is determined by the standard deviation, or spread of the data.



The 3 normal curves above show share the same mean but have different standard deviations. This indicative of the fact that the normal curve is actually a family of distributions that follows the equation below.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

There are an infinite number of normal curves because we have an infinite number of means and standard deviations.

The **normal distribution** can describe many things in the world such as scores on tests like the SAT. Characteristics in nature also fall along the normal curve, such as height or weight for adult men and women.

When the mean and standard deviation of a normal distribution are known, then several other values of the distribution can be calculated. It is important to be able to draw a curve and label the mean, as well as three standard deviations above and below the mean. Theoretically, almost all of the data should fall within 3 standard deviations (on each side) of the mean.

## Notes: Standardizing with Z-scores

At this point you might be thinking, "Great an infinite number of curves. How am I ever going to learn all of that? I should have been a student aid. Can I go see the counselor?"

Before you commit such an atrocity, you must consider that the normal curve is actually a family of curves and thus they share similar characteristics. Beyond the fact that they all follow the same general equation, they have the same distribution of data in that is distributed as follows:

- 68% of the data falls between plus or minus 1 standard deviations of the mean.
- 95% of the data falls between plus or minus 2 standard deviations of the mean.
- 99.7% of the data falls between plus or minus 3 standard deviations of the mean.

While that helps a great deal, there is still the issue of the fact that the normal curves have different standard deviations and means. For instance, how can I compare Terrence's ACT score to Sofia's SAT score? Answer we use Z-scores.

Recall that we briefly discussed that we can transform very different looking data using z-scores. So let's delve in a little deeper.

**Z-scores:** A z-score is a ratio that can be used to determine how many standard deviations a value lies from its mean by taking into account a measure of spread (the standard deviation of the distribution) and a measure of center (the mean of the distribution)

For individual values we use:  $Z = \frac{x - \mu}{\sigma}$

For sample means we use:  $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$  or  $Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$  where  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Z-scores can be computed for any shaped distribution. Consequently, we can use z-scores as a ruler to compare scores from different distributions. We can say that this persons scored 1.2 standard deviations above the mean and this person scored 1.8 standard deviations above the mean and we can do that regardless of the underlying distribution and regardless of the sample size.

**Scenario:** At a college the scores on the chemistry final exam are approximately normally distributed, with a mean of 75 and a standard deviation of 12. The scores on the calculus final are also approximately normally distributed, with a mean of 80 and a standard deviation of 8. A student scored 81 on the chemistry final and 84 on the calculus final. Relative to the students in each respective class, in which subject did this student do better?

<p><b>Chemistry</b>    <math>Z = \frac{x - \mu}{\sigma}</math></p> <p><math>\mu = 75</math></p> <p><math>\sigma = 12</math></p> <p><math>x = 81</math></p> <p><math>Z = \frac{81 - 75}{12}</math></p> <p><math>Z = .5</math></p>	<p><b>Calculus</b>    <math>Z = \frac{x - \mu}{\sigma}</math></p> <p><math>\mu = 80</math></p> <p><math>\sigma = 8</math></p> <p><math>x = 84</math></p> <p><math>Z = \frac{84 - 80}{8}</math></p> <p><math>Z = .5</math></p>
--	---

What would a 60 on the Chemistry final equate to on the calculus final?

<p><b>Chemistry</b>    <math>Z = \frac{x - \mu}{\sigma}</math></p> <p><math>\mu = 75</math></p> <p><math>\sigma = 12</math></p> <p><math>x = 60</math></p> <p><math>Z = \frac{60 - 75}{12}</math></p> <p><math>Z = -1.25</math></p>	<p><b>Calculus</b>    <math>Z = \frac{x - \mu}{\sigma}</math></p> <p><math>\mu = 80</math></p> <p><math>\sigma = 8</math></p> <p><math>x =</math></p> <p><math>-1.25 = \frac{x - 80}{8}</math></p> <p><math>x = 70</math></p>
---	---



## Notes: Introduction to the Standard Normal

With the normal distributions we don't have to just compare standard deviations, we can compare percentages, which is a good thing. I don't recall any student of mine ever telling someone that they were 1.2 standard deviations better. We always seem to compare percents. You will recall that we said that all normal curves have their data distributed as follows:

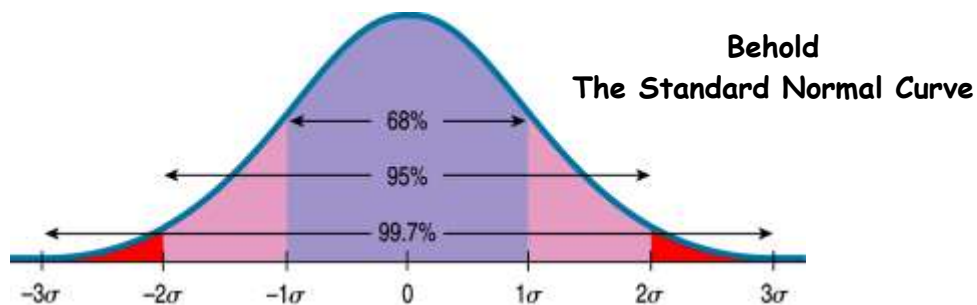
- 68% of the data falls between plus or minus 1 standard deviations of the mean.
- 95% of the data falls between plus or minus 2 standard deviations of the mean.
- 99.7% of the data falls between plus or minus 3 standard deviations of the mean.

We call this the **Empirical Rule**: you need to remember this.

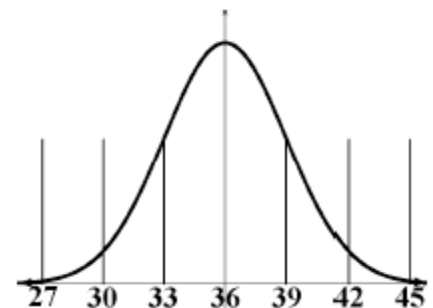
**Empirical Rule:** The Empirical Rule provides us with the amount of area under the normal curve or the percentage of data that lies between  $\pm 1$ ,  $\pm 2$ , and  $\pm 3$ , standard deviations. These percentages are true for all normal curves regardless of the mean or standard deviation of the normal curve.

**Caution:** The empirical rule only applies to normal curves. **Do not** use the empirical rule for any distribution that is not normal.

Because the family of normal curves all follow that distribution of data and because we have the z-score as a standardizing tool, we often use the standard normal which is a normal curve that has been scaled using z-scores and is centered at zero with a standard deviation of 1.



**Scenario:** Suppose the ages of adults taking a business course for the last 5 years at a local college are normally distributed as shown in the graph at right. Based on the 68-95-99.7 rule we know:

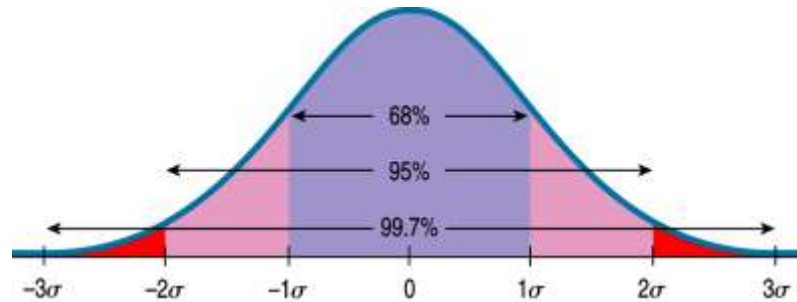


68% of all adults taking the business class are between the ages of **33 and 39**

95% of all adults taking the business class are between the ages of **30 and 42**

99.7% of all adults taking the business class are between the ages of **27 and 45**

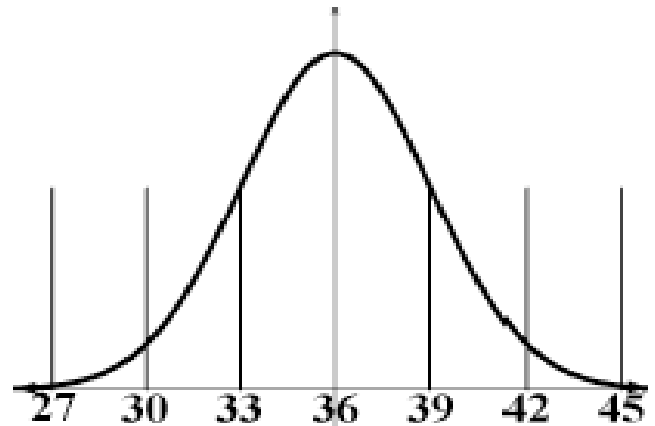
## Notes: Introduction to the Standard Normal



**Caution:** The percentages provided by the Empirical Rule are not percentile ranks. A Percentile rank provides the area from negative infinity to a specific value.

**Percentile Ranks:** A percentile rank provides us the amount of area or percent of data that is below a certain value. For a normal curve 68% of the data lies between  $\pm 1$  standard deviations. However, a z-score of 1 has a percentile rank of 84.13%. This means that 84.13% of the data fall below 1 standard deviation. The difference is that percentile ranks always measure the area from negative infinity up until the value.

**Scenario:** Suppose the ages of adults taking a business course for the last 5 years at a local college are normally distributed as shown in the graph at right. Based on the 68-95-99.7, calculate the following percentile ranks



27 \_\_\_\_\_

$$(1 - .997) \div 2 = .0015$$

30 \_\_\_\_\_

$$(1 - .95) \div 2 = .025$$

33 \_\_\_\_\_

$$(1 - .68) \div 2 = .16$$

36 \_\_\_\_\_

$$.50$$

39 \_\_\_\_\_

$$.5 + .68 \div 2 = .84$$

42 \_\_\_\_\_

$$.5 + .95 \div 2 = .975$$

45 \_\_\_\_\_

$$.5 + .997 \div 2 = .9985$$

Calculate the following. These are not percentile ranks.

30 to 36 \_\_\_\_\_

$$.50 - .025 = .475$$

36 to 42 \_\_\_\_\_

$$.975 - .50 = .475$$

39 to 45 \_\_\_\_\_

$$.9985 - .84 = .1585$$

27 to 33 \_\_\_\_\_

$$.16 - .0015 = .1585$$

27 to 30 \_\_\_\_\_

$$.025 - .0015 = .0235$$

42 to 45 \_\_\_\_\_

$$.9985 - .975 = .0235$$

## Notes: Normal Distribution Probability (Z-scores)

You should recall that a **z-score** is a ratio that can be used to determine how many standard deviations a value lies from its mean by taking into account a measure of spread (the standard deviation of the distribution) and a measure of center (the mean of the distribution)

For individual values we use:  $Z = \frac{x - \mu}{\sigma}$

For sample means we use:  $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$  or  $Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$  where  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Previously we discussed how Z-scores can be computed for any shaped distribution. Consequently, we can use z-scores as a ruler to compare scores from different distributions. We can say that this person **scored 1.2 standard deviations** above the mean and this person **scored 1.8 standard deviations** above the mean and we can do that regardless of the underlying distribution and regardless of the sample size.

We also discussed how we could compute exact percentages when we applied the Empirical rule to the normal distribution. However, not all z-scores are  $\pm 1$ ,  $\pm 2$ , and  $\pm 3$ . How can we find the percentile rank for associated with any Z-score. The answer is Calculus. We compute the area under the curve. Fortunately, that work is done for us either through the use of a table or the uses of our calculator.

As it turns out the students who scored 1.2 standard deviations and 1.8 standard deviations above the mean both came from normal populations. **The fact that they came from a normal population is critical.** If they had not come from normal populations, we would just have to shrug our shoulders because we don't have enough information.

**Consider the students who scored 1.2 and 1.8 Standard deviations above the mean.**

This is what we know:

- The population is normal - given
- The mean is 0 - we are working with a z-score from a normal population (*standard normal distribution*)
- The standard deviation is 1 (*standard normal distribution*)

### Steps for $\sigma$ 's 1.2 above

#### 1. Draw the graph

- Label the Mean
- Label the Upper & Lower
- Shade

#### 2. 2<sup>nd</sup> VARS $\downarrow$ 2:normalcdf (

- Lower: - infinity (-9999999)
- Upper: 1.2
- $\mu = 0$
- $\sigma = 1$

### Steps for $\sigma$ 's 1.8 above

#### 3. Draw the graph

- Label the Mean
- Label the Upper & Lower
- Shade

#### 4. 2<sup>nd</sup> VARS $\downarrow$ 2:normalcdf (

- Lower: - infinity (-9999999)
- Upper: **1.8**
- $\mu = 0$
- $\sigma = 1$

## Notes: Normal Distribution Probability (Z-scores)-Scenarios

### 1. Write the Equation

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ or } z = \frac{x - \mu}{\sigma}$$

$z$  = the number of standard deviations a value is from the mean

$\mu$  = the mean of the population or what is assumed to be true

$\bar{x}$  = the mean of the sample. **Note:**  $x$  = single value or score

$\sigma$  = the standard deviation of the population

$n$  = the size of the sample

### 2. Draw the graph

- Label the Mean
- Label the Upper and Lower
- Shade

### 3. Write the probability statement

### 4. 2<sup>nd</sup> VARS ↓ 2:normalcdf (

- Lower
- Upper
- $\mu$  =
- $\sigma$  =

Compute the z-scores and calculate the probabilities for the following: The life expectancy of a particular brand of lightbulb is normally distributed with a mean of 1500 hours and a standard deviation of 75 hours

a. What is the probability that the light bulb will last less than 1,410 hours?

$$z = \frac{x - \mu}{\sigma} \quad P(X < 1410) = P\left(z < \frac{1410 - 1500}{75}\right) \rightarrow P(z < -1.2) = .1151$$

2<sup>nd</sup> Vars Normal CDF  
Lower: -99999999  
Upper: -1.2  
 $\mu$ : 0 &  $\sigma$ : 1

b. What is the probability that the light bulb will last more than 1,583 hours?

$$z = \frac{x - \mu}{\sigma} \quad P(X < 1583) = P\left(z > \frac{1583 - 1500}{75}\right) \rightarrow P(z > 1.1067) = .1342$$

2<sup>nd</sup> Vars Normal CDF  
Lower: 1.1067  
Upper: 99999999  
 $\mu$ : 0 &  $\sigma$ : 1

c. What is the probability that the light bulb will last between 1,563 and 1,648 hours?

$$z = \frac{x - \mu}{\sigma} \quad P(1563 < X < 1648) = P\left(z < \frac{1648 - 1500}{75}\right) - P\left(z < \frac{1563 - 1500}{75}\right)$$

$$\rightarrow P(z < 1.973) - P(z < .84) = .1762$$

2<sup>nd</sup> Vars Normal CDF  
Lower: .84  
Upper: 1.973  
 $\mu$ : 0 &  $\sigma$ : 1

d. What is the probability that the light bulb will last between 1,416 and 1,677 hours?

$$z = \frac{x - \mu}{\sigma} \quad P(1416 < X < 1677) = P\left(z < \frac{1677 - 1500}{75}\right) - P\left(z < \frac{1416 - 1500}{75}\right)$$

$$\rightarrow P(z < 2.36) - P(z < -.12) = .8595$$

2<sup>nd</sup> Vars Normal CDF  
Lower: .84  
Upper: 1.973  
 $\mu$ : 0 &  $\sigma$ : 1

## Notes: Normal Distribution Probability (Actual Values)-Scenarios

Thus far we have been calculating probabilities for the normal using z-scores, however, the calculator permits us to go directly from the actual values to the answer all in one step. The only difference is that we now need to enter the actual mean, standard deviation and value(s) of interest.

### 1. Write the Equation

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ or } z = \frac{x - \mu}{\sigma}$$

$z$  = the number of standard deviations a value is from the mean

$\mu$  = the mean of the population or what is assumed to be true

$\bar{x}$  = the mean of the sample. **Note:**  $x$  = single value or score

$\sigma$  = the standard deviation of the population

$n$  = the size of the sample

### 2. Draw the graph

- Label the Mean
- Label the Upper and Lower
- Shade

### 3. Write the probability statement

### 4. 2<sup>nd</sup> VARS ↓ 2:normalcdf (

- Lower
- Upper
- $\mu$  =
- $\sigma$  =

Using the actual values calculate the probabilities for the following: A particular brand of lightbulb is **normally distributed** with a mean of 1500 hours and a standard deviation of 75 hours.

a. What is the probability that the light bulb will last less than 1,410 hours?

$$z = \frac{x - \mu}{\sigma} \quad P(X < 1410) = P\left(z < \frac{1410 - 1500}{75}\right) = .1151$$

2<sup>nd</sup> Vars Normal CDF

Lower: -99999999

Upper: 1410

$\mu$ : 1500 &  $\sigma$ : 75

b. What is the probability that the light bulb will last more than 1,583 hours?

$$z = \frac{x - \mu}{\sigma} \quad P(X > 1583) = P\left(z < \frac{1583 - 1500}{75}\right) = .1342$$

2<sup>nd</sup> Vars Normal CDF

Lower: 1583

Upper: 99999999

$\mu$ : 1500 &  $\sigma$ : 75

c. What is the probability that the light bulb will last between 1,563 and 1,648 hours?

$$z = \frac{x - \mu}{\sigma} \quad P(1563 < X < 1648) =$$

$$P\left(z < \frac{1648 - 1500}{75}\right) - P\left(z < \frac{1563 - 1500}{75}\right) = .1762$$

2<sup>nd</sup> Vars Normal CDF

Lower: 1563

Upper: 1648

$\mu$ : 1500 &  $\sigma$ : 75

d. What is the probability that the light bulb will last between 1,416 and 1,677 hours?

$$z = \frac{x - \mu}{\sigma} \quad P(1416 < X < 1677) =$$

$$P\left(z < \frac{1677 - 1500}{75}\right) - P\left(z < \frac{1416 - 1500}{75}\right) = .8595$$

2<sup>nd</sup> Vars Normal CDF

Lower: 1416

Upper: 1677

$\mu$ : 1500 &  $\sigma$ : 75

## Notes: Inverse Norm (Z-scores)

Thus far, we have been trying find the probability or a percentile rank given a z-score or actual value. We also have the ability to calculate a z-score or an actual value given a percentile rank **if we know that the underlying distribution is normal.**

### 1. Write the Equation

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ or } z = \frac{x - \mu}{\sigma}$$

$z$  = the number of standard deviations a value is from the mean

$\mu$  = the mean of the population or what is assumed to be true

$\bar{x}$  = the mean of the sample. **Note:**  $x$  = single value or score

$\sigma$  = the standard deviation of the population

$n$  = the size of the sample

### 2. Draw the graph

- Label the percent-be careful
- Label the Mean
- Label the Upper and Lower
- Shade

### 3. Write the probability statement

### 4. 2<sup>nd</sup> VARS ↓ 2:normalcdf (

- Lower
- Upper
- $\mu$  =
- $\sigma$  =

The Underlying population is **normal** for the examples that follow: **Flip back to page 98**

**Example:** A student claimed to have scored at the 99<sup>th</sup> percentile. What is the associated z-score?

**Answer: Z = 2.32**

2<sup>nd</sup> vars invNorm

Area: .99

$\mu$ : 0 &  $\sigma$ : 1

**Example:** A student claimed to have scored at the 25<sup>th</sup> percentile. What is the associated z-score?

**Answer: Z = -.67**

2<sup>nd</sup> vars invNorm

Area: .25

$\mu$ : 0 &  $\sigma$ : 1

**Example:** The upper 10% of students are associated with what z-score?

**Answer: Z = 1.28**      **Find the percentile rank 1 - .10 = .90**

2<sup>nd</sup> vars invNorm

Area: .90

$\mu$ : 0 &  $\sigma$ : 1

**Example:** The lowest 10% of students are associated with what z-score?

**Answer: Z = -1.28**

2<sup>nd</sup> vars invNorm

Area: .10

$\mu$ : 0 &  $\sigma$ : 1

**Example:** 80% of students scored above what z-score?

**Answer: Z = -.84**      **Find the percentile rank 1 - .80 = .20**

2<sup>nd</sup> vars invNorm

Area: .20

$\mu$ : 0 &  $\sigma$ : 1

## Notes: Inverse Norm (Actual Values)

Thus far, we have been trying find the probability or a percentile rank given a z-score or actual value. We also have the ability to calculate a z-score or an actual value given a percentile rank **if we know that the underlying distribution is normal.**

### 1. Write the Equation

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ or } z = \frac{x - \mu}{\sigma}$$

$z$  = the number of standard deviations a value is from the mean

$\mu$  = the mean of the population or what is assumed to be true

$\bar{x}$  = the mean of the sample. **Note:**  $x$  = single value or score

$\sigma$  = the standard deviation of the population

$n$  = the size of the sample

### 2. Draw the graph

- Label the percent-be careful
- Label the Mean
- Label the Upper and Lower
- Shade

### 3. Write the probability statement

### 4. 2<sup>nd</sup> VARS ↓ 2:normalcdf (

- Lower
- Upper
- $\mu$  =
- $\sigma$  =

For the following questions the population is normal with a mean of 79 and a standard deviation of 7.

**Example:** A student claimed to have scored at the 99<sup>th</sup> percentile. What is the score? **.99 < X**

$$z = \frac{x - \mu}{\sigma} \quad P\left(z = \frac{X - 79}{7}\right) = .99 \quad \begin{array}{l} 2^{\text{nd}} \text{ vars invNorm} \\ \text{Area: } .99 \\ \mu: 0 \ \& \ \sigma: 1 \end{array} \quad Z = 2.3263 \text{ so } \rightarrow 2.3262 = \frac{X - 79}{7} \quad \text{Solve: } X = 95.2844$$

**Example:** A student claimed to have scored at the 25<sup>th</sup> percentile. What is the score? **.25 < X**

$$z = \frac{x - \mu}{\sigma} \quad P\left(z = \frac{X - 79}{7}\right) = .25 \quad \begin{array}{l} 2^{\text{nd}} \text{ vars invNorm} \\ \text{Area: } .25 \\ \mu: 0 \ \& \ \sigma: 1 \end{array} \quad Z = -.6745 \text{ so } \rightarrow -.6745 = \frac{X - 79}{7} \quad \text{Solve: } X = 74.2786$$

**Ex** The upper 10% of students are associated with scores above? **X < 10% → .2 < X Complement Swap**

$$z = \frac{x - \mu}{\sigma} \quad P\left(z = \frac{X - 79}{7}\right) = .90 \quad \begin{array}{l} 2^{\text{nd}} \text{ vars invNorm} \\ \text{Area: } .90 \\ \mu: 0 \ \& \ \sigma: 1 \end{array} \quad Z = 1.2816 \text{ so } \rightarrow 1.2816 = \frac{X - 79}{7} \quad \text{Solve: } X = 87.9709$$

**Example:** The lowest 10% of students are associated with scores below what? **.10 < X**

$$z = \frac{x - \mu}{\sigma} \quad P\left(z = \frac{X - 79}{7}\right) = .10 \quad \begin{array}{l} 2^{\text{nd}} \text{ vars invNorm} \\ \text{Area: } .10 \\ \mu: 0 \ \& \ \sigma: 1 \end{array} \quad Z = -1.2816 \text{ so } \rightarrow -1.2816 = \frac{X - 79}{7} \quad \text{Solve: } X = 70.0291$$

**Example:** 80% of students scored above what? **X < 80% → .2 < X Complement Swap**

$$z = \frac{x - \mu}{\sigma} \quad P\left(z = \frac{X - 79}{7}\right) = .20 \quad \begin{array}{l} 2^{\text{nd}} \text{ vars invNorm} \\ \text{Area: } .20 \\ \mu: 0 \ \& \ \sigma: 1 \end{array} \quad Z = -.8416 \text{ so } \rightarrow -.8416 = \frac{X - 79}{7} \quad \text{Solve: } X = 73.1087$$

**Notes: Inverse Norm (Actual Values)**

**Scenario:** A packing machine is set to fill a cardboard box with a mean of 16.2 ounces of cereal. Suppose the amounts per box form a normal distribution with a standard deviation equal to 0.1 ounces.

- a. What percentage of the boxes will end up with at least 1 pound of cereal?

$$z = \frac{x-\mu}{\sigma} \quad P(X > 16) = P\left(z < \frac{16-16.2}{.1}\right) = .9772$$

- b. Ten percent of the boxes will contain less than what number of ounces?

**10% are less than  $x \rightarrow X < .10$**

$$z = \frac{x-\mu}{\sigma} \quad P\left(z = \frac{X-16.2}{0.1}\right) = .10 \quad \begin{array}{l} 2^{\text{nd}} \text{ vars invNorm} \\ \text{Area: } .10 \\ \mu: 16.2 \ \& \ \sigma: .1 \end{array} \quad Z = -1.2816 \text{ so } \rightarrow -1.2816 = \frac{X-16.2}{.1}$$

**Solve:  $X = 16.0718$**

- c. Eighty percent of the boxes will contain more than what number of ounces?

**80% are greater than  $X \rightarrow X < .80 \rightarrow .2 < X$  Complement Swap**

$$z = \frac{x-\mu}{\sigma} \quad P\left(z = \frac{X-16.2}{0.1}\right) = .20 \quad \begin{array}{l} 2^{\text{nd}} \text{ vars invNorm} \\ \text{Area: } .20 \\ \mu: 16.2 \ \& \ \sigma: .1 \end{array} \quad Z = -.8416 \text{ so } \rightarrow -.8416 = \frac{X-16.2}{.1}$$

**Solve:  $X = 16.1158$**

- c. 90% of the boxes will be below what weight?

**90% are less than  $X \rightarrow .90 < X$**

$$z = \frac{x-\mu}{\sigma} \quad P\left(z = \frac{X-16.2}{0.1}\right) = .90 \quad \begin{array}{l} 2^{\text{nd}} \text{ vars invNorm} \\ \text{Area: } .90 \\ \mu: 16.2 \ \& \ \sigma: .1 \end{array} \quad Z = 1.2816 \text{ so } \rightarrow 1.2816 = \frac{X-16.2}{.1}$$

**Solve:  $X = 16.3282$**

- d. 70% of the boxes are above what weight?

**70% are greater than  $X \rightarrow X < .70 \rightarrow .3 < X$  Complement Swap**

$$z = \frac{x-\mu}{\sigma} \quad P\left(z = \frac{X-16.2}{0.1}\right) = .30 \quad \begin{array}{l} 2^{\text{nd}} \text{ vars invNorm} \\ \text{Area: } .30 \\ \mu: 16.2 \ \& \ \sigma: .1 \end{array} \quad Z = -.5244 \text{ so } \rightarrow -.5244 = \frac{X-16.2}{.1}$$

**Solve:  $X = 16.1476$**

- e. 15% of the boxes are above what weight?

**15% are greater than  $X \rightarrow X < .15 \rightarrow .85 < X$  Complement Swap**

$$z = \frac{x-\mu}{\sigma} \quad P\left(z = \frac{X-16.2}{0.1}\right) = .85 \quad \begin{array}{l} 2^{\text{nd}} \text{ vars invNorm} \\ \text{Area: } .85 \\ \mu: 16.2 \ \& \ \sigma: .1 \end{array} \quad Z = 1.0364 \text{ so } \rightarrow 1.0364 = \frac{X-16.2}{.1}$$

**Solve:  $X = 16.3036$**



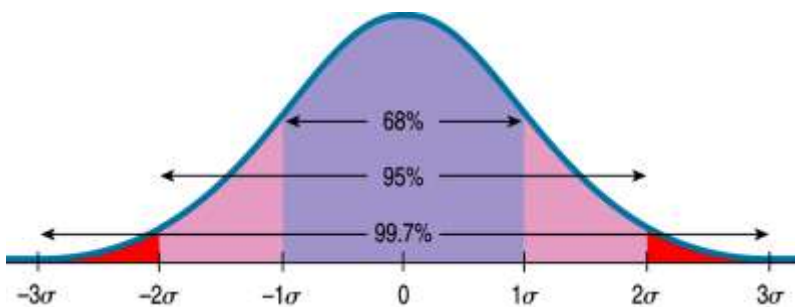
## Notes: Data Intervals for the Normal Distribution

We have learned to calculate probabilities based on any z-score that is associated with a normal distribution and we have expanded that ability to calculate the probabilities of actual values. We have also learned used the inverse norm function to work backwards from percentile ranking associated with a normal distribution to find the corresponding z-score or actual value.

Using the skills above and what we know about the Empirical rule we should be able to establish intervals centered about the mean that contain a specific percentage of the data.

Recall the Empirical rule states that normal curves have the following distribution:

- 68% of the data falls between plus or minus 1 standard deviations of the mean.
- 95% of the data falls between plus or minus 2 standard deviations of the mean.
- 99.7% of the data falls between plus or minus 3 standard deviations of the mean.



If via Algebra I rearrange  $z = \frac{x - \mu}{\sigma}$ ,

I arrive at the following:

$$x = \mu - z\sigma \quad \text{and} \quad x = \mu + z\sigma$$

which is most commonly expressed as:

$$x = \mu \pm z\sigma$$

Thus if I go back to my cereal box that has a mean of 16.2 ounces and a standard deviation of 1 ounce, I should be able to establish where certain weights of the boxes fall.

Use the equations above, to establish lower and upper bounds of an interval containing the following percentages:

- 68%

$$16.2 - (1)(.1) \quad \& \quad 16.2 + (1)(.1) \rightarrow 16.1 \text{ to } 16.3$$

- 95%

$$16.2 - (2)(.1) \quad \& \quad 16.2 + (2)(.1) \rightarrow 16.0 \text{ to } 16.4$$

- 80%

$$16.2 - (1.2815)(.1) \quad \& \quad 16.2 + (1.2815)(.1) \rightarrow 16.07185 \text{ to } 16.32815$$

$$\text{Lower Area: } \frac{(1 - \%)}{2}$$

$$\text{Upper Area: } \frac{(1 - \%)}{2} + \%$$

**Hint:** Draw and label the graph.

Calculate the associated z-scores.

- 98%

$$16.2 - (2.326)(.1) \quad \& \quad 16.2 + (2.326)(.1) \rightarrow 15.9674 \text{ to } 16.4326$$

- 99%

$$16.2 - (2.5758)(.1) \quad \& \quad 16.2 + (2.5758)(.1) \rightarrow 15.9424 \text{ to } 16.4576$$

## Notes: Normal Approximation of the Binomial

The normal distribution is continuous as opposed to the binomial which is discrete. In actuality, the number of people and animals are discrete (they really could be counted) However, the populations are so large that the binomial becomes difficult to work with and can exceed the limits of our calculator, as a consequence we use the normal distribution as an approximation of the binomial. So how large must a population be in order to justify using the normal approximation? To be honest the answer varies, but we will use:

- $np \geq 10$  number of successes is greater than 10
- $nq \geq 10$  number of failures is greater than 10

1. An archer with an 80% bull's eye rate will be shooting 200 arrows.

a) What are the mean and Standard deviation of the number of bull's eyes she might get?

$E(X) = np \rightarrow 200(.8) = 160$  on average we would expect her to hit the bull's eyes 160

$\sigma = \sqrt{npq} \rightarrow \sqrt{200(.8)(1-.8)} \rightarrow \sqrt{200(.8)(.2)} = \sqrt{32} = 5.656$  bull's eyes

b) Is a Normal Model appropriate here? Justify numerically

$$np = 200(.8) = 160 \geq 10$$

$$nq = 200(.2) = 40 \geq 10$$

Both conditions are met the normal approximation is appropriate

c) Use the Empirical rule to describe the distribution as to the number of bull's eyes she might get and draw the picture.

$$68\% = 160 \pm (1)5.656$$

$$154.344 \text{ to } 165.656$$

$$95\% = 160 \pm (2)5.656$$

$$148.688 \text{ to } 171.312$$

$$99.7\% = 160 \pm (3)5.656$$

$$143.032 \text{ to } 176.968$$

d) Would you be surprised if she made 140 bull's eyes or less?

Explain using both a binomial and the normal approximation.

$$P(X \leq 140) = P(X = 0) + \dots + P(X = 140)$$

$$z = \frac{x - \mu}{\sigma} \quad P(X \leq 140) = P\left(z < \frac{140 - 160}{5.656}\right)$$

$$\binom{200}{0} (.8)^0 (.2)^{200} + \dots + \binom{200}{140} (.8)^{140} (.2)^{60}$$

$$= .00020311 \quad (\text{normcdf})$$

$$= .0004977 \quad (\text{binomcdf } 140)$$

I would be surprised if she hit the target 140 or less times because the probability is low given her 80% target average.

2. The archer above purchases a new bow and hits 6 consecutive bull's eyes.

a) Is a normal approximation appropriate for the situation? Justify

$$np = 6(.8) = 4.8 < 10$$

$$nq = 6(.2) = 1.2 < 10$$

The conditions are not met. The normal approximation is NOT appropriate

b) Is this compelling evidence that the new bow has changed her bull's eye rate?

$$P(X = 6)$$

$$\binom{6}{6} (.8)^6 (.2)^0 = .2621 \quad (\text{binompdf } 6)$$

Given that she hits the target 80% of the time, we would expect her to hit it 6 times in a row 26% of the time.

This is not that unusual, I am not convinced the new bow has improved her target rate.

## Notes: Normal Approximation of the Binomial

The same archer with an 80% bull's eye rate is still shooting arrows.

3. With her new bow the archer hit 55-bull's eyes in 60 shots.

a) How many shots would you expect her to take to hit her first bulls-eye?

$$E(X) = \frac{1}{p} = \frac{1}{.8} = 1.25 \text{ On average, we would expect her to take 1.25 shots to hit the target}$$

b) Calculate the mean and standard deviation for the above scenario.

$$E(X) = np \rightarrow 60(.8) = 48 \text{ on average we would expect her to hit the bull's eye 48 times}$$

$$\sigma = \sqrt{npq} \rightarrow \sqrt{60(.8)(1-.8)} \rightarrow \sqrt{60(.8)(.2)} = \sqrt{9.6} = 3.0984 \text{ bull's eyes}$$

c) Is a normal approximation appropriate for the situation? Justify

$$np = 60(.8) = 48 \geq 10$$

$$nq = 60(.2) = 12 \geq 10$$

The conditions are met. The normal approximation is appropriate

d) Is this compelling evidence that the new bow has changed her bull's eye rate? **Use the Normal**

$$z = \frac{x-\mu}{\sigma} \quad P(X \geq 55) = P\left(z > \frac{55-48}{3.0984}\right) = .0119 \quad (\text{normcdf})$$

Given a success rate of 80%, there is only a 1.19% chance of hitting the target 55 times. That is pretty unlikely. It appears that her success rate may have changed.

4. Current data indicates that 16% of teens who try marijuana will become addicted. 800 adults who used marijuana as teens were sampled. 165 of these adults were found to be addicted to marijuana.

(a) Explain why a normal approximation is an appropriate for the situation?

$$np = 800(.16) = 128 \geq 10$$

$$nq = 800(.84) = 672 \geq 10$$

The conditions are met. The normal approximation is appropriate

(b) Calculate the mean and standard deviation of the distribution.

$$E(X) = np \rightarrow 800(.16) = 128 \text{ on average we would teens who tried marijuana to become addicted}$$

$$\sigma = \sqrt{npq} \rightarrow \sqrt{800(.16)(1-.16)} \rightarrow \sqrt{800(.16)(.84)} = \sqrt{107.52} = 10.37 \text{ addicts}$$

(c) How many would you expect to have to survey to find the first addict?

$$E(X) = \frac{1}{p} = \frac{1}{.16} = 6.25 \text{ On average we would expect to survey 6.25 adults who had tried marijuana as teenager to find our first addict.}$$

(d) Is this compelling evidence that the rate of addiction is different than previously believed? Use both the binomial and the normal approximation.

$$P(X \geq 165) = P(X = 165) + \dots + P(X = 800)$$

$$\binom{800}{165} (.16)^{165} (.84)^{635} + \dots + \binom{800}{800} (.16)^{800} (.84)^0$$

$$= .000325 \quad 1 - (\text{binomcdf } 164)$$

$$z = \frac{x-\mu}{\sigma} \quad P(X \geq 165) = P\left(z < \frac{165-128}{10.37}\right)$$

$$= .0001797 \quad (\text{normcdf})$$

It is unlikely that we would have such a large percentage of addicts if the actual rate was 16%. It appears that the rate of addiction is different than what is claimed

## Notes: Distribution of the Sample Mean-Normal Distribution

We have been focusing on the likelihood of getting a single value ( $n=1$ ) and for the Normal distribution we have been able to create the interval on which we expect the data to lie. Early on in this course we discussed that increasing sample size decreased our sampling error that is to say that the variation was lessened when we increased our sample size. To that end, we are going to begin looking at samples with  $n$  greater than 1.

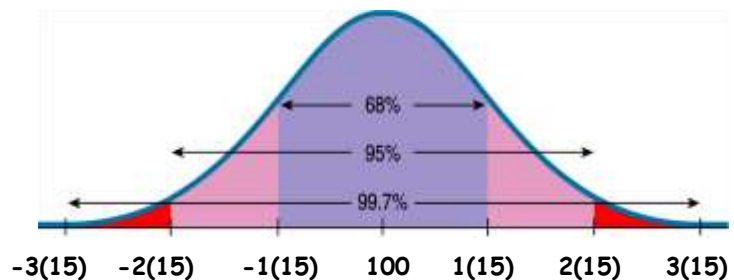
**We will begin by considering data that follow a Normal Distribution:** When we have a sample size greater than one our formula changes from  $z = \frac{x - \mu}{\sigma}$  to  $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ .

You will notice that the  $x$  changes to an  $\bar{x}$  because we have changed from working with a single value to working with an average of values. As we mentioned earlier, the sampling error diminishes as well and we are no longer working with just  $\sigma$  the standard deviation of the population.

**When we are working with the distribution of sample means the standard deviation of the means is:**  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$  which reads the standard deviation of the Means equals the standard deviation of the population divided by the square root of the sample size.

**So what does this mean and how do I apply it?**

**Scenario:** Consider that the Distribution of IQ's is normal with a mean of 100 and a standard deviation of 15. The graph of the Empirical Rule is:

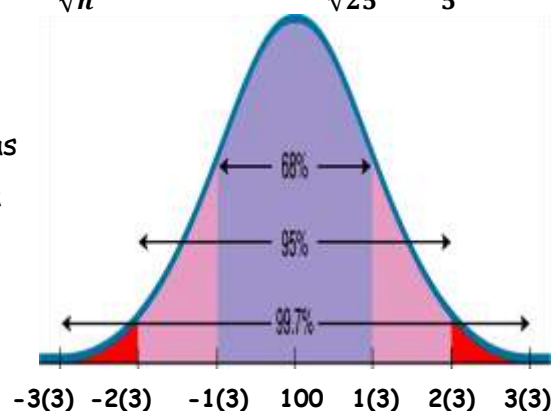


**Scenario:** Now consider the above situation, but with a sample of 25 as opposed to a single value. We would still expect the sample mean to be 100.

The **standard deviation** of the sample mean would be:  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$  we would find  $\frac{15}{\sqrt{25}} = \frac{15}{5} = 3$

While not to scale, the Graph of the Empirical Rule is narrower in width because of the smaller standard deviation. (Technically, the graph should be one fifth as wide  $\frac{1}{\sqrt{25}}$ ). Because the width is narrower and the area under the curve is still 1, the graph must be taller.

**Note:** The underlying distribution **Does Not Change**. The underlying distribution has a mean of 100 and a standard deviation of 15.



## Notes: Distribution of the Sample Mean-Normal Distribution

When we are dealing with an underlying distribution that is normal, the distribution of the sample means is always normal with  $\mu = \bar{x}$  and  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$  **regardless of sample size**

Unfortunately, not all distributions are normal. What if the distribution is highly skewed to right as are incomes? Fortunately, we have a solution the **Central Limit Theorem**.

