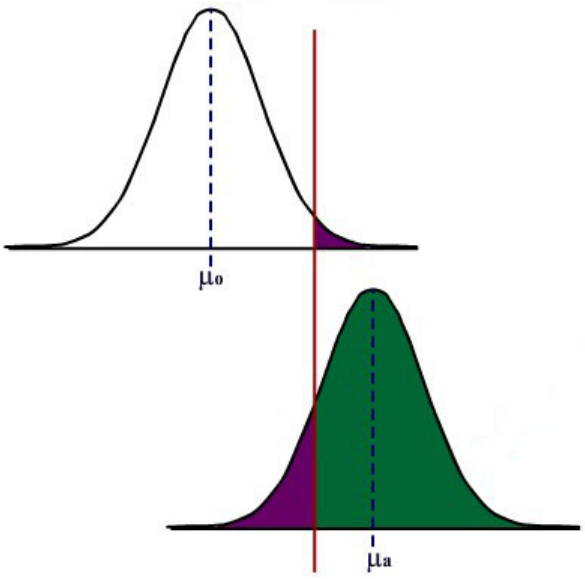
DE ANZA COLLEGE – DEPARTMENT OF MATHEMATICS

Inferential Statistics and Probability

A Holistic Approach

Maurice A. Geraghty

1/1/2017



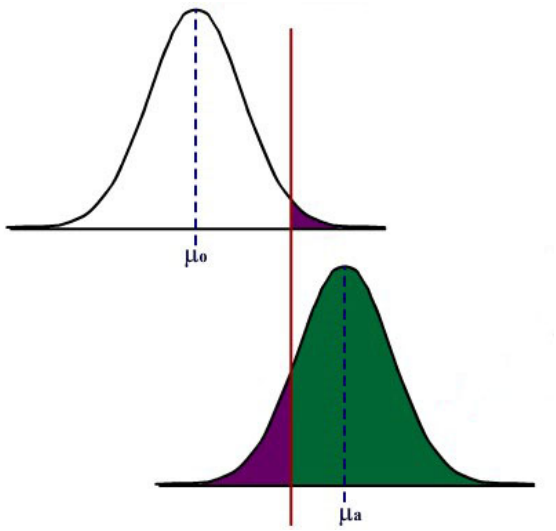
Material for an introductory lower division course in Probability and Statistics

Inference and Hypothesis Testing – A Holistic Approach

for an Introductory Lower Division Course in Probability and Statistics

Maurice A. Geraghty, De Anza College

January 1, 2017



0. Introduction - A Classroom Story and an Inspiration

Several years ago, I was teaching an introductory Statistics course at De Anza College where I had several achieving students who were dedicated to learn the material and who frequently asked me questions during class and office hours. Like many students, they were able to understand the material on descriptive statistics and interpreting graphs. Unlike many introductory Statistics students, they had excellent math and computer skills and went on to master probability, random variables and the Central Limit Theorem.

However, when the course turned to inference and hypothesis testing, I watched these students' performance deteriorate. One student asked me after class to again explain the difference between the Null and Alternative Hypotheses. I tried several methods, but it was clear these students never really understood the logic or the reasoning behind the procedure. These students could easily perform the calculations, but they had difficulty choosing the correct model, setting up the test, and stating the conclusion.

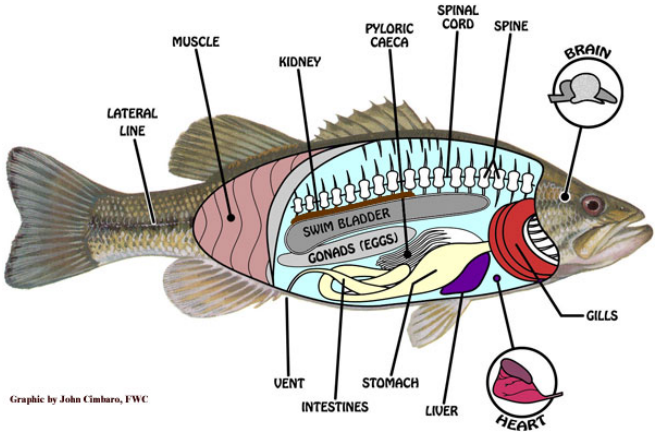
These students, (to their credit) continued to work hard; they wanted to understand the material, not simply pass the class. Since these students had excellent math skills, I went deeper into the explanation of Type II error and the statistical power function. Although they could compute power and sample size for different criteria, they still didn't conceptually understand hypothesis testing.