**Central Limit Theorem**-States that the sampling distribution of the expected values (means) of a population with a mean of  $\mu$  and a standard deviation of  $\sigma$  will be approximately normal for any population regardless of the shape of the underlying population if the **sample is large enough-(30 or larger)**

The distribution of the sample means is normal with  $\mu = \bar{x}$  and  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$  if n is 30 or larger

**Note:** please be aware that the underlying distribution is NOT changing. It is the distribution of the sample means that becomes normal.

**\*\*\*Critical Distinction:** Do not confuse the Law of Large numbers with the Central Limit Theorem\*\*\*

- **Law of Large numbers:** as the number of trials increases the percentage of successes moves closer to the expected number of successes-the theoretical number of successes.
- **Central Limit Theorem:** For a large sample the distribution of the means is normal with the following statistics  $\mu = \bar{x}$  and  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

**Sampling Distribution of a Sample Mean** The sampling distribution of the sample means is approximately normal with mean ( $\mu$ ) and standard deviation  $\frac{\sigma}{\sqrt{n}}$

### Conditions Necessary to use the Normal for Sample Means

1. Must have a simple random sample
2. The population must be normal **OR** n must be at least 30
3.  $\sigma$  (the standard deviation) must be known for the population  
Given  $\sigma$  (the standard deviation of the population) is known, the standard deviation of the sample means  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
4.  $\mu$ . (the population mean) is given

## Notes: Distribution of the Sample Means Scenario

**Scenario:** The distribution of household incomes in the United States is skewed to the right. The population mean is 71,500 and the standard deviation of the population is 26,200.

- a) A sample of 100 is taken. What is shape of the distribution of the sample values along with the mean and standard deviation of the distribution of the sample values?

**Shape: Skewed Right—Note:** The distribution of sample values always follows the distribution of the population.

$$\mu = 71,500$$

$$\sigma = 26,200$$

- b) A sample of 25 is taken. What is the shape of the distribution of sample means along with the mean and standard deviation of the distribution of the sample means?

**Shape: Unknown—Note:** The sample size is not large enough for Central Limit Theorem to apply. With this size sample, the shape is not normal enough.

$$\mu = 71,500$$

$$\sigma = 26,200$$

- c) A sample of 110 is taken. What is the shape of the distribution of the sample means along with the mean and standard deviation of the distribution of the sample means?

**Shape: Normal—Note:** The distribution of sample means is normal because  $n \geq 30$

$$\mu = 71,500$$

$$\sigma_{\bar{x}} = \frac{26200}{\sqrt{110}} = 2,498.07$$

- d) A sample of 50 is taken. What is the shape of the distribution of the  $\bar{x}$ 's along with the mean and standard deviation of the distribution of the  $\bar{x}$ 's ?

**Shape: Normal—Note:** The distribution of sample means is normal because  $n \geq 30$

$$\mu = 71,500$$

$$\sigma_{\bar{x}} = \frac{26200}{\sqrt{50}} = 3,705.24$$

- e) What is the probability that a household is randomly selected from the population and earns more than 95,000?

**The distribution of the population is skewed and I cannot use the distribution of means because my sample size is 1 which is less than 30**

- f) What percent of households earn between 55,000 and 85,000?

**The distribution of the population is skewed and I do not have enough information to determine the percentages. I cannot use the distribution of means because it asked about the distribution of the actual values.**

## Notes: Distribution of the Sample Means Scenarios

### 1. Write the Equation

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$z$  = the number of standard deviations a value is from the mean

$\mu$  = the mean of the population or what is assumed to be true

$\bar{x}$  = the mean of the sample. **Note:**  $x$  = single value or score

$\sigma$  = the standard deviation of the population

$n$  = the size of the sample

### 2. Draw the graph

- Label the percent-be careful
- Label the Mean
- Label the Upper and Lower
- Shade

### 3. Write the probability statement

### 4. 2<sup>nd</sup> VARS ↓ 2:normalcdf (

- Lower
- Upper
- $\mu$  =
- $\sigma$  =

**Scenario:** Suppose that the outstanding credit card balances for young couples are normally distributed with a mean of \$650 with a standard deviation of \$120.

- a. What is the probability that an individual has a credit card balance in excess of 700 dollars?

$$z = \frac{x - \mu}{\sigma} \quad P(X > 700) \rightarrow P\left(z > \frac{x - \mu}{\sigma}\right)$$

$$= P\left(z > \frac{700 - 650}{120}\right) = .3385$$

**Note:** If the population distribution had been skewed, I could NOT do this problem because the sample size was  $< 30$

- b. In an **SRS of 20 couples**, what is the probability that the mean outstanding credit card balance exceeds \$700?

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad P(X > 700) \rightarrow P\left(z > \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= P\left(z > \frac{700 - 650}{\frac{120}{\sqrt{20}}}\right) = .0312$$

**Note:** If the population distribution had been skewed, I could NOT do this problem because the sample size was  $< 30$  and the Central Limit Theorem would not apply.

- c. In an **SRS of 80 couples**, what is the probability that the mean outstanding credit card balance exceeds \$700?

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad P(X > 700) \rightarrow P\left(z > \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= P\left(z > \frac{700 - 650}{\frac{120}{\sqrt{80}}}\right) = .000097 \approx 0$$

**Note:** If the population distribution had been skewed, I could still do this problem because the Central Limit Theorem states that the distribution of means becomes approximately normal for  $n \geq 30$

## Notes: Distribution of the Sample Means Scenarios

**Scenario:** The Wechsler Adult Intelligence Scale (WAIS) is a common "IQ test" for adults. The distribution of WAIS scores for persons over 16 years of age is approximately normal with mean 100 and standard deviation 15.

- a) What is the probability that a randomly chosen individual has a WAIS score of 123 or higher?

$$z = \frac{x - \mu}{\sigma} \quad P(X > 123) \rightarrow P\left(z > \frac{x - \mu}{\sigma}\right)$$

$$= P\left(z > \frac{123 - 100}{15}\right) = .0626$$

- b) What is the shape of the distribution of sample means along with the mean and standard deviation of the distribution for a SRS of 6 people?

**Shape: Normal**--The underlying population distribution is Normal and thus the distribution of the sample means is also normal regardless of the sample size

$$\mu = 100$$

$$\sigma_{\bar{x}} = \frac{15}{\sqrt{6}} = 6.124$$

- c) What is the probability that a SRS of 6 people has an average score of 123 or higher?

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad P(X > 123) \rightarrow P\left(z > \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= P\left(z > \frac{123 - 100}{\frac{15}{\sqrt{6}}}\right) = .000086 \approx 0$$

- d) A SRS of 40 people was taken. What is the shape of the distribution of the 40 test scores along with the mean and standard deviation?

**Shape: Normal**--The sample values shapes always follows the underlying population distribution and is therefore normal

$$\mu = 100$$

$$\sigma = 15$$

- e) What is the probability that a SRS of 9 people have an average score between 105 and 125?

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad P(105 < \bar{X} < 125)$$

$$= P\left(z < \frac{125 - 100}{\frac{15}{\sqrt{9}}}\right) - P\left(z < \frac{105 - 100}{\frac{15}{\sqrt{9}}}\right) = .1586$$

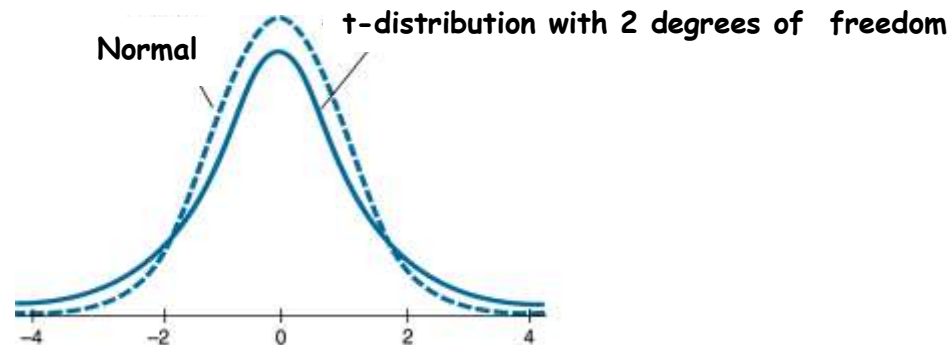
2<sup>nd</sup> Vars Normal CDF  
Lower: 105  
Upper: 125  
 $\mu: 100$  &  $\sigma: \frac{15}{\sqrt{9}}$



## Notes: T-Distribution & Distribution of Means

One of the requirements for using the normal distribution is that we know the standard deviation of the population which we rarely do. Enter the t distribution (aka, Student t distribution) is the sampling distribution of the t statistic.

The distribution of the t statistic is similar to the distribution of a standard score/Z-statistic. Both distributions are symmetrical with a mean of zero. Also, both distributions are bell-shaped, although the t-distribution has a larger variance, has more area in the tails, and has less data near the mean.



There are actually many different t-distributions. The particular model or form of the t-distribution is determined by its number of degrees of freedom which is one less than the sample size.

T distribution formula with n-1 degrees of freedom:  $t_{df} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$  or  $t_{df} = \frac{\bar{x} - \mu}{S_{\bar{x}}}$  where  $S_{\bar{x}} = \frac{s}{\sqrt{n}}$

$\bar{x}$  is the sample mean,  $\mu$  is the population mean,  $s$  is the standard deviation of the sample, and  $n$  is the sample size.

### Properties of the t-distribution

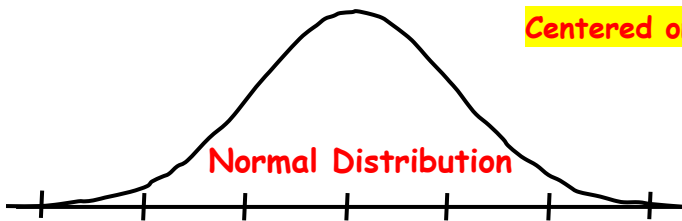
1. The t-distribution is different for different degree of freedom. **Degrees of Freedom equals n-1**
2. The t-distribution is centered at zero and is symmetric about zero.
3. The area under the curve is one.
4. As the sample size increases, the density curve of t-gets closer to the standard normal density curve.
5. The T-distribution has more area in the tails than does the normal.
6. The t-distribution has a greater variance than does the normal

### When to use the T-Distribution

1. When the data is **nearly normal (unimodal & symmetric)** and  $\sigma$  the standard deviation of the pop. is not known and there is no clear skewness or outliers it is ok to have a sample size less than 30.
2. When the  $\sigma$  standard deviation of the population is not known and the underlying data does not follow a normal curve the sample size must be 30 or larger.

**Note:** The t-distribution should *not* be used with small samples  $n < 30$  from populations unless the underlying distribution is approximately normal. **To check your data for normality, you will have to draw boxplots.**

## Notes: Comparing the Distribution of Means of the Normal and T-distributions

Centered on a mean ( $\bar{x}$ ) of 0

$$\bar{x} \pm z^* \times \frac{\sigma}{\sqrt{n}}$$

Area = 0.68

$$\sigma = 5 \quad 0 \pm .9944 \times \frac{5}{\sqrt{4}}$$

n = 4

$$(-2.486, 2.486)$$

Area = 0.95

$$\sigma = 5 \quad 0 \pm 1.96 \times \frac{5}{\sqrt{4}}$$

n = 4

$$(-4.9, 4.9)$$

Area = 0.997

$$\sigma = 5 \quad 0 \pm 2.967 \times \frac{5}{\sqrt{4}}$$

n = 4

$$(-7.419, 7.419)$$

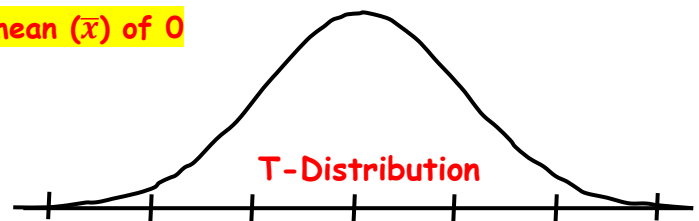
Calculator:  
2<sup>nd</sup> VARS

↓3: invNorm(

Press ENTER

area: (1-%)÷2

df: n-1



$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

Area = 0.68

$$s = 5 \quad 0 \pm 1.889 \times \frac{5}{\sqrt{4}}$$

n = 4

$$(-2.972, 2.972)$$

Area = 0.95

$$s = 5 \quad 0 \pm 3.182 \times \frac{5}{\sqrt{4}}$$

n = 4

$$(-7.956, 7.956)$$

Area = 0.997

$$s = 5 \quad 0 \pm 8.891 \times \frac{5}{\sqrt{4}}$$

n = 4

$$(-22.23, 22.23)$$

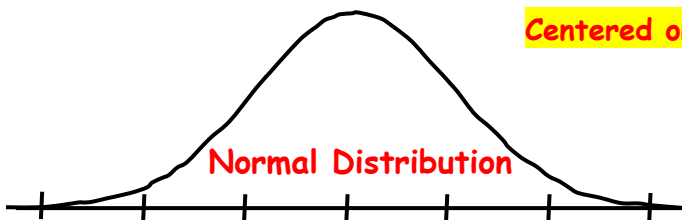
Calculator:  
2<sup>nd</sup> VARS

↓4: invT(

Press ENTER

area: (1-%)÷2

df: n-1

Centered on a mean ( $\bar{x}$ ) of 3

$$\bar{x} \pm z^* \times \frac{\sigma}{\sqrt{n}}$$

Area = 0.68

$$\sigma = 5 \quad 3 \pm .9944 \times \frac{5}{\sqrt{4}}$$

n = 4

$$(0.5138, 5.486)$$

Area = 0.95

$$\sigma = 5 \quad 3 \pm 1.96 \times \frac{5}{\sqrt{4}}$$

n = 4

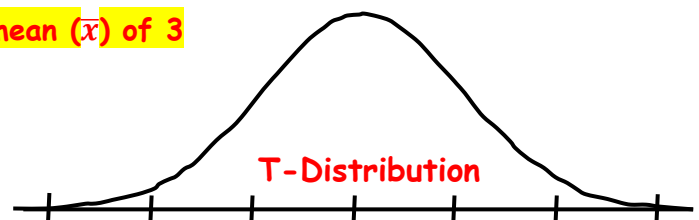
$$(-1.9, 7.9)$$

Area = 0.997

$$\sigma = 5 \quad 3 \pm 2.967 \times \frac{5}{\sqrt{4}}$$

n = 4

$$(-4.419, 10.419)$$



$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

Area = 0.68

$$s = 5 \quad 3 \pm 1.889 \times \frac{5}{\sqrt{4}}$$

n = 4

$$(.0277, 5.972)$$

Area = 0.95

$$s = 5 \quad 3 \pm 3.182 \times \frac{5}{\sqrt{4}}$$

n = 4

$$(-4.956, 10.956)$$

Area = 0.997

$$s = 5 \quad 3 \pm 8.891 \times \frac{5}{\sqrt{4}}$$

n = 4

$$(-19.23, 25.23)$$

## Notes: Comparing the Distribution of Means of the Normal and T-distributions

### Distribution of Sample Means for the T-distribution:

- The set of all sample means is normally distributed
- The mean of the set of sample means **equals** the population mean
- The standard deviation of the sample mean ( $\bar{x}$ ) is **approximately**  $S_{\bar{x}} = \frac{s}{\sqrt{n}}$

**Scenario 1:** A college's data about the incoming freshman indicates that the mean of their high school GPA's was 3.4. The students are randomly assigned to freshman writing seminars in groups of 25. Assume that GPA of the students follow normal distribution.

- a. What is the probability that the Mean GPA of the group is less than **3.3** given that the standard deviation of the population is known and is .35?

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad P(\bar{x} < 3) \rightarrow P\left(z < \frac{3.3 - 3.4}{\frac{.35}{\sqrt{25}}}\right)$$

$$= P(z < -1.4286) = .0766$$

2<sup>nd</sup> Vars Normal CDF  
Lower: - infinity  
Upper: 3.3  
 $\mu: 3.4$  &  $\sigma: \frac{.35}{\sqrt{25}}$

What is the probability that the Mean GPA of the group is less than 3 given that the standard deviation of the sample is .35?

**Calculator:**

2nd VARS

↓ 6:tcdf(

Press ENTER

Lower:

Upper:

df: n-1

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad P(\bar{x} < 3) \rightarrow P\left(t < \frac{3.3 - 3.4}{\frac{.35}{\sqrt{25}}}\right)$$

$$= P(t < -1.4286) = .0830$$

2<sup>nd</sup> Vars tCDF  
Lower: - infinity  
Upper: -1.4286  
df: 25-1

**Scenario 2:** Assume that the duration of human pregnancies can be described as a normal model with mean 266 days.

- a. Suppose a certain obstetrician is currently providing prenatal care to 60 pregnant women. What's the probability that the mean duration of these patients' pregnancies will be less than 260 days if the standard deviation of the population is 16 days?

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad P(\bar{x} < 260) \rightarrow P\left(z < \frac{260 - 266}{\frac{16}{\sqrt{60}}}\right)$$

$$= P(z < -2.905) = .0018$$

2<sup>nd</sup> Vars Normal CDF  
Lower: - infinity  
Upper: 260  
 $\mu: 266$  &  $\sigma: \frac{16}{\sqrt{60}}$

- b. Suppose a certain obstetrician is currently providing prenatal care to 60 pregnant women. What's the probability that the mean duration of these patients' pregnancies will be less than 260 days if the standard deviation of sample is 16 days?

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad P(\bar{x} < 260) \rightarrow P\left(t < \frac{260 - 266}{\frac{16}{\sqrt{60}}}\right)$$

$$= P(t < -2.905) = .0026$$

2<sup>nd</sup> Vars tCDF  
Lower: - infinity  
Upper: -2.905  
df: 60-1

## Notes: Confidence Intervals 1-Sample T

The reason we collect data is we usually don't know something about a population and we are forced to take a sample to use as a proxy or estimate of the population parameter of interest.

If we only have sample data, we are never able to say what the exact population mean is. However, we can claim, with a certain level of confidence, that the population mean lies within a specified interval or range. That range is called a confidence interval and is comprised of 2 components the estimate and a margin of error. The estimate or best guess for the population mean is the sample mean. The margin of error creates a range below and above the estimate. The range is based on the sample standard deviation of the sample and the confidence level.

**Confidence Intervals:** estimate  $\pm$  margin of error  $\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$

**Estimate:** The sample mean ( $\bar{x}$ ) estimates the population parameter  $\mu$ .

**Margin of error:** For a given confidence level, the maximum distance a sample statistic may lie from the true population parameter

- The **margin of error** is the (t-score) $\times$ (the standard error of  $\bar{x}$   $SE_{\bar{x}}$ ) or  $t^* \times \frac{s}{\sqrt{n}}$
- S is the standard deviation of the sample and n is the sample size

Because we only have sample data, we are never able to say what the exact population mean is. We can claim, with a certain level of confidence, that the population mean lies within a specified interval.

- We are \_\_\_\_\_% confident that the true population mean lies within the interval \_\_\_\_\_
- In repeated testing, we expect this method will capture the true population mean \_\_\_\_\_% of the time.

**We Cannot Claim:**

- \_\_\_\_\_% of the values lie within the interval—*the interval may be wrong*
- \_\_\_\_\_% chance that a randomly selected \_\_\_\_\_ will lie within the interval—*the interval may be wrong*
- \_\_\_\_\_% of samples will result in this interval—*the interval may be wrong*
- \_\_\_\_\_% probability that the true population mean lies within the interval

**(do not use probability; use confidence)**

**We Can Claim:** \_\_\_\_\_% Confident that the true population mean lies within the interval

**We Can Claim:** (\_\_\_\_\_% of the intervals collected by this method will capture the true population mean)

**Note:** We are making a claim about the population mean and not the sample mean.

We do not need to make a claim about the sample mean. The sample mean will be in the center of the interval 100% of the time.

**Note:**

- As we increase the level of confidence the margin of error or spread increases.
- As we increase alpha the margin of error decreases and our confidence decreases.
- As the sample size increases the margin of error or spread decreases.

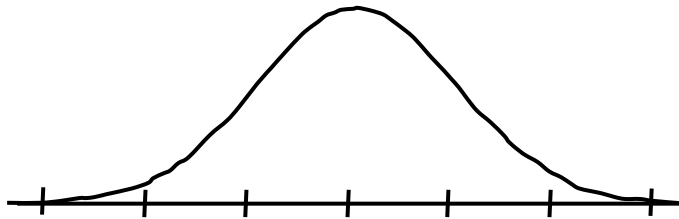
## Notes: Confidence Intervals when $\sigma$ is known

What is a confidence interval anyway?

Consider the empirical rule graph below

(Remember the empirical rule applies the Normal Distribution and not the T-distribution)

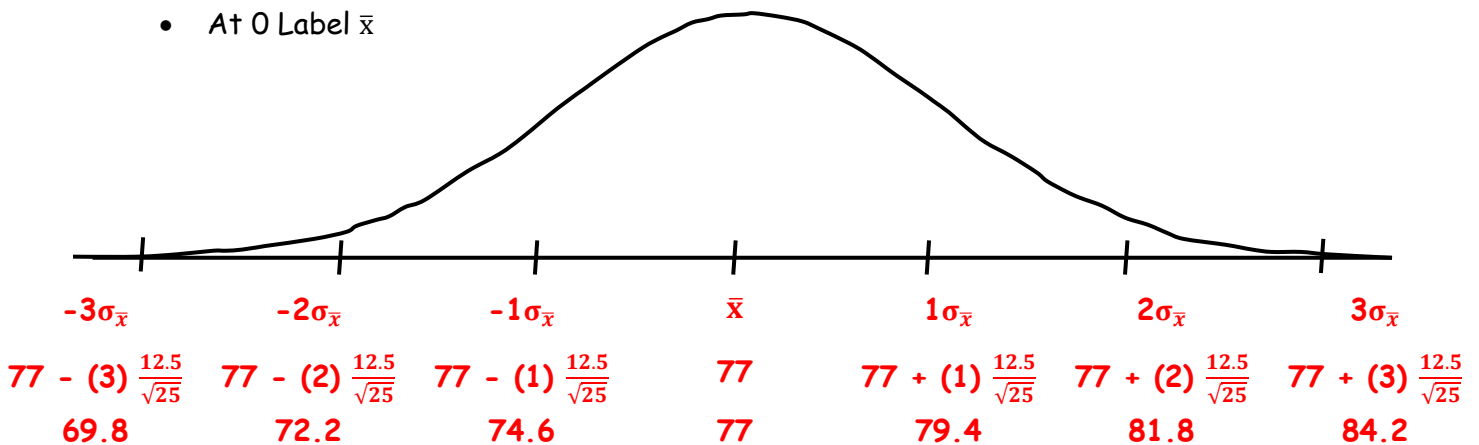
- Label the standard deviations  $-3\sigma, -2\sigma, \dots, 3\sigma$
- At 0 Label  $\bar{x}$
- If we claim that we are 95% confident, we believe that the true population mean is located somewhere between  $-2$  and  $+2$  standard deviations below or above  $\bar{x}$



**Making it Real (confidence intervals for a normal distribution)**

Consider the empirical rule graph below

- Label the standard deviations  $-3\sigma_{\bar{x}}, -2\sigma_{\bar{x}}, \dots, 3\sigma_{\bar{x}}$  etc.
- At 0 Label  $\bar{x}$



The population standard deviation for the above is known ( $\sigma = 12.5$ ). The Sample mean ( $\bar{x}$ ) is 77  $n=25$

- Below  $\bar{x}$  record the sample mean
- Calculate  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12.5}{\sqrt{25}}$
- Below  $-3\sigma_{\bar{x}}$  write  $-3(\frac{12.5}{\sqrt{25}})$ ; Below  $-2\sigma_{\bar{x}}$  write  $-2(\frac{12.5}{\sqrt{25}})$ ; and continue all the way to  $3\sigma_{\bar{x}}$
- Multiply the values and record below each

**Create a 68% confidence interval**

- Write  $\bar{x} \pm 1\sigma_{\bar{x}}$
- Plug in the values
- Solve

**(74.6, 79.4)**

**Create a 95% confidence interval**

- Write  $\bar{x} \pm 2\sigma_{\bar{x}}$
- Plug in the values
- Solve

**(72.2, 81.8)**

**Create a 99.7% confidence interval**

- Write  $\bar{x} \pm 3\sigma_{\bar{x}}$
- Plug in the values
- Solve

**(69.8, 84.2)**

## Notes: Confidence Intervals 1-Sample T $\sigma$ is not known

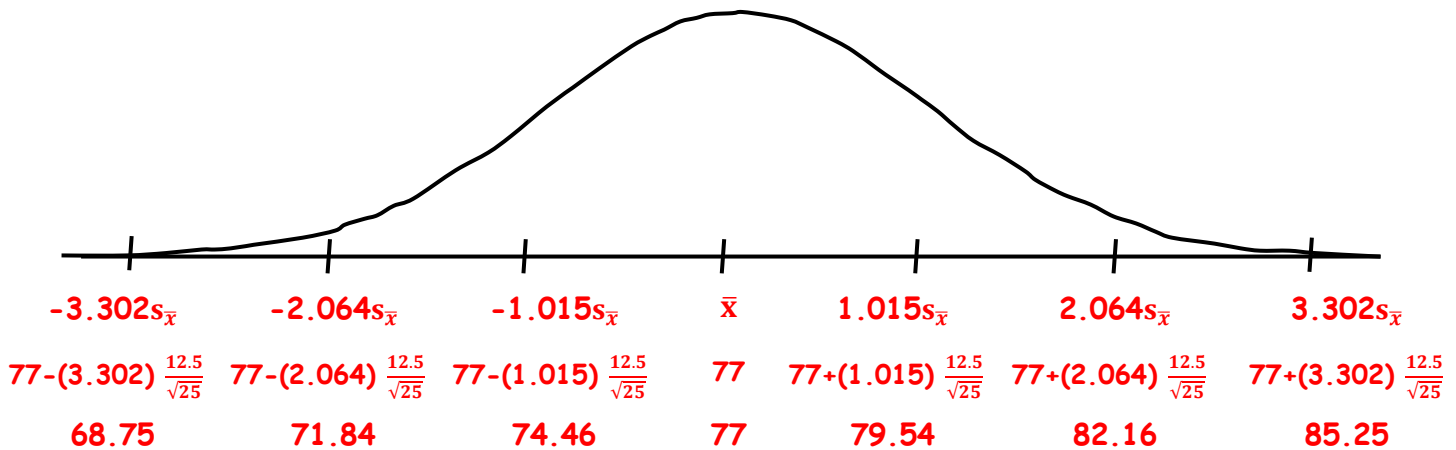
**Remember:** the empirical rule does not apply for t-distributions because there are many different t-distributions. There is a separate t-distribution for each degree of freedom.

### Consider the Following T-distribution

- Sample mean  $\bar{x} = 77$
- Sample Standard deviation  $s = 12.5$
- Sample size  $n = 25$  (how many degrees of freedom?)

### Making it Real (confidence intervals for a t-distribution)

- In the center place the value of  $\bar{x}$
- Below each tick mark record the value of the standard error of  $\bar{x}$  or  $SE_{\bar{x}} = \frac{s}{\sqrt{n}} =$



You will need to use the following calculator commands to calculate ( $t^*$ )

- 2<sup>nd</sup> Vars
  - Inverse  $t$
  - Area =  $\frac{(1 - \text{Confidence level})}{2}$
  - $df = n - 1$
- Calculate the t-value ( $t^*$ ) associated with a 68% area for a t-distribution  $n = 25$   **$\pm(1.015)$**
  - Calculate the t-value ( $t^*$ ) associated with a 95% area for a t-distribution  $n = 25$   **$\pm(2.064)$**
  - Calculate the t-value ( $t^*$ ) associated with a 99.7% area for a t-distribution  $n = 25$   **$\pm(3.302)$**

### Create a 68% confidence interval

- Write  $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$
- Plug in the values
- Solve

**(74.46, 79.54)**

### Create a 95% confidence interval

- Write  $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$
- Plug in the values
- Solve

**(71.84, 82.16)**

### Create a 99.7% confidence interval

- Write  $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$
- Plug in the values
- Solve

**(68.75, 85.25)**

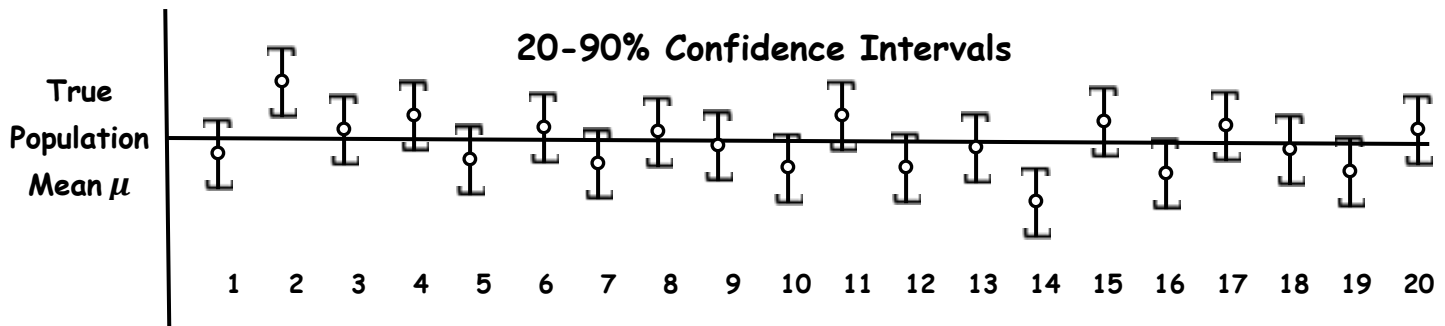
### Confidence Intervals for Means $\sigma$ is known: Z-distribution

The population for the confidence intervals in the chart below is highly skewed. The sample size for each confidence interval is 50, which is greater than 30, therefore the central limit theorem applies. Thus, the distribution of the sample means is approximately normal for each confidence interval.

$\sigma$ , the standard deviation of the population is known which means that Z-scores may be used for the critical values. The formula for a confidence interval where  $\sigma$  is known is as follows:

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \text{ yields } \text{sample mean} \pm (1.645) \frac{\sigma}{\sqrt{50}}$$

**Note:** Because  $\sigma$ , the standard deviation of the population, is known, the margin of error for each 90% confidence interval is the same size. While sampling error will not affect the width of the interval when  $\sigma$  is known, sampling error does affect the sample mean.



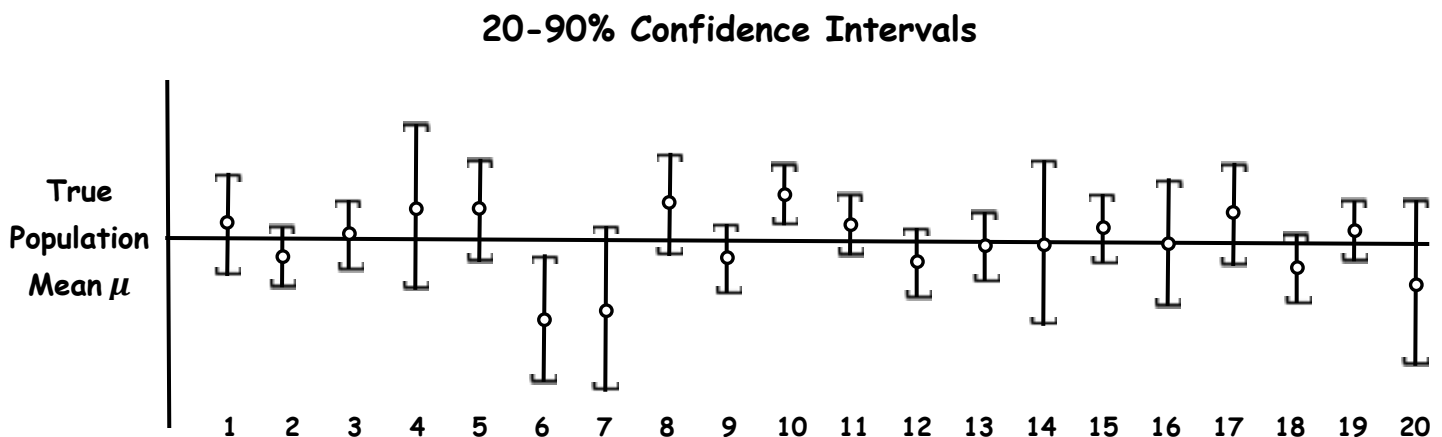
### Confidence Intervals for Means $\sigma$ is *NOT* known $s$ is used: T-distribution

The population for the confidence intervals in the chart below is highly skewed. The sample size for each confidence interval is 50, which is greater than 30, therefore the central limit theorem applies. Thus, the distribution of the sample means is approximately normal for each confidence interval.

$\sigma$ , the standard deviation of the population is not known and  $s$  the standard deviation of the sample is used which means that t-scores will need to be used for the critical values. The formula for a confidence interval is as follows:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} \quad t^* \text{ with 49 degrees of freedom} = 1.677 \text{ yields } \text{sample mean} \pm (1.677) \frac{s}{\sqrt{50}}$$

**Note:** Because  $s$ , the standard deviation of the population, is being used, the margin of error for each 90% confidence interval varies due to sampling error as does the sample mean.



## Recipe for Success: 1 Sample T-Distribution (Confidence Interval)

1. Define the parameter  $\mu$  in context

2. Write the Conditions

1. Random Sample

2.  $n < 10\%$  of the population

3. Normal Population or  $n \geq 30$

*If data is given, draw a boxplot or histogram-to show normality*

3. Write the Equation

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

$t$  = the number of standard deviations a value is from the mean

$\bar{x}$  = the mean of the sample.

$s$  = the standard deviation of the sample

$n$  = the size of the sample

4. Graph and Shade

5. Enter the Data if Given

- Stat Edit

- Enter Data in column L1

- 2nd Quit

- Stat Calc

- 1-Var Stats:  $\bar{x}$ ,  $s_x$ ,  $n$

6. Identify and label all inputs

- $s$  comes from the problem or the data

- $\bar{x}$  comes from the problem or the data

- $n$  comes from the problem or the data

- $df = n-1$

7. Calculate  $t^*$

- 2<sup>nd</sup> Vars

- Inverse  $t$

- Area =  $\frac{(1-\text{Confidence level})}{2}$

- $df = n-1$

8. Plug in the values

9. Calculate the Interval

- Stat Tests

- **8:T Interval**

- Highlight **Stats** (*Highlight Data if data is given*)

- $\bar{x}$  comes from the problem or the data

- $s_x$  come from the problem or the data

- $n$  comes from the problem or the data

- Inter the confidence level

10. Write the Interval

11. Write the Conclusion

We are \_\_\_\_\_% confident that the true population mean for \_\_\_\_\_  
lies within the interval \_\_\_\_\_.

*Restate the definition of the mean*

12. Explain the meaning of the confidence level-if asked

In repeated sampling, we expect that this method will capture the true population mean

\_\_\_\_\_percent of the time.

*Restate the Confidence Level*



## Notes: 1 Sample T-Distribution Confidence Interval (Scenario)

**Scenario 1:** A publishing company has just published a new college textbook. Before the company decides the price at which to sell textbook, it wants to know the average price of such textbook in the market. The research department at the company took a **random sample** of 36 textbook and collected information on their prices. The information produced a mean price of \$70.50 for this sample with a sample standard deviation of 4.50. Construct a 90% confidence interval for the price of this type of textbook.

**Let  $\mu$  equal the true population mean price of college textbooks of this type**

### Conditions:

1. Random Sample: Given
2. It is reasonable to assume  $36/.10 = 360$  that there are at least 360 textbooks of this type.
3.  $n = 36$  the distribution of means is normal by central limit theorem

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

$$70.50 \pm 1.69 \left( \frac{4.5}{\sqrt{36}} \right)$$

$$(69.23, 71.77)$$

**We are 90% confident that the true population mean price for college textbooks lies with the interval (69.23, 71.77)**

**In repeated sampling, we expect that this method will capture the true population mean price of textbooks 90 percent of the time.**

Using the above problem calculate a 95% confidence interval for the type of textbook.

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

$$70.50 \pm 2.03 \left( \frac{4.5}{\sqrt{36}} \right) \quad (68.98, 72.02)$$

Using the above problem calculate a 99% confidence interval for the type of textbook

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

$$70.50 \pm 2.72 \left( \frac{4.5}{\sqrt{36}} \right) \quad (68.46, 72.54)$$

What do you notice about the 3 different confidence intervals? Explain the differences.

**As the level of confidence increases, the t-statistic increases and the interval width increases.**

## Recipe for Success: 1 Sample T-Distribution (Confidence Interval)

1. Define the parameter  $\mu$  in context
2. Write the Conditions

1. Random Sample
2.  $n < 10\%$  of the population
3. Normal Population or  $n \geq 30$

*If data is given, draw a boxplot or histogram-to show normality*

3. Write the Equation

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

$t$  = the number of standard deviations a value is from the mean

$\bar{x}$  = the mean of the sample.

$s$  = the standard deviation of the sample

$n$  = the size of the sample

4. Graph and Shade

5. Enter the Data if Given

- Stat Edit
- Enter Data in column L1
- 2nd Quit
- Stat Calc
- 1-Var Stats:  $\bar{x}$ ,  $S_x$ ,  $n$

6. Identify and label all inputs

- $s$  comes from the problem or the data
- $\bar{x}$  comes from the problem or the data
- $n$  comes from the problem or the data
- $df = n-1$

7. Calculate  $t^*$

- 2<sup>nd</sup> Vars
- Inverse  $t$
- Area =  $\frac{(1-\text{Confidence level})}{2}$
- $df = n-1$

8. Plug in the values

9. Calculate the Interval

- Stat Tests
- **8:T Interval**
- Highlight **Stats** (*Highlight Data if data is given*)
- $\bar{x}$  comes from the problem or the data
- $S_x$  come from the problem or the data
- $n$  comes from the problem or the data
- Inter the confidence level

10. Write the Interval

11. Write the Conclusion

We are \_\_\_\_\_% confident that the true population mean for \_\_\_\_\_  
lies within the interval \_\_\_\_\_.

*Restate the definition of the mean*

12. Explain the meaning of the confidence level-if asked

In repeated sampling, we expect that this method will capture the true population mean  
\_\_\_\_\_percent of the time.

*Restate the Confidence Level*

**Notes: 1 Sample T-Distribution Confidence Interval (Scenario)**

**Scenario 2:** A professor at a large university believes that students take an average of 15 credit hours per term. A random sample of 24 students in her class of 250 students reported the following number of credit hours that they were taking:

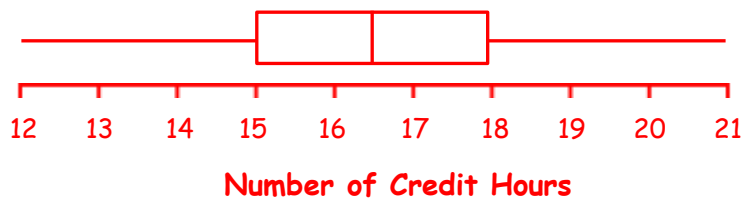
12	13	14	14	15	15	15	16	16	16	16	16
17	17	17	18	18	18	18	19	19	19	20	21

Find a 95% confidence interval for the number of credit hours taken by the students in the professor's class. Interpret the interval.

**Let  $\mu$  equal the true population mean number of credit hours taken by students for this professor**

**Conditions:**

1. Random Sample: Given
2.  $24/.10 = 240$ . The sample size of 24 is less than 10% of the population of 250 Students.
3. The boxplot is reasonably symmetric with no outliers. The distribution appears to be approximately normal

**Sample Statistics:**

$$\bar{x} = 16.625$$

$$s = 2.222$$

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

$$16.625 \pm 2.0686 \left( \frac{2.222}{\sqrt{24}} \right)$$

$$(15.686, 17.564)$$

**We are 95% confident that the true population mean number of credit hours taken by students for this professor lies with the interval (15.686, 17.564)**

**In repeated sampling, we expect that this method will capture the true population mean number of credit hours taken by students 95 percent of the time.**

Does this sample indicate that students are taking more credit hours than the professor believes?