On my long drive home, I was listening to National Public Radio's *Talk of the Nation*[1] where there was a discussion on the difference between the reductionist and holistic approaches to the sciences, which the commentator described as the western tradition vs. the eastern tradition. The reductionist or western method of analyzing a problem, mechanism or phenomenon is to look at the component pieces of the system being studied. For example, a nutritionist breaks a potato down into vitamins, minerals, carbohydrates, fats, calories, fiber and proteins. Reductionist analysis is prevalent in all the sciences, including Inferential Statistics and Hypothesis Testing.

Holistic or eastern tradition analysis is less concerned with the component parts of a problem, mechanism or phenomenon but instead how this system operates as a whole, including its surrounding environment. For example, a holistic nutritionist would look at the potato in its environment: when it was eaten, with what other foods, how it was grown, or how it was prepared. In holism, the potato is much more than the sum of its parts.

Consider these two renderings of fish:

The first image is a drawing of fish anatomy by John Cimbaro used by the La Crosse Fish Health Center.[2] This drawing tells us a lot about how a fish is constructed, and where the vital organs are located. There is much detail given to the scales, fins, mouth and eyes.



The second image is a watercolor by the Chinese artist Chen Zheng-Long[3]. In this artwork, we learn very little about fish anatomy seeing only minimalistic eyes, scales and fins. However, the artist shows how fish are social creatures, how their fins move to swim and the type of plants they like. Unlike the first drawing, we learn much more about the interaction of the fish in its surrounding environment and much less about how a fish is built.

This illustrative example shows the difference between reductionist and holistic analyses. Each rendering teaches something important about the fish: The reductionist drawing of the fish anatomy helps explain how a fish is built and the holistic watercolor helps explain how a fish relates to its environment. Both the reductionist and holistic methods add to knowledge and understanding, and both philosophies are important. Unfortunately, much of Western science has been dominated by the reductionist philosophy, including the backbone of the scientific method, Inferential Statistics.

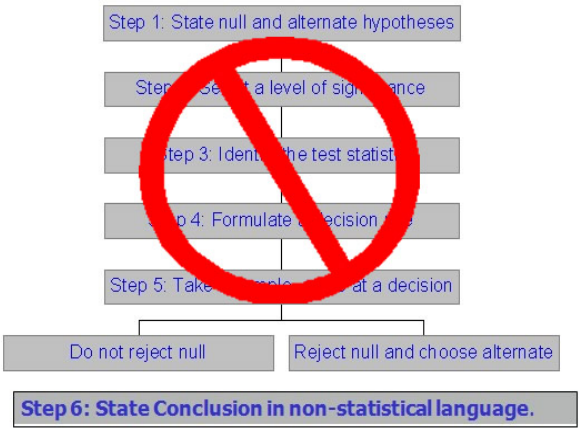
Although science has traditionally been reluctant to embrace, often hostile to including holistic philosophy in the scientific method, there have been many who now support a multicultural or multi-philosophical approach. In his book Holism and Reductionism in Biology and Ecology[4], Looijen claims that "holism and reductionism should be seen as mutually dependent, and hence co-operating research programs than as conflicting views of nature or of relations between sciences." Holism develops the "macro-laws" that reductionism needs to "delve deeper" into understanding or explaining a concept or phenomena. I believe this claim applies to the study of Statistics as well.

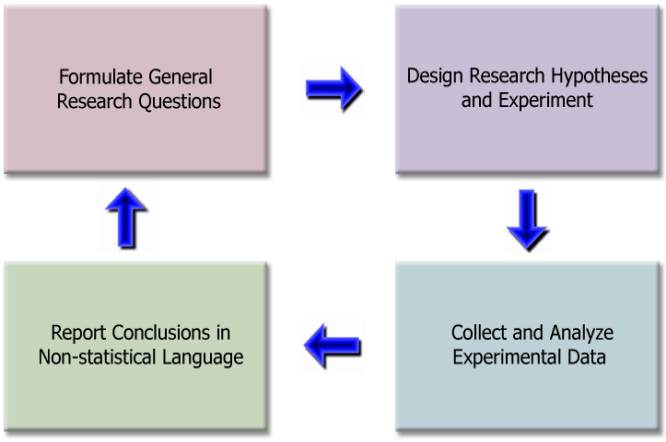
I realize that the problem of my high-achieving students being unable to comprehend hypothesis testing could be cultural – these were international students who may have been schooled under a more holistic philosophy. The Introductory Statistics curriculum and most texts give an incomplete explanation of the logic of Hypothesis Testing, eliminating or barely explaining such topics as Power, the consequence of Type II error or Bayesian alternatives. The problem is how to supplement an Introductory Statistics course with a holistic philosophy without depriving the students of the required reductionist course curriculum – all in one quarter or semester!

I believe it is possible to teach the concept of Inferential Statistics holistically. This course material is a result of that inspiration, which was designed to supplement, not replace, a traditional course textbook or workbook. This supplemental material includes:

- Examples of deriving research hypotheses from general questions and explanatory conclusions consistent with the general question and test results.
- An in-depth explanation of statistical power and type II error.

- Techniques for checking that validity of model assumptions and indentifying potential outliers using graphs and summary statistics.
- Replacement of the traditional step-by-step "cookbook" for hypothesis testing with interrelated procedures.
- De-emphasis of algebraic calculations in favor of a conceptual understanding using computer software to perform tedious calculations.
- Interactive Flash animations to explain the Central Limit Theorem, inference, confidence intervals, and the general hypothesis testing model including Type II error and power.
- PowerPoint Slides of the material for classroom demonstration.
- Excel Data sets for use with computer projects and labs.





This material is limited to one population hypothesis testing but could easily be extended to other models. My experience has been that once students understand the logic of hypothesis testing, the introduction of new models is a minor change in the procedure.

The Blind Man and the Elephant

This old story from China or India was made into the poem *The Blind Man and the Elephant* by John Godfrey Saxe[5]. Six blind men find excellent empirical evidence from different parts of the elephant and all come to reasoned inferences that match their observations. Their research is flawless and their conclusions are completely wrong, showing the necessity of including holistic analysis in the scientific process.

Here is the poem in its entirety:

> It was six men of Indostan, to learning much inclined,
> who went to see the elephant (Though all of them were blind),
> that each by observation, might satisfy his mind.

The first approached the elephant, and, happening to fall,
against his broad and sturdy side, at once began to bawl:
"God bless me! but the elephant, is nothing but a wall!"

The second feeling of the tusk, cried: "Ho! what have we here,
so very round and smooth and sharp? To me tis mighty clear,
this wonder of an elephant, is very like a spear!"

The third approached the animal, and, happening to take,
the squirming trunk within his hands, "I see," quoth he,
the elephant is very like a snake!"

The fourth reached out his eager hand, and felt about the knee:
"What most this wondrous beast is like, is mighty plain," quoth he;
"Tis clear enough the elephant is very like a tree."

The fifth, who chanced to touch the ear, Said; "E'en the blindest man
can tell what this resembles most; Deny the fact who can,
This marvel of an elephant, is very like a fan!"

The sixth no sooner had begun, about the beast to grope,
than, seizing on the swinging tail, that fell within his scope,
"I see," quothe he, "the elephant is very like a rope!"

And so these men of Indostan, disputed loud and long,
each in his own opinion, exceeding stiff and strong,
Though each was partly in the right, and all were in the wrong!

So, oft in theologic wars, the disputants, I ween,
tread on in utter ignorance, of what each other mean,
and prate about the elephant, not one of them has seen!

-John Godfrey Saxe

What can go wrong in research - two stories

The first story is about a drug that was thought to be effective in research, but was pulled from the market when it was found to be ineffective in practice.

FDA Orders Trimethobenzamide Suppositories Off the market[6]

FDA today ordered makers of unapproved suppositories containing trimethobenzamide hydrochloride to stop manufacturing and distributing those products.

Companies that market the suppositories, according to FDA, are Bio Pharm, Dispensing Solutions, G&W Laboratories, Paddock Laboratories, and Perrigo New York. Bio Pharm also distributes the products, along with Major Pharmaceuticals, PDRX Pharmaceuticals, Physicians Total Care, Qualitest Pharmaceuticals, RedPharm, and Shire U.S. Manufacturing.

FDA had determined in January 1979 that trimethobenzamide suppositories lacked "substantial evidence of effectiveness" and proposed withdrawing approval of any NDA for the products.

"There's a variety of reasons" why it has taken FDA nearly 30 years to finally get the suppositories off the market, Levy said.

At least 21 infant deaths have been associated with unapproved carbinoxamine-containing products, Levy noted.

Many products with unapproved labeling may be included in widely used pharmaceutical reference materials, such as the *Physicians' Desk Reference*, and are sometimes advertised in medical journals, he said.

Regulators urged consumers using suppositories containing trimethobenzamide to contact their health care providers about the products.

The second story is about promising research that was abandoned because the test data showed no significant improvement for patients taking the drug.