**Because our interval does not contain 15 and lies above that number, this sample indicates that the students are on average taking more than 15 credit hours.**

**Notes: 1 Sample T-Distribution Confidence Interval (Scenario)**

**Scenario 3:** Insurance companies track life expectancy information to assist in determining the cost of life insurance policies. The insurance company knows that, last year, the life expectancy of its policyholders was 77 years. They want to know if their clients this year have a longer life expectancy, on average, so the company randomly samples some of the recently paid policies to see if the mean life expectancy of policyholders has increased. The insurance company will only change their premium structure if there is evidence that people who buy their policies are living longer than before.

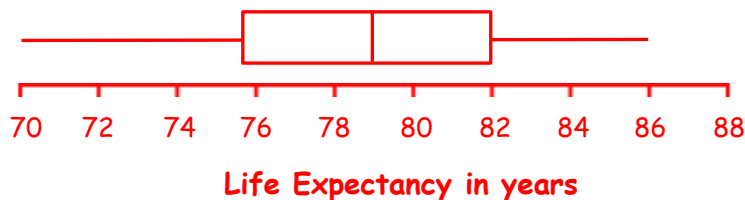
86	75	83	84	81	77	78	79	79	81
76	85	70	76	79	81	73	74	72	83

Does this sample indicate that the insurance company should change its premiums because life expectancy is now different? Calculate a 90% confidence interval and interpret the level.

**Let  $\mu$  equal the true population mean of policy holder's life expectancy in years**

**Conditions:**

1. Random Sample: Given
2.  $20/.10 = 200$ . Reasonable to assume that the company has more than 200 policy holders. Therefore we can assume that the sample size of 20 is less than 10% of the total population of policy holders
3. The boxplot is reasonably symmetric with no outliers. The distribution appears to be approximately normal

**Sample Statistics:**

$$\bar{x} = 78.6$$

$$S = 4.477$$

$$\bar{x} \pm t^* \left( \frac{S}{\sqrt{n}} \right)$$

$$78.6 \pm 1.729 \left( \frac{4.477}{\sqrt{20}} \right)$$

$$(76.869, 80.331)$$

We are 90% confident that the true population mean life expectancy for policy holders lies with the interval (76.869, 80.331)

In repeated sampling, we expect that this method will capture the true population mean life expectancy of policy holders 90 percent of the time.

## Notes: Hypothesis Tests

**Hypotheses:** an assumption about a population parameter. This assumption may or may not be true

**Hypothesis Testing:** The formal procedures used to accept or reject a **statistical hypotheses**. A hypothesis test makes a yes/no decision about the plausibility of a parameter value.

- In Statistics, a hypothesis proposes a model for the world. Then we look at the data.
- If the data are consistent with that model, we have no reason to disbelieve the hypothesis.
  - Data consistent with the model *lend support* to the hypothesis, **but do not prove it**.
  - If the data is sufficiently inconsistent, **we can reject the model**.

**Consider the logic of Jury Trials:**

- We assume a defendant is innocent. (this is our hypothesis)
- We retain that hypotheses unless the facts make that claim unlikely beyond a reasonable doubt.
- Then, and only then, we reject the hypothesis of innocence and declare the person guilty.

**The same logic used in jury trials is used in statistical tests of hypotheses:**

- We begin by assuming that a hypothesis is true.
- Next we consider whether the data are consistent with the hypothesis.
- If data is consistent with the original claim (the null hypothesis), we retain/accept that the null as true. **We never prove the null.** (we didn't gather evidence to prove it; only to disprove it)
- If the data is not consistent with the original claim (the null hypothesis), then like a jury, we determine whether we believe beyond a reasonable doubt that the original hypothesis is false.

**P-Value:** In statistics, we are able to quantify our level of doubt. The p-value is a **conditional probability** and gives the likelihood that we would get a test statistic as extreme or more extreme given that the original claim/null hypothesis is true.

- **Given that the null is true, there is a \_\_\_\_% chance that we would calculate a test statistic as extreme or more extreme in favor of the alternative hypothesis.**
- When the data are consistent with the model from the null hypothesis, the P-value is high and we are unable to reject the null hypothesis.
  - In that case, we have to "retain" the null hypothesis we started with.
  - We can't claim to have proved it; instead we "*fail to reject the null hypothesis*" when the data are consistent with the null hypothesis model and are within the range of what we would expect from natural sampling variability.
- If the P-value is low enough, we'll "*reject the null hypothesis,*" since what we observed would be very unlikely were the null model true.

**Testing Hypotheses**

- The null hypothesis, which we denote  $H_0$ , specifies a population model parameter of interest and proposes a value for that parameter.
  - We might have, for example,  $H_0: \mu = 5$
- We want to compare our data to what we would expect given that  $H_0$  is true.
  - We can do this by finding out how many standard deviations away from the claimed null value we are.
- We then ask how likely it is to get results like we did if the null hypothesis were true.

## Notes: Hypothesis Tests

### A Trial as a Hypothesis Test

- Hypothesis testing is very much like a court trial.
- The null hypothesis is that the defendant is innocent.
- We then present the evidence—collect data.
- Then we judge the evidence—“Could these data plausibly have happened by chance if the null hypothesis were true?”
  - If they were very unlikely to have occurred, then the evidence raises more than a reasonable doubt in our minds about the null hypothesis.
- Ultimately, we must make a decision. How unlikely is unlikely?
- Some people advocate setting rigid standards—1 time out of 20 (0.05) or 1 time out of 100 (0.01).
- But if you have to make the decision, you must judge for yourself in any particular situation whether the probability is small enough to constitute “reasonable doubt.”

### What to Do with an “Innocent” Defendant

- If the evidence is not strong enough to reject the presumption of innocent, the jury returns with a verdict of “not guilty.”
  - The jury does not say that the defendant is innocent.
  - All it says is that there is not enough evidence to convict, to reject innocence.
  - The defendant may, in fact, be innocent, but the jury has no way to be sure
- Said statistically, we will *fail to reject* the null hypothesis.
  - We never declare the null hypothesis to be true, because we simply do not know whether it's true or not.
  - Sometimes in this case we say that the *null hypothesis has been retained*.
- In a trial, the burden of proof is on the prosecution.
- In a hypothesis test, the burden of proof is on the unusual claim.
- The null hypothesis is the ordinary state of affairs, so it's the alternative to the null hypothesis that we consider unusual (and for which we must find sufficient evidence).

### There are four basic parts to a hypothesis test:

1. Hypotheses:
  - State the Null: **Parameter =**
  - Alternative hypothesis: **Parameter <** or **Parameter >** or **Parameter ≠** ...(2-tailed)
2. Model and Conditions:
  - Give the Name or Formula for the test
  - List the conditions
3. Mechanics:
  - Plug the values into the equation
  - Calculate the test statistic
  - Calculate the p-value
4. Conclusion
  - State whether we reject or fail to reject the null hypothesis
  - The conclusion should be stated in context
  - The conclusion should give the actions that should be taken

## Notes: Hypothesis Test 1-Sample T-test

### Vocabulary:

- **Hypothesis:** An assumption about a characteristic about a population. In this case the population parameter that we are testing is ( $\mu$ ) the population mean
- **Hypothesis testing:** A method of determining the validity of a hypothesis. Is the claim likely to be true? (*Note: we never prove that the claim is true or false.*)
- **Null Hypothesis:** This is the claim about the population parameter.  
It is always expressed as  $H_0$ : the population parameter = the assumed claim. **It is always =**  
So for means it would appear something like  $H_0: \mu = .81$  (assuming .81 is the claim)  
**(Note: the null is based on the concept of innocent until proven guilty-We assume innocence and we assume that the claim is true, & will continue to do so until there is evidence to the contrary)**
- **Alternative Hypothesis:** What must concluded if the null hypothesis is found to be unlikely.  
The alternative Hypothesis can be expressed in 3 different manners depending on the question.  
  - $H_A: \mu \neq$  the claim-the population proportion is **different than** the claim
  - $H_A: \mu >$  the claim-the population proportion is **greater than** the claim
  - $H_A: \mu <$  the claim-the population proportion is **less than** the claim**Regardless of how  $H_A$ : is expressed  $H_0$ : will always have an = sign**
- **2 tail tests-** a test in which the rejection region is in both tails  
**always associated with  $H_A: \mu \neq$  the claim**
- **1 tail test-** a test in which the rejection region is in only one of the tails  
  - Upper Tail Test:  $H_A: \mu >$  the claim**
  - Lower Tail Test:  $H_A: \mu <$  the claim**
- **Alpha-(the level of significance)is a predetermined amount of risk or probability of committing a Type I error that the tester is willing to accept and is the level of significance of the test.** The alpha level is the amount of area in the rejection region/the area in the tail(s)
- **Critical Value-the value that is associated with a given alpha level.** These are the actual values that define the rejection regions. They are **predetermined based on a selected alpha.**
- **test statistic- the value that is computed from our sample data.**
- **p-value-A conditional probability that tells us the likelihood that our data would occur given that the null hypothesis is true.** In other words given the null really is actually correct, how likely are we to find the **test statistic** from the sample of the population.  
**The p-value is what tells us whether to reject or not.**

**We reject if the p-value is less than alpha. This is an always and forever amen.**

**We reject if the p-value is less than alpha. This is an always and forever amen.**

## Recipe for Success: 1 Sample T-Distribution (Hypothesis Test)

### 1. Write your Hypothesis

- Null  $H_0: \mu =$
- Alternative  $H_A: \mu \neq$  or  $<$  or  $>$

### 2. Define the parameter $\mu$ in context

### 3. Write the Conditions

1. Random Sample
2.  $n < 10\%$  of the population
3. Normal Population or  $n > 30$

*If data is given, draw a boxplot or histogram-to show normality*

### 4. Write the Equation

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$t$  = the number of standard deviations a value is from the mean

$\mu$  = the mean of the population or what is assumed to be true

$\bar{x}$  = the mean of the sample.

$s$  = the standard deviation of the sample

$n$  = the size of the sample

### 5. Draw the graph and Shade

### 6. Enter Data (if given)

- Stat Edit
- Enter Data in column L1
- 2nd Quit
- Stat Calc
- 1-Var Stats:  $\bar{x}$ ,  $s_x$ ,  $n$

### 7. List & Label all of input values

### 8. Plug values into the equation

### 9. Calculate the $t$ and the p-value

$$df = n - 1$$

*(df is the degrees of freedom)*

- Stat Tests
- **2:T-Test Enter**
- Highlight **Stats** (*Highlight Data if data is given*)
- $\mu$  comes from the problem
- $\bar{x}$  comes from the problem or the data
- $s_x$  come from the problem or the data
- $n$  comes from the problem or the data
- Choose  $\neq$  or  $<$  or  $>$
- (use Shaded graph of  $H_A$ )

### 10. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 11. Write the Conclusion

Reject the Null: Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population mean for \_\_\_\_\_ is \_\_\_\_\_

*Restate  $H_A \neq$  or  $<$  or  $>$  mean*

*Restate the definition of the*

Fail to Reject the Null: Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population mean for \_\_\_\_\_ is \_\_\_\_\_

\_\_\_\_\_ is \_\_\_\_\_

*Restate the definition of the mean*

*Restate  $H_A \neq$  or  $<$  or  $>$  mean*



**Notes: 1 Sample T-Distribution Hypothesis Test (Scenario)**

**Scenario 1:** The EPA reports that the exhaust emissions for a certain car model has a normal distribution with a mean of 1.45 grams of nitrous oxide per mile. The car manufacturer claims their new process reduces the mean level of exhaust emitted for this car model. A SRS of 28 cars is taken. The mean level of exhaust emitted for this sample is 1.27 grams with a standard deviation of 0.4. Is there statistical evidence at an alpha level of .05 to support the car manufacturer's claim?

$$H_0: \mu = 1.45 \text{ grams}$$

$$H_A: \mu < 1.45 \text{ grams}$$

**Let  $\mu$  equal the true population mean of nitrous oxide emissions in grams per mile**

**Conditions:**

1. Random Sample-Given
2.  $28/.10 = 280$  It is reasonable to assume that there are at least 280 cars of this model.
3. The population is Normal- Given

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad P(\bar{x} < 1.27) \rightarrow P\left(t < \frac{1.27 - 1.45}{\frac{0.4}{\sqrt{28}}}\right)$$

$$= P(t < -2.381) = .0123$$

P-value	alpha
.0123	.05

**Less than**

**We reject the  $H_0$**

**Our p-value is .0123. We reject the null. There is sufficient evidence at alpha = .05 to suggest that the true population mean for nitrous oxide emissions is less than 1.45 grams.**

## Recipe for Success: 1 Sample T-Distribution (Hypothesis Test)

### 1. Write your Hypothesis

- Null  $H_0: \mu =$
- Alternative  $H_A: \mu \neq$  or  $<$  or  $>$

### 2. Define the parameter $\mu$ in context

### 3. Write the Conditions

1. Random Sample
2.  $n < 10\%$  of the population
3. Normal Population or  $n > 30$

*If data is given, draw a boxplot or histogram-to show normality*

### 4. Write the Equation

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$t$  = the number of standard deviations a value is from the mean

$\mu$  = the mean of the population or what is assumed to be true

$\bar{x}$  = the mean of the sample.

$s$  = the standard deviation of the sample

$n$  = the size of the sample

### 5. Draw the graph and Shade

### 6. Enter Data (if given)

- Stat Edit
- Enter Data in column L1
- 2nd Quit
- Stat Calc
- 1-Var Stats:  $\bar{x}$ ,  $s_x$ ,  $n$

### 7. List & Label all of input values

### 8. Plug values into the equation

### 9. Calculate the $t$ and the p-value

$$df = n - 1$$

*(df is the degrees of freedom)*

- Stat Tests
- **2:T-Test Enter**
- Highlight **Stats** (*Highlight Data if data is given*)
- $\mu$  comes from the problem
- $\bar{x}$  comes from the problem or the data
- $s_x$  come from the problem or the data
- $n$  comes from the problem or the data
- Choose  $\neq$  or  $<$  or  $>$
- (use Shaded graph of  $H_A$ )

### 10. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 11. Write the Conclusion

Reject the Null: Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population mean for \_\_\_\_\_ is \_\_\_\_\_

*Restate  $H_A \neq$  or  $<$  or  $>$  mean*

*Restate the definition of the*

Fail to Reject the Null: Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population mean for \_\_\_\_\_ is \_\_\_\_\_

\_\_\_\_\_ is \_\_\_\_\_

*Restate the definition of the mean*

*Restate  $H_A \neq$  or  $<$  or  $>$  mean*

**Notes: 1 Sample T-Distribution Hypothesis Test (Scenario)**

**Scenario 2:** A credit card company wondered whether giving frequent flyer miles for every purchase would increase card usage, which has a current mean of \$2500 per year. They gave free miles to a SRS of 37 credit card customers and found the sample mean to be \$2542 with a standard deviation of \$109. Is there statistical evidence at an  $\alpha$  level of .01 to show that card usage has increased?

$$H_0: \mu = 2,500 \text{ dollar}$$

$$H_A: \mu > 2,500$$

Let  $\mu$  equal the true population mean credit card usage

**Conditions:**

1. Random Sample-Given
2.  $37/.10 = 370$  It is reasonable to assume that there are at least 370 credit card users.
3.  $n = 37$  which is greater than 30. The distribution of means is normal by central limit theorem

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad P(\bar{x} > 2,542) \rightarrow P\left(t < \frac{2542 - 2500}{\frac{109}{\sqrt{37}}}\right)$$

$$= P(t > -2.344) = .0124$$

P-value	alpha
.0124	.01

We fail to reject the  $H_0$

Our p-value is .0124. We fail to reject the null. There is not sufficient evidence at  $\alpha = .01$  to suggest that the true population mean for credit card usage is greater than 2,500 dollars.

**Notes: 1 Sample T-Distribution Hypothesis Test (Scenario)**

**Scenario 3:** A steel supplier claims that their bars have elongation of 40 percent under a stress of 10 metric tons. A random sample of 12 steel bars was taken from a warehouse and the elongation measured in percent per 2 inches determined.

Observed values of **39 41 35 33 35 39 33 37 36 40 41 42** were obtained.

Is there evidence at a significance level of .05 to demonstrate that the elongation for the steel bars was different from 40?

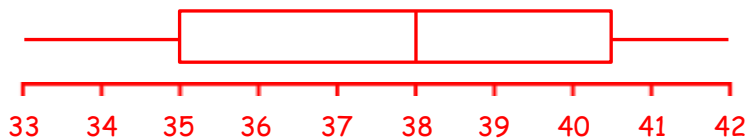
$H_0: \mu = 40$  percent

$H_A: \mu \neq 40$  percent

Let  $\mu$  equal the true population mean for the percent of steel bar elongation

Conditions:

1. Random Sample-Given
2.  $12/.10 = 120$  It is reasonable to assume that at least 120 steel bars are produced.
3. The boxplot is reasonably symmetric with no outliers. The distribution appears to be approximately normal



$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad 2 \times P(\bar{x} < 37.58) \rightarrow 2 \times P\left(t < \frac{37.58 - 40}{\frac{3.175}{\sqrt{12}}}\right)$$

$$= 2 \times P(t < -2.636) = .0231$$

P-value	alpha
.0231	Less than .05

We reject the  $H_0$

Our p-value is .0231. We reject the null. There is sufficient evidence at  $\alpha = .05$  to suggest that the true population mean for the percent of steel bar elongation is different than 40 percent.

## Notes: Hypothesis Tests (Review)

**Hypotheses:** an assumption about a population parameter. This assumption may or may not be true

**Hypothesis Testing:** The formal procedures used to accept or reject a **statistical hypotheses**. A hypothesis test makes a yes/no decision about the plausibility of a parameter value.

- In Statistics, a hypothesis proposes a model for the world. Then we look at the data.
- If the data are consistent with that model, we have no reason to disbelieve the hypothesis.
  - Data consistent with the model *lend support* to the hypothesis, **but do not prove it**.
  - If the data is sufficiently inconsistent, **we can reject the model**.

**Consider the logic of Jury Trials:**

- We assume a defendant is innocent. (this is our hypothesis)
- We retain that hypotheses unless the facts make that claim unlikely beyond a reasonable doubt.
- Then, and only then, we reject the hypothesis of innocence and declare the person guilty.

**The same logic used in jury trials is used in statistical tests of hypotheses:**

- We begin by assuming that a hypothesis is true.
- Next we consider whether the data are consistent with the hypothesis.
- If data is consistent with the original claim (the null hypothesis), we retain/accept that the null as true. **We never prove the null.** (we didn't gather evidence to prove it; only to disprove it)
- If the data is not consistent with the original claim (the null hypothesis), then like a jury, we determine whether we believe beyond a reasonable doubt that the original hypothesis is false.

**P-Value:** In statistics, we are able to quantify our level of doubt. The p-value is a **conditional probability** and gives the likelihood that we would get a test statistic as extreme or more extreme given that the original claim/null hypothesis is true.

- **Given that the null is true, there is a \_\_\_\_\_% chance that we would calculate a test statistic as extreme or more extreme in favor of the alternative hypothesis.**
- When the data are consistent with the model from the null hypothesis, the P-value is high and we are unable to reject the null hypothesis.
  - In that case, we have to "retain" the null hypothesis we started with.
  - We can't claim to have proved it; instead we "*fail to reject the null hypothesis*" when the data are consistent with the null hypothesis model and are within the range of what we would expect from natural sampling variability.
- If the P-value is low enough, we'll "*reject the null hypothesis*," since what we observed would be very unlikely were the null model true.

**Testing Hypotheses**

- The null hypothesis, which we denote  $H_0$ , specifies a population model parameter of interest and proposes a value for that parameter.
  - We might have, for example,  $H_0: \mu = 5$
- We want to compare our data to what we would expect given that  $H_0$  is true.
  - We can do this by finding out how many standard deviations away from the claimed null value we are.
- We then ask how likely it is to get results like we did if the null hypothesis were true.

## Notes: Hypothesis Tests (Review)

### A Trial as a Hypothesis Test

- Hypothesis testing is very much like a court trial.
- The null hypothesis is that the defendant is innocent.
- We then present the evidence—collect data.
- Then we judge the evidence—“Could these data plausibly have happened by chance if the null hypothesis were true?”
  - If they were very unlikely to have occurred, then the evidence raises more than a reasonable doubt in our minds about the null hypothesis.
- Ultimately, we must make a decision. How unlikely is unlikely?
- Some people advocate setting rigid standards—1 time out of 20 (0.05) or 1 time out of 100 (0.01).
- But if you have to make the decision, you must judge for yourself in any particular situation whether the probability is small enough to constitute “reasonable doubt.”

### What to Do with an “Innocent” Defendant

- If the evidence is not strong enough to reject the presumption of innocent, the jury returns with a verdict of “not guilty.”
  - The jury does not say that the defendant is innocent.
  - All it says is that there is not enough evidence to convict, to reject innocence.
  - The defendant may, in fact, be innocent, but the jury has no way to be sure
- Said statistically, we will *fail to reject* the null hypothesis.
  - We never declare the null hypothesis to be true, because we simply do not know whether it's true or not.
  - Sometimes in this case we say that the *null hypothesis has been retained*.
- In a trial, the burden of proof is on the prosecution.
- In a hypothesis test, the burden of proof is on the unusual claim.
- The null hypothesis is the ordinary state of affairs, so it's the alternative to the null hypothesis that we consider unusual (and for which we must find sufficient evidence).

### There are four basic parts to a hypothesis test:

1. Hypotheses:
  - State the Null: **Parameter =**
  - Alternative hypothesis: **Parameter <** or **Parameter >** or **Parameter ≠** ...(2-tailed)
2. Model and Conditions:
  - Give the Name or Formula for the test
  - List the conditions
3. Mechanics:
  - Plug the values into the equation
  - Calculate the test statistic
  - Calculate the p-value
4. Conclusion
  - State whether we reject or fail to reject the null hypothesis
  - The conclusion should be stated in context
  - The conclusion should give the actions that should be taken

## Notes: Paired T-Distribution

### What is Paired Data?

Data are paired when the observations are collected in pairs (2 samples) or the observations in one group are naturally related to observations in the other group.

### How does Paired Data Arise? (How do I recognize that the 2 samples are paired?)

- A subject is measured Before and After treatments and the difference is computed for each subject and then the average of the differences is computed
- A subject is tested twice with different treatments and the difference is computed for each subject and then the average of the differences is computed
- Pairing may also be the comparison of naturally occurring **Couples or Pairs**
  - Twins-One twin would receive treatment 1 and the other twin would receive treatment 2 and the difference in responses would be computed for each set of twins.
  - Spouses- The husband is compared to his wife and the differences between the husband's and wife's response are computed.
  - Siblings- A sibling is compared to his/her sibling and the differences between each sibling pair's response is computed.

**Note:** For a matched pair design we are looking at the difference between each pair. Those differences will be summed and then an average of those differences will be computed to create our sample average ( $\bar{x}_d$ ).

**In order to be a matched pair test the list of the two samples must be of equal length—obviously.**

### We have worked with Paired Data before:

Recall when we created matched pairs designed experiments for waterproof boots and for mosquito repellent. Why did we do that?

### What a Paired T-Test Does

A paired t-test looks at the difference between paired values in two samples in cases where each value in one sample has a natural partner in the other. In a paired t-test, we calculate the differences for each pair. In doing so, we have taken two samples and created a single sample of differences. It is that single sample of differences that we use for all of our statistical calculations.

**Paired Data is very powerful and should be used anytime that it is available.** Why because paired data controls for variation within the pairs. In other words, we recognize that if data has a pair then an association exists and the samples are not independent. However, the distribution of the differences can still be independent.

### Caution:

If the data is paired, **we do not take the mean of each sample and then compare the means.** That type of test is a 2-sample test for means which will be discussed later. **We do take the mean of the paired differences**

## Recipe for Success: Hypothesis Test for a Paired T-test

### 1. Write your Hypothesis

- Null  $H_0: \mu_d = 0$
- Alternative  $H_A: \mu_d \neq$  or  $<$  or  $> 0$

### 2. Define parameter $\mu_d$ in context & write the conditions

1. Random Sample
2.  $n < 10\%$  of the population
3. Differences are independent
4. The differences are Normal or  $n > 30$

*If data is given, draw a boxplot or histogram-to show normality*

### 3. Write the Equation

$$t = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

$t$  = the number of standard deviations a value is from the mean

$$\mu_d = 0$$

$\bar{x}_d$  = the mean of the sample differences

$s_d$  = the standard deviation of the sample differences

$n$  = the number of sample differences

### 4. Enter Data (if given)

- Stat Edit
- Enter Data in columns  $L_1$  &  $L_2$
- At the top of  $L_3$  type  $2^{nd}$   $L_1 - 2^{nd}$   $L_2$

### 5. Find $\bar{x}_d$

- Stat Calc
- 1-Var Stats press Enter
- List type  $2^{nd}$   $L_3$

### 6. List & Label all of input values

$n, \bar{x}_d, s_d$

$\mu_d = 0$  &  $df = n - 1$

### 7. Plug values into the equation

### 8. Calculate the $t$ and the p-value

$$\mu_d = 0$$

$$df = n - 1$$

- Stat Tests
- 2:T-Test Enter
- Highlight **Data** if data is used otherwise highlight **STATS**
- $s_d$  come from the problem or the data
- $\bar{x}_d$  comes from the problem or the data
- $n$  comes from the problem or the data
- Choose  $\neq$  or  $<$  or  $>$  (look for key words)

### 9. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 10. Write the Conclusion

Reject the Null: Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at  $\alpha =$  \_\_\_\_\_ to suggest that the true population mean differences for \_\_\_\_\_ is \_\_\_\_\_

*Restate the definition of the mean differences      Restate  $H_A \neq$  or  $<$  or  $>$  mean differences*

Fail to Reject the Null: Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at  $\alpha =$  \_\_\_\_\_ to suggest that the true population mean difference for \_\_\_\_\_ is \_\_\_\_\_

*Restate the definition of the mean difference      Restate  $H_A \neq$  or  $<$  or  $>$  mean difference*



## Notes: Hypothesis Test for a Paired T-test (Scenario)

**Scenario 1:** One indicator of physical fitness is resting pulse rate. Fourteen men volunteered to test an exercise device advertised on television by using it three times a week for 20 minutes. Their resting pulse rate beats per minute (bpm) were measured before the test began, and then again after six weeks. Results are shown in the table below.

Subject		Alec	Tripp	Matt	Peter	Chris	Cory	Sebastian	Luke	Ture	Winston	Kevin	Proloy	Alex	Tim
Beats per min.	Before	73	84	83	85	87	91	87	99	85	82	83	79	80	77
	After	73	82	79	81	86	87	84	91	83	79	84	76	77	77

Is there evidence that this kind of exercise can reduce resting pulse rates?

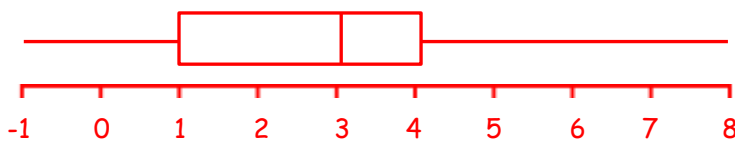
$$H_0: \mu_d = 0$$

$$H_A: \mu_d > 0$$

Let  $\mu_d$  = equal the true population mean difference in resting heart rate (before - after) the 6 week exercise program.

### Conditions:

1. Random Sample ( this is not a random sample but we assume that the men are representative of the population)
2.  $14/.10 = 140$ . There are a few more than 140 men in the population
3. The differences in pulse rate are independent in that it is reasonable to assume that change in one person's pulse rate does not affect that of another.
4. The boxplot is reasonably symmetric with no outliers. The distribution appears to be approximately normal



$$\bar{x}_d = 2.57 \quad P(\bar{x}_d > 2.57)$$

$$s_d = 2.24 \quad df = 14 - 1$$

$$t = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}} \rightarrow P\left(t > \frac{2.57 - 0}{\frac{2.24}{\sqrt{14}}}\right) = P(t > 4.289) = .0004$$

P-value                      alpha  
 .0004    Less than    .05

We reject the  $H_0$

Our p-value is .0004. We reject the null. There is sufficient evidence at  $\alpha = .05$  to suggest that the true population mean difference in resting heart rate (before - after) is greater than zero.

## Recipe for Success: Confidence Interval for a Paired T-test

1. Define parameter  $\mu_d$  in context

2. Write the Conditions

1. Random Sample
2.  $n < 10\%$  of the population
3. Differences are independent
4. The differences are Normal or  $n > 30$

*If data is given, draw a boxplot/ histogram-to show normality*

3. Write the formula for the Test

$$\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}}$$

$t^*$  = the number of standard deviations a value is from the mean & is based on the Confidence Level

$$\mu_d = 0$$

$\bar{x}_d$  = the mean of the sample differences

$s_d$  = the standard deviation of the sample differences

$n$  = the number of sample differences

4. Enter the Data if Given

- Stat Edit
- Enter Data in columns  $L_1$  &  $L_2$

5. Find  $\bar{x}_d$

- Stat Edit
- At the top of  $L_3$  type  $2^{nd}$   $L_1 - 2^{nd}$   $L_2$

6. Identify & label all inputs.

- $s_d$  come from the problem or the data
- $\bar{x}_d$  comes from the problem or the data
- $n$  comes from the problem or the data
- $df = n - 1$  (*df is the degrees of freedom*)

7. Calculate  $t^*$

- $2^{nd}$  Vars
- Inverse t
- Area =  $\frac{(1 - \text{Confidence level})}{2}$
- $df = n - 1$

8. Plug in and calculate the Confidence Interval

- STAT
- TESTS
- 8: T Interval

9. Write the Interval

10. Write the Conclusion

We are \_\_\_\_\_% confident that the true population mean difference for \_\_\_\_\_ lies within the interval \_\_\_\_\_.

*Restate the definition of the mean differences*

11. Explain the meaning of the confidence level-if asked

In repeated sampling we expect this method to capture the true population mean difference for \_\_\_\_\_% of the time

*Restate the definition of the mean differences*

### Notes: Confidence Interval for a Paired T-test (Scenario)

**Scenario 1:** One indicator of physical fitness is resting pulse rate. Fourteen men volunteered to test an exercise device advertised on television by using it three times a week for 20 minutes. Their resting pulse rate beats per minute (bpm) were measured before the test began, and then again after six weeks. Results are shown in the table below.

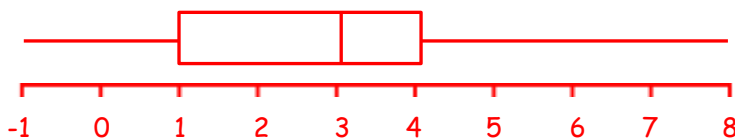
Subject		Alec	Tripp	Matt	Peter	Chris	Cory	Sebastian	Luke	Ture	Winston	Kevin	Proloy	Alex	Tim
Beats per min.	Before	73	84	83	85	87	91	87	99	85	82	83	79	80	77
	After	73	82	79	81	86	87	84	91	83	79	84	76	77	77

Create a 95% confidence interval for the mean difference and explain the meaning of the confidence level?

Let  $\mu_d$  = equal the true population mean difference in resting heart rate (before - after) the 6 week exercise program.

#### Conditions:

1. Random Sample ( this is not a random sample but we assume that the men are representative of the population)
2.  $14/.10 = 140$ . There are a few more than 140 men in the population
3. The differences in pulse rate are independent in that it is reasonable to assume that change in one person's pulse rate does not affect that of another.
4. The boxplot is reasonably symmetric with no outliers. The distribution appears to be approximately normal



$$\bar{x}_d = 2.57$$

$$s_d = 2.24 \quad df = 14-1$$

$$\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}} \rightarrow 2.57 \pm 2.16 \frac{2.24}{\sqrt{14}}$$

$$(1.276, 3.867)$$

We are 95% confident that the true population mean difference in resting heart rate (before - after) lies with the interval (1.276, 3.867)

In repeated sampling, we expect that this method will capture the true population mean difference in resting heart rate (before - after) 95 percent of the time.

**Note:** 0 does not lie within the interval. Therefore we are 95 percent confident that there is a difference in resting heart rate after 6 weeks of exercise

## Recipe for Success: Hypothesis Test for a Paired T-test

### 1. Write your Hypothesis

- Null  $H_0: \mu_d = 0$
- Alternative  $H_A: \mu_d \neq$  or  $<$  or  $> 0$

### 2. Define parameter $\mu_d$ in context & write the conditions

#### 1. Random Sample

#### 2. $n < 10\%$ of the population

#### 3. Differences are independent

#### 4. The differences are Normal or $n > 30$

*If data is given, draw a boxplot or histogram-to show normality*

### 3. Write the Equation

$$t = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

$t$  = the number of standard deviations a value is from the mean

$\mu_d = 0$

$\bar{x}_d$  = the mean of the sample differences

$s_d$  = the standard deviation of the sample differences

$n$  = the number of sample differences

### 4. Enter Data (if given)

- Stat Edit

- Enter Data in columns  $L_1$  &  $L_2$

- At the top of  $L_3$  type  $2^{nd}$   $L_1 - 2^{nd}$   $L_2$

### 5. Find $\bar{x}_d$

- Stat Calc

- 1-Var Stats press Enter

- List type  $2^{nd}$   $L_3$

### 6. List & Label all of input values

$n, \bar{x}_d, s_d$

$\mu_d = 0$  &  $df = n-1$

### 7. Plug values into the equation

### 8. Calculate the $t$ and the p-value

$\mu_d = 0$

$df = n-1$

- Stat Tests

- 2:T-Test Enter

- Highlight Data if data is used otherwise highlight **STATS**

- $s_d$  come from the problem or the data

- $\bar{x}_d$  comes from the problem or the data

- $n$  comes from the problem or the data

- Choose  $\neq$  or  $<$  or  $>$  (look for key words)

### 9. State the Decision

- The p-value is \_\_\_\_\_

- If the p-value is less than alpha, Reject the Null

- If the p-value is greater than alpha, Fail to reject the Null

### 10. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population mean differences for

\_\_\_\_\_ is \_\_\_\_\_

*Restate the definition of the mean differences      Restate  $H_A \neq$  or  $<$  or  $>$  mean differences*

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population mean difference for

\_\_\_\_\_ is \_\_\_\_\_

*Restate the definition of the mean difference      Restate  $H_A \neq$  or  $<$  or  $>$  mean difference*

**Notes: Hypothesis Test for a Paired T-test (Scenario)**

**Scenario 2:** Professor Andy Neill measured the time (in seconds) required to catch a falling meter stick for 12 randomly selected students' dominant hand and non-dominant hand. Professor Neill claims that the reaction time in an individual's dominant hand is less than the reaction time in their non-dominant hand.

Student	1	2	3	4	5	6	7	8	9	10	11	12
Dominant Hand	0.177	0.210	0.186	0.189	0.198	0.194	0.160	0.163	0.166	0.152	0.190	0.172
Non-dominant	0.179	.202	.208	.184	.215	.193	.194	.160	.209	.164	.210	.197

Test an appropriate hypothesis and state your conclusion

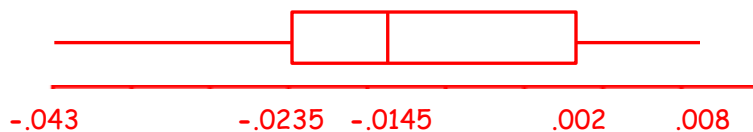
$$H_0: \mu_d = 0$$

$$H_A: \mu_d < 0$$

Let  $\mu_d$  = equal the true population mean difference in reaction time (dominant - non-dominant)

**Conditions:**

1. Random Sample - Given
2.  $12/10 = 120$ . Reasonable to assume that there are at least 120 students
3. The differences in reaction times are independent in that it is reasonable to assume that one person's reaction time does not impact the reaction time of others
4. The boxplot is reasonably symmetric with no outliers. The distribution appears to be approximately normal



$$\bar{x}_d = -.0132 \quad P(\bar{x}_d < -.0132)$$

$$s_d = .0164 \quad df = 12-1$$

$$t = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}} \rightarrow P(t > \frac{-0.0132 - 0}{\frac{0.0164}{\sqrt{12}}}) = P(t < -2.776) = .009$$

P-value                      alpha  
 .009      Less than      .05

We reject the  $H_0$

Our p-value is .009. We reject the null. There is sufficient evidence at  $\alpha = .05$  to suggest that the true population mean difference in reaction time (dominant - non-dominant) is less than zero. The dominant hand is faster

## Recipe for Success: Confidence Interval for a Paired T-test

1. Define parameter  $\mu_d$  in context

2. Write the Conditions

1. Random Sample
2.  $n < 10\%$  of the population
3. Differences are independent
4. The differences are Normal or  $n > 30$

*If data is given, draw a boxplot/ histogram-to show normality*

3. Write the formula for the Test

$$\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}}$$

$t^*$  = the number of standard deviations a value is from the mean & is based on the Confidence Level

$$\mu_d = 0$$

$\bar{x}_d$  = the mean of the sample differences

$s_d$  = the standard deviation of the sample differences

$n$  = the number of sample differences

4. Enter the Data if Given

- Stat Edit
- Enter Data in columns  $L_1$  &  $L_2$

5. Find  $\bar{x}_d$

- Stat Edit
- At the top of  $L_3$  type  $2^{nd} L_1 - 2^{nd} L_2$

6. Identify & label all inputs.

- $s_d$  come from the problem or the data
- $\bar{x}_d$  comes from the problem or the data
- $n$  comes from the problem or the data
- $df = n - 1$  (*df is the degrees of freedom*)

7. Calculate  $t^*$

- $2^{nd}$  Vars
- Inverse t
- Area =  $\frac{(1 - \text{Confidence level})}{2}$
- $df = n - 1$

8. Plug in and calculate the Confidence Interval

- STAT
- TESTS
- 8: T Interval

9. Write the Interval

10. Write the Conclusion

We are \_\_\_\_\_% confident that the true population mean difference for \_\_\_\_\_ lies within the interval \_\_\_\_\_.

*Restate the definition of the mean differences*

11. Explain the meaning of the confidence level-if asked

In repeated sampling we expect this method to capture the true population mean difference

for \_\_\_\_\_ % of the time

*Restate the definition of the mean differences*

## Notes: Confidence Interval for a Paired T-test (Scenario)

**Scenario 2:** Professor Andy Neill measured the time (in seconds) required to catch a falling meter stick for 12 randomly selected students' dominant hand and non-dominant hand. Professor Neill claims that the reaction time in an individual's dominant hand is less than the reaction time in their non-dominant hand.

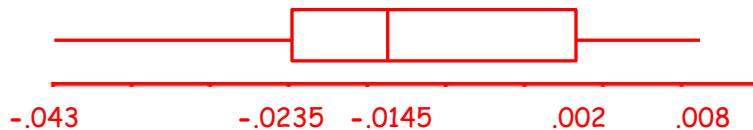
Student	1	2	3	4	5	6	7	8	9	10	11	12
Dominant Hand	0.177	0.210	0.186	0.189	0.198	0.194	0.160	0.163	0.166	0.152	0.190	0.172
Non-dominant	0.179	.202	.208	.184	.215	.193	.194	.160	.209	.164	.210	.197

Create a 97% confidence interval for the mean difference and explain the confidence level?

Let  $\mu_d$  = equal the true population mean difference in reaction time (dominant minus non-dominant)

### Conditions:

1. Random Sample - Given
2.  $12/10 = 120$ . Reasonable to assume that there are at least 120 students
3. The differences in reaction times are independent in that it is reasonable to assume that one person's reaction time does not impact the reaction time of others
4. The boxplot is reasonably symmetric with no outliers. The distribution appears to be approximately normal



$$\bar{x}_d = -.0132$$

$$s_d = .0164 \quad df = 12-1$$

$$\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}} \rightarrow -.0132 \pm 2.49 \frac{.0164}{\sqrt{12}}$$

$$(-.025, -.0014)$$

We are 97% confident that the true population mean difference in reaction time (dominant minus non-dominant) lies within the interval  $(-.025, -.0014)$

In repeated sampling, we expect that this method will capture the true population mean difference in reaction time (dominant minus non-dominant) 97 percent of the time.