Drug Found Ineffective Against Lung Disease[7]

Treatment with interferon gamma-1b (Ifn-g1b) does not improve survival in people with a fatal lung disease called idiopathic pulmonary fibrosis, according to a study that was halted early after no benefit to participants was found.

Previous research had suggested that Ifn-g1b might benefit people with idiopathic pulmonary fibrosis, particularly those with mild to moderate disease.

The new study included 826 people, ages 40 to 79, who lived in Europe and North America. They were given injections of either 200 micrograms of Ifn-g1b (551 people) or a placebo (275) three times a week.

After a median of 64 weeks, 15 percent of those in the Ifn-g1b group and 13 percent in the placebo group had died. Symptoms such as flu-like illness, fatigue, fever and chills were more common among those in the Ifn-g1b group than in the placebo group. The two groups had similar rates of serious side effects, the researchers found.

"We cannot recommend treatment with interferon gamma-1b since the drug did not improve survival for patients with idiopathic pulmonary fibrosis, which refutes previous findings from subgroup analyses of survival in studies of patients with mild-to-moderate physiological impairment of pulmonary function," Dr. Talmadge E. King Jr., of the University of California, San Francisco, and colleagues wrote in the study published online and in an upcoming print issue of *The Lancet.*

The negative findings of this study "should be regarded as definite, [but] they should not discourage patients to participate in one of the several clinical trials currently underway to find effective treatments for this devastating disease," Dr. Demosthenes Bouros, of the Democritus University of Thrace in Greece, wrote in an accompanying editorial.

> Bouros added that people deemed suitable "should be enrolled early in the transplantation list, which is today the only mode of treatment that prolongs survival."

Although these are both stories of failures in using drugs to treat diseases, they represent two different aspects of hypothesis testing. In the first story, the suppositories were thought to effective in treatment from the initial trials, but were later shown to be ineffective in the general population. This is an example of what statisticians call **Type I Error**, supporting a hypothesis (the suppositories are effective) that later turns out to be false.

In the second story, researchers chose to abandon research when the interferon was found to be ineffective in treating lung disease during clinical trials. Now this may have been the correct decision, but what if this treatment was truly effective and the researchers just had an unusual group of test subjects? This would be an example of what statisticians call **Type II Error**, failing to support a hypothesis (the interferon is effective) that later turns out to be true. Unlike the first story, we will never get to find out the answer to this question since the treatment will not be released to the general public.

In a traditional Introductory Statistics course, very little time is spent analyzing the potential error shown in the second story. However, both types of error are important and will be explored in this course material.

1. Populations and Sampling

Not available yet - refer to slides and other textbooks

2. Displaying and Analyzing Data with Graphs

Not available yet - refer to slides and other textbooks

3. Descriptive Statistics

Not available yet - refer to slides and other textbooks

4. Probability

4.1 What is Probability?

Rather than defining probability, let me give some real life examples:

> The Golden State Warriors are trailing the Cleveland Cavaliers by one point late in an important NBA game. Cleveland forward LeBron James fouls Golden State guard Stephen Curry with 1.4 seconds left in the game, meaning Curry will get to shoot 2 free throws. What is the probability the Warriors will win the game?
>
> Thuy is an actress and auditions for a starring role in a Broadway musical. The audition goes extremely well and the director says she did a great job, sings beautifully, and is perfect for the role. He promises to call her back the next day after auditions are completed. What is the probability Thuy will get the role in the musical?
>
> Robert is a student taking a Statistics class for the second time, after dropping the class in the prior quarter. He has a lot of math anxiety, but needs to pass the class to be able to transfer to San Jose State and continue his dream of becoming a psychologist. What is the probability he will successfully pass the class?
>
> Lupe goes to the doctor after having some pain in her lower back. Her family has a history of kidney problems, so the doctor decides to run some additional tests. What is the probability that Lupe has a kidney disorder requiring treatment?

In all of these examples, it is uncertain or unknown what the actual outcomes will be, however, we can make a guess as to whether each **outcome** is either more likely or less likely. We can quantify this by a value between 0 and 1, or between 0% and 100%. For example, maybe we say The Warriors have a good chance of winning the game since Curry is one of the best free throw shooters in the NBA, say 0.7 or 70% . Maybe Thuy (from her experience in auditioning) is less likely of getting the starring role, say 0.2 or 20%. These quantities are called **probabilities.**

> **Probability** is the measure of the **likelihood** that an **event A** will occur. This measure is a quantity between 0 (never) and 1 (always) and will be expressed as **P(A)** – read as "The probability event A occurs."