**Note:** 0 does not lie within the interval. Therefore we are 97 percent confident that there is a difference in reaction time (dominant minus non-dominant)

## Recipe for Success: Hypothesis Test for a Paired T-test

### 1. Write your Hypothesis

- Null  $H_0: \mu_d = 0$
- Alternative  $H_A: \mu_d \neq$  or  $<$  or  $>$  0

### 2. Define parameter $\mu_d$ in context & write the conditions

1. Random Sample
2.  $n < 10\%$  of the population
3. Differences are independent
4. The differences are Normal or  $n > 30$

*If data is given, draw a boxplot or histogram-to show normality*

### 3. Write the Equation

$$t = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

$t$  = the number of standard deviations a value is from the mean  
 $\mu_d = 0$

$\bar{x}_d$  = the mean of the sample differences

$s_d$  = the standard deviation of the sample differences

$n$  = the number of sample differences

### 4. Enter Data (if given)

- Stat Edit

• Enter Data in columns  $L_1$  &  $L_2$

• At the top of  $L_3$  type  $2^{nd}$   $L_1 - 2^{nd}$   $L_2$

### 5. Find $\bar{x}_d$

- Stat Calc

• 1-Var Stats press Enter

• List type  $2^{nd}$   $L_3$

### 6. List & Label all of input values

$n, \bar{x}_d, s_d$

$\mu_d = 0$  &  $df = n - 1$

### 7. Plug values into the equation

### 8. Calculate the $t$ and the p-value

$\mu_d = 0$

$df = n - 1$

- Stat Tests

• 2:T-Test Enter

• Highlight Data if data is used otherwise highlight **STATS**

•  $s_d$  come from the problem or the data

•  $\bar{x}_d$  comes from the problem or the data

•  $n$  comes from the problem or the data

• Choose  $\neq$  or  $<$  or  $>$  (look for key words)

### 9. State the Decision

• The p-value is \_\_\_\_\_

• If the p-value is less than alpha, Reject the Null

• If the p-value is greater than alpha, Fail to reject the Null

### 10. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population mean differences for

\_\_\_\_\_ is \_\_\_\_\_

*Restate the definition of the mean differences      Restate  $H_A \neq$  or  $<$  or  $>$  mean differences*

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population mean difference for

\_\_\_\_\_ is \_\_\_\_\_

*Restate the definition of the mean difference      Restate  $H_A \neq$  or  $<$  or  $>$  mean difference*





### Notes: Confidence Interval for a Paired T-test (Scenario)

**Scenario 3: 2007 Problem 4** Investigators at the U.S. Department of Agriculture wished to compare methods of determining the level of *E. coli* bacteria contamination in beef. Two different methods (A and B) of determining the level of contamination were used to each of ten randomly selected specimens of a certain type of beef. The data obtained, in milli-microbes/liter of ground beef, for each of the methods are shown in the table below.

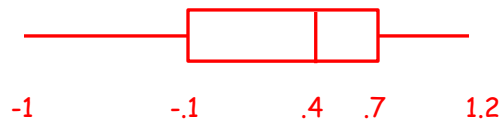
	Specimen									
	1	2	3	4	5	6	7	8	9	10
A	22.7	23.6	24.0	27.1	27.4	27.8	34.4	35.2	40.4	46.8
B	23.0	23.1	23.7	26.5	26.6	27.1	33.2	35.0	40.5	47.8

Create a 90% Confidence Interval for the mean difference. Does your confidence interval indicate that the detection methods are different? Justify your response

Let  $\mu_d$  = equal the true population mean difference in detection methods for ecoli (A - B)

#### Conditions:

1. Random Sample- Given
2.  $10/10 = 100$ . There are a few more than 100 pieces of beef.
3. It is reasonable to assume that the test differences for ecoli are independent in that one test should not impact another.
4. The boxplot is reasonably symmetric with no outliers. The distribution appears to be approximately normal



$$\bar{x}_d = -.29$$

$$s_d = .6297 \quad df = 10-1$$

$$\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}} \rightarrow -.29 \pm 1.83 \frac{.6297}{\sqrt{10}}$$

$$(-.075, .655)$$

We are 90% confident that the true population mean difference in detection methods for ecoli (A - B) lies with the interval (-.075, .655)

In repeated sampling, we expect that this method will capture the true mean difference in detection methods for ecoli 90 percent of the time.

Note: 0 does lie within the interval. Therefore we are unable to determine which method is better at detecting ecoli.

## Notes: Type I & Type II Errors

### Review of Concepts:

#### How to Think About P-Values

- A P-value is a conditional probability
  - The probability of getting a test statistic as extreme or more extreme given that the null hypothesis is true.
  - There is a \_\_\_\_\_% chance that we would get a test statistic this extreme in favor of \_\_\_\_\_ when in fact \_\_\_\_\_ is true.
- $P\text{-value}$   
*Restate  $H_A$*                       *Restate  $H_0$*

**Note:** The P-value is NOT the probability that the null hypothesis is true.

#### How to Think About Alpha Levels

- Sometimes we need to make a firm decision about whether or not to reject the null hypothesis.
- When the P-value is small, it tells us that our data **are rare given the null hypothesis**.
- We can define a "rare event" by arbitrarily setting a threshold for our P-value.
  - If our P-value falls below that point, we will reject the null in favor of the alternative. We call such results statistically significant.
  - The threshold is called an alpha level, denoted by  $\alpha$ .
- The alpha level is also called the significance level. When we reject the null hypothesis, we say that the test is "significant at that level."
- What can you say if the P-value is large and does not fall below  $\alpha$ ?
  - You should say: "The data have failed to provide sufficient evidence to reject the null hypothesis."
  - Do not say that you "accept the null hypothesis."
  - Recall that, in a jury trial, if we do not find the defendant guilty, we say the defendant is "not guilty"—we don't say that the defendant is "innocent."

#### Sampling Variation and Alpha

- The alpha level measures the risk that we are willing to take in rejecting a true null.
  - An alpha level of .05 means that we expect to incorrectly reject the null hypothesis 5% of the time.
  - This error is due to the natural variations that comes from sampling—sampling variation
  - Obviously, different people and different situations will result in different significance levels or alpha levels
    - What you consider to be statistically significant might not be the same as what someone else considers statistically significant.

**Note:** Different alpha levels may be used, but each test will give only one P-value.

Why Wouldn't We Choose a very small alpha to minimize the chance of mistakenly rejecting a true null?

- Answer: Because it becomes very difficult to correctly reject a false null.
- Selecting alpha is a balancing act between 2 types of errors:
  1. **Rejecting a true null and**
  2. **Failing to reject a false null**

## Notes: Type I & Type II Errors

### Making Errors

- Here's some shocking news: nobody's perfect. Even with lots of evidence we can still make the wrong decision. This error is due to sampling variation
- When we perform a hypothesis test, we can make mistakes in *two* ways:
  1. The null hypothesis is true, but we mistakenly reject it. (Type I error or alpha- $\alpha$ )
  2. The null hypothesis is false, but we fail to reject it. (Type II error or Beta- $\beta$ )
- Which type of error is more serious depends on the situation at hand. In other words, the gravity of the error is context dependent.
- Here's an illustration of the four situations in a hypothesis test:

		The Truth	
		$H_0$ True	$H_0$ False
My Decision	Reject $H_0$	Type I Error	Correct Decision
	Retain $H_0$	Correct Decision	Type II Error

**Note: If we reject, the only type of error that we can commit is a Type I Error**

**Note: If we fail to reject, the only type of error that we can commit is a Type II Error**

**Alpha is the probability of committing a Type I Error.**

- Alpha is a conditional probability
- The probability of rejecting the null given that the null is in fact true

**How often will a Type I error occur?**

- Since a Type I error is rejecting a true null hypothesis (this is a mistake), the probability of a Type I error is our  $\alpha$  level—we get to choose how often we are wrong

**Question:** Why not choose a very small alpha level to minimize our chances of rejecting a true null and being guilty of committing a Type I Error?

**Answer:** Because it becomes very difficult to correctly reject a false null which leads to a different type of error.

**Beta is the probability of committing a Type II Error.**

- Beta is a conditional probability
- The probability of failing to reject the null given that the null is in fact false

**How often will a Type II error occur?**

- Very difficult to answer because we do not know the actual population parameter
- As our alpha level increases our beta level decreases

**Note:** alpha plus beta do not equal one

- We can reduce Type II Error by increasing the sample size

**Power of the test ( $1-\beta$ ):**

- The probability of correctly rejecting a False null hypothesis (this is correct)
- Power is a conditional probability:
 

The probability of rejecting the null hypothesis given that it is false.
- We can increase the power of the test by increasing the sample size
- There is an inverse relationship between power and Type II Error
- If we increase our alpha level our power goes up

## Notes: Type I & Type II Errors

**Type I Error**-the probability of rejecting the null given that the null is true-this is measured by **alpha**.

- A Type I Error cannot be committed if the null is not rejected
- A Type I Error may have been committed if the null is rejected (but only if the null is true)

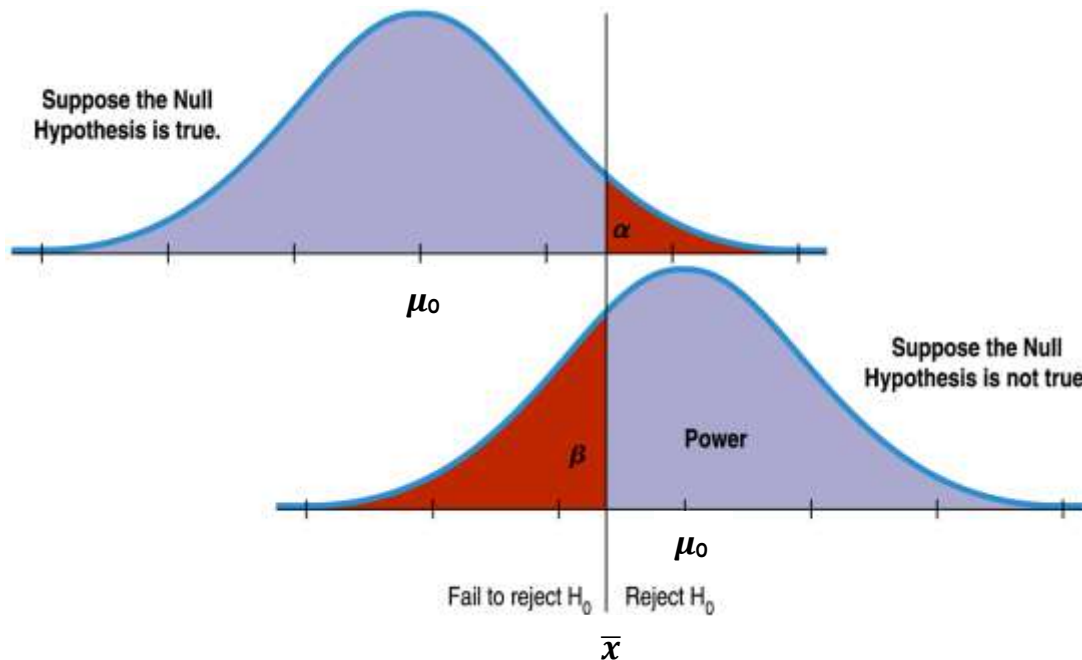
**Type II Error**-the probability of accepting the null given that the null is in fact false-this is measured by **beta**.

- A Type II Error cannot be committed if the null is rejected
- A Type II Error may have been committed if the null is not rejected (but only if the null is false)

**Power of the test**  $1-\beta$  represents the probability rejecting the null hypothesis given that it is false or correctly rejecting a false null.

The diagram below demonstrates the relationship between Type I and Type II Errors.

**Memorize and be able to draw this diagram**



**As alpha increases, Beta decrease and power increases**

- More likely to reject the null hypothesis
- More likely to commit a Type I error.
- Less likely to commit a Type II error.

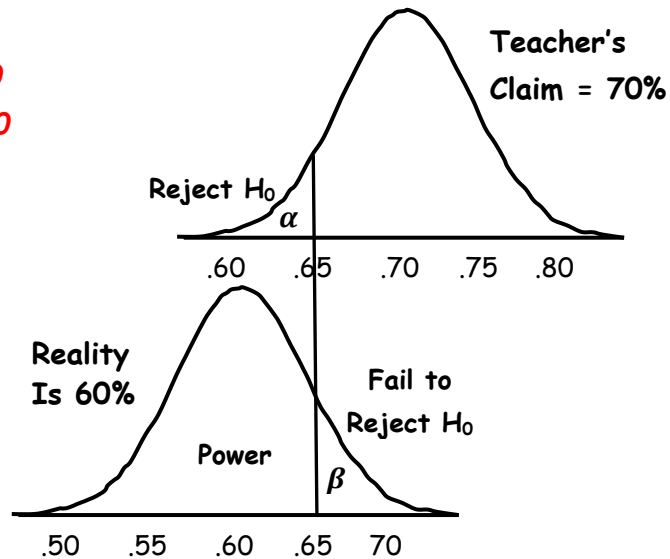
**As alpha decrease, beta increases and Power decreases**

- Less likely to reject the null hypothesis
- Less likely to commit a Type I error.
- More likely to commit a Type II error.

## Notes: Type I & Type II Errors

**Example:** A teacher claims that the test average is 70%, but you believe that the true figure is lower, you plan to gather an SRS and will reject the claim if in your sample the test average is 65% or less. The truth is that the test average is a 60%.

Null  $H_0: \mu_0 = 70$   
 Alternative  $H_A: \mu_0 < 70$



The upper graph shows the null hypothesis model with the claim that  $\mu_0 = .70$  and the plan to reject the null if  $\mu_0 < .65$ . The lower graph shows the true model with  $\mu_0 = .60$ .

When will we fail to reject the null hypothesis that  $\mu_0 = .70$ ?

**We Fail to reject the teacher's claim for any sample with an average greater than .65. This would result in a Type II Error because we know that the actual average is .60**

When will we rightly conclude that the null hypothesis is incorrect?

**We will correctly reject the null hypothesis for any sample with an average less than .65. The probability that we would correctly reject the null is known as power.**

Power gives the probability of avoiding a Type II error. **Power = (1 -  $\beta$ )**

Power has a different value for different possible correct values of the population parameter.

Choosing a smaller alpha results in a higher risk of committing a Type II error.

The greater the distance between the null and the true value of the population, the smaller the risk of committing a Type II error and the greater the power.

**Notes: Type I & Type II Errors**

**Example:** What reason did the Trump administration give to justify the need for extreme vetting of Syrian Refugees who were seeking political asylum?

Write the Hypothesis

Null  $H_0$ : The person immigrating is not a terrorist

Alternative  $H_A$ : The person immigrating is a terrorist

Fill in the table		Actual Situation	
		The person is not Terrorist	The person is a Terrorist
Administration/ Immigration Decision	The person is not Terrorist	Correct Decision	Type II Error We failed to reject $H_0$
	The person is a Terrorist	Type I Error We Rejected $H_0$	Correct Decision

Answer the questions in context:

What is a Type I Error in this situation and what are the consequences:

A Type I Error is rejecting the Null by mistake when in fact it is true.

In this case, we would be labeling a Syrian Refugee as a terrorist when they are not.

As a consequence, the innocent Syrian refugee will not be allowed to immigrate and stands a very real chance of being killed in their home country.

What is a Type II Error in this situation and what are the consequences

A Type II Error is accepting the Null by mistake when in fact it is false.

In this case, we would mistakenly believe that a Syrian Refugee is not a terrorist when they are.

As a consequence, the Syrian Refugee-terrorist would be allowed to immigrate placing American lives at risk.

## Recipe for Success: Type I Errors

**Type I Error**-The probability of Rejecting the null given that the null is True.

1. **Write the Hypothesis**
  - Null  $H_0$ :
  - Alternative  $H_A$ :

$\alpha$  = the probability of committing a Type I error  
 $\beta$  = the probability of committing a Type II error  
 $1 - \beta$  = the power of the test
2. **Define parameter ( $\mu$  or  $p$ ) in context**
3. **Define a Type I Error  $\alpha$** 

**Definition:** The probability of rejecting the Null given that the Null is true.

**Remember:** To commit a **Type I Error**, we must have **rejected the null** and were incorrect

**simplified definition:** Rejecting a true Null and Accepting a False Alternative
4. **Explain a Type I Error in Context of the problem**

In this case, the probability of rejecting \_\_\_\_\_ in favor of \_\_\_\_\_

*Restate  $H_0$*

\_\_\_\_\_ given the fact \_\_\_\_\_ is true.

*Restate  $H_A$*                       *Restate  $H_0$*
5. **Explain the consequences of Committing a Type I Error**

**The consequences for committing at Type I Error are...**

Explanation must be in the context and in simplified language.

**Methods of Decreasing Type I Errors- $\alpha$**

  1. **Decrease  $\alpha$  - the level of significance**
    - Increases  $\beta$ -the probability of a Type II Error
    - More likely to accept a false null-(this is an error)
    - Power Decreases
  2. **Decrease Power**
    - More likely to accept a false null—(Type II increases: negative)
    - Less likely to reject a true null—(Type I decreases: positive)
  3. **Increase Sample Size**
    - **Decreases Type I Error**
    - Decreases Type II Error
    - Increases Power
    - (Costs money and Time)

**In General:**

As  $\alpha \downarrow$ , power  $\downarrow$ , &  $\beta \uparrow$

And

As  $\alpha \uparrow$ , power  $\uparrow$ , &  $\beta \downarrow$

### P-value

1. **Write the Hypothesis**
  - Null  $H_0$ :
  - Alternative  $H_A$ :
2. **Define parameter ( $\mu$  or  $p$ ) in context**
3. **Define P-value**

P-value is the probability of getting a test statistic as extreme or more extreme given that the null is true.
4. **Explain P-value in the Context of the problem**

There is a \_\_\_\_\_% chance that we would get a test statistic \_\_\_\_\_

*P-value*

this extreme in favor of \_\_\_\_\_ when in fact \_\_\_\_\_ is true

*Restate  $H_A$*                       *Restate  $H_0$*



## Notes: Type I & Type II Errors (Scenario)

**Scenario: 2008 Form B Question 4.** A researcher wants to conduct a study to test whether listening to soothing music for 20 minutes help to reduce diastolic blood pressure in patients with high blood pressure, compared to simply sitting quietly in a noise-free environment for 20 minutes. One hundred patients with high blood pressure at a large medical clinic are available to participate in this study.

- (a) Propose a design for this study to compare these two treatments.
- (b) The null hypothesis for this study is that there is no difference in the mean reduction of diastolic blood pressure for the two treatments and the alternative hypothesis is that the mean reduction in diastolic blood pressure is greater for the music treatment. If the null hypothesis is rejected, the clinic will offer this music therapy as a free service to their patients with high blood pressure. Describe a Type I error and the consequences in the context of this study
- (c) A p-value of .0017 is calculated. Explain the meaning of the p-value in context and explain what type of error may have been committed.

**(a) Write the names of the 100 volunteers on 100 individual slips of paper. Place the slips in a bin and mix thoroughly. Draw 50 names without replacement. The first 50 names will be assigned to listening to soothing music for 20 minutes. The remaining 50 patients will sit quietly in a noise free environment. Prior to administering the treatment the diastolic blood pressure will be measured for each volunteer. After the 20 minute treatment has passed the patients will again measure their diastolic pressure and compare diastolic pressure differences (pre-post) to determine which treatment resulted in the greatest decrease.**

- (b) Null                     $H_0: \mu_1 = \mu_2$   
Alternative             $H_A: \mu_1 > \mu_2$**

**Let  $\mu_1$  equal the true population mean for diastolic blood pressure drop for those listening to music**

**Let  $\mu_2$  equal the true population mean for diastolic blood pressure drop for those sitting in a quiet room**

**A Type I error is mistakenly rejecting the null when in fact it is true.**

**In this case, we would mistakenly believe that listening to music would have a greater impact on lowering diastolic blood pressure than sitting quietly in a room when in fact it did not.**

**As a result the clinic would be providing and paying for a treatment for patients that wasn't any more effective than sitting in a quiet room.**

- (c) P-value is the probability of getting a test statistic as extreme or more extreme given that then null is true. In this case, there is .17% chance that we would get a test statistic this extreme in favor music having a greater reduction in diastolic blood pressure than sitting in a quiet room when in fact the two treatments are equal in their effectiveness.**

**A p-value of .0017 is less than alpha = .05 thus we would have rejected the null.**

**If we reject the null, the only type of error that we could have committed is a Type I Error.**

## Recipe for Success: Type II Errors

**Type II Error**-The probability of **Accepting the null** given that the null is **False**.

1. **Write the Hypothesis**
  - Null  $H_0$ :
  - Alternative  $H_A$ :

$\alpha$  = the probability of committing a Type I error  
 $\beta$  = the probability of committing a Type II error  
 $1 - \beta$  = the power of the test
2. **Define parameter ( $\mu$  or  $p$ ) in context**
3. **Define a Type II Error  $\beta$** 

**Definition:** The probability of accepting the Null given that it is false.

**simplified definition:** Accepting a False Null and Rejecting a True Alternative

**Remember:** to commit a **Type II Error** we failed to reject the null and were incorrect.
4. **Explain Type II Error in Context of the problem**

In this case, the probability of accepting \_\_\_\_\_ given that \_\_\_\_\_ is false and \_\_\_\_\_ is true.

*Restate  $H_0$*                       *Restate  $H_A$*
5. **Explain the consequences of Committing a Type II Error**

**The consequences for committing a Type II Error are...**

Explanation must be in the context and in simplified language.

### Methods of Decreasing Type II Errors- $\beta$

As  $\beta \downarrow$ , power  $\uparrow$ , &  $\alpha \uparrow$

And

As  $\beta \uparrow$ , power  $\downarrow$ , &  $\alpha \downarrow$

1. **Increase  $\alpha$  - the level of significance**
  - Increases-the probability of a Type I Error  $\alpha$
  - More likely to reject a True Null—(this is an error)
  - Power Increases
2. **Increase Power**
  - More likely to reject a true null—(Type I increases: negative)
  - Less likely to accept a false null—(Type II decreases: positive)
4. **Increase Sample Size**
  - **Decreases Type I Error**
  - Decreases Type II Error
  - Increases Power
  - (Costs money and Time)

### P-value

1. **Write the Hypothesis**
  - Null  $H_0$ :
  - Alternative  $H_A$ :
2. **Define parameter ( $\mu$  or  $p$ ) in context**
3. **Define P-value**

P-value is the probability of getting a test statistic as extreme or more extreme given that the null is true.
4. **Explain P-value in the Context of the problem**

There is a \_\_\_\_\_% chance that we would get a test statistic \_\_\_\_\_ this extreme in favor of \_\_\_\_\_ when in fact \_\_\_\_\_ is true.

*P-value*                      *Restate  $H_A$*                       *Restate  $H_0$*

## Notes: Type I & Type II Errors (Scenario)

**Scenario: 2008 Form B Question 4.** A researcher wants to conduct a study to test whether listening to soothing music for 20 minutes help to reduce diastolic blood pressure in patients with high blood pressure, compared to simply sitting quietly in a noise-free environment for 20 minutes. One hundred patients with high blood pressure at a large medical clinic are available to participate in this study.

- (a) The null hypothesis for this study is that there is no difference in the mean reduction of diastolic blood pressure for the two treatments and the alternative hypothesis is that the mean reduction in diastolic blood pressure is greater for the music treatment. If the null hypothesis is rejected, the clinic will offer this music therapy as a free service to their patients with high blood pressure. Describe a Type II errors and its consequences in the context of this study, Also, discuss in the context of this study which is more serious a Type I or Type II error.
- (b) A p-value of .247 is calculated. Explain the meaning of the p-value in context and explain what type of error may have been committed.

(a) Null  $H_0: \mu_1 = \mu_2$   
 Alternative  $H_A: \mu_1 > \mu_2$

Let  $\mu_1$  equal the true population mean for diastolic blood pressure drop for those listening to music

Let  $\mu_2$  equal the true population mean for diastolic blood pressure drop for those sitting in a quiet room

A Type II Error is mistakenly failing to reject the null when in fact it is false.

In this case, we would mistakenly believe that there was no difference in listening to music and sitting in a quiet room on lowering diastolic blood pressure when in fact listening to music was more effective.

As a result the clinic would not provide an effective treatment (listening to music) for lowering diastolic blood pressure.

**Clinic:** A Type I Error is more serious because the clinic is wasting money on an ineffective treatment and telling patients they are improving when they are not.

**Patients:** A Type II Error is more serious because they are not receiving an effective treatment (the music) and will be forced to treat their condition with drugs or some other more invasive treatment.

- (b) P-value is the probability of getting a test statistic as extreme or more extreme given that then null is true. In this case, there is 24.7% chance that we would get a test statistic this extreme in favor music having a greater reduction in diastolic blood pressure than sitting in a quiet room when in fact the two treatments are equal in their effectiveness.

A p-value of .247 is greater than alpha = .05 thus we would have failed to reject the null.

If we fail to reject the null, the only type of error that we could have committed is a Type II Error.

**Multiple Choice Practice Errors**

- Which of the following statements is incorrect?  
(A) The significance level of a test is the probability of a Type II error.  
(B) Given a particular alternative, the power of a test against that alternative is 1 minus the probability of the Type II error associated with that alternative.  
(C) If the significance level remains fixed, increasing the sample size will reduce the probability of a Type II error.  
(D) If the significance level remains fixed, increasing the sample size will raise the power.  
(E) Holding the sample size fixed, increasing the significance level will decrease the probability of a Type II error.
- A manufacturer of heart-lung machines periodically checks a sample of its product and performs a major recalibration if readings are sufficiently off target. Similarly, a rug factory periodically checks the sizes of its throw rugs coming off an assembly line and halts production if measurements are sufficiently off target. In both situations, we have the null hypothesis that the production equipment is performing satisfactorily. For each situation, which is the more serious concern, a Type I or Type II error?  
(A) Machine producer: Type I error, carpet manufacturer: Type I error  
(B) Machine producer: Type I error, carpet manufacturer: Type II error  
(C) Machine producer: Type II error, carpet manufacturer: Type I error  
(D) Machine producer: Type II error, carpet manufacturer: Type II error  
(E) This is impossible to answer without making an expected value judgment between human life and accurate throw rug sizes.
- Which of following is incorrect?  
(A) The power of a test concerns its ability to detect a true alternative hypothesis.  
(B) The significance level of a test is the probability rejecting a true null hypothesis.  
(C) The probability of a Type I error plus the probability of a Type II error always equals 1.  
(D) Power equals 1 minus the probability of failing to reject a false null hypothesis.  
(E) Anything that makes a null hypothesis harder to reject will increase the probability of committing a Type II error.
- Suppose  $H_0: p = .6$ ,  $H_A: p > .6$ , and against the alternative  $p = .7$ , the power is  $.8$ . Which of the following is a valid conclusion?  
(A) The probability of committing a Type I error is  $.1$ .  
(B) If  $p = .7$  is true, the probability of failing to reject  $H_0$  is  $.2$ .  
(C) The probability of committing a Type II error is  $.3$ .  
(D) All of the above are valid conclusions.  
(E) None of the above are valid conclusions.
- If all other variables remain constant, which of the following will not increase the power of a hypothesis test?  
(A) Increasing the sample size  
(B) Increasing the significance level  
(C) Increasing the probability of a Type II error  
(D) Decreased variability in the data  
(E) Increased distance between the true and the hypothesized parameter.

## Notes: Review of Rules for Random Variables

### Rules for Adding, Subtracting & Multiplying by Constants

#### Addition:

- Constants **can be** added to measures of center (mean & median)
- Constants **can NOT be** added to measures of spread (range, IQR, variance & standard deviation)

#### Subtraction:

- Constants **can be** subtracted to measures of center (mean & median)
- Constants **can NOT be** subtracted to measures of spread (range, IQR, variance & standard deviation)

#### Multiplication:

- Constants **can be** multiplied to measures of center (mean & median)
- Constants **can be** multiplied to measures of spread (range, standard deviation & IQR)
- Constants **must be squared** first and the result can be multiplied by a variance

#### Division:

- Measures of center (mean & median) **can be** divided by a constant
- Measures of spread (range, IQR & standard deviation) **can be** divided by the constant
- Variances can be divided by the square of the constant

#### Example:

Random Variable	Mean	Median	Range	IQR	Standard Deviation	Variance
(X)	23.5	21	15	6.3	2.7	7.29

Calculate the mean, median, range, IQR, standard deviation and variance for the following:

1.  $X + 4$

Random Variable	Mean	Median	Range	IQR	Standard Deviation	Variance
(X)	27.5	25	15	6.3	2.7	7.29

2.  $X - 20$

Random Variable	Mean	Median	Range	IQR	Standard Deviation	Variance
(X)	3.5	1	15	6.3	2.7	7.29

3.  $3X$

Random Variable	Mean	Median	Range	IQR	Standard Deviation	Variance
(X)	70.5	63	45	18.9	8.1	65.61

4.  $X \div 4$

Random Variable	Mean	Median	Range	IQR	Standard Deviation	Variance
(X)	5.875	5.25	3.75	1.575	.675	.455625

5.  $.8X + 2$

Random Variable	Mean	Median	Range	IQR	Standard Deviation	Variance
(X)	20.8	18.8	12	5.04	2.16	4.6656

## Notes: Review of Rules for Random Variables

### Random Variables: Rules for Combining Random Variables

#### Addition:

- **Means:** can be summed directly
- **Variances:** can be summed directly
- **Standard deviations:** can NOT be added. Must be converted to variances summed and then the square root of the sum must be taken to convert the summed variance to a standard deviation

#### Subtraction:

- **Means:** can be subtracted directly
- **Variances:** can NOT be subtracted and must be summed directly
- **Standard deviations:** can NOT be subtracted. Must be converted to variances summed and the square root of the sum must be taken to convert the summed variance to a standard deviation

#### Example:

Random Variable	Mean	Standard Deviation	Variance
(X)	23	4	16
(Y)	29	5	25

For the above independent Random Variables calculate the mean, standard deviation and variance for the following:

1.  $X + Y$

Mean	Standard Deviation	Variance
$23+29=52$	$(4^2 + 5^2)^{.5} = 6.403$	$16 + 25 = 41$

2.  $X - Y$

Mean	Standard Deviation	Variance
$23-29=5-3$	$(4^2 + 5^2)^{.5} = 6.403$	$16 + 25 = 41$

3.  $.4X - .6Y$

Mean	Standard Deviation	Variance
$23-29=5-3$	$(4^2 + 5^2)^{.5} = 6.403$	$16 + 25 = 41$

4.  $(X + X + X) \div 3$

Mean	Standard Deviation	Variance
$(23+23+23) \div 3 = 23$	$((4^2+4^2+4^2)^{.5} \div 3) = 2.309$	$((4^2+4^2+4^2)^{.5} \div 3)^2 = 5.333$

5.  $(Y + Y + Y + Y) \div 4$

Mean	Standard Deviation	Variance
$(29+29+29+29) \div 4 = 29$	$((5^2+5^2+5^2+5^2)^{.5} \div 4) = 2.5$	$((5^2+5^2+5^2+5^2)^{.5} \div 4)^2 = 6.25$

6.  $(X + X + X) \div 3 + (Y + Y + Y + Y) \div 4$

Mean	Standard Deviation	Variance
$(23+29) = 52$	$(2.309^2+2.5^2)^{.5} = 3.403$	$5.333 + 6.25 = 11.583$

## Notes: Paired T-test Versus a 2-Sample T-test

**Purpose:** A 2-sample t-test is used to determine whether or not there exists a difference between the means of 2 independent groups.

**When to Use:** We often want to know how two groups differ, whether a treatment is better than a placebo control, or whether this year's results are better than last year's.

- **Paired T-tests: Samples are dependent a relationship exists**
  - Same subject tested before and after treatment is given
  - Same subject tested twice (receives both treatments)
  - Tests on Twins and the difference between each pair is computed
  - Tests on Spouses and the difference between each couple is computed
  - Tests on Siblings and the difference between each sibling pair is computed
  - Data groups must be the exact same length
- **2-samples T-tests: Samples are independent**
  - Treatments are randomly assigned and the average of one group is compared to another.
  - Data groups do not need to be the same length

### Determining whether to use a Paired T-test or a 2-Sample T-test

1. An experiment measures people's lung capacity before and then after an exercise program to see if their fitness has improved.
  - Which T-test would you use? Paired T-test or independent 2 sample t-test
  - How many tails does this test have? 1-tailed or 2 tailed
2. A different experiment measures the lung capacity of one group who took one exercise program and another group who took a different exercise program to determine if a difference in outcomes existed.
  - Which T-test would you use? Paired T-test or independent 2 sample t-test
  - How many tails does this test have? 1-tailed or 2 tailed
3. You want to see if an educational program will help raise test scores. Your experimental hypothesis is "Post-training test scores will be significantly higher than Pre-training test scores."
  - Which T-test would you use? Paired T-test or independent 2 sample t-test
  - How many tails does this test have? 1-tailed or 2 tailed
4. For Each Scenario determine whether it would be more appropriate to use a 2-sample T-test or Paired T-test?

#### Example Scenario

- Comparing the average height of men and women
- Comparing the weight of spouses
- Comparing the BMI of teen agers to that adults
- Comparing Cholesterol levels of a child to their parent
- Comparing the blood pressures of patients receiving Drug A to those receiving Drug B
- Comparing the weight of dieters before and after a diet.

## Recipe for Success: 2-Sample T Hypothesis Test (difference of Means)

### 1. Write the Hypothesis

- Null  $H_0: \mu_1 = \mu_2$
- Alternative  $H_A: \mu_1 \neq \mu_2$  or  $\mu_1 < \mu_2$  or  $\mu_1 > \mu_2$

### 2. Define $\mu_1$ & $\mu_2$ in context

### 3. Write the Conditions

1. Both samples are random
2.  $n < 10\%$  of the population
3. Populations are independent
4. Normal Population or  $n > 30$

*If data is given, draw a boxplot or histogram-to show normality*

### 4. Write the Equation

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$t$  = the number of standard deviations from the mean

$\mu_1$  &  $\mu_2$  = the means of the population (may be assumed)

$\bar{x}_1$  &  $\bar{x}_2$  = the means of the samples

$s_1$  &  $s_2$  = the standard deviation of the sample

$n_1$  &  $n_2$  = the size of the sample

### 5. Enter Data (if given)

- Stat Edit
- Enter Data in columns  $L_1$  &  $L_2$
- 2<sup>nd</sup> Quit
- Stat Calc
- 1-Var Stats  $L_1: \bar{x}_1, s_1, n_1$  and  $L_2: \bar{x}_2, s_2, n_2$

### 6. List & Label all of input values

$\bar{x}_1, s_1, n_1, \bar{x}_2, s_2, n_2$

df (comes from the calculator)

- Stat Tests
- 4: 2-SampTTest Enter
- Highlight **Data** if data is used otherwise highlight **STATS**
- $s_1$  &  $s_2$  comes from the problem or the data
- $\bar{x}_1$  &  $\bar{x}_2$  comes from the problem or the data
- $n_1$  &  $n_2$  comes from the problem or the data
- **pooled** highlight no
- **Choose  $\neq$  or  $<$  or  $>$**  (look for key words)

### 7. Plug values into the equation

### 8. Write the $t$ and the $p$ -value

The  $T$  and the  $p$ -value are calculated in step 5

### 9. State the Decision

- The  $p$ -value is \_\_\_\_\_
- compare to alpha:  $p$ -value ( $<$  or  $>$ ) **alpha**
- If the  $p$ -value is **less** than alpha, **Reject the Null**
- If the  $p$ -value is **greater** than alpha, **Fail to reject the Null**

### 10. Write the Conclusion

**Reject the Null:** Our  $p$ -value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at  $\alpha =$  \_\_\_\_\_ to suggest that the difference in the true population mean for

\_\_\_\_\_ is \_\_\_\_\_  
 Restate the definition of the 1<sup>st</sup> mean      Restate  $H_A \neq$  or  $<$  or  $>$  2<sup>nd</sup> mean definition

**Fail to Reject the Null:** Our  $p$ -value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at  $\alpha =$  \_\_\_\_\_ to suggest that the difference in true population mean for

\_\_\_\_\_ is \_\_\_\_\_  
 Restate the definition of the 1<sup>st</sup> mean      Restate  $H_A \neq$  or  $<$  or  $>$  2<sup>nd</sup> mean definition



## Notes: 2-Sample T Hypothesis Test (Scenario)

**Scenario 1:** The average ACT score of an independent random sample of 16 non-athletes at Blinn College was compared to that of an independent random sample of 7 athletes. The school's guidance counselor was curious to know whether or not Blinn was admitting athletes with lower ACT scores than non-athletes. Perform an appropriate test and state your conclusion. Assume alpha = .05

Null  $H_0: \mu_1 = \mu_2$   
 Alternative  $H_A: \mu_1 > \mu_2$

Let  $\mu_1$  equal the true population mean ACT Score for non-athletes at Blinn College

Let  $\mu_2$  equal the true population mean ACT Score for athletes at Blinn College

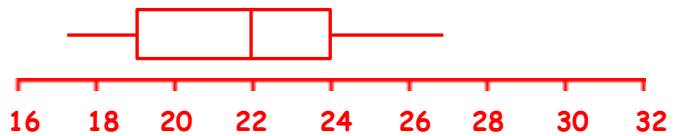
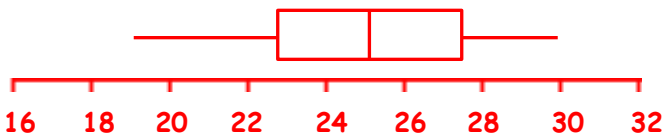
Composite ACT Score		
Non-Athletes		Athletes
25	21	22
22	27	21
19	29	24
25	26	27
24	30	19
25	27	23
24	26	17
23	23	

### Conditions: Non-Athletes

1. Random Sample-Given
2.  $16/.10 = 160$  It is reasonable to assume that there are at least 160 students at Blinn who are not athletes
3. The samples are independent-Given
4. The boxplot is reasonably symmetric with no outliers. The distribution appears to be approximately normal

### Conditions: Athletes

- Random Sample-Given  
 $7/.10 = 70$  It is reasonable to assume that there are at least 70 students at Blinn who are not athletes  
 The samples are independent-Given  
 The boxplot is reasonably symmetric with no outliers. The distribution appears to be approximately normal



$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad df = 10.12 \quad P(t > \frac{24.75 - 21.86}{\sqrt{\frac{(2.84)^2}{16} + \frac{(3.29)^2}{7}}})$$

$\bar{x}_1 = 24.75$      $\bar{x}_2 = 21.86$      $= P(t > 2.021) = .0352$

$s_1 = 2.84$      $s_2 = 3.29$     P-value                      alpha

$n_1 = 16$      $n_2 = 7$     .0352    Less than    .05

We reject the  $H_0$

Our p-value is .0352. We reject the null. There is sufficient evidence at alpha = .05 to suggest that the differences in the true population mean of ACT scores for non-athletes is greater than that of athletes.