4.2 Types of Probability

Classical probability (also called Mathematical Probability) is determined by counting or by using a mathematical formula or model.

Examples:

The probability of getting a "Heads" when tossing a fair coin is 0.5 or 50%.

The probability of rolling a 4 on a fair six-sided die is 1/6, since all numbers are equally likely.

Empirical probability is based on the relative frequencies of historical data, studies or experiments.

Examples:

The probability Stephen Curry make a free throw is 90.8% based on the frequency of successes from all prior free throws.

The probability of a random student getting an A in a statistics class taught by Professor Nguyen is 22.8%, because grade records show that of the 1000 students who took her class in the past, 228 received an A.

In a study of 832 adults with colon cancer, an experimental drug reduced tumors in 131 patients. The probability that the experimental drug reduces colon cancer tumors is 131/832 or 15.7%.

Subjective probability is a “one-shot” educated guess based on anecdotal stories, intuition or a feeling as to whether an event is likely, unlikely or “50-50”. Subjective probability is often inaccurate.

Examples:

Although Robert is nervous about retaking the Statistics course after dropping the prior quarter, he is 90% sure he will pass the class because the website ratemyprofessor.com gave the instructor very positive reviews.

Jasmine believes that she will probably not like a new movie that is coming out soon because she is not a fan of the actor who is starring in the film. She is about 20% sure she will like the new movie.

No matter how probability is initially derived, the laws and rules of probability will be treated the same.

4.3 How to Calculate Classical Probability

We can use counting methods to determine classical probability. However, we need to be careful in our methods to be sure to get the correct answer.

An **Event** is a result of an experiment, usually referred to with a capital letter A, B, C, etc. Consider the experiment of flipping two coins. Then use the letter A to refer to the event xactlof getting exactly one head.

An **Outcome** is a result of the experiment that cannot be broken down into smaller events. Consider event A, getting exactly one head. Note there are two ways or outcomes to get one head in two tosses, by first getting a head then a tail, or by first getting a tail, then a head. Let’s write these distinct outcomes as HT and TH.

The **Sample Space** is the set of all possible outcomes of an experiment. In the experiment of flipping two coins, there are 4 possible outcomes which will be expressed in set notation.

Sample Space = { HH, HT, TH, TT }

We can now redefine an **Event** of an experiment to be a subset of the Sample Space. If event A is getting exactly one head in two coin tosses, then

A = { HT, TH }

After carefully listing the outcomes of the Sample Space and the outcomes of the event, we can then calculate the **probability** the event occurs.

Probability Event Occurs = number of outcomes in Event / number of outcomes in Sample Space

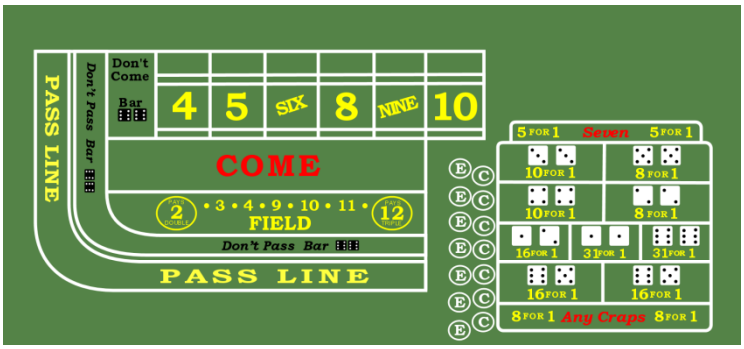
We will use the notation P(A) to mean the probability event A occurs.

In the example, the probability of getting exactly 1 head in two coin tosses is 2 out of 4 or 50%.

P(A) = 2/4 = 0.5 = 50%

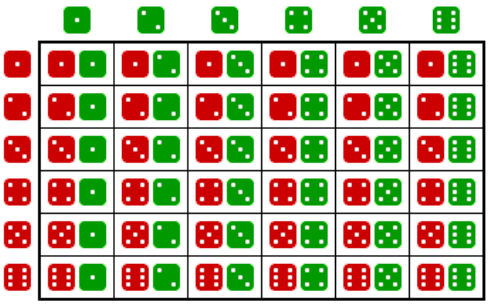
Example

In the casino game of craps, two dice are rolled and then totaled. There are many bets in craps, so let us consider **the Field bet**. In this bet, the player will win even money if a total of 3, 4, 9, 10 or 11 is rolled. If a total of 2 is rolled, the player will win double the original bet and if a total of 12 is rolled, the player will win the triple the original bet. If a total of 5, 6, 7 or 8 is rolled, the player loses the original bet.