On average, non-Athletes have higher ACT Scores than athletes

## Recipe for Success: 2 Sample T-Confidence Intervals

1. Define  $\mu_1$  &  $\mu_2$  in context
2. Write the Conditions

1. Both samples are random
2.  $n < 10\%$  of the population
3. Populations are independent
4. Normal Population or  $n > 30$

*If data is given, draw a boxplot or histogram-to show normality*

3. Write the formula for the Test

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$\bar{x}_1$  &  $\bar{x}_2$  = the means of the samples

$s_1$  &  $s_2$  = the standard deviation of the sample

$n_1$  &  $n_2$  = the size of the sample

4. Graph and Shade
5. Enter the Data if Given

- Stat Edit
- Enter Data in columns  $L_1$  &  $L_2$
- 2<sup>nd</sup> Quit
- Stat Calc
- 1-Var Stats  $L_1$ :  $\bar{x}_1, s_1, n_1$  and  $L_2$ :  $\bar{x}_2, s_2, n_2$

6. List & Label all of input values

$\bar{x}_1, s_1, n_1, \bar{x}_2, s_2, n_2, df$

$df$  (comes from the calculator)

- Stat Tests
- 0:2-Samp T Int
- Highlight **Data** if data is used otherwise highlight **STATS**
- $s_1$  &  $s_2$  comes from the problem or the data
- $\bar{x}_1$  &  $\bar{x}_2$  comes from the problem or the data
- $n_1$  &  $n_2$  comes from the problem or the data
- **pooled** highlight no

7. Calculate  $t^*$

- 2<sup>nd</sup> Vars
- Inverse t
- Area =  $\frac{(1-\text{Confidence level})}{2}$
- $df$  (comes from the calculator in the step above)

8. Plug in the values

9. Write the interval

10. Write the Conclusion

We are \_\_\_\_\_% confident that the true population mean difference for \_\_\_\_\_

*Restate the definition of the  $\mu_1$*

and \_\_\_\_\_ lies within the interval \_\_\_\_\_

*Restate the definition of the  $\mu_2$*

11. Explain the meaning of the confidence level-if asked

In repeated sampling we expect this method to capture the true population mean difference

for \_\_\_\_\_ and \_\_\_\_\_% of the time.

*Restate the definition of the  $\mu_1$*       *Restate the definition of the  $\mu_2$*

**Notes: 2-Sample T Confidence Interval (Scenario)**

**Scenario 1:** The average ACT score of an independent random sample of 16 non-athletes at Blinn College was compared to that of an independent random sample of 7 athletes. The school's guidance counselor was curious to know whether or not Blinn was admitting athletes with lower ACT scores than non-athletes. Create and interpret both the 95% confidence interval and confidence level. Is the 95% confidence interval consistent with your hypothesis test? Explain.

Null  $H_0: \mu_1 = \mu_2$

Alternative  $H_A: \mu_1 > \mu_2$

**See Page 161 for  
Conditions**

Let  $\mu_1$  equal the true population mean ACT Score for non-athletes at Blinn College

Let  $\mu_2$  equal the true population mean ACT Score for athletes at Blinn College

Composite ACT Score		
Non-Athletes		Athletes
25	21	22
22	27	21
19	29	24
25	26	27
24	30	19
25	27	23
24	26	17
23	23	

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad df = 10.12 \rightarrow 24.75 - 21.86 \pm 2.22 \sqrt{\frac{(2.84)^2}{16} + \frac{(3.29)^2}{7}}$$

$$\bar{x}_1 = 24.75 \quad \bar{x}_2 = 21.86$$

$$s_1 = 2.84 \quad s_2 = 3.29$$

$$n_1 = 16 \quad n_2 = 7$$

Interval:

$$(-.2907, 6.0765)$$

We are 95% confident that the true population difference in means of ACT scores between non-athletes and athletes at Blinn College lies with the interval  $(-.2907, 6.0765)$

In repeated sampling, we expect that this method will capture the true population difference in means of ACT scores between non-athletes and athletes at Blinn College 95 percent of the time.

Note: 0 does lie within the interval. Therefore we are unable to determine whether or not athletes or non-athletes have a higher overall average ACT score

The hypothesis test is not consistent with the confidence interval. The hypothesis test found that the ACT scores of non-athletes was higher than that of athletes while the confidence interval found no difference.

The difference in results is due to the fact that the hypothesis test was one tailed with the rejection area being in only 1 tail and the confidence interval is a 2-tailed test with the rejection area being split between 2 tails. Remember that when going from one tail to two tail you must double your p-value.

Note: For consistent results run the hypothesis test as a two-tail test

## Recipe for Success: 2-Sample T Hypothesis Test (difference of Means)

### 1. Write the Hypothesis

- Null  $H_0: \mu_1 = \mu_2$
- Alternative  $H_A: \mu_1 \neq \mu_2$  or  $\mu_1 < \mu_2$  or  $\mu_1 > \mu_2$

### 2. Define $\mu_1$ & $\mu_2$ in context

### 3. Write the Conditions

1. Both samples are random
2.  $n < 10\%$  of the population
3. Populations are independent
4. Normal Population or  $n > 30$

*If data is given, draw a boxplot or histogram-to show normality*

### 4. Write the Equation

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$t$  = the number of standard deviations from the mean

$\mu_1$  &  $\mu_2$  = the means of the population (may be assumed)

$\bar{x}_1$  &  $\bar{x}_2$  = the means of the samples

$s_1$  &  $s_2$  = the standard deviation of the sample

$n_1$  &  $n_2$  = the size of the sample

### 5. Enter Data (if given)

- Stat Edit
- Enter Data in columns  $L_1$  &  $L_2$
- 2<sup>nd</sup> Quit
- Stat Calc
- 1-Var Stats  $L_1: \bar{x}_1, s_1, n_1$  and  $L_2: \bar{x}_2, s_2, n_2$

### 6. List & Label all of input values

$\bar{x}_1, s_1, n_1, \bar{x}_2, s_2, n_2$

df (comes from the calculator)

- Stat Tests
- 4: 2-SampTTest Enter
- Highlight **Data** if data is used otherwise highlight **STATS**
- $s_1$  &  $s_2$  comes from the problem or the data
- $\bar{x}_1$  &  $\bar{x}_2$  comes from the problem or the data
- $n_1$  &  $n_2$  comes from the problem or the data
- **pooled** highlight no
- **Choose  $\neq$  or  $<$  or  $>$**  (look for key words)

### 7. Plug values into the equation

### 8. Write the $t$ and the $p$ -value

The  $T$  and the  $p$ -value are calculated in step 5

### 9. State the Decision

- The  $p$ -value is \_\_\_\_\_
- compare to alpha:  $p$ -value ( $<$  or  $>$ ) **alpha**
- If the  $p$ -value is **less** than alpha, **Reject the Null**
- If the  $p$ -value is **greater** than alpha, **Fail to reject the Null**

### 10. Write the Conclusion

**Reject the Null:** Our  $p$ -value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at  $\alpha =$  \_\_\_\_\_ to suggest that the difference in the true population mean for

\_\_\_\_\_ is \_\_\_\_\_  
*Restate the definition of the 1<sup>st</sup> mean      Restate  $H_A \neq$  or  $<$  or  $>$  2<sup>nd</sup> mean definition*

**Fail to Reject the Null:** Our  $p$ -value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at  $\alpha =$  \_\_\_\_\_ to suggest that the difference in true population mean for

\_\_\_\_\_ is \_\_\_\_\_  
*Restate the definition of the 1<sup>st</sup> mean      Restate  $H_A \neq$  or  $<$  or  $>$  2<sup>nd</sup> mean definition*

## Notes: 2-Sample T Hypothesis Test (Scenario)

**Scenario 2:** In June 2002, the Journal of Applied Psychology reported to a study that examined whether the content of TV shows influenced the ability of viewers to recall brand names of items featured in the commercials. The researchers randomly assigned volunteers to watch one of three programs, each containing the same commercials. One of the programs had violent content, another sexual content, and the third neutral content. After the shows ended, the subjects were asked to recall the brands of products that were advertised. Results are summarized below:

	Program Type		
	Violent	Sexual	Neutral
<b>Sample Size</b>	108	108	108
<b>Mean number of brands recalled</b>	2.08	1.71	3.17
<b>Standard Deviation</b>	1.87	1.76	1.77

Is there a difference in ad memory between programs with **neutral** content and those with violent content? Test an appropriate hypothesis and state your conclusion. What decision should be made?

**Null**  $H_0: \mu_1 = \mu_2$

**Alternative**  $H_A: \mu_1 \neq \mu_2$

Let  $\mu_1$  equal the true population mean number of brands recalled while watching neutral content

Let  $\mu_2$  equal the true population mean number of brands recalled while watching violent content

**Conditions:**

1. **Random Assignment**-Given for Both
2.  $108/.10 = 1080$  It is reasonable to assume that there are at least 1080 people watching shows with neutral content and 1080 watching shows with violent content
3. **Independent Populations:** Reasonable to assume that the number of ads recalled by one group is not affected by the other group
4.  $n = 108$  which is greater than 30 for both samples: The distributions of means is normal by central limit theorem for both

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad df = 213.36 \quad 2 \times P(t > \frac{3.17 - 2.08}{\sqrt{\frac{(1.87)^2}{108} + \frac{(1.77)^2}{108}}})$$

$$= 2 \times P(t > 4.399) \approx .0$$

P-value alpha

.0 Less than .05

We reject the  $H_0$

Our p-value is 0. We reject the null. There is sufficient evidence at  $\alpha = .05$  to suggest that the differences in the true population mean of brands recalled while watching neutral content is greater than while watching violent content.

On average, more brands are recalled while watching neutral content rather than violent

## Recipe for Success: 2 Sample T-Confidence Intervals

1. Define  $\mu_1$  &  $\mu_2$  in context
2. Write the Conditions

1. Both samples are random
2.  $n < 10\%$  of the population
3. Populations are independent
4. Normal Population or  $n > 30$

*If data is given, draw a boxplot or histogram-to show normality*

3. Write the formula for the Test

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$\bar{x}_1$  &  $\bar{x}_2$  = the means of the samples

$s_1$  &  $s_2$  = the standard deviation of the sample

$n_1$  &  $n_2$  = the size of the sample

4. Graph and Shade
5. Enter the Data if Given

- Stat Edit
- Enter Data in columns  $L_1$  &  $L_2$
- 2<sup>nd</sup> Quit
- Stat Calc
- 1-Var Stats  $L_1$ :  $\bar{x}_1$ ,  $s_1$ ,  $n_1$  and  $L_2$ :  $\bar{x}_2$ ,  $s_2$ ,  $n_2$

6. List & Label all of input values

$\bar{x}_1$ ,  $s_1$ ,  $n_1$ ,  $\bar{x}_2$ ,  $s_2$ ,  $n_2$ ,  $df$

$df$  (comes from the calculator)

- Stat Tests
- 0:2-Samp T Int
- Highlight **Data** if data is used otherwise highlight **STATS**
- $s_1$  &  $s_2$  comes from the problem or the data
- $\bar{x}_1$  &  $\bar{x}_2$  comes from the problem or the data
- $n_1$  &  $n_2$  comes from the problem or the data
- **pooled** highlight no

7. Calculate  $t^*$

- 2<sup>nd</sup> Vars
- Inverse t
- Area =  $\frac{(1-\text{Confidence level})}{2}$
- $df$  (comes from the calculator in the step above)

8. Plug in the values

9. Write the interval

10. Write the Conclusion

We are \_\_\_\_\_% confident that the true population mean difference for \_\_\_\_\_

*Restate the definition of the  $\mu_1$*

and \_\_\_\_\_ lies within the interval \_\_\_\_\_

*Restate the definition of the  $\mu_2$*

11. Explain the meaning of the confidence level-if asked

In repeated sampling we expect this method to capture the true population mean difference

for \_\_\_\_\_ and \_\_\_\_\_% of the time.

*Restate the definition of the  $\mu_1$       Restate the definition of the  $\mu_2$*

## Notes: 2-Sample T Confidence Interval (Scenario)

**Scenario 2:** In June 2002, the Journal of Applied Psychology reported to a study that examined whether the content of TV shows influenced the ability of viewers to recall brand names of items featured in the commercials. The researchers randomly assigned volunteers to watch one of three programs, each containing the same commercials. One of the programs had violent content, another sexual content, and the third neutral content. After the shows ended, the subjects were asked to recall the brands of products that were advertised. Results are summarized below:

	Program Type		
	Violent	Sexual	Neutral
<b>Sample Size</b>	108	108	108
<b>Mean number of brands recalled</b>	2.08	1.71	3.17
<b>Standard Deviation</b>	1.87	1.76	1.77

The company can afford to run ads during one TV show, and has decided not to sponsor a show with sexual content. Create a 95% confidence interval for the difference in the mean number of brand names remembered between groups watching neutral shows and those watching violent shows. What is your recommendation? Justify

**See Page 165 for  
Conditions**

**Null**             $H_0: \mu_1 = \mu_2$   
**Alternative**    $H_A: \mu_1 \neq \mu_2$

Let  $\mu_1$  equal the true population mean number of brands recalled while watching neutral content

Let  $\mu_2$  equal the true population mean number of brands recalled while watching violent content

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad df = 213.36 \rightarrow 3.17 - 2.08 \pm 1.97 \sqrt{\frac{(1.77)^2}{108} + \frac{(2.08)^2}{108}}$$

**Interval: (.6016, 1.572)**

We are 95% confident that the true population difference for the mean number of brands recalled between shows with neutral content and those with violent content lies with the interval (.6016, 1.572)

In repeated sampling, we expect that this method will capture the true population difference for the mean number of brands recalled between shows with neutral content and those with violent content 95 percent of the time.

**Note:** 0 does not lie within the interval. Therefore we are able to determine that more brands are recalled while watching shows with neutral content than with violent content.

**Recommendation:** Based on our samples, we expect that on average viewers will remember between .60 and 1.57 brands more while watching shows with neutral content than with violent content. Advertisers should place their ads during programs with neutral content as the population is bored and watching the ads and not the shows,

## Recipe for Success: 2-Sample T Hypothesis Test (difference of Means)

### 1. Write the Hypothesis

- Null  $H_0: \mu_1 = \mu_2$
- Alternative  $H_A: \mu_1 \neq \mu_2$  or  $\mu_1 < \mu_2$  or  $\mu_1 > \mu_2$

### 2. Define $\mu_1$ & $\mu_2$ in context

### 3. Write the Conditions

1. Both samples are random
2.  $n < 10\%$  of the population
3. Populations are independent
4. Normal Population or  $n > 30$

*If data is given, draw a boxplot or histogram-to show normality*

### 4. Write the Equation

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$t$  = the number of standard deviations from the mean

$\mu_1$  &  $\mu_2$  = the means of the population (may be assumed)

$\bar{x}_1$  &  $\bar{x}_2$  = the means of the samples

$s_1$  &  $s_2$  = the standard deviation of the sample

$n_1$  &  $n_2$  = the size of the sample

### 5. Enter Data (if given)

- Stat Edit
- Enter Data in columns  $L_1$  &  $L_2$
- 2<sup>nd</sup> Quit
- Stat Calc
- 1-Var Stats  $L_1: \bar{x}_1, s_1, n_1$  and  $L_2: \bar{x}_2, s_2, n_2$

### 6. List & Label all of input values

$\bar{x}_1, s_1, n_1, \bar{x}_2, s_2, n_2$

df (comes from the calculator)

- Stat Tests
- **4:** 2-SampTTest Enter
- Highlight **Data** if data is used otherwise highlight **STATS**
- **$s_1$  &  $s_2$**  comes from the problem or the data
- **$\bar{x}_1$  &  $\bar{x}_2$**  comes from the problem or the data
- **$n_1$  &  $n_2$**  comes from the problem or the data
- **pooled** highlight no
- **Choose  $\neq$  or  $<$  or  $>$**  (look for key words)

### 7. Plug values into the equation

### 8. Write the $t$ and the $p$ -value

The  $T$  and the  $p$ -value are calculated in step 5

### 9. State the Decision

- The  $p$ -value is \_\_\_\_\_
- compare to alpha:  $p$ -value ( $<$  or  $>$ ) **alpha**
- If the  $p$ -value is **less** than alpha, **Reject the Null**
- If the  $p$ -value is **greater** than alpha, **Fail to reject the Null**

### 10. Write the Conclusion

**Reject the Null:** Our  $p$ -value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at  $\alpha =$  \_\_\_\_\_ to suggest that the difference in the true population mean for

\_\_\_\_\_ is \_\_\_\_\_  
 Restate the definition of the 1<sup>st</sup> mean      Restate  $H_A \neq$  or  $<$  or  $>$  2<sup>nd</sup> mean definition

**Fail to Reject the Null:** Our  $p$ -value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at  $\alpha =$  \_\_\_\_\_ to suggest that the difference in true population mean for

\_\_\_\_\_ is \_\_\_\_\_  
 Restate the definition of the 1<sup>st</sup> mean      Restate  $H_A \neq$  or  $<$  or  $>$  2<sup>nd</sup> mean definition



## Notes: 2-Sample T Hypothesis Test (Scenario)

**Scenario 3:** A statistics teacher wanted to know if males could navigate through a maze faster than females. A random sample of male and female Reagan students was selected and the times it took them to complete the maze were recorded below. Determine whether or not there exist a difference between completion times for men and women.

Null  $H_0: \mu_1 = \mu_2$

Alternative  $H_A: \mu_1 \neq \mu_2$

Let  $\mu_1$  equal the true population mean time for females to complete the maze

Let  $\mu_2$  equal the true population mean time for males to complete the maze

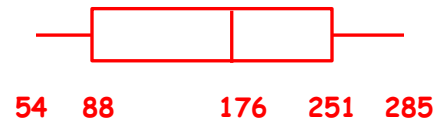
### Conditions:

1. Random Samples-Given for both
2.  $31/.10 = 310$  There are more than 310 females at Reagan
3. Samples are independent-Reasonable to assume that female maze times are not affected by males
4.  $n = 31 > 30$  The distribution of means is normal by Central Limit Theorem for female times

$19/.10 = 190$  there are more than 190 males at Reagan

The boxplot of male times is reasonably symmetric with no outliers. The distribution appears to be approximately normal

Maze Completion Times				
Female Times			Male Times	
120	335	185	285	285
88	172	131	251	76
147	217	277	220	54
160	178	140	94	285
223	264	177	75	77
80	230	121	180	257
79	108	364	204	112
256	205	94	188	119
97	180	125	88	103
85	122		176	
337	136			



$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad df = 37.29 \quad 2 \times P(t > \frac{175.26 - 164.68}{\sqrt{\frac{(79.28)^2}{31} + \frac{(81.75)^2}{19}}})$$

$$\bar{x}_1 = 175.26 \quad \bar{x}_2 = 164.68 \quad = 2 \times P(t > .449) = .656$$

$$s_1 = 79.28 \quad s_2 = 81.75 \quad P\text{-value} \quad \alpha$$

$$n_1 = 31 \quad n_2 = 19 \quad .656 \quad \text{Greater than } .05$$

We fail to reject the  $H_0$

Our p-value is .656. We fail to reject the null. There is Not sufficient evidence at alpha = .05 to suggest that the differences in the true population mean maze for females and males are different.

On average, the female and male maze times are not statistically different.

## Recipe for Success: 2 Sample T-Confidence Intervals

1. Define  $\mu_1$  &  $\mu_2$  in context
2. Write the Conditions

1. Both samples are random
2.  $n < 10\%$  of the population
3. Populations are independent
4. Normal Population or  $n > 30$

*If data is given, draw a boxplot or histogram-to show normality*

3. Write the formula for the Test

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$\bar{x}_1$  &  $\bar{x}_2$  = the means of the samples

$s_1$  &  $s_2$  = the standard deviation of the sample

$n_1$  &  $n_2$  = the size of the sample

4. Graph and Shade
5. Enter the Data if Given

- Stat Edit
- Enter Data in columns  $L_1$  &  $L_2$
- 2<sup>nd</sup> Quit
- Stat Calc
- 1-Var Stats  $L_1$ :  $\bar{x}_1, s_1, n_1$  and  $L_2$ :  $\bar{x}_2, s_2, n_2$

6. List & Label all of input values

$\bar{x}_1, s_1, n_1, \bar{x}_2, s_2, n_2, df$

$df$  (comes from the calculator)

- Stat Tests
- 0:2-Samp T Int
- Highlight **Data** if data is used otherwise highlight **STATS**
- $s_1$  &  $s_2$  comes from the problem or the data
- $\bar{x}_1$  &  $\bar{x}_2$  comes from the problem or the data
- $n_1$  &  $n_2$  comes from the problem or the data
- **pooled** highlight no

7. Calculate  $t^*$

- 2<sup>nd</sup> Vars
- Inverse t
- Area =  $\frac{(1-\text{Confidence level})}{2}$
- $df$  (comes from the calculator in the step above)

8. Plug in the values

9. Write the interval

10. Write the Conclusion

We are \_\_\_\_\_% confident that the true population mean difference for \_\_\_\_\_

*Restate the definition of the  $\mu_1$*

and \_\_\_\_\_ lies within the interval \_\_\_\_\_

*Restate the definition of the  $\mu_2$*

11. Explain the meaning of the confidence level-if asked

In repeated sampling we expect this method to capture the true population mean difference

for \_\_\_\_\_ and \_\_\_\_\_% of the time.

*Restate the definition of the  $\mu_1$*       *Restate the definition of the  $\mu_2$*

**Notes: 2-Sample T Confidence Interval (Scenario)**

**Scenario 3:** A statistics teacher wanted to know if males could navigate through a maze faster than females. A random sample of male and female Reagan students was selected and the times it took them to complete the maze were recorded below.

Create a 90% confidence interval for the scenario and explain your findings in context.

**See Page 169 for  
Conditions**

Null  $H_0: \mu_1 = \mu_2$

Alternative  $H_A: \mu_1 \neq \mu_2$

Let  $\mu_1$  equal the true population mean time for females to complete the maze

Let  $\mu_2$  equal the true population mean time for males to complete the maze

$$\bar{x}_1 = 175.26 \quad \bar{x}_2 = 164.68$$

$$s_1 = 79.28 \quad s_2 = 81.75$$

$$n_1 = 31 \quad n_2 = 19$$

Maze Completion Times				
Female Times			Male Times	
120	335	185	285	285
88	172	131	251	76
147	217	277	220	54
160	178	140	94	285
223	264	177	75	77
80	230	121	180	257
79	108	364	204	112
256	205	94	188	119
97	180	125	88	103
85	122		176	
337	136			

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad df = 37.29 \rightarrow 175.26 - 164.68 \pm 1.687 \sqrt{\frac{(79.28)^2}{31} + \frac{(81.75)^2}{19}}$$

Interval: (-.29.15, 50.29)

We are 90% confident that the true population difference of mean maze times between females and males lies with the interval (-.29.15, 50.29)

In repeated sampling, we expect that this method will capture the true population difference of mean maze times between females and males 90 percent of the time.

Note: 0 does lie within the interval. Therefore, there is not a statistically significant difference between female and male maze times.

Based on our samples, it appears that both female and male maze times are the same.

## Notes: Recalling the Binomial Distribution

**Bernoulli Trials:** a random experiment with exactly two possible outcomes, "success" and "failure", in which the probability of success is the same every time the experiment is conducted.

4. two possible outcomes (success "p" and failure "q").
5. the probability of success, "p", is constant.
6. the trials are independent

**Note:** When we don't have an infinite population, the trials are not independent. However it is still okay to proceed as long as the sample is smaller than 10% of the population. **M&M example** ( $n < 10\%$  of the population)

**The Binomial Model:** Gives us the probability for a random variable that counts the number of successes in a fixed number of Bernoulli trials.

### Characteristics:

- Has 2 possible outcomes
- Each trial is independent
- the probability of success, "p", is constant
- The trials are independent (subject to the note above)
- The number of trials is predetermined

**Equation:** 
$$\sum_{x=0}^n \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

- **n** is the number of trials
- **x** is the number of success
- **p** is the probability or **proportion** of successes for the population
- **q (1-p)** the probability or **proportion** of failures for the population

**Mean ( $\mu$ ) or  $E(X)=np$**

**Variance or  $\text{Var}(X) \sigma^2 = npq$**

**Standard deviation  $\sigma = \sqrt{npq}$**

### Binomials are usually phrased as:

- What is the probability of some number of successes in a given number of trials? (Binomial PDF)
- What is the probability of no more than some number of successes in a given number of trials? (Binomial CDF)
- What is the probability of at least some number of successes in a given number of trials? (Binomial CDF) usually 1-Binomial CDF

**Scenario: Suppose that the probability of having a defective light bulb is .15**

- Given a box of 30 lightbulbs what is the probability that 2 are defective? **Trials: 30**

$$P(X = 2) \rightarrow \binom{30}{2} (.15)^2 (.7)^{28} = .1003$$

**2<sup>nd</sup> Vars Binom PDF**

**P: .15**

**Xvalue: 2**

- Given a box of 30 lightbulbs, what is the probability that no more than 2 are defective? **Trials: 30**

**P: .15**

$$P(X \leq 2) \rightarrow \binom{30}{0} (.15)^0 (.85)^{30} + \dots + \binom{30}{2} (.15)^2 (.85)^{28} = .1514 \rightarrow \text{2<sup>nd</sup> Vars Binom CDF}$$

**Xvalue: 2**

- Given a box of 30 lightbulbs, what is the probability that at least 28 **are not** defective?

$$P(X \geq 28) \rightarrow \binom{30}{28} (.15)^2 (.85)^{28} + \dots + \binom{30}{30} (.85)^{30} (.15)^0 = .8486$$

**Trials: 30**

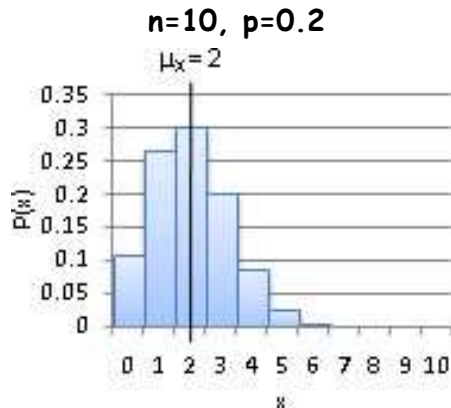
**P: .15**

**1 - 2<sup>nd</sup> Vars Binom CDF**

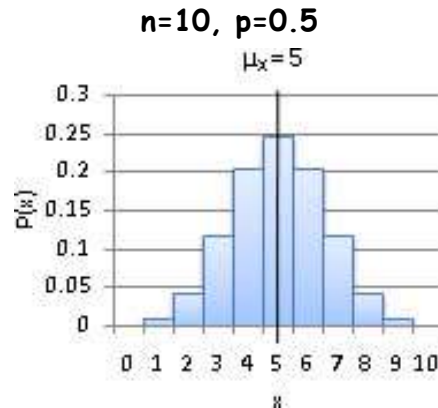
**Xvalue: 27**

## Notes: The Normal Approximation of the Binomial

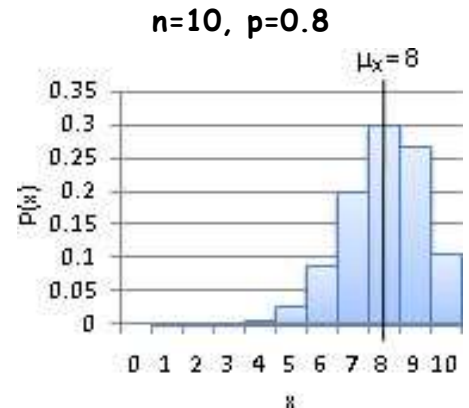
The best way to understand the effect of  $n$  and  $p$  on the shape of a binomial probability distribution is to look at some histograms, so let's look at some possibilities.



As  $P$  approaches 0 the binomial becomes progressively more skewed to the right.

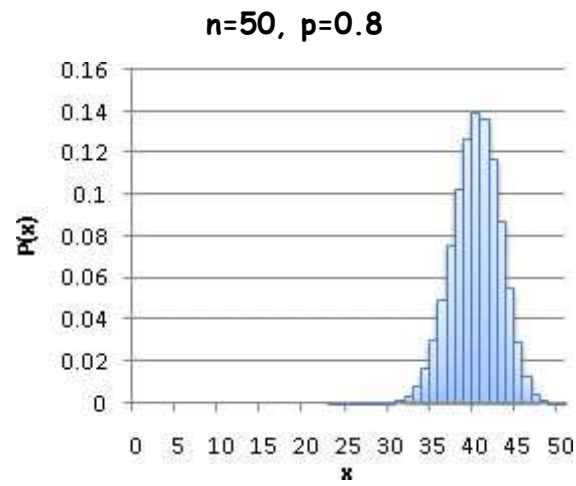
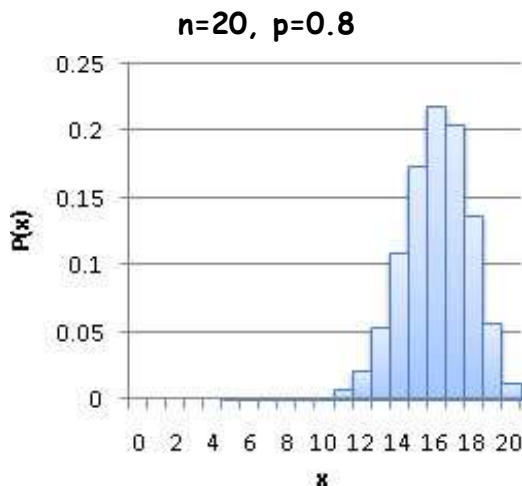


As  $P$  approaches 0.5 the binomial becomes more normal



As  $P$  approaches 1 the binomial becomes progressively more skewed to the left.

Based on the above, it would appear that the distribution is symmetric only if  $p=0.5$ , but this isn't actually true. Watch what happens as the number of trials,  $n$ , increases:



Because of this property which is an essentially the application of the central limit theorem, we can use the normal model to approximate the binomial if the sample size is large enough.

**Why would we do such a thing?** The binomial becomes difficult to work with and can exceed the limits of our calculator if we have a substantial number trials.

**Requirements: (sample size must be large enough)**

1.  $np \geq 10$  number of successes is greater than 10
2.  $nq \geq 10$  number of failures is greater than 10

**Scenario:** An archer with an 80% bull's eye rate will be shooting 2000 arrows.

- a) What are the mean and Standard deviation of the number of bull's eyes she might get?

$$E(X)=np \rightarrow E(X) = 2000(.8) = 1,600 \quad \sigma = \sqrt{npq} \rightarrow \sigma = \sqrt{2000(.8)(.2)} = 17.89$$

- b) Is a Normal Model appropriate here? Justify numerically

$$np \geq 10 \rightarrow 2000(.80) = 1,600 \text{ Satisfied} \quad \text{Both } np \text{ \& } nq \text{ are greater than 10 the sample is}$$

$$nq \geq 10 \rightarrow 2000(.20) = 400 \text{ Satisfied} \quad \text{large enough to utilize the normal distribution}$$

## Notes: Distribution of Sample Proportions

**Central Limit Theorem for Means**-States that the sampling distribution of the expected values (means) of a population with a mean of  $\mu$  and a standard deviation of  $\sigma$  will be approximately normal for any population regardless of the shape of the underlying population if the **sample is large enough**.

- **Distribution of Sample Means:** Follows a normal distribution where  $E(\bar{x})$  or  $\mu = \bar{x}$  and  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$  or  $s_{\bar{x}} = \frac{s}{\sqrt{n}}$  when  $\sigma$  is unknown for any shaped distribution if the sample is large enough  $n \geq 30$

**Central Limit Theorem for Proportions**-States that the sampling distribution of the expected proportion "P" for a population with a standard deviation of  $\sigma$  will be approximately normal for any population regardless of the shape of the underlying population if the **sample is large enough**.

**Conditions:**

1. **Random Sample**
2.  **$n < 10\%$  of the population** (*degree of dependence is minimal, relatively independent-remember we are sampling without replacement*)
3. **Sample is Large Enough** (necessary to claim normality by Central Limit Theorem)
  - $n\hat{p} > 10$  where  $n$  is the sample size and  $\hat{p}$  is the sample proportion
  - $n\hat{q} > 10$  where  $n$  is the sample size and  $\hat{q}$  is  $(1 - \hat{p})$

**Distribution of Sample proportions:** Follows a normal distribution with  $z = \frac{\hat{p} - p}{\sqrt{\frac{(p)(q)}{n}}}$

- **The Center:**
  - $E(x)$  or "P" is the population proportion
  - $\hat{P} = \frac{x}{n}$  "P" is the population proportion and  $\hat{P}$  the sample proportion
    - $x$  is the number of observed successes (outcomes of interest)
    - $n$  is the sample size
- **The spread:**
  - $\sigma_p = \sqrt{\frac{(p)(q)}{n}}$  (standard error of the population proportion)-use for hypothesis test because we assume the claim of the null is true.
  - $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$  (standard error of the sample proportion)-use for confidence intervals because we are looking for an estimate of the population proportion

**Scenario:** Assume that 27% of the students at Reagan wear contacts or other corrective eyewear and that we randomly select a sample of 110 students.

- A. What is the distribution of  $\hat{P}$ ? Specify the name of the distribution the mean and standard deviation and verify that all conditions are met. **The distribution is binomial, however we can use the normal approximation because  $n\hat{p} = 110(.27) = 29.7 > 10$  &  $n\hat{q} = 110(.73) = 29.7 > 10$**

**with  $E(X) = .27$  and  $\sigma_p = \sqrt{\frac{(.27)(.73)}{110}}$   $\sigma_p = .0423$**

- B. What's the probability that more than 30% of the sample use corrective eyewear?

**$P(\hat{p} > .30) \rightarrow P(z > \frac{.30 - .27}{\sqrt{\frac{(.27)(.73)}{110}}}) \rightarrow P(z > .7087) = .2392$**

## Notes: Hypothesis Tests for 1-Sample Proportions

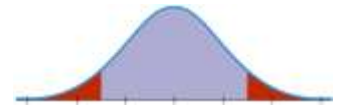
**Hypothesis Tests:** Remember the purpose of a hypothesis test is to determine whether the data support a claim or not. We always begin with the assumption that the null is true. Unless we have strong evidence against the null we retain or accept the null but we do not prove it. If we have strong evidence against the null we reject it in favor of the alternative. We do not prove the alternative.

- **We gather evidence against a claim**
- **We must judge plausibility of the evidence or data**
  - How likely are we to get data like this if the null really were true
  - We never have 100% certainty, but we can quantify our level belief in the evidence
  - The p-value tells us how surprised we are to see our results
- **We never prove a claim:** we are forced to accept the claim because we failed to reject it
- **We never prove the alternative:** we accept the alternative because we rejected the null and the alternative is our only remaining choice

### The Hypothesis:

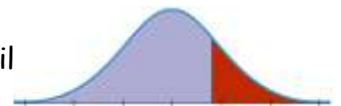
- The null-  $H_0: P = P_0$  ( $P$  is the true population proportion &  $P_0$  is the claimed population proportion)
- The alternative:
  - $H_A: P \neq P_0$  the true population is different than the claimed (2-tailed)

**Note:** The alpha level is divided equally into both tails  
If our level of significance was 10%, 5% would be in each tail



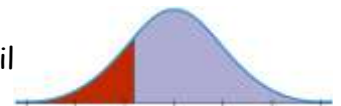
- $H_A: P > P_0$  the true population proportion is greater than the claimed (1-tailed, upper)

**Note:** The alpha level is in the upper tail  
If our level of significance was 10%, all of it would be in the Upper tail



- $H_A: P < P_0$  the true population proportion is less than the claimed (1-tailed, lower)

**Note:** The alpha level is in the upper tail  
If our level of significance was 10%, all of it would be in the Upper tail



**P-Value:** the probability we would find results as extreme or more extreme given that the null is true.  
How unlikely is too unlikely?

- **We decide based on the situation.**
- **Our threshold (alpha) should be determined prior to running our test.**

The formula: 
$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{(p_0)(q_0)}{n}}}$$

**Notice that this is a z-test.** The reason this is a z-test is because  $P_0$  is assumed to be the true population proportion and will continue to be so unless we find sufficient evidence to reject that claim.

## Recipe for Success: 1 Sample Proportion Hypothesis Test

### 1. Write the Hypothesis

- Null  $H_0: p_0 =$
- Alternative  $H_A: p_0 \neq$  or  $<$  or  $>$

### 2. Define parameter $p_0$ in context

### 3. Write the Conditions

- Simple Random Sample
- $n\hat{p} \geq 10$
- $n\hat{q} \geq 10$
- $n$  is less than 10% of the population  $\frac{n}{.1}$

### 4. Write the Equation

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{(p_0)(q_0)}{n}}}$$

$z$  = the number of standard deviations a value is from the center

$p_0$  = the population **proportion** or what is assumed to be true

$q_0 = 1 - p_0$  ( $q_0$  = the expected proportion of failures)

$n$  = the size of the sample

$x$  = the number of successes

$\hat{p} = \frac{x}{n}$  sample proportion of successes-those that met criteria

$\hat{q} = 1 - \hat{p}$  the sample proportion of failures

### 5. Draw the graph and Shade $H_A$

### 6. List & Label all of input values

- $p_0$  should be given
- $q_0 = 1 - p_0$
- $x$  = the number of successes from the sample
- $n$  = the sample size
- $\hat{p} = \frac{x}{n}$
- $\hat{q} = 1 - \hat{p}$

### 7. Plug values into the equation

### 8. Calculate the $z$ and the $p$ -value

- Stat Tests
- 1-proportion  $z$ -test
- $p_0$  comes from the problem
- $x$  comes from the problem or the data
- $n$  comes from the problem or the data
- Choose  $\neq$  or  $<$  or  $>$  (using Shaded graph of  $H_A$ )

### 9. State the Decision

- The  $p$ -value is \_\_\_\_\_
- If the  $p$ -value is less than alpha, Reject the Null
- If the  $p$ -value is greater than alpha, Fail to reject the Null

### 10. Write the Conclusion

Reject the Null: Our  $p$ -value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population proportion for \_\_\_\_\_

is \_\_\_\_\_ *Restate the definition of the  $p_0$*   
*Restate  $H_A \neq$  or  $<$  or  $>$   $p_0$*

Fail to Reject the Null: Our  $p$ -value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population mean for \_\_\_\_\_

\_\_\_\_\_ is \_\_\_\_\_  
*Restate the definition of the  $p_0$*  *Restate  $H_A \neq$  or  $<$  or  $>$   $p_0$*



## Notes: 1 Sample Proportions Hypothesis Test (Scenario)

**Scenario 1:** A union Spokesperson claims that 75% of union members will support a strike if their basic demands are not met. A company negotiator believes the true percentage is lower and runs a hypothesis test at the 10% significance level. What is the conclusion if 87 out of an SRS of 125 union members say they will strike?

$$H_0: p_0 = 75\%$$

$$H_A: p_0 < 75\%$$

$p_0$  is the true population proportion of union members who will strike if their demands are not met.

**Conditions:**

1. **Random Sample:** Given

2. **N is less than 10% of the population:**  $125/.1 = 1,250$ . It is reasonable to assume that there are at least 1,250 union members

3.  $n\hat{p} \geq 10 = 87$

$n\hat{q} \geq 10 = 38$  The sample is large enough to use the normal approximation

$$\hat{p} = \frac{x}{n} \rightarrow \frac{87}{125} = .696$$

$$\hat{q} = 1 - \hat{p} \rightarrow 1 - .696 = .304 \quad z = \frac{\hat{p} - p_0}{\sqrt{\frac{(p_0)(q_0)}{n}}} \rightarrow z = \frac{.696 - .75}{\sqrt{\frac{(.75)(.25)}{125}}} \rightarrow P(z < -1.39) = .0816$$

P-value

alpha

.0816

Less than

.10

We reject the Null

Our p-value is .0816. We reject the Null. There is sufficient evidence at  $\alpha = .10$  to suggest that the true population proportion of union members who will strike is less than 75%

We believe that less than 75% of the members will strike. We do not believe the union's claim.

Explain the meaning of the p-value in context and the type of Error that may have been committed.

P-value is the probability that we would get a test statistics this extreme given that the null is true.

There is an 8.16% chance that we would get a test statistic this extreme in favor the true proportion of union strikers being less than 75% when in fact the null is true and 75% of the union would strike.

We rejected the null so the only type of error that we could have committed was a Type I Error.

A type I error is rejecting a true null.

In this case committing a Type I Error means that we mistakenly be believed that less than 75% of the union when strike when in fact 75% of the union would strike. The consequence is that we will give bad advice to those negotiating a labor deal

## Notes: Confidence Intervals for 1-Sample Proportions

**Confidence Intervals:** Remember that the purpose of a confidence interval is to estimate a range of plausible values for the parameter of interest and follow the form:

**estimate  $\pm$  margin of error**

- All confidence intervals are 2-tailed

**Note:** alpha--  $\alpha$  --is the combined area in the tails

- As we increase the level of confidence the margin of error (the interval width increases)
- As we increase alpha the margin of error & our confidence decreases (the interval width decreases)
- As we increase the sample size the margin of error decreases (the interval width decreases)
- As we increase the standard deviation the margin of error increases (the interval width increases)

**Note:** Because we only have sample data, we are never able to say what the exact population proportion or mean is. We can claim, with a certain level of confidence, that the population proportion or mean lies within a specified interval.

**Confidence Intervals of proportions:**  $\hat{p} \pm z^* \sqrt{\frac{(\hat{p})(\hat{q})}{n}}$

- **Estimate:** The sample proportion " $\hat{P}$ " is the estimate for the population proportion "P"
- **Margin of Error:** the maximum distance a sample statistic may lie from the true population parameter or vice versa

- The **margin of error** is the (z-score) $\times$ (the standard error) or  $z^* \times \sqrt{\frac{\hat{p}\hat{q}}{n}}$
- **Z-score-** the confidence level-provides the maximum number of standard of deviations that we believe the sample proportion may lie away from the true population proportion

**Note:** The standard error is often referred to as the standard deviation

- **standard error**  $\sqrt{\frac{\hat{p}\hat{q}}{n}}$  the standard deviation of the sample proportion

### Confidence Interval Claims

**We Can Claim:** \_\_\_\_\_% Confident that the true population proportion lies within the interval

**We Can Claim:** \_\_\_\_\_% of the intervals collected by this method will capture the population proportion

**Note:** We are making a claim about the population proportion and not the sample proportion.

We do not need to make a claim about the sample proportion. The sample proportion will be in the center of the interval 100% of the time. The sample proportion is the estimate.

**We Cannot Claim:**

- \_\_\_\_\_% of the values lie within the interval—*the interval may be wrong*
- \_\_\_\_\_% chance that a randomly selected \_\_\_\_\_ will lie within the interval—*the interval may be wrong*
- \_\_\_\_\_% of samples will result in this interval—*the interval may be wrong*
- \_\_\_\_\_% probability that the true population proportion lies within the interval

**(do not use probability; use confidence)**

**Remember:** The purpose of the interval is to provide a range of values for the parameter it is not designed to capture a given percentage of the data and as the sample size increases the amount of the population that lies with the interval decreases.

**Note:** If the claim (the null) lies within the interval, there is not enough evidence to reject the null. The result is not significant.

## Confidence Intervals

**Confidence Interval Hints:** Please remember that although the population does not change every sample is going to be different with a different mean and a different standard deviation. As a consequence every confidence interval will have a different point estimator ( $\bar{x}$  or  $\hat{p}$ ) and different end points. We do not have any confidence in the endpoints, we are confident in the method and believe with a given level of confidence that we captured the true population parameter.

We only make claims about the population. We never make claims about the sample because we know all of the values of the sample. We know the exact center and spread of the sample, but we use those values to make inferences/claims about the population:

1. To determine the average spent on entertainment during a year in college, a simple random sample of 35 students is interviewed, showing a mean of \$825 with a standard deviation of \$240. Which of the following is the best interpretation of a 90 percent confidence interval estimate for the average spent on entertainment during a year in college?
  - (A) 90 percent of college students spend between \$756 and \$894 on entertainment yearly.
  - (B) 90 percent of college students spend a mean dollar amount on entertainment yearly that is between \$756 and \$894.
  - (C) We are 90 percent confident that college students spend between \$756 and \$894 on entertainment yearly.
  - (D) We are 90 percent confident that college students spend a mean dollar amount between \$756 and \$894 on entertainment yearly.
  - (E) We are 90 percent confident that in the chosen sample, the mean dollar amount spent on entertainment yearly by college students is between \$756 and \$894.
2. A planning board in Elm County is interested in estimating the proportion of its residents that are in favor of offering incentives to high-tech industries to build plants in that county. A random sample of Elm County residents was selected. All of the selected residents were asked, "Are you in favor of offering incentives to high-tech industries to build plants in your county?" A 95 percent confidence interval for the proportion of residents in favor of offering incentives was calculated to be  $0.54 \pm 0.05$ . Which of the following statements is correct?
  - (A) At the 95% confidence level, the estimate of 0.54 is within 0.05 of the true proportion of county residents in favor of offering incentives to high-tech industries to build plants in the county.
  - (B) At the 95% confidence level, the majority of area residents are in favor of offering incentives to high-tech industries to build plants in the county.
  - (C) In repeated sampling, 95% of sample proportions will fall in the interval (0.49, 0.59)
  - (D) In repeated sampling, the true proportion of county residents in favor of offering incentives to high-tech industries to build plants in the county will fall in the interval (0.49, 0.59).
  - (E) In repeated sampling, 95% of the time the true proportion of county residents in favor of offering incentives to high-tech industries to build plants in the county will be equal to 0.54.

## Recipe for Success: 1 Sample Proportions (Confidence Interval)

1. Define parameter  $p_0$  (the population proportion) in context
2. Write the Conditions
  - Simple Random Sample
  - $n\hat{p} \geq 10$
  - $n\hat{q} \geq 10$
  - $n$  is less than 10% of the population  $\frac{n}{.1}$

3. Write the Equation

$$\hat{p} \pm z^* \sqrt{\frac{(\hat{p})(\hat{q})}{n}}$$

4. List the Values

$\hat{p} = \frac{x}{n}$  sample proportion of successes-those that met criteria  
 $\hat{q} = 1 - \hat{p}$  the sample proportion of failures  
 $z^*$ = the number of standard deviations a value is from the center  
 $x$  = the number of successes or measured outcomes of interest  
 $n$  = the size of the sample

5. Calculate  $z^*$

- 2<sup>nd</sup> Vars
- Inverse Norm
- Area =  $\frac{(1-\text{Confidence level})}{2}$
- $\mu = 0$  and  $\sigma = 1$

6. Plug in the values

7. Calculate the Interval

- Stat Tests
- 1-PropZInt
- $x$  comes from the problem or the data
- $n$  comes from the problem or the data
- **C-Level** Confidence level comes from the problem

8. Write the interval

9. Write the Conclusion

We are \_\_\_\_\_% confident that the true population proportion for \_\_\_\_\_ lies within the interval \_\_\_\_\_.  
*Restate the definition of the  $p_0$*

10. Explain the meaning of the confidence level-if asked

In repeated sampling, we expect that this method will capture the true population proportion \_\_\_\_\_ percent of the time.  
*Restate the Confidence Level*

**Notes: 1-Sample Proportion Confidence Intervals (Scenarios)**

**Scenario 1:** The local news has taken a SRS of union members and found that 87 out of 125 union members surveyed said that they will strike. Calculate a 95% confidence level and provide a conclusion.

$$H_0: p_0 = 75\%$$

$$H_A: p_0 < 75\%$$

$p_0$  is the true population proportion of union members who will strike if their demands are not met.

**Conditions:**

1. **Random Sample:** Given

2. **N is less than 10% of the population:**  $125/.1 = 1,250$ . It is reasonable to assume that there are at least 1,250 union members

3.  $n\hat{p} = 87 \geq 10$

$n\hat{q} = 38 \geq 10$  The sample is large enough to use the normal approximation

$$\hat{p} = \frac{x}{n} \rightarrow \frac{87}{125} = .696$$

$$\hat{q} = 1 - \hat{p} \rightarrow 1 - .696 = .304$$

**Calculate  $Z^*$**

2<sup>nd</sup> Vars Inverse Norm

$$\text{Area} = \frac{(1-.95)}{2}$$

$$\mu = 0 \text{ and } \sigma = 1$$

$$Z^* = 1.96$$

$$\hat{p} \pm z^* \sqrt{\frac{(\hat{p})(\hat{q})}{n}} \rightarrow .696 \pm 1.96 \sqrt{\frac{(.696)(.304)}{125}}$$

$$(.6153, .7766)$$

We are 95% confident that the true population proportion of union members who will strike lies within the interval (.6153, .7766).

Because 75% lies within the interval, we do not have sufficient evidence to dismiss the claim that 75% of union members will strike.

In repeated sampling, we expect that this method will capture the true population proportion of union members who will strike 95% of the time.

## Recipe for Success: 1 Sample Proportion Hypothesis Test

### 1. Write the Hypothesis

- Null  $H_0: p_0 =$
- Alternative  $H_A: p_0 \neq$  or  $<$  or  $>$

### 2. Define parameter $p_0$ in context

### 3. Write the Conditions

- Simple Random Sample
- $n\hat{p} \geq 10$
- $n\hat{q} \geq 10$
- $n$  is less than 10% of the population  $\frac{n}{.1}$

### 4. Write the Equation

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{(p_0)(q_0)}{n}}}$$

$z$  = the number of standard deviations a value is from the center

$p_0$  = the population **proportion** or what is assumed to be true

$q_0 = 1 - p_0$  ( $q_0$  = the expected proportion of failures)

$n$  = the size of the sample

$x$  = the number of successes

$\hat{p} = \frac{x}{n}$  sample proportion of successes-those that met criteria

$\hat{q} = 1 - \hat{p}$  the sample proportion of failures

### 5. Draw the graph and Shade $H_A$

### 6. List & Label all of input values

- $p_0$  should be given
- $q_0 = 1 - p_0$
- $x$  = the number of successes from the sample
- $n$  = the sample size
- $\hat{p} = \frac{x}{n}$
- $\hat{q} = 1 - \hat{p}$

### 7. Plug values into the equation

### 8. Calculate the $z$ and the $p$ -value

- Stat Tests
- 1-proportion z-test
- $p_0$  comes from the problem
- $x$  comes from the problem or the data
- $n$  comes from the problem or the data
- Choose  $\neq$  or  $<$  or  $>$  (using Shaded graph of  $H_A$ )
- The  $p$ -value is \_\_\_\_\_
- If the  $p$ -value is less than alpha, Reject the Null
- If the  $p$ -value is greater than alpha, Fail to reject the Null

### 9. State the Decision

### 10. Write the Conclusion

Reject the Null: Our  $p$ -value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population proportion for \_\_\_\_\_ is \_\_\_\_\_  
*Restate the definition of the  $p_0$*   
*Restate  $H_A \neq$  or  $<$  or  $> p_0$*

Fail to Reject the Null: Our  $p$ -value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population mean for \_\_\_\_\_ is \_\_\_\_\_  
*Restate the definition of the  $p_0$*  *Restate  $H_A \neq$  or  $<$  or  $> p_0$*

**Notes: 1 Sample Proportions Hypothesis Test (Scenario)**

**Scenario 2:** For a particular shopping mall, 60% of the visitors spend over twenty-five dollars when they visit the mall. The mall was recently modeled and the shops are trying to determine if the percent of visitors spending at least twenty-five dollars has changed. A SRS of 550 people leaving the mall showed that 64% had spent over \$25.00. Provide statistical evidence to answer the vendor's question using a hypothesis test.

$$H_0: p_0 = 60\%$$

$$H_A: p_0 \neq 60\%$$

$p_0$  is the true population proportion of mall shoppers who spend at least 25 dollars.

**Conditions:**

1. **Random Sample:** Given

2. **N is less than 10% of the population:**  $550/.1 = 5,550$ . It is reasonable to assume that there are at least 5,550 shoppers at the mall

3.  $n\hat{p} = 550(.64) = 352 \geq 10$

$n\hat{q} = 550(.36) = 198 \geq 10$  The sample is large enough to use the normal approximation

$$\hat{p} = \frac{x}{n} \rightarrow \frac{x}{550} = .64 \rightarrow x = 352$$

$$\hat{q} = 1 - \hat{p} \rightarrow 1 - .64 = .36 \quad z = \frac{\hat{p} - p_0}{\sqrt{\frac{(p_0)(q_0)}{n}}} \rightarrow z = \frac{.64 - .60}{\sqrt{\frac{(.60)(.40)}{550}}} \rightarrow 2P(z > 1.915) = .0555$$

P-value

alpha

.0555 greater than .05 We fail to reject the Null

Our p-value is .0555. We fail to reject the Null. There is not sufficient evidence at alpha = .05 to suggest that the true population proportion of shoppers who spend more than 25 dollars has changed from the original 60% prior to the remodel

We cannot say that the remodel changed the percentage of shoppers who spend at least \$25.00.

Explain the meaning of the p-value in context and the type of Error that may have been committed.

**P-value:** The probability that we would get a test statistics this extreme given the null is true.

There is an 5.55% chance that we would get a test statistic this extreme favoring that the percentage of mall shoppers had changed when in fact the null is true and the percentage of shoppers spending 25 dollars is still 60%.

We Failed to reject the null. The only type of error that we could have committed is a Type II.

A type II error is failing to reject a false null.

In this context, we would believe that the percentage of shoppers spending 25 dollars had not changed when in fact it had.

As a consequence, the mall owners will not be able to charge the shops higher rents.

## Notes: 1 Sample Proportions Hypothesis Test (Scenario)

### Recipe for Success: 1 Sample Proportions (Confidence Interval)

1. Define parameter  $p_0$  (the population proportion) in context
2. Write the Conditions
  - Simple Random Sample
  - $n\hat{p} \geq 10$
  - $n\hat{q} \geq 10$
  - $n$  is less than 10% of the population  $\frac{n}{.1}$

3. Write the Equation

$$\hat{p} \pm z^* \sqrt{\frac{(\hat{p})(\hat{q})}{n}}$$

4. List the Values

$\hat{p} = \frac{x}{n}$  sample proportion of successes-those that met criteria  
 $\hat{q} = 1 - \hat{p}$  the sample proportion of failures  
 $z^*$ = the number of standard deviations a value is from the center  
 $x$  = the number of successes or measured outcomes of interest  
 $n$  = the size of the sample

5. Calculate  $z^*$

- 2<sup>nd</sup> Vars
- Inverse Norm
- Area =  $\frac{(1-\text{Confidence level})}{2}$
- $\mu = 0$  and  $\sigma = 1$

6. Plug in the values

7. Calculate the Interval

- Stat Tests
- 1-PropZInt
- $x$  comes from the problem or the data
- $n$  comes from the problem or the data
- **C-Level** Confidence level comes from the problem

8. Write the interval

9. Write the Conclusion

We are \_\_\_\_\_% confident that the true population proportion for \_\_\_\_\_ lies within the interval \_\_\_\_\_.  
*Restate the definition of the  $p_0$*

10. Explain the meaning of the confidence level-if asked

In repeated sampling, we expect that this method will capture the true population proportion \_\_\_\_\_ percent of the time.  
*Restate the Confidence Level*



**Notes: 1-Sample Proportion Confidence Intervals (Scenarios)**

**Scenario 2:** A SRS of 550 people leaving the mall showed that 64% had spent over \$25.00. Calculate a 99% confidence level and provide a conclusion.

$$H_0: p_0 = 60\%$$

$$H_A: p_0 \neq 60\%$$

$p_0$  is the true population proportion of mall shoppers who spend at least 25 dollars.

Conditions:

1. Random Sample: Given
2. N is less than 10% of the population:  $550/.1 = 5,550$ . It is reasonable to assume that there are at least 5,550 shoppers at the mall
3.  $n\hat{p} = 550(.64) = 352 \geq 10$

$$n\hat{q} = 550(.36) = 198 \geq 10 \quad \text{The sample is large enough to use the normal approximation}$$

$$\hat{p} = \frac{x}{n} \rightarrow \frac{x}{550} = .64 \rightarrow x = 352$$

Calculate  $Z^*$

2<sup>nd</sup> Vars Inverse Norm

$$\text{Area} = \frac{(1-.99)}{2}$$

$$\mu = 0 \text{ and } \sigma = 1$$

$$Z^* = 2.576$$

$$\hat{p} \pm z^* \sqrt{\frac{(\hat{p})(\hat{q})}{n}} \rightarrow .696 \pm 2.576 \sqrt{\frac{(.696)(.304)}{125}}$$

$$(.587, .693)$$

We are 99% confident that the true population proportion of mall shoppers spending more than 25 dollars lies within the interval (.587, .693).

Because 60% lies within the interval, we do not have sufficient evidence say that the remodel changed the percentage of shoppers spending more than 25 dollars at the mall

In repeated sampling, we expect that this method will capture the true population proportion of mall shoppers spending more than 25 dollars 95% of the time.

**Notes: 1 Sample Proportions Hypothesis Test (Scenario)**

**Scenario 3: 1998 Question 5.** A large university provides housing for 10 percent of its graduate students to live on campus. The university's housing office thinks that the percentage of graduate students looking for housing campus may be more than 10 percent. The housing office decides to survey a random sample of graduate students, and 62 of the 481 respondents say that they are looking for housing on campus.

- (a) On the basis of the survey data, would you recommend that the housing office consider increasing the amount of housing on campus available to graduate students? Give appropriate evidence to support your recommendation.
- (b) In addition to the 481 graduate students who responded to the survey, there were 19 who did not respond. If these 19 had responded, is it possible that your recommendation would have changed? Explain.

$$H_0: p_0 = 10\%$$

$$H_A: p_0 > 10\%$$

$p_0$  is the true population proportion of graduate students looking for on campus housing

Conditions:

1. Random Sample: Given

2. N is less than 10% of the population:  $481/.1 = 4,810$ . It is reasonable to assume that there are at least 4,810 graduate students

3.  $n\hat{p} = 62 \geq 10$

$n\hat{q} = 419 \geq 10$  The sample is large enough to use the normal approximation

$$\hat{p} = \frac{x}{n} \rightarrow \frac{62}{481} = .1289$$

$$\hat{q} = 1 - \hat{p} \rightarrow 1 - .1289 = .8711 \quad z = \frac{\hat{p} - p_0}{\sqrt{\frac{(p_0)(q_0)}{n}}} \rightarrow z = \frac{.1289 - .10}{\sqrt{\frac{(.10)(.90)}{481}}} \rightarrow P(z > 2.11) = .0173$$

P-value

.0173

alpha

.05

Less than

We reject the Null

Our p-value is .0173. We reject the Null. There is sufficient evidence at  $\alpha = .05$  to suggest that the true population proportion of graduate students who would like on campus housing is greater than 10%.

(b) Since we have already rejected, we need to find out if we would have failed to reject if all 19 had not been interested in on-campus housing.

$$\hat{p} = \frac{x}{n} \rightarrow \frac{62}{500} = .124 \quad \hat{q} = 1 - \hat{p} \rightarrow 1 - .124 = .876 \quad z = \frac{\hat{p} - p_0}{\sqrt{\frac{(p_0)(q_0)}{n}}} \rightarrow z = \frac{.124 - .10}{\sqrt{\frac{(.10)(.90)}{500}}} \rightarrow$$

$$P(z > 1.789) = .0356$$

P-value

.0356

alpha

.05

Less than

We reject the Null

The decision would not have changed

## Sample Size Calculations

Given the confidence level and the standard deviation of the population, the sample size that will produce a predetermined margin of error **ME** of the confidence interval estimate of **P** the true population proportion is given by  $n = \frac{(z^*)^2 \hat{p}\hat{q}}{ME^2}$

**Note:** You can memorize this or you can use the confidence interval equation & solve using basic algebra.

1. An experiment finds that 27% of 53 subjects report improvement after using a new medicine. What sample size is necessary to create a margin of error of plus or minus 3% with a 98% confidence level?

$$\hat{p} \pm z^* \sqrt{\frac{(\hat{p})(\hat{q})}{n}} \quad \text{and Margin of Error} = z^* \sqrt{\frac{(\hat{p})(\hat{q})}{n}}$$

2<sup>nd</sup> Vars Inverse Norm       $p = .27 \quad \& \quad q = 1 - .27 = .73$

Area =  $\frac{(1-.98)}{2}$       M.E. = .03

$\mu = 0$  and  $\sigma = 1$        $Z^* = 2.326$

$$.03 = 2.326 \sqrt{\frac{(.27)(.73)}{n}} \rightarrow \frac{.03}{2.326} = \frac{2.326}{2.326} \sqrt{\frac{.1971}{n}} \rightarrow .012898 = \sqrt{\frac{.1971}{n}}$$

$$.012898^2 = \left( \sqrt{\frac{.1971}{n}} \right)^2 \rightarrow .00016635 = \frac{.1971}{n} \rightarrow .00016635n = .1971 \rightarrow 1185$$

2. What sample size is necessary to estimate the outcome of an election with a margin of error of plus or minus 3% with 95% confidence? **Hint: What information am I missing** and yes, we can do this problem by making an assumption.

$$\hat{p} \pm z^* \sqrt{\frac{(\hat{p})(\hat{q})}{n}} \quad \text{and Margin of Error} = z^* \sqrt{\frac{(\hat{p})(\hat{q})}{n}}$$

2<sup>nd</sup> Vars Inverse Norm       $p = .50 \quad \& \quad q = 1 - .50 = .50$

Area =  $\frac{(1-.95)}{2}$       M.E. = .03

$\mu = 0$  and  $\sigma = 1$        $Z^* = 1.96$

$$.03 = 1.96 \sqrt{\frac{(.50)(.50)}{n}} \rightarrow \frac{.03}{1.96} = \frac{1.96}{1.96} \sqrt{\frac{.25}{n}} \rightarrow .015306 = \sqrt{\frac{.25}{n}}$$

$$.015306^2 = \left( \sqrt{\frac{.25}{n}} \right)^2 \rightarrow .000234277 = \frac{.25}{n} \rightarrow .000234277n = .25 \rightarrow 1068$$

## Sample Size of Confidence Intervals

- For a given large sample size, which of the following gives the smallest margin of error in calculating a confidence interval for a population proportion?
  - 90 percent confidence with  $\hat{p} = .15$
  - 95 percent confidence with  $\hat{p} = .15$
  - 99 percent confidence with  $\hat{p} = .15$
  - 90 percent confidence with  $\hat{p} = .23$
  - 95 percent confidence with  $\hat{p} = .23$
- Two 95 percent confidence interval estimates are obtained: 1<sup>st</sup> (78.5, 84.5) and 2<sup>nd</sup> (80.3, 88.2).
  - If the sample sizes are the same, which has the larger standard deviation?
  - If the sample standard deviations are the same, which has the larger sample size?
  - a. 1<sup>st</sup>    b. 1<sup>st</sup>
  - a. 1<sup>st</sup>    b. 2<sup>nd</sup>
  - a. 2<sup>nd</sup>    b. 1<sup>st</sup>
  - a. 2<sup>nd</sup>    b. 2<sup>nd</sup>
  - More information is needed to answer these questions.
- Which of the following would result in the widest confidence interval?
  - Small sample size and 95 percent confidence
  - Small sample size and 99 percent confidence
  - Large sample size and 95 percent confidence
  - Large sample size and 99 percent confidence
  - This cannot be answered without knowing an appropriate standard deviation
- In a New York Times poll measuring a candidate's popularity, the newspaper claimed that in 19 of 20 cases its poll results should be no more than three percentage points off in either direction. What confidence level are pollsters working with, and what size sample should they have obtained?
  - 3%, 20
  - 6%, 20
  - 6%, 100
  - 95%, 33
  - 95%, 1068
- A t-statistic was used to conduct a test of the null hypothesis  $H_0: \mu = 0$  against the alternative  $H_a: \mu \neq 0$ . The p-value was 0.056. A two-sided confidence interval for  $\mu$  is to be constructed. Of the following, which is the largest level of confidence for which the confidence interval will NOT contain 0?
  - 90% confidence
  - 93% confidence
  - 95% confidence
  - 98% confidence
  - 99% confidence

**Sample Size of Confidence Intervals**

5. The school superintendent wants to know what percentage of property owners are willing to support an increase in school taxes. What size sample should be obtained to determine with 90 percent confidence the support level to within 5 percent?
- (A) 17  
(B) 33  
**(C) 271**  
(D) 289  
(E) 1,083
6. A political action group wishes to learn the government approval rating on the environment. From a past study, they know that they will have to poll 270 people for their desired level of confidence. If they want to keep the same confidence interval and have a margin of error one third the size, how many people will they have to poll?
- (A) 30  
(B) 90  
(C) 468  
(D) 810  
**(E) 2,430**
7. In a random survey of 450 adults, 28 percent said that they felt their credit card debt is too high. With what degree of confidence can the pollster say that  $28 \pm 4$  percent of adults believe that their credit card debt is too high?
- (A) 70.0 percent  
(B) 91.0 percent  
**(C) 94.1 percent**  
(D) 95.0 percent  
(E) 96.0 percent
8. A guidance counselor wishes to determine the mean number of changes in academic major by college students to within  $\pm 0.1$  at a 90 percent confidence level. What sample size should be chosen if it is known that the standard deviation is 0.45?
- (A) 8  
(B) 54  
**(C) 55**  
(D) 78  
(E) 110

**Sample Size of Confidence Intervals**

9. Two confidence interval estimates from the same sample are (72.2, 77.8) and (71.3, 78.7). One estimate is at the 95 percent level, and the other is at the 99 percent level. Which is which?
- (A) (72.2, 77.8) is the 95 percent level.
  - (B) (72.2, 77.8) is the 99 percent level.
  - (C) This question cannot be answered without knowing the sample size.
  - (D) This question cannot be answered without knowing the sample standard deviation
  - (E) This question cannot be answered without knowing both the sample size and standard deviation.
10. In general, how does doubling the sample size change the confidence interval size?
- (A) Doubles the interval size
  - (B) Halves the interval size
  - (C) Multiplies the interval size by 1.414
  - (D) Divides the interval size by 1.414
  - (E) This question cannot be answered without knowing the sample size.
11. A 90 percent confidence interval is to be created to estimate the proportion of television viewers in a certain area who favor moving the broadcast of the late weeknight news to an hour earlier than it is currently. Initially, the confidence interval will be created using a sample of 9,000 viewers in the area. Assuming that the sample proportion does not change, what would be the relationship between the width of the original confidence interval that is created based on a sample of only 1,000 viewers in the area?
- (A) The second confidence interval would be 9 times as wide as the original confidence interval.
  - (B) The second confidence interval would be 3 times as wide as the original confidence interval.
  - (C) The width of the second confidence interval would be equal to the width of the original confidence interval.
  - (D) The second confidence interval would be  $\frac{1}{3}$  times as wide as the original confidence interval.
  - (E) The second confidence interval would be  $\frac{1}{9}$  times as wide as the original confidence interval.

## Notes: 2-Sample Proportion Confidence Intervals

**Purpose:** A 2-sample z-interval for proportions is used to create a plausible range for the difference between the proportions of 2 independent populations.

**Confidence Intervals:** estimate  $\pm$  margin of error

- **Estimate:** The difference between 2 sample proportions.
- **Margin of error:** For a given confidence level, the maximum distance a sample statistic may lie from the true population parameter and equals  $z^* \times$  (Standard error)
  - $z^*$  the maximum number of standard deviations for a given confidence level
  - **Standard error** - the standard deviation of the combined variables
- **Formula:**  $\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$  estimate  $\pm$  margin of error.

**Remember:** variances and standard deviations must be independent in order to combine. The standard deviation of a 2-sample proportion is  $\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$  and the variance is  $\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}$

### Confidence Interval Claims

**We Can Claim:** \_\_\_\_\_% Confident that the true population proportion lies within the interval

**We Can Claim:** \_\_\_\_\_% of the intervals collected by this method will capture the population proportion

**Note:** We are making a claim about the population proportion and not the sample proportion.

We do not need to make a claim about the sample proportion. The sample proportion will be in the center of the interval 100% of the time. The sample proportion is the estimate.

**We Cannot Claim:**

- \_\_\_\_\_% of the values lie within the interval—*the interval may be wrong*
- \_\_\_\_\_% chance that a randomly selected \_\_\_\_\_ will lie within the interval—*the interval may be wrong*
- \_\_\_\_\_% of samples will result in this interval—*the interval may be wrong*
- \_\_\_\_\_% probability that the true population proportion lies within the interval

**(do not use probability; use confidence)**

**Remember:** The purpose of the interval is to provide a range of values for the parameter it is not designed to capture a given percentage of the data and as the sample size increases the amount of the population that lies with the interval decreases.

**Note:** If the claim (the null) lies within the interval, there is not enough evidence to reject the null. The result is not significant.

**Changes in Interval width**

- All confidence intervals are 2-tailed **Note:** alpha--  $\alpha$  --is the combined area in the tails
- As we increase the level of confidence the margin of error (the interval width increases)
- As we increase alpha the margin of error & our confidence decreases (the interval width decreases)
- As we increase the sample size the margin of error decreases (the interval width decreases)
- As we increase the standard deviation the margin of error increases (the interval width increases)

**Note:** Because we only have sample data, we are never able to say what the exact difference in population proportion or means is. We can claim with a certain level of confidence that the population proportion or mean difference lies within a specified interval.

## Recipe for Success: Confidence Interval-2 Sample Proportions

1. Define  $p_1$  &  $p_2$  in context  
(the population proportions)

$p_1$  = the population **proportion** for the 1<sup>st</sup> proportion  
 $p_2$  = the population **proportion** for the 2<sup>nd</sup> proportion

2. Write the Conditions  
(must be for both sets of data)

- Independent Random Samples
- $n\hat{p} \geq 10$
- $n\hat{q} \geq 10$
- $n$  is less than 10% of the population  $\frac{n}{.1}$

3. Write the Equation

$$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

4. List the Values

$z$  = the number of standard deviations a value is from the center  
 $n_1$  = the size of the sample of the 1<sup>st</sup> proportion  
 $x_1$  = the number of outcomes of interest of the 1<sup>st</sup> proportion  
 $\hat{p}_1 = \frac{x_1}{n_1}$  1st sample proportion of interest  
 $n_2$  = the size of the sample of the 2<sup>nd</sup> proportion  
 $x_2$  = the number of outcomes of interest of the 2<sup>nd</sup> proportion  
 $\hat{p}_2 = \frac{x_2}{n_2}$  2nd sample proportion of interest

5. Calculate  $z^*$

- 2<sup>nd</sup> Vars
- Inverse Norm
- Area =  $\frac{(1 - \text{Confidence level})}{2}$
- $\mu = 0$  and  $\sigma = 1$

6. Plug in the values

7. Calculate the Interval

- Stat Tests
- **2-PropZInt**
- $x_1$  comes from the problem or the data
- $n_1$  comes from the problem or the data
- $x_2$  comes from the problem or the data
- $n_2$  comes from the problem or the data
- **C-Level** Confidence level comes from the problem

8. Write the interval

9. Write the Conclusion

We are \_\_\_\_\_% confident that the true population proportion difference between \_\_\_\_\_ and \_\_\_\_\_ lies within the interval \_\_\_\_\_.

*Restate the definition of the  $p_1$*       *Restate the definition of the  $p_2$*

10. Determining significance.

- If 0 is in the interval—There is **No** significant difference
- If 0 is not in the interval—There **Is** a significant difference



## Notes: Confidence Intervals-2 Sample Proportions

**Scenario 1:** A grocery store manager notes in an SRS of 85 people going through the express line that only 10 paid with checks. In an SRS of 92 customers passing through the regular line 37 paid with checks. Construct a 95% confidence interval for the difference.

$$H_0: p_1 = p_2$$

$$H_A: p_1 \neq p_2$$

$p_1$  = the true population proportion of people paying by check in the express line

$p_2$  = the true population proportion of people paying by check in the regular line

Conditions:

1. Random sample for both is Given

Reasonable to assume independence as one line should not impact the other

2.  $n <$  than 10% of the population for both  
 $85/.1 = 850$  reasonable to assume that more than 850 customers use the express line

$92/.1 = 920$  reasonable to assume that more than 920 customers use the regular line

3.  $n_1\hat{p}_1 = 10 \geq 10$  and  $n_2\hat{p}_2 = 37 \geq 10$   
 $n_1\hat{q}_1 = 75 \geq 10$   $n_2\hat{q}_2 = 55 \geq 10$

Both samples are large enough to use the normal approximation

$$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\hat{p}_1 = \frac{x_1}{n_1} \rightarrow \frac{10}{85} = .1176$$

$$\hat{p}_2 = \frac{x_2}{n_2} \rightarrow \frac{37}{92} = .4022$$

$$\hat{q}_1 = 1 - \hat{p}_1 \rightarrow 1 - .1176 = .8824$$

$$\hat{q}_2 = 1 - \hat{p}_2 \rightarrow 1 - .4022 = .5978$$

Calculate  $Z^*$

2<sup>nd</sup> Vars Inverse Norm

$$\text{Area} = \frac{(1-.95)}{2}$$

$$\mu = 0 \text{ and } \sigma = 1$$

$$Z^* = 1.96$$

$$.1176 - .4022 \pm 1.96 \sqrt{\frac{(.1176)(.8824)}{85} + \frac{(.4022)(.5978)}{92}}$$

$$\text{Interval: } (-.4059, -.1632)$$

We are 95% confident that the true population proportion difference between those paying by check in the express line and the regular line lies within the interval  $(-.4059, -.1632)$ .

In repeated sampling, we expect that our method will capture the true population proportion difference between those paying by check in the express line and the regular line 95% of the time.

**Note:** zero does not lie within the interval. The difference in the proportion of people paying by check in the two lines is significant. Because the interval is negative we know that there is a greater proportion of people paying by check in the regular line. Those who pay by check in the express line are inconsiderate jerks.

## Notes: Confidence Intervals-2 Sample Proportions

**Scenario 2:** 84% of an SRS of 125 nurses express job satisfaction working 7am-3pm shifts, while 72% of a SRS of 150 nurses express satisfaction working 11pm-7am shifts. Establish a 90% confidence interval for the difference

$$H_0: p_1 = p_2$$

$$H_A: p_1 \neq p_2$$

$p_1$  = the true population proportion of satisfied nurses in the daytime shift

$p_2$  = the true population proportion of satisfied nurses in the night time shift

Conditions:

1. Random sample for both is Given
2. Reasonable to assume that one shifts job satisfaction does not impact the other
3.  $n <$  than 10% of the population for both  
 $125/.1 = 1250$  reasonable to assume that there are more than 1250 nurses working daytime shifts.       $150/.1 = 1500$  reasonable to assume that more than 1500 nurses working night-time shifts
4.  $n_1\hat{p}_1 = 105 \geq 10$  and  $n_2\hat{p}_2 = 108 \geq 10$       Both samples are large enough to use the normal approximation  
 $n_1\hat{q}_1 = 20 \geq 10$        $n_2\hat{q}_2 = 42 \geq 10$

$$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\hat{p}_1 = \frac{x_1}{n_1} \rightarrow \frac{x_1}{125} = .84 \quad x_1 = 105$$

$$\hat{p}_2 = \frac{x_2}{n_2} \rightarrow \frac{x_2}{150} = .72 \quad x_2 = 108$$

$$\hat{q}_1 = 1 - \hat{p}_1 \rightarrow 1 - .84 = .16$$

$$\hat{q}_2 = 1 - \hat{p}_2 \rightarrow 1 - .72 = .28$$

Calculate  $Z^*$

2<sup>nd</sup> Vars Inverse Norm

$$\text{Area} = \frac{(1-.90)}{2}$$

$$\mu = 0 \text{ and } \sigma = 1 \quad Z^* = 1.645$$

$$.84 - .72 \pm 1.645 \sqrt{\frac{(.84)(.16)}{125} + \frac{(.72)(.28)}{150}}$$

$$\text{Interval: } (.0391, .2009)$$

We are 90% confident that the true population proportion difference between nurses satisfied with their jobs in the day time versus night time shift lies within the interval (.0391, .2009).

In repeated sampling, we expect that our method will capture the true population proportion difference between nurses satisfied with their jobs in the day time versus night time shift 90% of the time.

Note: zero does not lie within the interval. The difference in the proportions of satisfied nurses is significant. Because the interval is positive, we know that there is a greater proportion of daytime shift nurses satisfied with their jobs when compared to night time shifts.

## Notes: 2-Sample Proportion Hypothesis Tests

**Purpose:** A 2-sample hypothesis test for proportions is used to determine whether there is a difference between the proportions of 2 independent populations.

### Pooling

The Null hypothesis is  $H_0: p_1 = p_2$  or  $H_0: p_1 - p_2 = 0$  indicating a claim of no difference

We believe there is no difference between the two proportions for each population. The standard deviation of a 2 sample proportion is  $\sqrt{\frac{\hat{p}_c \hat{q}_c}{n_1} + \frac{\hat{p}_c \hat{q}_c}{n_2}}$  and is determined by the proportion.

Because we believe that both proportions are equal and because standard deviation is solely determined by the proportion it follows that the standard deviations of both populations must be equal.

Consequently we combine or pool the data counts to get one overall proportion using the formula:

$$\hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2} \text{ or } \hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2}$$

**We only pool for 2-sample proportion hypothesis tests.**

We do NOT pool for 2-sample confidence intervals for proportions

We do NOT pool for test of means

We do NOT pool for confidence intervals for means.

**We only pool for 2-sample proportion hypothesis tests.**

**Formula for the test statistic:** 
$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{(\hat{p}_c)(\hat{q}_c)}{n_1} + \frac{(\hat{p}_c)(\hat{q}_c)}{n_2}}}$$

Where:  $\hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2}$

$z$  = the number of standard deviations a value is from the center

$n_1$  = the size of the sample of the 1<sup>st</sup> proportion

$x_1$  = the number of outcomes of interest of the 1<sup>st</sup> proportion

$\hat{p}_1 = \frac{x_1}{n_1}$  1st sample proportion of interest

$n_2$  = the size of the sample of the 2<sup>nd</sup> proportion

$x_2$  = the number of outcomes of interest of the 2<sup>nd</sup> proportion

$\hat{p}_2 = \frac{x_2}{n_2}$  2nd sample proportion of interest

$\hat{p}_c$  = the 2 combined or pooled proportion successes

$\hat{q}_c = 1 - \hat{p}_c$  the 2 combined or pooled proportion successes

**Conditions:** Both intervals and hypothesis tests for 2 sample proportions are based on the following conditions that must be listed and checked:

- Simple Random Sample or Random Assignment
- $n\hat{p} \geq 10$  &  $n\hat{q} \geq 10$  Making certain the sample is large enough to apply the central limit theorem
- $n$  is less than 10% of the population  $\frac{n}{N} < .1$  Making certain the probability does not change substantially.