At first glance, this looks like a winning bet or the player since the player wins on 7 different numbers and the casino only wins on 4 different numbers. However, we know that a casino always designs games so they have the advantage. Let us carefully use the counting methods to calculate the probability of a player winning the Field bet.

Let's first consider the task of listing the sample space of possible outcomes. Since there are two dice rolled, we can consider each outcome to be an ordered pair. There are 6 possible values for the first die and 6 possible ways for the second die, meaning there are 36 ordered pairs or outcomes. In the diagram, the red die is the first roll and the green die is the second roll.



$$Sample\ Space = \begin{Bmatrix} (1,1),(1,2),(1,3),(1,4),(1,5),(1,6), \\ (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \\ (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), \\ (6,1),(6,2),(6,3),(6,4),(6,5),(6,6) \end{Bmatrix}$$

Now define the event W to be the winning pairs of numbers in the Field bet, the pairs that add up to 2, 3, 4, 9, 10, 11 or 12. The winning pairs of numbers are shown in blue and the losing pairs are shown in red.

$$Sample\ Space = \begin{Bmatrix} (1,1),(1,2),(1,3),(1,4),(1,5),(1,6), \\ (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \\ (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), \\ (6,1),(6,2),(6,3),(6,4),(6,5),(6,6) \end{Bmatrix} \qquad W = \begin{Bmatrix} (1,1),(1,2),(1,3), \\ (2,1),(2,2), \\ (3,1),(3,6), \\ (4,5),(4,6), \\ (5,4),(5,5),(5,6), \\ (6,3),(6,4),(6,5),(6,6) \end{Bmatrix}$$

This means there are 16 outcomes out of 36 where the player wins. It's now easy to see the probability of winning is less than 50%, as the casino took the numbers that occur the most frequently.

$$P(W) = \frac{16}{36} = \frac{4}{9} \approx 44.4\%$$

As a final note on this example, you might recall that the casino pays double if the player rolls (1,1) or triple if the player rolls (6,6). Even taking this extra bonus into account, if a player makes 36 $100 bets, the casino will expect to win $2000 (20 numbers x $100) and the player will expect to win $1900 (16 numbers x $100, plus $100 extra for the 2 and $200 extra for the 12), meaning the player loses $100 for every $3600 bet, a house (casino) advantage of 2.78% .

Field Bet – Summary of 36 possible rolls	Amount won On $100 bets
$(1,1)$ (pays double)	+$200
$(6,6)$ (pays triple)	+$300
$(1,2),(1,3),(2,1),(2,2),(3,1),(3,6),(4,5),$ $(4,6),(5,4),(5,5),(5,6),(6,3),(6,4),(6,5)$	+$1400
$(1,4),(1,5),(1,6),(2,3),(2,4),(2,5),(2,6),(3,2),(3,3),(3,4),$ $(3,5),(4,1),(4,2),(4,3),(4,4),(5,1),(5,2),(5,3),(6,1),(6,2)$	-$2000
Overall expected result of 36 rolls ($3600 bet)	-$100

Just remember in the long run, the casino always wins.

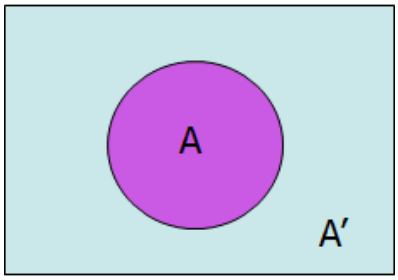
4.4 Rule of Complement

Sometimes it is difficult to calculate the probability that an event occurs, but it is much easier to calculate the probability that an event **does not** occur.

For example you may want to determine the probability that a student at California State University – East Bay majors in something other than Business. Instead of adding up all the non-Business major probabilities, it would be much easier to find the chance that a student at CSUEB majors in Business, say 21%. Then you would determine the probability that a student does not major in Business (all other students) is the remaining 79%

A' (read as "A-complement") is the event that event A does not occur. In that case, the **Rule of Complement** is:

$$P(A)+P(A')=1 \qquad P(A)=1-P(A') \qquad P(A')=1-P(A)$$



Example:

In a game, you must keep rolling a six-sided die until you get a six. What is the probability that you would need 2 or more rolls to get a six?

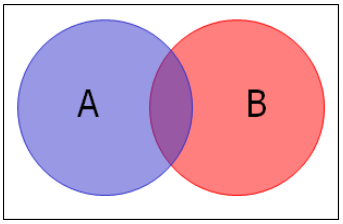
The event A is "2 or more rolls to get a six" which would be a very difficult probability to calculate - it's actually an infinite sum!

The event A' is "do not take 2 or more rolls to get a six" which is the same as saying "get a six on the first roll." That's a much easier probability to calculate, P(A') = 1/6.

So P(A) = 1 – P(A') = 1 – 1/6 = 5/6.

Therefore, the probability of needing two or more rolls to get a six is 5/6 or about 83.3%

4.5 Joint Probability and Additive Rule



Two or more events can be combined into **joint events** by using "or" statements or "and" statements.

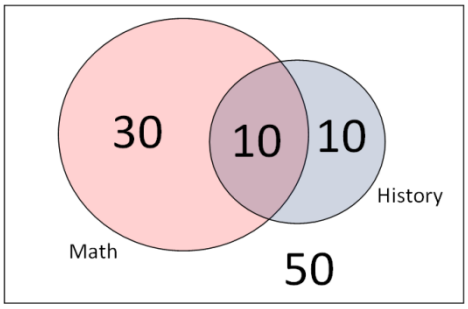
The **Union** of two events A and B is that either event A or B occur (or both). (the blue, red and purple parts of the Venn diagram shown to the right.)

The **Intersection** of two events A and B is that both events A or B occur. (the purple overlap of the Venn diagram shown to the right.)

Marginal Probability means the probability of a single event occurring.

Joint Probability means the probability of the union or intersection of multiple events occurring.

Example: In a group of 100 students a total 40 students take Math, a total of 20 students take History, and 10 students take both Math and History. (Note that these 10 students were already counted twice as being Math students and History students.) Find the marginal and joint probabilities.



Marginal Probabilities:

P(Math) = 40/100 = 0.4
P(History) = 20/100 = 0.2

Joint Probabilities:

P(Math and History) = 10/100 = 0.1 (this is the intersection of the two events)
P(Math or History) = 50/100 = 0.5 (this is the union of the two events)

We can make a rule for relating joint and marginal probabilities but noticing that we are double counting the outcomes in the intersection of two events when combining marginal probabilities from event each event. This is called the **Additive Rule**

The Additive Rule for Probability

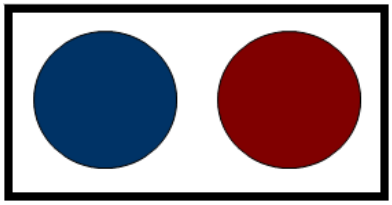
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Example:

Calculate the probability that a student is taking Math or History using the additive rule. Compare to the direct calculation in the prior example.

P(Math or History) = P(Math) + P(History) – P(Math or History)
P(Math or History) = 0.4 + 0.2 – 0.1 = 0.5

Mutually Exclusive means that two events A, B cannot both occur, that the intersection of two events has no possible outcomes.



The Additive Rule for Mutually Exclusive Events

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

Example:

500 students at a community are taking Spanish 1A in the Fall Quarter this year. 32 students are in Section 11 and 30 students are in Section 12. Find the probability a Spanish 1A student is in Sections 11 or 12.

Since students cannot be in two sections of the same class, the events Section 11 and Section 12 are mutually exclusive. P(Sec 11 or 12) = P(Sec 11) + P(Sec 12) = 32/500 + 30 /500 = 62/500 = 0.124.

4.6 Conditional Probability

Conditional Probability means the probability of an event A occurring given that another event B has already occurred. This probability is written as **P(A|B)** which is read as **P(A given B).**

Example:

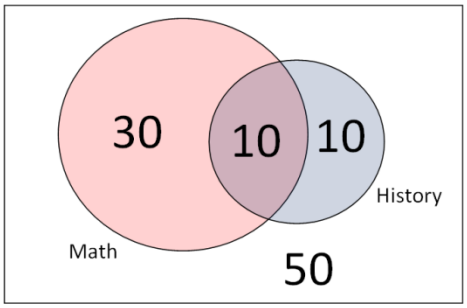
In the 2016 United States presidential election, Donald Trump received 46% of the total vote, Hillary Clinton received 48%, and other candidates received 6%. (Note: although Clinton received about 3 million more votes than Trump, the Electoral College determined the actual winner to be Trump.)