## Recipe for Success: Hypothesis Test: 2 Sample Proportions

### 1. Write your Hypothesis

- Null  $H_0: p_1 = p_2$
- Alternative  $H_A: p_1 \neq$  or  $<$  or  $>$   $p_2$
- $p_1$  = the population **proportion** for the 1<sup>st</sup> proportion
- $p_2$  = the population **proportion** for the 2<sup>nd</sup> proportion

### 2. Define $p_1$ & $p_2$ in context

### 3. Write the Conditions

(must be for both sets of data)

- **Independent Random Samples**
- **n is less than 10% of the population**  $\frac{n}{.1}$
- $n\hat{p} \geq 10$
- $n\hat{q} \geq 10$

### 4. Write the Equations

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{(\hat{p}_c)(\hat{q}_c)}{n_1} + \frac{(\hat{p}_c)(\hat{q}_c)}{n_2}}}$$

$$\hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2}$$

$z$  = the number of standard deviations a value is from the center

$n_1$  = the size of the sample of the 1<sup>st</sup> proportion

$x_1$  = the number of outcomes of interest of the 1<sup>st</sup> proportion

$\hat{p}_1 = \frac{x_1}{n_1}$  1st sample proportion of interest

$n_2$  = the size of the sample of the 2<sup>nd</sup> proportion

$x_2$  = the number of outcomes of interest of the 2<sup>nd</sup> proportion

$\hat{p}_2 = \frac{x_2}{n_2}$  2nd sample proportion of interest

$\hat{p}_c$  = the 2 combined or pooled proportion successes

$\hat{q}_c = 1 - \hat{p}_c$  the 2 combined or pooled proportion successes

### 5. List & Label all of input values

Calculate  $\hat{p}_1$  &  $\hat{p}_2$  &  $\hat{p}_c$  &  $\hat{q}_c$

### 6. Plug values into the equation

### 7. Calculate the z and the p-value

- Stat Tests
- 2-proportion z-test
- $x_1$  &  $x_2$  comes from the problem or the data
- $n_1$  &  $n_2$  comes from the problem or the data
- Choose  $\neq$  or  $<$  or  $>$

### 8. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 9. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population proportion \_\_\_\_\_ is \_\_\_\_\_

\_\_\_\_\_ true population proportion \_\_\_\_\_  
 Restate  $H_A \neq$  or  $<$  or  $>$  \_\_\_\_\_ Restate the definition of the  $p_2$

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population proportion for

\_\_\_\_\_ is \_\_\_\_\_ than the true population proportion for  
 Restate the definition of the  $p_1$  \_\_\_\_\_ Restate  $H_A \neq$  or  $<$  or  $>$  \_\_\_\_\_  
 \_\_\_\_\_ Restate the definition of the  $p_2$

## Notes: Hypothesis Tests-2 Sample Proportions (Scenario)

**Scenario 1:** Suppose that in an early election campaign a telephone poll of 800 registered voters shows 460 in favor of a particular candidate. Just before the election, a second poll shows only 520 of 1,000 registered voters expressing the same preference. At the 10% significance level is there sufficient evidence to suggest that the candidate's popularity has decreased?

$$H_0: p_1 = p_2$$

$$H_A: p_1 > p_2$$

$p_1$  = the true population proportion of voters supporting a candidate early in the campaign

$p_2$  = the true population proportion of voters supporting a candidate late in the campaign

### Conditions:

1. Both are Random samples: not given Must assume that the two telephone polls represent the voting populations
2. Reasonable to assume that the polls are independent as one poll result should not impact the others
3.  $n <$  than 10% of the population for both  
 $800/.1 = 8000$  reasonable to assume that there are more than 8000 voters  
 $1000/.1 = 10000$  reasonable to assume that more than 10000 voters
4.  $n_1\hat{p}_1 = 460 \geq 10$  and  $n_2\hat{p}_2 = 520 \geq 10$   
 $n_1\hat{q}_1 = 340 \geq 10$  and  $n_2\hat{q}_2 = 480 \geq 10$   
 Both samples are large enough to use the normal approximation

$$\hat{p}_1 = \frac{x_1}{n_1} \rightarrow \frac{460}{800} = .575$$

$$\hat{p}_2 = \frac{x_2}{n_2} \rightarrow \frac{520}{1000} = .52$$

$$\hat{q}_1 = 1 - \hat{p}_1 \rightarrow 1 - .575 = .425$$

$$\hat{q}_2 = 1 - \hat{p}_2 \rightarrow 1 - .52 = .48$$

$$\hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2} \rightarrow \frac{460 + 520}{800 + 1000} \rightarrow \hat{p}_c = .5444$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{(\hat{p}_c)(\hat{q}_c)}{n_1} + \frac{(\hat{p}_c)(\hat{q}_c)}{n_2}}} \rightarrow \frac{.575 - .52}{\sqrt{\frac{(.5444)(.4556)}{800} + \frac{(.5444)(.4556)}{1000}}} \rightarrow P(Z > 2.328) = .00995$$

P-value

alpha

.01

Less than

.10

We reject the Null

Our p-value is .01. We reject the Null. There is sufficient evidence at alpha = .10 to suggest that the true population proportion of voters in favor of the candidate was greater earlier in the campaign than it was at the end of the campaign.

It appears that the candidate is losing support.

## Recipe for Success: Hypothesis Test: 2 Sample Proportions

### 1. Write your Hypothesis

- Null  $H_0: p_1 = p_2$
- Alternative  $H_A: p_1 \neq$  or  $<$  or  $>$   $p_2$

### 2. Define $p_1$ & $p_2$ in context

- $p_1$  = the population **proportion** for the 1<sup>st</sup> proportion
- $p_2$  = the population **proportion** for the 2<sup>nd</sup> proportion

### 3. Write the Conditions

(must be for both sets of data)

- **Independent Random Samples**
- **n is less than 10% of the population**  $\frac{n}{.1}$
- $n\hat{p} \geq 10$
- $n\hat{q} \geq 10$

### 4. Write the Equations

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{(\hat{p}_c)(\hat{q}_c)}{n_1} + \frac{(\hat{p}_c)(\hat{q}_c)}{n_2}}}$$

$$\hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2}$$

$z$  = the number of standard deviations a value is from the center

$n_1$  = the size of the sample of the 1<sup>st</sup> proportion

$x_1$  = the number of outcomes of interest of the 1<sup>st</sup> proportion

$\hat{p}_1 = \frac{x_1}{n_1}$  1st sample proportion of interest

$n_2$  = the size of the sample of the 2<sup>nd</sup> proportion

$x_2$  = the number of outcomes of interest of the 2<sup>nd</sup> proportion

$\hat{p}_2 = \frac{x_2}{n_2}$  2nd sample proportion of interest

$\hat{p}_c$  = the 2 combined or pooled proportion successes

$\hat{q}_c = 1 - \hat{p}_c$  the 2 combined or pooled proportion successes

### 5. List & Label all of input values

Calculate  $\hat{p}_1$  &  $\hat{p}_2$  &  $\hat{p}_c$  &  $\hat{q}_c$

### 6. Plug values into the equation

### 7. Calculate the z and the p-value

- Stat Tests
- 2-proportion z-test
- $x_1$  &  $x_2$  comes from the problem or the data
- $n_1$  &  $n_2$  comes from the problem or the data
- Choose  $\neq$  or  $<$  or  $>$

### 8. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 9. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population proportion \_\_\_\_\_ is \_\_\_\_\_

*Restate the definition of the  $p_1$*

\_\_\_\_\_ true population proportion \_\_\_\_\_.

*Restate  $H_A \neq$  or  $<$  or  $>$*

*Restate the definition of the  $p_2$*

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population proportion for

\_\_\_\_\_ is \_\_\_\_\_ than the true population proportion for \_\_\_\_\_

*Restate the definition of the  $p_1$  Restate  $H_A \neq$  or  $<$  or  $>$*

\_\_\_\_\_ *Restate the definition of the  $p_2$*

## Notes: Hypothesis Tests-2 Sample Proportions (Scenario)

**Scenario 2:** An automobile manufacturer randomly assigns cars to 2 distinct assembly procedures. In a sample of 350 cars coming off the line using the 1<sup>st</sup> procedure there are 28 with major defects, while a sample of 500 autos from the 2<sup>nd</sup> shows 32 defects. Is this significant at the 10% level?

$$H_0: p_1 = p_2$$

$$H_A: p_1 \neq p_2$$

$p_1$  = the true population proportion of cars with defects with the 1<sup>st</sup> assembly procedure

$p_2$  = the true population proportion of cars with defects with the 2<sup>nd</sup> assembly procedure

Conditions:

1. Random Assignment given

2. Reasonable to assume independence as the treatments were randomly assigned and one assemble procedure should not impact the others

3.  $n_1\hat{p}_1 = 28 \geq 10$     and     $n_2\hat{p}_2 = 32 \geq 10$     Both samples are large enough to use the  
 $n_1\hat{q}_1 = 322 \geq 10$      $n_2\hat{q}_2 = 468 \geq 10$     normal approximation

$$\hat{p}_1 = \frac{x_1}{n_1} \rightarrow \frac{28}{350} = .08$$

$$\hat{p}_2 = \frac{x_2}{n_2} \rightarrow \frac{32}{500} = .064$$

$$\hat{q}_1 = 1 - \hat{p}_1 \rightarrow 1 - .08 = .92$$

$$\hat{q}_2 = 1 - \hat{p}_2 \rightarrow 1 - .064 = .936$$

$$\hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2} \rightarrow \frac{28 + 32}{350 + 500} \rightarrow \hat{p}_c = .0706$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{(\hat{p}_c)(\hat{q}_c)}{n_1} + \frac{(\hat{p}_c)(\hat{q}_c)}{n_2}}} \rightarrow \frac{.08 - .064}{\sqrt{\frac{(.0706)(.9294)}{350} + \frac{(.0706)(.9294)}{500}}} \rightarrow 2 \times P(Z > .8963) = .3700$$

P-value

alpha

.37

Greater than

.10

We Fail to reject the Null

Our p-value is .3700. We fail to reject the Null. There is Not sufficient evidence at  $\alpha = .10$  to suggest that the true population proportion of defects has changed from the 1<sup>st</sup> assembly method to the second.

From our study, both assembly methods appear to produce the same proportion of major defects.

The cars will fail and people will die, but at least we will know that the proportion will be the same regardless of assemble method.

## Notes: Hypothesis Tests-2 Sample Proportions (Scenario)

**Scenario 3: 2016 Question 5:** A polling agency showed the following two statements to a random sample of 1,048 adults in the United States.

**Environment Statement:** Protection of the environment should be given priority over economic growth.

**Economy Statement:** Economic growth should be given priority over protection of the environment.

The order in which the statements were shown was randomly selected for each person in the sample. After reading the statements, each person was asked to choose the statement that was most consistent with his or her opinion. The results are shown in the table.

	Environment Statement	Economy Statement	No Preference
Percent of Sample	58%	37%	5%

- (A) Assume the conditions for inference have been met. Construct and interpret a 95% confidence interval for the proportion of all adults in the United States who would have chosen the economy statement.
- (B) One of the conditions for inference that was met is that the number who chose the economy statement and the number who did not choose the economy statement are both greater than 10. Explain why it is necessary to satisfy that condition.
- (C) A suggestion was made to use a two-sample Z-interval for a difference between proportions to investigate whether the difference in proportions between adults in United States who would have chosen the environment state and adults in the United States who would have chosen the economy statement is statistically significant. Is the two-sample z-interval for a difference between proportions an appropriate procedure to investigate the difference? Justify your answer.

$p_0$  is the true population proportion of adults who would have chosen the economy statement.

$$\hat{p} = \frac{x}{n} \rightarrow .37 = \frac{x}{1048} \rightarrow x = 388; \quad \hat{q} = 1 - \hat{p} \rightarrow 1 - .37 = .63$$

**Calculate Z\***

2<sup>nd</sup> Vars Inverse Norm

$$\text{Area} = \frac{(1-.95)}{2}$$

$$\mu = 0 \text{ and } \sigma = 1$$

$$Z^* = 1.96$$

$$\hat{p} \pm z^* \sqrt{\frac{(\hat{p})(\hat{q})}{n}} \rightarrow .37 \pm 1.96 \sqrt{\frac{(.37)(.63)}{1048}} \rightarrow (.3427, .4013)$$

- (a) We are 95% confident that the true population proportion of adults who would have chosen the economy statement lies within the interval (.3427, .4013)
- (b) The condition is necessary to ensure that the sample is large enough to use the normal approximation for the binomial.
- (c) A 2-sample procedure is not appropriate because the data was collected from a single sample rather than 2 independent samples.



## Notes: Scatterplots

Previously we talked about how this course is divided into 4 large sections:

1. **The Collection** of Data (methods and cautions)
2. **The Display & Description** of the Data that was collected
3. **The Probability** of Obtaining the Results we observed
4. **Making Inferences or Decisions** based on the Likelihood of the results we obtained

When we were studying the display and description of data, there was one type of data that we avoided discussing in any detail and that is bivariate data. Our primary focus was on univariate data. For a Univariate graph one axis (typically the x) listed all possible outcome/response of the variable of interest while the y-axis kept track of the counts ie. the number of times a particular x-variable outcome/response occurred.

While both variables must be quantitative, bivariate data does not keep track of counts. Instead bivariate data provides the distinct measurements of a response variable (y) for a given value of the explanatory variable (x). The paired measurement is plotted as a single point on a scatterplot.

- Explanatory variable—the x variable that is believed to be the variable that drives the relationship (inputs)
- Response variable—the y variable which is the result of a change in the x-variable (outputs)

The plotting of multiple responses on a scatterplot allow us to determine what type of relationship exists between the two variables. We are able to determine how strong of a relationship exists and how predictive a pattern may be.

Scatterplots allow us to see the strength or lack thereof of the relationship/association between 2 variables. They also allow us to see patterns and unusual features such as clusters, gaps, outliers and points of influence.

**Example:** Husband's age versus wife's age. In general, the older the husband is, the older the wife is. But is the relationship perfect? Can we make some predictions?



### Notes: Creating Scatterplots

**1. Turn on Stat Diagnostics**

- Press **MODE**
- ↓ **STATDIAGNOSTICS:**
- → Highlight **ON**
- Press **ENTER**
- Press **2<sup>nd</sup> Mode/Quit**

**2. Input the Data**

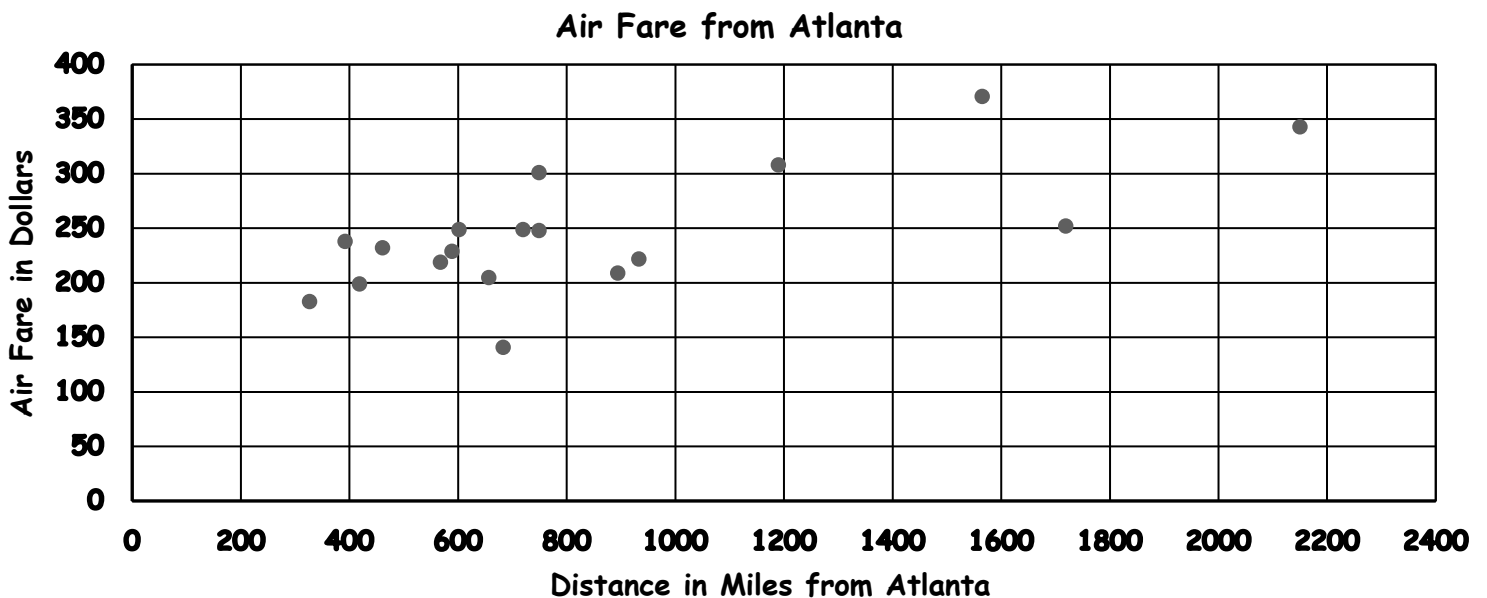
- Press **STAT**
- Highlight **EDIT**
- Press **ENTER**
- Enter "x" values in **L<sub>1</sub>**
- Enter "y" values in **L<sub>2</sub>**
- Press **2<sup>nd</sup> Mode/Quit**

**3. Graphing: Scatter Plot**

- Press **2<sup>nd</sup> STAT PLOT**
- Highlight **1: Plot 1** Press **ENTER**
- Highlight **On** Press **ENTER**
- ↓ Highlight **First Graph** Press **ENTER**
- ↓ **XList:** Press **2<sup>nd</sup> L<sub>1</sub>** Enter
- ↓ **YList:** Press **2<sup>nd</sup> L<sub>2</sub>** Enter
- Press **ZOOM 9**

Atlanta to:	Distance L1	Fare L2
Baltimore	568	219
Boston	933	222
Dallas	720	249
Denver	1190	308
Detroit	602	249
Kansas City	683	141
Las Vegas	1719	252
Miami	589	229
Memphis	327	183
Minneapolis	894	209
New Orleans	419	199
New York	749	248
Oklahoma City	749	301
Orlando	392	238
Philadelphia	657	205
St. Louis	461	232
Salt Lake	1565	371
Seattle	2150	343

Summary Statistics		
Atlanta to:	Distance L1	Fare L2
Means	853.7	244.33
Standard Deviation	497.8	56.37



**Notes: Linear Regression**

## Notes: Linear Regression

**Linear regression** -an attempt to model the relationship between two variables by fitting a linear equation to observed data. (the line is often referred to as the line of best fit) One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. In our example graph the explanatory variable (x) is the distance in miles from Atlanta and our dependent variable (y) is the fare.

The algorithm for creating the line of best fit is designed to minimize the sum of the overall actual distances from the predictive line of best fit.

The linear regression equation:  $\hat{y} = \beta_0 + \beta_1x + \epsilon$  which is essentially  $y = b + mx$  or  $y = mx + b$

$\hat{y}$  is the value that the model predicts for y

$\beta_0$  is the value of the y-intercept of the regression equation which is the value of the equation when x is zero

$\beta_1$  is the slope of the regression line (units of y per units of x)

$\epsilon$  is the **error term or residual**. It is how far an actual value is from a predicted value and is found by subtracting the **observed - expected** or  $y - \hat{y}$ . The sum of the errors for a regression line is zero.

### 1. Calculate the Regression Statistics

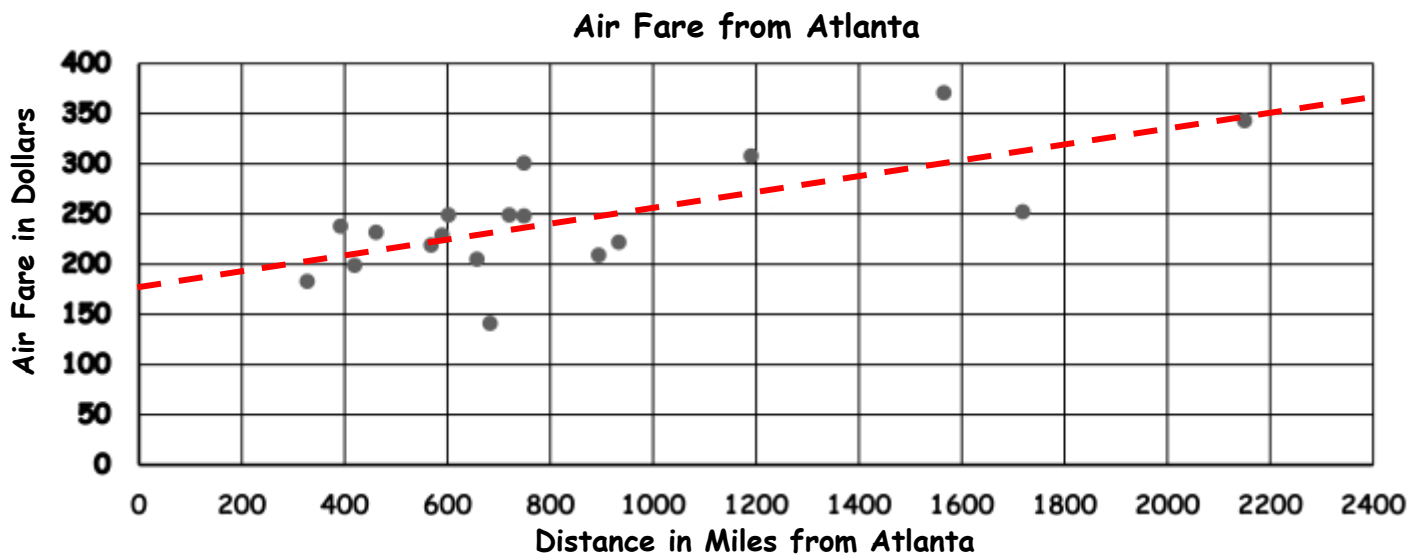
- Regression Equation  $y = a + bx$
- Slope:  $B_1 = b$
- Y-intercept:  $B_0 = a$
- Correlation Coefficient: r
- Coefficient of Determination:  $r^2$

- Press **STAT**
- → Highlight **CALC**
- ↓ **LinReg(a + bx)**
- ↓ XList: Press **2<sup>nd</sup> L<sub>1</sub>** Enter
- ↓ YList: Press **2<sup>nd</sup> L<sub>2</sub>** Enter
- ↓ Store RegEQ: Press **2<sup>nd</sup> ALPHA TRACE ENTER**
- Press **2<sup>nd</sup> Mode/Quit**

### 2. Graphing the Regression Equation

- Press **Zoom 9**

$$B_0 = a = 177.21, \quad B_1 = b = .0786 \quad \hat{y} = 177.21 + .0786x \quad r^2 = .4820 \quad r = .6943$$



## Notes: Residuals

As much as we would like there to be a perfect linear model with all of the data points situated on the predictive line that almost always never happens. The vertical distance the actual point is from the predictive line is known as the error term or residual.

$\epsilon$  is the **error term or residual**. It is how far an actual value is from a predicted value and is found by subtracting the **observed - expected** or  $y - \hat{y}$  or **actual - predicted**

$$\begin{aligned} \text{Residual} &= \text{Actual value} - \text{Model Value} \\ \text{Error} &= \text{Data} - \text{Predicted} \end{aligned}$$

$$\begin{aligned} \text{Actual value} &= \text{Model value} + \text{Residual} \\ \text{Data} &= \text{Predicted} + \text{Error} \end{aligned}$$

The sum of the errors for a regression line is zero.

For a linear equation to be appropriate, the graph of the Errors/Residuals must be random and without a pattern on a residual plot.

### • Calculating Residuals (actual - predicted)

- Press **STAT**
- Highlight **EDIT**; Press **ENTER**
- $\rightarrow$  Column **L<sub>3</sub>**  $\uparrow$  Highlight **L<sub>3</sub>**
- Press **2<sup>nd</sup> STAT/LIST**
- $\downarrow$  Highlight **7 RESID**
- Press **ENTER** twice

*The residuals are in L<sub>3</sub>*

A **positive** residual means the **actual is greater** than the predicted-**above the regression line**

A **negative** residual means the **actual is less** than the predicted-**below the regression line**

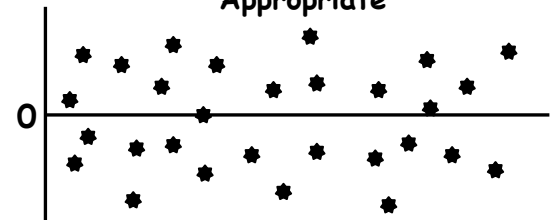
### 2. Calculate the Standard Deviation of the Residuals "S"

- Press **STAT** $\rightarrow$ Highlight **CALC**
- $\downarrow$  **1-VAR STATS** Press **2<sup>nd</sup> 3 (L<sub>3</sub>)**
- $\downarrow \downarrow$  Press **Enter**
- Write the values for  $\Sigma x^2$  & **n**
- Divide:  $\Sigma x^2$  by  $(n-2)$  so  $\frac{\Sigma x^2}{(n-2)}$
- Press  $\wedge .5$  so  $\sqrt{\frac{\Sigma x^2}{(n-2)}}$

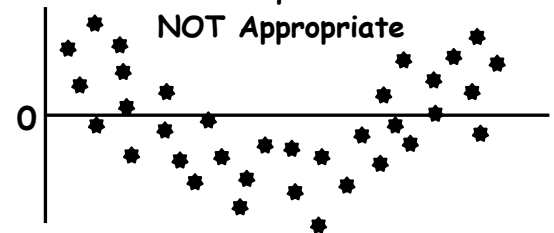
### 3. Graphing Residuals (actual - predicted)

- Press **2<sup>nd</sup> STAT PLOT**
- Highlight **1: Plot 1** Press **ENTER**
- Highlight **On** Press **ENTER**
- $\downarrow$  Highlight **First Graph** Press **ENTER**
- $\downarrow$  **XList**: Press **2<sup>nd</sup> L<sub>1</sub>** Enter
- $\downarrow$  **YList**: Press **2<sup>nd</sup> L<sub>3</sub>** Enter
- Press **ZOOM 9**

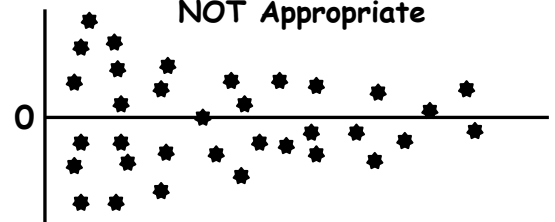
Linear Equation Is  
Appropriate



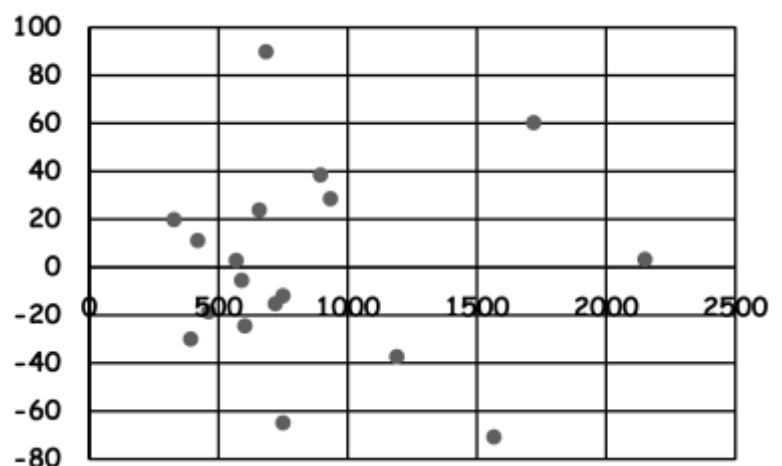
Linear Equation Is  
NOT Appropriate



Linear Equation Is  
NOT Appropriate



Residual Plot for Airfare vs. Distance



## Notes: Correlation

The residual plot may indicate that a line is appropriate, but it does not tell us how strong the linear relationship is between the response and explanatory variable. In order to determine the strength of a linear relationship we calculate "**r**" the **correlation coefficient**. 5

**Note:** we only talk about a correlation coefficient when we are discussing linear relationships.

When we calculate the regression Statistics, the calculator automatically give us "r" assuming STATS Diagnostics is turned on.

The **correlation coefficient** measures how closely the points in a scatter diagram are spread around the regression line. The value of the correlation coefficient always lies in the range of  $-1$  to  $1$  that is,  $-1 \leq \rho \leq 1$  and  $-1 \leq r \leq 1$  where  $\rho$  is the correlation coefficient calculated for the population data and  $r$  is the correlation coefficient calculated for the sample data.

If  $r = 1$  (*perfect positive linear correlation*) The slope is positive and the data points lie directly on the line of regression.

If  $r = -1$  (*perfect negative linear correlation*) The slope is negative and the data points lie directly on the line of regression.

If  $r$  is **close to zero**; (*there is no **Linear** correlation between the two variables*). However the variables may have a very high correlation.

### Rules of Thumb:

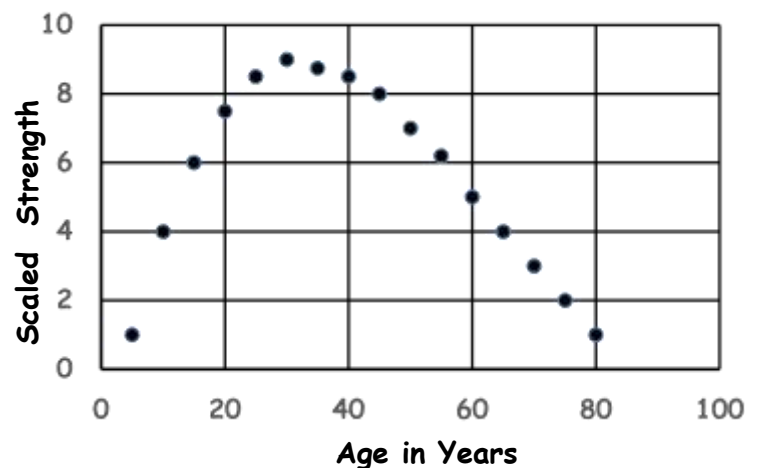
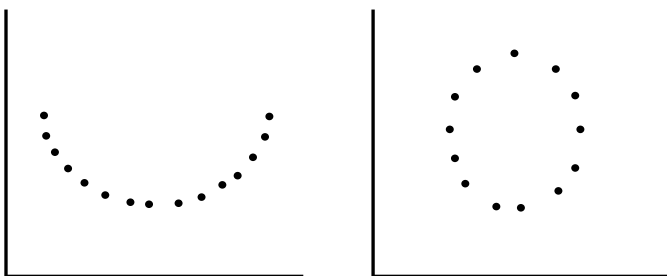
If  $|r| \geq .75$  there is a strong linear association

If  $.40 < |r| < .75$  there is a moderate linear association

If  $|r| < .40$  there is a weak linear association

### Caution: A high correlation or association does not mean causation

**Note:** Just because  $r$  is low does not mean that the variables are not highly correlated. Consider the following example graphs.



**Recipe for Success: Definition Template for Regression**

**Slope  $\beta_1 = m = \frac{\Delta y}{\Delta x}$**  (using  $y = m x + b$  notation)

\_\_\_\_\_ is the expected amount of change in \_\_\_\_\_  
*Value of the slope* *Restate the definition of Y*

for a 1 unit change in \_\_\_\_\_ .  
*Restate the definition of x*

**Slope Intercept  $\beta_0 = b = \text{constant}$**  (using  $y = m x + b$  notation)

The amount of \_\_\_\_\_ that you begin with or would have if  
*Restate the definition of Y*

the amount of \_\_\_\_\_ = zero.  
*Restate the definition of x*

**Correlation Coefficient ( $r$ ) =  $\sqrt{R^2}$**

•  $|r| \geq .75$  There is a strong \_\_\_\_\_ linear relationship between  
 (+ or - )--use the sign of the slope  
 \_\_\_\_\_ and \_\_\_\_\_  
*Restate the definition of Y* *Restate the definition of x*

•  $.40 < |r| < .75$  There is a moderately strong \_\_\_\_\_ linear  
 (+ or - )--use the sign of the slope  
 relationship between \_\_\_\_\_ and \_\_\_\_\_  
*Restate the definition of Y* *Restate the definition of x*

•  $|r| < .40$  There is a weak \_\_\_\_\_ linear relationship between  
 (+ or - )--use the sign of the slope  
 \_\_\_\_\_ and \_\_\_\_\_  
*Restate the definition of Y* *Restate the definition of x*

**Coefficient of Determination ( $R^2$ )**

\_\_\_\_\_ % of the variation in the \_\_\_\_\_ can be explained by  
*Restate the definition of Y*

changes in the \_\_\_\_\_  
*Restate the definition of x*

**S** The **standard deviation of the residuals** is \_\_\_\_\_ and measures the variance in

\_\_\_\_\_ for a given amount of \_\_\_\_\_  
*Restate the definition of Y* *Restate the definition of x*

**Standard Error of the Slope:** The standard error of the slope is \_\_\_\_\_. Because the slope is estimated from the sample, other samples are likely to have differing slopes. The standard error of the slope quantifies the amount of variation in sample slopes that could be expected from different samples.

## Notes: Linear Regression Definitions

Atlanta to:	Distance L1	Fare L2
Baltimore	568	219
Boston	933	222
Dallas	720	249
Denver	1190	308
Detroit	602	249
Kansas City	683	141
Las Vegas	1719	252
Miami	589	229
Memphis	327	183
Minneapolis	894	209
New Orleans	419	199
New York	749	248
Oklahoma City	749	301
Orlando	392	238
Philadelphia	657	205
St. Louis	461	232
Salt Lake	1565	371
Seattle	2150	343

Summary Statistics		
Atlanta to:	Distance L1	Fare L2
Means	853.7	244.33
Standard Deviation	497.8	56.37

- Define  $x$ :  **$X$  is the distance traveled in miles**
- Define  $y$ :  **$Y$  is the fare charged in dollars for a given trip**
- Write the equation of the linear model.  
 **$\hat{y} = 177.21 + .0786x$**
- Find the slope of the regression line and explain what the slope means in this context.  
**The slope is  $\frac{.0786 \text{ dollars}}{1 \text{ mile}}$**   
**For every 1 mile increase in distance there is a .0786 dollar increase in fare.**
- Find the  $y$ -intercept of the regression line and explain what the  $y$ -intercept means in this context.  
**The  $Y$ -intercept is 177.21.**  
**We would expect our fare to be 177.21 if we boarded the plane and flew a total of zero miles.**
- Calculate " $S$ " and explain what it means in context.  
**The standard deviation of the residuals is 41.82 and measures the variance in fare for a given distance.**

7. Find  $r$  and explain and discuss its strength.

**$r$  is .694. There is a moderately strong positive linear relationship between air fare and distance traveled.**

8. Find  $r^2$  Explain what  $r^2$  means in this context.

**$r^2 = .482$  and is the coefficient of determination. 48.2% of the variation in air fare can be explained by changes in the distance of the flight**

9. Estimate the fare for a 750-mile flight.

$$\hat{y} = 177.21 + .0786(750) \rightarrow \hat{y} = 236.18 \text{ dollars}$$

10. Estimate the fare for a 392-mile flight.

$$\hat{y} = 177.21 + .0786(392) \rightarrow \hat{y} = 208.03 \text{ dollars}$$

11. Based on your answer in "10" calculate the residual for a 392 mile flight.

$$\text{Observed} - \text{Expected} \rightarrow y - \hat{y} \rightarrow 238 - 208.03 = 29.7$$

## Recipe for Success: Hypothesis Test for Linear Regression (using data)

1. Define X & Y in context

2. Write your Hypothesis

- Null  $H_0: \beta_1 = 0$
- Alternative  $H_A: \beta_1 \neq 0$

$H_0$ : There is no linear relationship between x & y

$H_A$ : There is a linear relationship between x & y

3. Define  $\beta_1$  in context

$\frac{\Delta y}{\Delta x}$  = the change in y for every 1 unit change in x

4. Write the Conditions

- Random Sample
- Linear Scatterplot
- No pattern in residual plot
- Distribution of the residuals is normal

5. Write the Equations

$$t = \frac{b}{s_b} \quad s_b = \frac{\sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}}{\sqrt{\sum(x-\bar{x})^2}}$$

$\beta_1$  = the population slope of the regression line

t = the number of standard deviations from the mean

b = the slope of the regression line from the sample

$s_b$  = the standard error of the regression line

n = the sample size

df = n-2

6. Enter Data

- Stat Edit
- Enter X into  $L_1$
- Enter Y into  $L_2$

7. Calculate the p-value

- Stat Tests
- LinRegTTest → Enter
- X List-  $L_1$
- Y List-  $L_2$
- Freq 1
- $\beta_1 \neq$  (or whatever your hypothesis is)
- ReqEQ Alpha Trace

8. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

9. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population slope is **NOT** zero. There is a linear relationship between \_\_\_\_\_ and \_\_\_\_\_.  
*Restate the definition of x*                      *Restate the definition of y.*

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that the true population slope is different than zero. No linear relationship exists between \_\_\_\_\_ and \_\_\_\_\_.  
*Restate the definition of y.*                      *Restate the definition of x*



## Notes: Hypothesis Test for the Slope of the Regression Line (Data Scenario)

**Scenario 1:** Using the Data from the air fare and distance scenario, perform a hypothesis test with an alpha level of 0.05 to determine whether or not the slope is significant.

**Let X = the distance traveled in miles**

**Let Y = the fare charged in dollars**

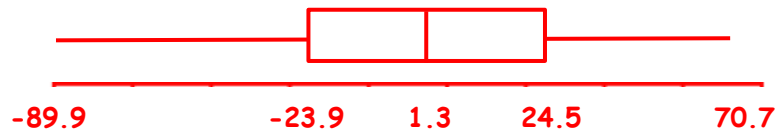
$H_0: \beta_1 = 0$      $H_0$ : There is no linear relationship between x & y

$H_A: \beta_1 \neq 0$      $H_A$ : There is a linear relationship between x & y

$\beta_1$  = the expected increase in air fare for every 1 mile increase in distance traveled

**Conditions:**

1. Random Sample—Not given; we must assume that the flight information is representative of domestic commercial flights
2. Scatterplot Appears linear see page 197
3. Residual plot is random see page 198
4. The boxplot of the residuals shows no outliers and is reasonably symmetric. Okay to assume that the residuals are approximately normally distributed.



$$s_b = \frac{\sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}}{\sqrt{\sum(x-\bar{x})^2}} = \frac{\sqrt{\frac{27980.6}{18-2}}}{\sqrt{4212525.6}} = \frac{\sqrt{1748.79}}{\sqrt{4212525.6}} = \frac{41.82}{2052.44} = .02037$$

**df: 18-2 = 16**

$$t = \frac{b}{s_b} = \frac{.0786}{.02037} = 3.86$$

**2×P(t>3.86) = .0014**

P-value	alpha
.0014	.05
Less than	We reject the Null

Our p-value is .0014. We reject the Null. There is sufficient evidence at alpha = .05 to suggest that the true population slope is NOT zero.

There is a linear relationship between air fare and distance traveled.

## Recipe for Success: Confidence Intervals for Linear Regression (using data)

1. Define X & Y in context

2. Define  $\beta_1$  in context

$\frac{\Delta y}{\Delta x}$  = the change in y for every 1 unit change in x

3. Write the Conditions

- Random Sample
- Linear Scatterplot
- No pattern in residual plot
- Distribution of the residuals is normal

3. Write the formula for the Test

$$b \pm t^* s_b$$

$$s_b = \frac{\sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}}{\sqrt{\sum(x-\bar{x})^2}}$$

$\beta_1$  = the population slope of the regression line

t = the number of standard deviations from the mean

b = the slope of the regression line from the sample

$s_b$  = the standard error of the regression line

n = the sample size

df = n-2

4. Calculate  $t^*$

- 2<sup>nd</sup> Vars
- Inverse t
- Area =  $\frac{(1-\text{Confidence level})}{2}$
- df = n-2

5. Enter Data

- Stat Edit
- Enter X into L<sub>1</sub>
- Enter Y into L<sub>2</sub>

6. Calculate Interval

- Stat Tests
- LinRegTTest → Enter
- X List- L<sub>1</sub>
- Y List- L<sub>2</sub>
- Freq 1
- C-level (confidence Level)
- ReqEQ Alpha Trace

7. Write the interval

8. Write the Conclusion

We are \_\_\_\_\_% confident that the true population slope for \_\_\_\_\_  
relative to \_\_\_\_\_ lies within the interval \_\_\_\_\_.  
*Restate the definition of y* *Restate the definition of x*

9. Explain the meaning of the confidence level-if asked

In repeated sampling, we expect that this method will capture the true population slope for \_\_\_\_\_  
relative to \_\_\_\_\_ percent of the time.  
*Restate the definition of y* *Restate the definition of x*

## Notes: Confidence Interval for the Slope of the Regression Line (Data Scenario)

**Scenario 1:** Using the Data form the air fare and distance scenario create a 96% confidence interval for the slope of the regression line.

Let  $X$  = the distance traveled in miles

Let  $Y$  = the fare charged in dollars

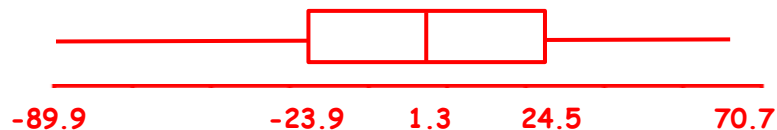
$H_0: \beta_1 = 0$      $H_0$ : There is no linear relationship between  $x$  &  $y$

$H_A: \beta_1 \neq 0$      $H_A$ : There is a linear relationship between  $x$  &  $y$

$\beta_1$  = the expected increase in air fare for every 1 mile increase in distance traveled

Conditions:

1. Random Sample—Not given; we must assume that the flight information is representative of domestic commercial flights
2. Scatterplot Appears linear see page 197
3. Residual plot is random see page 198
4. The boxplot of the residuals shows no outliers and is reasonably symmetric. Okay to assume that the residuals are approximately normally distributed.



$$b \pm t^* s_b \rightarrow .0786 \pm 2.11(.02037)$$

$$b = .0786 \qquad \qquad \qquad (.0354, .1218)$$

$$t^* = 2.11 \quad \text{df: } 18-2 = 16$$

$$s_b = \frac{\sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}}{\sqrt{\sum(x-\bar{x})^2}} = \frac{\sqrt{\frac{27980.6}{18-2}}}{\sqrt{4212525.6}} = \frac{\sqrt{1748.79}}{\sqrt{4212525.6}} = \frac{41.82}{2052.44} = .02037$$

We are 95% confident that the true population slope for air fare relative to distance traveled lies within the interval (.0354, .1218).

In repeated sampling, we expect that this method will capture the true population slope for air fare relative to distance traveled 95 percent of the time.

**Notes: Computer Print-outs for Regression**

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	26037.4	1	26037.4	14.9
Residual	27980	16	1748.79	
Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	177.215	19.99	8.86	≤0.0001
Distance	0.0786	0.0204	3.86	≤0.0014
R-squared = 48.2%		R-squared (adjusted) = 45.0%		
s = 41.82				

1. What is the **explanatory variable?** **Distance;** What is the response variable? **Air Fare**
2. Identify the slope of the regression line **.0786 dollars per mile**
3. Identify the y-intercept **177.215 dollars**
4. Write the Regression equation  $\hat{y} = 177.215 + .0786(x)$
5. Calculate the sample size. **Degrees of freedom of the residual plus 2**  $(16 + 2) = 18$
6. Identify the standard deviation of the residuals. How could it be calculated using the values from the printout?

$$\sqrt{\frac{\text{Sum of the Squares of the Residual}}{\text{Degrees of Freedom of the Residual}}} = \sqrt{\frac{27980}{16}} = \sqrt{1748.79} = 41.82 \rightarrow s = 41.82$$

7. How can the t-statistic for the slope be calculated from the table values?

$$t = \frac{\text{Distance Coefficient}}{\text{Distance s.e. of Coefficient}} = \frac{.0786}{.0204} = 3.86 \rightarrow t = 3.86$$

8. What is the p-value of the test? How can you verify with your calculator utilizing the information from the print-out?

$$2 \times P(t > 3.86) \rightarrow 2^{\text{nd}} \text{ Vars } \text{tcdf} \quad P(t > 3.86) = .000693$$

$$\text{Lower: } 3.86, \text{ Upper: infinity} \rightarrow 2 \times (.000693) = .0014$$

$$\text{df: } 16 \quad \text{p-value} = .0014$$

9. What is the correlation coefficient?

**Correlation Coefficient = the sign of the slope ×  $\sqrt{\text{Coefficient of Determination}}$**

**Correlation Coefficient = the sign of the slope ×  $\sqrt{\text{R-squared}}$**

$$r = (\text{positive}) \times \sqrt{\text{R-squared}} \rightarrow r = \sqrt{.482} \rightarrow .6943$$

## Notes: Chi-Square Goodness of Fit

### Facts About the Chi-Square Distribution

1. The curve is nonsymmetrical and skewed to the right.
2. There is a different chi-square curve for each degree freedom (this is like the t-distribution)
3. The test statistic for any test is always greater than or equal to zero.
4. When  $df > 90$ , the chi-square curve approximates the normal.
5. The mean,  $\mu$ , is located just to the right of the peak.

### Types of Chi-Square Tests:

1. **Chi-Square Goodness of Fit:** Tests to see if a sample distribution matches a known distribution.  
**Key Words:** *distribution, difference, historical, previous, model, matches*
2. **Chi-Square Test of Independence:** Tests to determine whether or not two or more variables from a single sample are associated/related/independent.  
**Key words:** *independence, association, relation*
3. **Chi-Square Test of Homogeneity:** Tests to see if a single variable sampled from 2 or more populations or groups has the **same proportion**.  
**Key words:** *proportions, ratios*

### Chi Square Goodness of Fit Test

The **Chi Square Goodness of Fit test** is a hypothesis test used to determine if the data "fit" a particular distribution or not. **The null and the alternate hypotheses for this test may be written in sentences or may be stated as equations or inequalities. The goodness-of-fit test is almost always right tailed.**

The test statistic for a goodness-of-fit test is:  $\chi^2 = \sum_1^n \frac{(O - E)^2}{E}$

where:

- $O$  = observed values (data)
- $E$  = expected values (from theory)
- $n$  = the number of different data cells or categories
- Degrees of freedom are  $df = (\text{number of categories} - 1)$ .

**Chi-Square Goodness of Fit:** Tests to see if a sample distribution matches some known distribution.

( $\chi^2$  GOF has only one variable and one sample)  $df = (\text{number of categories} - 1)$ .

$H_0$ : The \_\_\_\_\_ data fits/follows or matches a \_\_\_\_\_ distribution

$H_A$ : The \_\_\_\_\_ data does not fit/follow or match a \_\_\_\_\_ distribution

**Key words:** *distribution, difference, historical, previous, model, matches*

## Recipe for Success: Chi-Square Goodness-of-Fit Hypothesis Test

### 1. Write your Hypothesis

- Null  $H_0$ : The \_\_\_\_\_ data follows a \_\_\_\_\_ distribution
  - Alternative  $H_A$ : The \_\_\_\_\_ data does not follow a \_\_\_\_\_ distribution
- Key words: distribution, difference, historical, previous, model, matches**

### 2. Write the Conditions

1. Random Sample (1-sample with 1-variable)
2. All expected values  $> 5$
3. Counted data for observed

### 3. Write the Equation

$$\chi^2 = \frac{(O - E)^2}{E}$$

$\chi^2$  = How far away a distribution is from the expected/Null  
 $O$  = observed Value-the value from the sample  
 $E$  = the expected or predicted value.  
 $n$  = the number of categories

### 4. Calculate the Expected Values

- Sum the number observations in the sample
  - Multiply the number of observations by the Historical %
- Note: For the Expected value of a Uniform Distribution:**  
 Divide the sum of the sample observations by the number of categories

### 5. List the Expected Values

### 6. Enter Data

- Stat Edit
- Enter Sample Data in column L<sub>1</sub>
- Enter Expected Data in column L<sub>2</sub>
- 2<sup>nd</sup> Quit

### 7. Plug values into the equation (1<sup>st</sup> and last Expected values)

$$\chi^2 = \frac{(O-E)^2}{E} + \dots + \frac{(O-E)^2}{E}$$

### 8. Calculate the $\chi^2$ and the p-value

- Stat Tests
- $\chi^2$  GOF-Test Enter
- Observed: Press 2<sup>nd</sup> L<sub>1</sub>
- Expected: Press 2<sup>nd</sup> L<sub>2</sub>
- **df= n-1** (df is the degrees of freedom)

### 9. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 10. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_ to suggest that the \_\_\_\_\_.  
*Restate  $H_A$*

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_ to suggest that the \_\_\_\_\_.  
*Restate  $H_A$*

## Notes: Chi-Square Goodness of Fit

**Scenario 1:** Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Three statistics instructors wondered whether the absentee rate was the **same** for every day of the school week. They took a **random** sample of absent students from three of their statistics classes during one week of the term. The results of the survey appear in the table.

Class Days	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Number of absent students	28	22	18	20	32	120
Expected Values 120/5	24	24	24	24	24	120

Run an appropriate hypothesis test to answer the professor's question.

$H_0$ : The absentee data follows a uniform distribution

$H_A$ : The absentee data does not follow a uniform distribution

Conditions:

1. Random Sample: Not given but the sample is assumed to be representative of all of the statistics students
2. Expected values are all greater than 5 see table above
3. All observed values are counted data (whole numbers)

$$\chi^2 = \frac{(O-E)^2}{E} \quad \text{df: } 5-1 = 4$$

$$\chi^2 = \frac{(28-24)^2}{24} + \frac{(22-24)^2}{24} + \frac{(18-24)^2}{24} + \frac{(20-24)^2}{24} + \frac{(32-24)^2}{24} = 5.67$$

$$P(\chi^2 > 5.67) = .225$$

P-value	alpha
.225	.10
Greater than	We Fail to reject the Null

Our p-value is .225. We Fail to reject the Null. There is not sufficient evidence at alpha = .10 to suggest that the distribution of data does not follow the uniform distribution

We do not reject the null. Absenteeism follows the uniform distribution and is the same for all days of the week.

$$CNTRB = \{.6667, .1667, 1.5, .6667, 2.6667\}$$

Friday's absenteeism contributes the most to the Chi-square test and differs from the expected value more than the other days of the week.

## Recipe for Success: Chi-Square Goodness-of-Fit Hypothesis Test

### 1. Write your Hypothesis

- Null  $H_0$ : The \_\_\_\_\_ data follows a \_\_\_\_\_ distribution
  - Alternative  $H_A$ : The \_\_\_\_\_ data does not follow a \_\_\_\_\_ distribution
- Key words: distribution, difference, historical, previous, model, matches**

### 2. Write the Conditions

1. Random Sample (1-sample with 1-variable)
2. All expected values  $> 5$
3. Counted data for observed

### 3. Write the Equation

$$\chi^2 = \frac{(O - E)^2}{E}$$

$\chi^2$  = How far away a distribution is from the expected/Null  
 $O$  = observed Value-the value from the sample  
 $E$  = the expected or predicted value.  
 $n$  = the number of categories

### 4. Calculate the Expected Values

- Sum the number observations in the sample
  - Multiply the number of observations by the Historical %
- Note: For the Expected value of a Uniform Distribution:**  
 Divide the sum of the sample observations by the number of categories

### 5. List the Expected Values

### 6. Enter Data

- Stat Edit
- Enter Sample Data in column L<sub>1</sub>
- Enter Expected Data in column L<sub>2</sub>
- 2<sup>nd</sup> Quit

### 7. Plug values into the equation (1<sup>st</sup> and last Expected values)

$$\chi^2 = \frac{(O-E)^2}{E} + \dots + \frac{(O-E)^2}{E}$$

### 8. Calculate the $\chi^2$ and the p-value

- Stat Tests
- $\chi^2$  GOF-Test Enter
- Observed: Press 2<sup>nd</sup> L<sub>1</sub>
- Expected: Press 2<sup>nd</sup> L<sub>2</sub>
- **df= n-1** (df is the degrees of freedom)

### 9. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 10. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the is \_\_\_\_\_.  
*Restate  $H_A$*

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that is \_\_\_\_\_.  
*Restate  $H_A$*



## Notes: Chi-Square Goodness of Fit

**Scenario 2:** A grocery store manager wishes to determine whether a certain product will sell equally well in any of five locations in the store. Five displays are set-up, one in each location, and the resulting numbers of the product sold are noted.

Location	1	2	3	4	5	Total
Number of Items Sold	43	29	52	34	48	206
Expected Values $206/5$	41.2	41.2	41.2	41.2	41.2	206

Is there evidence that location makes a difference? Test at the 10% level of significance.

$H_0$ : The product location sales data follows a uniform distribution

$H_A$ : The product location sales data does not follow a uniform distribution

**Conditions:**

1. Random Sample: Not given but the sample is assumed to be representative of sales data
2. Expected values are all greater than 5 see table above
3. All observed values are counted data (whole numbers)

$$\chi^2 = \frac{(O-E)^2}{E} \quad \text{df: } 5-1 = 4$$

$$\chi^2 = \frac{(43-41.2)^2}{41.2} + \dots + \frac{(48-41.2)^2}{41.2} = 8.903$$

$$P(\chi^2 > 8.903) = .0636$$

P-value	alpha
.0636	.10
Less than	We reject the Null

Our p-value is .0636. We reject the Null. There is sufficient evidence at alpha = .10 to suggest that the distribution of data does not follow the uniform distribution

We do reject the null. Product placement is the same for all five locations in the store

CNTRB = {.0786, 3.612, 1.5, 2.831, 1.258}

Product location 2 differs the most from the expected and is below what is expected  
 Product location 3 sells the most above what is expected.

## Recipe for Success: Chi-Square Goodness-of-Fit Hypothesis Test

### 1. Write your Hypothesis

- Null  $H_0$ : The \_\_\_\_\_ data follows a \_\_\_\_\_ distribution
  - Alternative  $H_A$ : The \_\_\_\_\_ data does not follow a \_\_\_\_\_ distribution
- Key words: distribution, difference, historical, previous, model, matches**

### 2. Write the Conditions

1. Random Sample (1-sample with 1-variable)
2. All expected values  $> 5$
3. Counted data for observed

### 3. Write the Equation

$$\chi^2 = \frac{(O - E)^2}{E}$$

$\chi^2$  = How far away a distribution is from the expected/Null  
 $O$  = observed Value-the value from the sample  
 $E$  = the expected or predicted value.  
 $n$  = the number of categories

### 4. Calculate the Expected Values

- Sum the number observations in the sample
  - Multiply the number of observations by the Historical %
- Note: For the Expected value of a Uniform Distribution:**  
 Divide the sum of the sample observations by the number of categories

### 5. List the Expected Values

### 6. Enter Data

- Stat Edit
- Enter Sample Data in column L<sub>1</sub>
- Enter Expected Data in column L<sub>2</sub>
- 2<sup>nd</sup> Quit

### 7. Plug values into the equation (1<sup>st</sup> and last Expected values)

$$\chi^2 = \frac{(O-E)^2}{E} + \dots + \frac{(O-E)^2}{E}$$

### 8. Calculate the $\chi^2$ and the p-value

- Stat Tests
- $\chi^2$  GOF-Test Enter
- Observed: Press 2<sup>nd</sup> L<sub>1</sub>
- Expected: Press 2<sup>nd</sup> L<sub>2</sub>
- **df= n-1** (df is the degrees of freedom)

### 9. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 10. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the is \_\_\_\_\_.  
*Restate  $H_A$*

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that is \_\_\_\_\_.  
*Restate  $H_A$*

**Notes: Chi-Square Goodness of Fit**

**Scenario 3:** In a recent year, at the 6pm time slot, television channels 2, 3, 4, and 5 captured the entire audience with 30%, 25%, 20%, and 25%, respectively. During the first week of the next session, 500 viewers are interviewed. The table records those observations.

Television Channels	2	3	4	5	Total
Number of viewers	139	138	112	111	500
Expected Values	$(.30)(500)$ = 150	$(.25)(500)$ = 125	$(.20)(500)$ = 100	$(.25)(500)$ = 125	500

Did the viewing preferences have a statistically significant change?

$H_0$ : Television audience viewership follows the historical distribution

$H_A$ : Television audience viewership does not follow the historical distribution

Conditions:

1. Random Sample: Not given but the sample is assumed to be representative of all television viewers
2. Expected values are all greater than 5 see table above
3. All observed values are counted data (whole numbers)

$$\chi^2 = \frac{(O-E)^2}{E} \quad \text{df: } 4-1 = 3$$

$$\chi^2 = \frac{(139-150)^2}{150} + \dots + \frac{(111-125)^2}{125} = 5.167$$

$$P(\chi^2 > 5.67) = .1599$$

P-value	alpha
.1599	.10
Greater than	We Fail to reject the Null

Our p-value is .1599. We fail to reject the Null. There is not sufficient evidence at alpha = .10 to suggest that the television audience viewership does not follow the historical distribution

We do not reject the null. Television viewership proportions have not changed

$$\text{CNTRB} = \{.8067, 1.352, 1.44, 1.568\}$$

Channel 2 has the least change in viewership

Channel 4 has the greatest proportion increase in viewership

Channel 5 has the greatest decrease in viewership and the greatest overall change in viewership

## Recipe for Success: Chi-Square Goodness-of-Fit Hypothesis Test

### 1. Write your Hypothesis

- Null  $H_0$ : The \_\_\_\_\_ data follows a \_\_\_\_\_ distribution
  - Alternative  $H_A$ : The \_\_\_\_\_ data does not follow a \_\_\_\_\_ distribution
- Key words: distribution, difference, historical, previous, model, matches**

### 2. Write the Conditions

1. Random Sample (1-sample with 1-variable)
2. All expected values  $> 5$
3. Counted data for observed

### 3. Write the Equation

$$\chi^2 = \frac{(O - E)^2}{E}$$

$\chi^2$  = How far away a distribution is from the expected/Null  
 $O$  = observed Value-the value from the sample  
 $E$  = the expected or predicted value.  
 $n$  = the number of categories

### 4. Calculate the Expected Values

- Sum the number observations in the sample
  - Multiply the number of observations by the Historical %
- Note: For the Expected value of a Uniform Distribution:**  
 Divide the sum of the sample observations by the number of categories

### 5. List the Expected Values

### 6. Enter Data

- Stat Edit
- Enter Sample Data in column L<sub>1</sub>
- Enter Expected Data in column L<sub>2</sub>
- 2<sup>nd</sup> Quit

### 7. Plug values into the equation (1<sup>st</sup> and last Expected values)

$$\chi^2 = \frac{(O-E)^2}{E} + \dots + \frac{(O-E)^2}{E}$$

### 8. Calculate the $\chi^2$ and the p-value

- Stat Tests
- $\chi^2$  GOF-Test Enter
- Observed: Press 2<sup>nd</sup> L<sub>1</sub>
- Expected: Press 2<sup>nd</sup> L<sub>2</sub>
- **df= n-1** (df is the degrees of freedom)

### 9. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 10. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the is \_\_\_\_\_.  
*Restate  $H_A$*

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that is \_\_\_\_\_.  
*Restate  $H_A$*



## Notes: Review of Independence

**Independent Events**-Events that do not rely on one another. The outcome of one does not impact the outcome of the other. If independent  $P(A \cap B) = P(A) \times P(B)$ -Independent events are not associated, Independent events are not correlated and are not related.

- **Multiplication Principal**-if two events are independent their probabilities can be multiplied to find the likelihood that both events will occur at the same time.  
Equation  $P(A \cap B) = P(A) \times P(B)$  this reads the probability of the intersection of A and B, often stated as A and B, is equal to the probability of A times the probability of B
- **Graphically:** The corresponding segments of a segment will be the same size because they have the same proportions.

**Scenario: 2011 Problem 2** The table below shows the political party registration by gender of all 500 registered voters in Franklin Township.

PARTY REGISTRATION - FRANKLIN TOWNSHIP				
	Party W	Party X	Party Y	Total
Female	60	120	120	300
Male	28	124	48	200
Total	88	244	168	500

(b) Among the registered voters of Franklin Township, are the events "is a male" and "is registered for Party Y" independent? **Justify your answer based on probabilities calculated from the table above.**

If independent, then the  $P(\text{Male}) \times P(\text{Party Y}) = P(\text{Male} \cap \text{Party Y})$

$$\frac{200}{500} \times \frac{168}{500} = \frac{84}{625} \text{ or } .1344$$

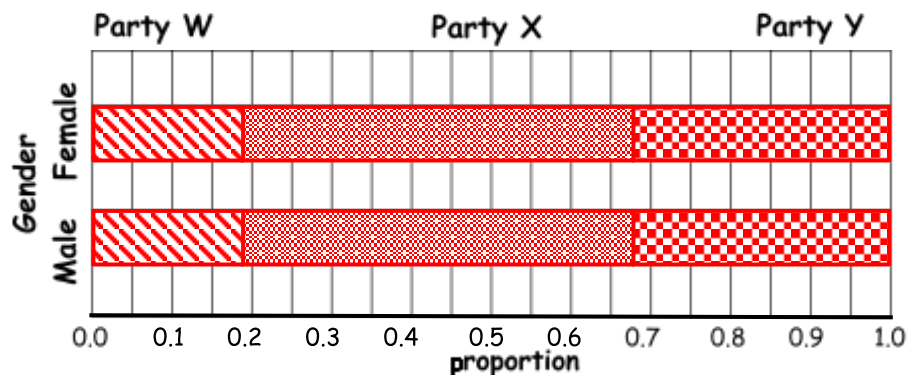
$$\text{and } .1344 \neq \frac{48}{500} \text{ or } .096 \therefore \text{not independent}$$

c) In Lawrence Township, the proportions of all registered voters for Parties W, X, and Y are the same as for Franklin Township, and party registration is independent of gender. Complete the graph below to show the distributions of party registration by gender in Lawrence Township

$$\text{Party w} = \frac{88}{500} = .176$$

$$\text{Party x} = \frac{244}{500} = .488$$

$$\text{Party y} = \frac{168}{500} = .336$$



## Notes: Review of Independence

**Scenario 2:** Suppose A = a speeding violation in the last year and B = a car phone user. If A and B are independent then  $P(A \text{ AND } B) = P(A) \times P(B)$ . A AND B is the event that a driver received a speeding violation last year and is also a car phone user. Suppose, in a study of drivers who received speeding violations in the last year and who use car phones, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 were car phone users and 450 were not.

If A and B are independent, then  $P(A \text{ AND } B) = P(A) \times P(B)$ . Complete the table

	Car Phone Users	No Car Phone	Total
Speeding Ticket	$(305)(70) \div 755 = 28.28$	$(450)(70) \div 755 = 41.72$	70
No ticket	$(305)(685) \div 755 = 276.72$	$(450)(685) \div 755 = 408.28$	685
Total	305	450	755

About 28 people from the sample are expected to be car phone users and to receive speeding violations. In a test of independence, we state the null and alternate hypotheses in words. Since the contingency table consists of **two factors**, the null hypothesis states that the factors are **independent** and the alternate hypothesis states that they are **not independent (dependent)**. If we do a test of independence using the example above, then the null hypothesis is:

**H<sub>0</sub>:** Being a car phone user and receiving a speeding violation are independent events.

**H<sub>A</sub>:** Being a car phone user and receiving a speeding violation are not independent events or are associated.

If the null hypothesis were true, we would expect about 28 people to be car phone users and to receive a speeding violation.

### Chi Square Test of Independence

The Chi Square Test of independence is a hypothesis test used to determine whether two factors are independent or not—whether or not an association exists. **The null and the alternate hypotheses for this test may be written in sentences. The test of independence is always right tailed.**

**Chi-Square Test of Independence:** Tests to see if 2 or more variables from a single sample are associated/related/independent.  $df = (r - 1)(c - 1)$ .  $r$ =row &  $c$ =columns

**H<sub>0</sub>:** The \_\_\_\_\_ and \_\_\_\_\_ are independent  
Variables being compared

**H<sub>A</sub>:** The \_\_\_\_\_ and \_\_\_\_\_ are not independent  
Variables being compared

**Key words:** *independence, association, relation*

The test statistic for a test of independence is:  $\chi^2 = \sum_1^n \frac{(O - E)^2}{E}$

where:

- O= observed values (data)
- E = expected values (from theory)      $E = \frac{(\text{row total})(\text{column total})}{\text{total number}}$
- n = the number of different data cells or categories
- **Degrees of freedom are  $df = (r - 1)(c - 1)$ .**

## Recipe for Success: Chi-Square Test of Independence

### 1. Write your Hypothesis

- Null  $H_0$ : The \_\_\_\_\_ and \_\_\_\_\_ are independent  
Variables being compared
- Alternative  $H_A$ : The \_\_\_\_\_ and \_\_\_\_\_ are associated and are not independent.  
Variables being compared

**Key words: independence, association, relation**

### 2. Write the Conditions

1. Random Sample (1-sample with 2-variable)
2. All expected values  $> 5$
3. Counted data for observed

### 3. Write the Equation

$$\chi^2 = \frac{(O - E)^2}{E}$$

$\chi^2$  = How far away a distribution is from the expected/Null  
 $O$  = observed Value-the value from the sample  
 $E$  = the expected or predicted value.  
 $n$  = the number of categories

### 4. Calculate & List Expected Values

- 2<sup>nd</sup> Matrix → Edit → [A]
- Enter Matrix size (row x column) do not include totals.
- Enter observed Data
- 2<sup>nd</sup> Quit
- STAT TESTS
- C:  $\chi^2$ -Test
- 2nd matrix → Edit ↓ [B] (these are the expected values)

### 5. Plug values into the equation (1<sup>st</sup> and last Expected values)

$$\chi^2 = \frac{(O-E)^2}{E} + \dots + \frac{(O-E)^2}{E}$$

### 6. Calculate the $\chi^2$ and the p-value

- STAT TESTS
- C:  $\chi^2$ -Test
- **df = (r-1)(c-1)** (r = rows & c = columns)

### 7. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 8. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_ to suggest that the \_\_\_\_\_ are \_\_\_\_\_ are associated and are not independent. *Restate Variables*

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_ to suggest that the \_\_\_\_\_ and \_\_\_\_\_ are not independent. *Restate Variables*



## Notes: Chi-Square Test of Independence (Scenario)

**Scenario 1:** In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. The following table is a **random sample** of the adult volunteers and the number of hours they volunteer per week.

**Number of Hours Worked Per Week by Volunteer Type (Observed)**

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours	Row Total
Community College Students	111	96	48	255
4-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

Are the number of hours volunteered **independent** of the type of volunteer?

**Number of Hours Worked Per Week by Volunteer Type (Expected Values)**

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours	Row Total
Community College Students	90.572	115.19	49.237	255
4-Year College Students	103	131	55.995	290
Nonstudents	104.42	132.81	56.768	294
Column Total	298	379	162	839

**Null  $H_0$ : The number of hours worked and the type of volunteer are independent.**

**Alternative  $H_A$ : The number of hours worked and the type of volunteer are Not independent.**

**Conditions:**

1. **Random Sample Given**
2. **All expected values > 5 See table**
3. **Counted data for observed all are whole numbers**

$$\chi^2 = \frac{(O-E)^2}{E} \quad df = (3-1)(3-1) = 4$$

$$\chi^2 = \frac{(111-90.572)^2}{90.572} + \dots + \frac{(53-56.768)^2}{56.768} = 12.99 \rightarrow P(\chi^2 > 12.99) = .011$$

P-value		alpha
.011	less than	.05
<b>We reject the Null</b>		

Our p-value is about .011. We reject the Null. There is sufficient evidence at alpha = .05 to suggest that the number of hours worked and the type of volunteer are Not independent.

**Conclusion:** Number of hours worked and type of volunteer are associated.

## Recipe for Success: Chi-Square Test of Independence

### 1. Write your Hypothesis

- Null  $H_0$ : The \_\_\_\_\_ and \_\_\_\_\_ are independent  
Variables being compared
- Alternative  $H_A$ : The \_\_\_\_\_ and \_\_\_\_\_ are associated and are not independent.  
Variables being compared

**Key words: independence, association, relation**

### 2. Write the Conditions

1. Random Sample (1-sample with 2-variable)
2. All expected values  $> 5$
3. Counted data for observed

### 3. Write the Equation

$$\chi^2 = \frac{(O - E)^2}{E}$$

$\chi^2$  = How far away a distribution is from the expected/Null  
 $O$  = observed Value-the value from the sample  
 $E$  = the expected or predicted value.  
 $n$  = the number of categories

### 4. Calculate & List Expected Values

- 2<sup>nd</sup> Matrix → Edit → [A]
- Enter Matrix size (row x column) do not include totals.
- Enter observed Data
- 2<sup>nd</sup> Quit
- STAT TESTS
- C:  $\chi^2$ -Test
- 2nd matrix → Edit ↓ [B] (these are the expected values)

### 5. Plug values into the equation (1<sup>st</sup> and last Expected values)

$$\chi^2 = \frac{(O-E)^2}{E} + \dots + \frac{(O-E)^2}{E}$$

### 6. Calculate the $\chi^2$ and the p-value

- STAT TESTS
- C:  $\chi^2$ -Test
- **df = (r-1)(c-1)** (r = rows & c = columns)

### 7. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 8. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_ to suggest that the \_\_\_\_\_ are \_\_\_\_\_ are associated and are not independent. *Restate Variables*

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_ to suggest that the \_\_\_\_\_ and \_\_\_\_\_ are not independent. *Restate Variables*



### Notes: Chi-Square Test of Independence (Scenario)

**Scenario 3 2003 Question 5** A random sample of 200 students was selected from a large college in the United States. Each selected student was asked to give his or her opinion about the following statement.

"The most important quality of a person who aspires to be the President of the United States is a knowledge of foreign affairs."

Each response was recorded in one of five categories. The gender of each selected student was noted. The data are summarized in the table below:

	Response Category				
	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
Male	10/ 13.5	15/ 18	15/ 18	25/ 22.5	25/ 18
Female	20/ 16.5	25/ 22	25/ 22	25/ 27.5	15/ 22

Is there sufficient evidence to indicate that the response is dependent on gender? Provide statistical evidence to support your conclusion.

**Null  $H_0$ : Response category and gender are independent**

**Alternative  $H_A$ : Response category and gender are Not independent.**

**Conditions:**

1. **Random Sample Given**
2. **All expected values > 5 See table**
3. **Counted data for observed all are whole numbers**

$$\chi^2 = \frac{(O-E)^2}{E} \quad df = (2-1)(5-1) = 4$$

$$\chi^2 = \frac{(10-13.5)^2}{13.5} + \dots + \frac{(15-22)^2}{22} = 8.923 \rightarrow P(\chi^2 > 8.923) = .063$$

**P-value**

**0.063**

**alpha**

**Greater than**

**.05**

**We fail to reject the Null**

Our p-value is about .063. We fail to reject the Null. There is not sufficient evidence at alpha = .05 to suggest that response category and gender are not independent

**Conclusion: Response category and gender are independent.**



## Recipe for Success: Chi-Square Test of Homogeneity

### 1. Write your Hypothesis

- Null  $H_0$ : The distribution of \_\_\_\_\_ and \_\_\_\_\_ are the same.  
1<sup>st</sup> sample 2<sup>nd</sup> sample
- Alternative  $H_A$ : The distribution of \_\_\_\_\_ and \_\_\_\_\_ are not the same  
1<sup>st</sup> sample 2<sup>nd</sup> sample

**Key words:** *distribution are the same; same*

### 2. Write the Conditions

1. Random Sample (2-sample with 1-variable)
2. All expected values  $> 5$
3. Counted data for observed

### 3. Write the Equation

$$\chi^2 = \frac{(O - E)^2}{E}$$

$\chi^2$  = How far away a distribution is from the expected/Null  
 $O$  = observed Value-the value from the sample  
 $E$  = the expected or predicted value.  
 $n$  = the number of categories

### 4. Calculate & List Expected Values

- 2<sup>nd</sup> Matrix → Edit → [A]
- Enter Matrix size (row x column) do not include totals.
- Enter observed Data
- 2<sup>nd</sup> Quit
- STAT TESTS
- C:  $\chi^2$ -Test
- 2nd matrix → Edit ↓ [B] (these are the expected values)

### 5. Plug values into the equation (1<sup>st</sup> and last Expected values)

$$\chi^2 = \frac{(O-E)^2}{E} + \dots + \frac{(O-E)^2}{E}$$

### 6. Calculate the $\chi^2$ and the p-value

- STAT TESTS
- C:  $\chi^2$ -Test
- df =  $(r-1)(c-1)$  ( $r$  = rows &  $c$  = columns)

### 7. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 8. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the distributions of \_\_\_\_\_ and \_\_\_\_\_ are the same.  
2<sup>nd</sup> sample 1<sup>st</sup> sample

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that the distributions of \_\_\_\_\_ and \_\_\_\_\_ are not the same.  
2<sup>nd</sup> sample 1<sup>st</sup> sample

## Notes: Chi-Square Test of Homogeneity

**Scenario 1:** In a large city a group of AP Statistics student work together on a project to determine which group of school employees has the greatest proportion who are satisfied with their jobs. In independent simple random samples of 100 teachers, 60 administrators, 45 custodians, and 55 secretaries, the numbers satisfied with their jobs were found to be 82, 38, 34, and 36 respectively. Is there evidence that employee job satisfaction varies by job category?

The observed counts are as follows:

	Satisfied	Not Satisfied	Totals
Teachers	82/73.077	18/26.923	100
Administrators	38/43.846	22/16.154	60
Custodians	34/32.885	11/12.115	45
Secretaries	36/40.192	19/14.808	55
Totals	190	70	260

**Null  $H_0$ :** The proportion of satisfied employees is the same for each job type

**Alternative  $H_A$ :** The proportion of satisfied employees is Not the same for each job type

**Conditions:**

1. **Independent Random Samples Given**
2. **All expected values > 5 See table**
3. **Counted data for observed all are whole numbers**

$$\chi^2 = \frac{(O-E)^2}{E} \quad df = (4-1)(2-1) = 3$$

$$\chi^2 = \frac{(73.077-82)^2}{73.077} + \dots + \frac{(19-14.808)^2}{14.808} = 8.707 \rightarrow P(\chi^2 > 8.707) = .0334$$

P-value                      alpha  
0.0334    less than    .05    We reject the Null

Our p-value is .0334. We reject the Null. There is sufficient evidence at  $\alpha = .05$  to suggest that the proportion of satisfied employees is Not the same for each job type.

**Conclusion:** We believe that the proportion of satisfied employees is Not the same for each job type

## Recipe for Success: Chi-Square Test of Homogeneity

### 1. Write your Hypothesis

- Null  $H_0$ : The distribution of \_\_\_\_\_ and \_\_\_\_\_ are the same.  
1<sup>st</sup> sample 2<sup>nd</sup> sample
- Alternative  $H_A$ : The distribution of \_\_\_\_\_ and \_\_\_\_\_ are not the same  
1<sup>st</sup> sample 2<sup>nd</sup> sample

**Key words:** *distribution are the same; same*

### 2. Write the Conditions

1. Random Sample (2-sample with 1-variable)
2. All expected values  $> 5$
3. Counted data for observed

### 3. Write the Equation

$$\chi^2 = \frac{(O - E)^2}{E}$$

$\chi^2$  = How far away a distribution is from the expected/Null  
 $O$  = observed Value-the value from the sample  
 $E$  = the expected or predicted value.  
 $n$  = the number of categories

### 4. Calculate & List Expected Values

- 2<sup>nd</sup> Matrix → Edit → [A]
- Enter Matrix size (row x column) do not include totals.
- Enter observed Data
- 2<sup>nd</sup> Quit
- STAT TESTS
- C:  $\chi^2$ -Test
- 2nd matrix → Edit ↓ [B] (these are the expected values)

### 5. Plug values into the equation (1<sup>st</sup> and last Expected values)

$$\chi^2 = \frac{(O-E)^2}{E} + \dots + \frac{(O-E)^2}{E}$$

### 6. Calculate the $\chi^2$ and the p-value

- STAT TESTS
- C:  $\chi^2$ -Test
- df =  $(r-1)(c-1)$  ( $r$  = rows &  $c$  = columns)

### 7. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 8. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the distributions of \_\_\_\_\_ and \_\_\_\_\_ are the same.  
2<sup>nd</sup> sample 1<sup>st</sup> sample

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that the distributions of \_\_\_\_\_ and \_\_\_\_\_ are not the same.  
2<sup>nd</sup> sample 1<sup>st</sup> sample





## Recipe for Success: Chi-Square Test of Homogeneity

### 1. Write your Hypothesis

- Null  $H_0$ : The distribution of \_\_\_\_\_ and \_\_\_\_\_ are the same.  
1<sup>st</sup> sample 2<sup>nd</sup> sample
- Alternative  $H_A$ : The distribution of \_\_\_\_\_ and \_\_\_\_\_ are not the same  
1<sup>st</sup> sample 2<sup>nd</sup> sample

**Key words:** *distribution are the same; same*

### 2. Write the Conditions

1. Random Sample (2-sample with 1-variable)
2. All expected values  $> 5$
3. Counted data for observed

### 3. Write the Equation

$$\chi^2 = \frac{(O - E)^2}{E}$$

$\chi^2$  = How far away a distribution is from the expected/Null  
 $O$  = observed Value-the value from the sample  
 $E$  = the expected or predicted value.  
 $n$  = the number of categories

### 4. Calculate & List Expected Values

- 2<sup>nd</sup> Matrix → Edit → [A]
- Enter Matrix size (row x column) do not include totals.
- Enter observed Data
- 2<sup>nd</sup> Quit
- STAT TESTS
- C:  $\chi^2$ -Test
- 2nd matrix → Edit ↓ [B] (these are the expected values)

### 5. Plug values into the equation (1<sup>st</sup> and last Expected values)

$$\chi^2 = \frac{(O-E)^2}{E} + \dots + \frac{(O-E)^2}{E}$$

### 6. Calculate the $\chi^2$ and the p-value

- STAT TESTS
- C:  $\chi^2$ -Test
- **df = (r-1)(c-1)** (r = rows & c = columns)

### 7. State the Decision

- The p-value is \_\_\_\_\_
- If the p-value is less than alpha, Reject the Null
- If the p-value is greater than alpha, Fail to reject the Null

### 8. Write the Conclusion

**Reject the Null:** Our p-value is \_\_\_\_\_. We reject the Null. There is sufficient evidence at alpha = \_\_\_\_\_ to suggest that the distributions of \_\_\_\_\_ and \_\_\_\_\_ are the same.  
2<sup>nd</sup> sample 1<sup>st</sup> sample

**Fail to Reject the Null:** Our p-value is \_\_\_\_\_. We Fail to reject the Null. There is not sufficient evidence at alpha = \_\_\_\_\_ to suggest that the distributions of \_\_\_\_\_ and \_\_\_\_\_ are not the same.  
2<sup>nd</sup> sample 1<sup>st</sup> sample

## Notes: Chi-Square Test of Homogeneity

**Scenario 3:** One **random sample** indicates that the number of televisions that American families have is distributed (this is the **given** distribution for the American population) as follows:

Number of Televisions	Percent	Frequency	
0	10	60	63
1	16	96	107.5
2	55	330	335
3	11	66	63
4 or more	8	48	31.5

A second **random sample** in the far western United States resulted in the following data:

Number of Televisions	Frequency	
0	66	63
1	119	107.5
2	340	335
3	60	63
4 or more	15	31.5

At the 1% significance level, does it appear that the distributions of televisions are the same in the far western United States and the American population as a whole?

**Null  $H_0$ :** The US and the far west have the same proportion of televisions per household

**Alternative  $H_A$ :** The US and the far west do not have the same proportion of televisions per household

**Conditions:**

1. 2 Random Samples Given
2. All expected values > 5 See table
3. Counted data for observed all are whole numbers

$$\chi^2 = \frac{(O-E)^2}{E} \quad df = (5-1)(2-1) = 4$$

$$\chi^2 = \frac{(60-63)^2}{63} + \dots + \frac{(15-31.5)^2}{31.5} = 20.47 \rightarrow P(\chi^2 > 20.47) = .0004$$

**P-value**                      **alpha**  
**0.0004**    less than    **.01**    **We reject the Null**

Our p-value is .0004. We reject the Null. There is sufficient evidence at  $\alpha = .01$  to suggest that the US and the far west do not have the same proportion of televisions per household

**Conclusion:** We believe that the US and the far west do not have the same proportion of televisions per household.