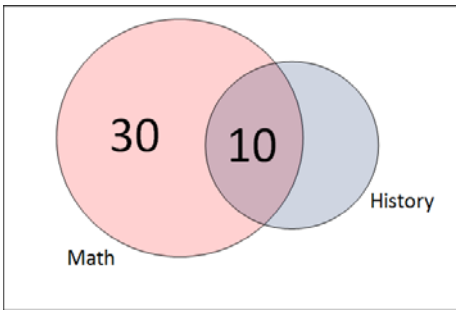
CNN conducted exit polls to determine how people voted based on demographic statistics, such as gender.[8] These exit polls showed that 53% of the voters were female and 47% of the voters were male. These two values are examples of marginal probabilities.

The polls also showed that Donald Trump received 41% of the female vote and 52% of the male vote. These two values are examples of conditional probability, where the condition is knowing the gender of the voter.

Events	Marginal Probabilities	Conditional Probabilities
T = Voter chooses Trump	P(T) = 0.46	P(T|F) = 0.41
F = Voter is Female	P(F) = 0.53	P(T|M) = 0.52
M = Voter is Male	P(M) = 0.47	

In calculating the probability of A given B, we only need to consider the elements of Event B instead of the entire sample space. Let us revisit the example of students taking Math and History. Suppose we wanted to calculate the probability a student who is taking math is also taking history. In this case we only need to consider the 40 students taking math as the sample space and the 10 students taking both math and history as the conditional event occurring.





P(History) = 20/100 = 0.20

P(History|Math) = 10/40 = 0.25

In this example, we used classical counting probability rules, but conditional probability can be calculating directly using known marginal and conditional probabilities.

Rules for Conditional Probability

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \qquad P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

P(History) = 10/100 = 0.10
P(Math and History) = 10/100 = 0.10
P(History|Math) = 0.10/0.40 = 0.25

Example:

Of all cell phone users in the US, 15% have a smart phone with AT&T. 25% of all cell phone users use AT&T. Given a selected cell phone user has AT&T, find the probability the user also has a smart phone.

Let A = AT&T subscriber. Let B = Smart Phone User

$$P(A) = 0.25 \quad P(A \text{ and } B) = 0.15 \quad P(A|B) = \frac{0.15}{0.25} = 0.60$$

This means 60% of all AT&T subscribers have smart phones.

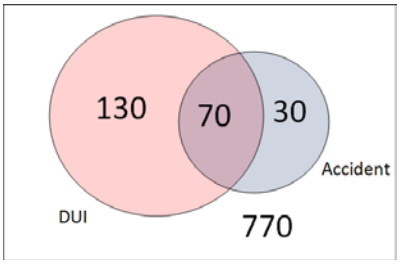
4.7 Contingency (Two-way) Tables

Contingency Tables, also known as cross tabulations, crosstabs or two-way tables, is a method of displaying the counts of the responses of two categorical variables from data.

Example:

1000 drivers were asked if they were involved in accident in the last year. They were also asked if during this time, they were DUI, driving under the influence of alcohol or drugs. The totals are summarized in a contingency table:

	Accident	No Accident	Total
DUI	70	130	200
Non- DUI	30	770	800
Total	100	900	1000



In the table, each column represents a choice for the accident question and each row represents a choice for the DUI question.

Marginal Probabilities can be determined form the contingency table by using the outside total values for each event divided by the total sample size.

- Probability a driver had an accident = P(A) = 100/1000 = 0.10
- Probability a driver was not DUI = P(D') = 1 - P(D) = 1 - 200/100 = 0.80

Joint Probabilities can be determined from the contingency table by using the inside values of the table divided by the total sample size.

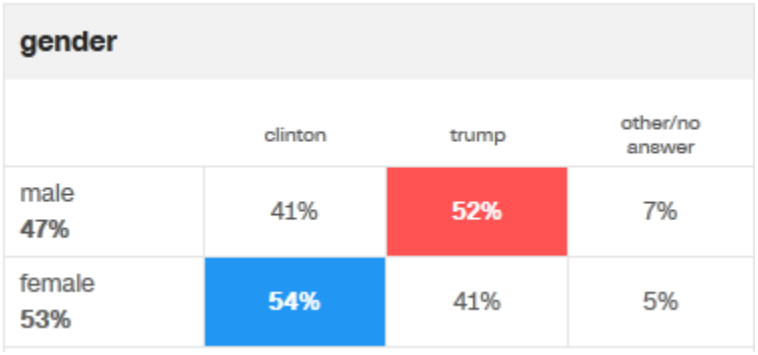
- Probability a driver had an accident **and** was DUI= P(A and D) = 70/1000 = 0.07
- Probability a driver had an accident **or** was DUI= P(A or D) = (100+200-70)/1000 = 0.23

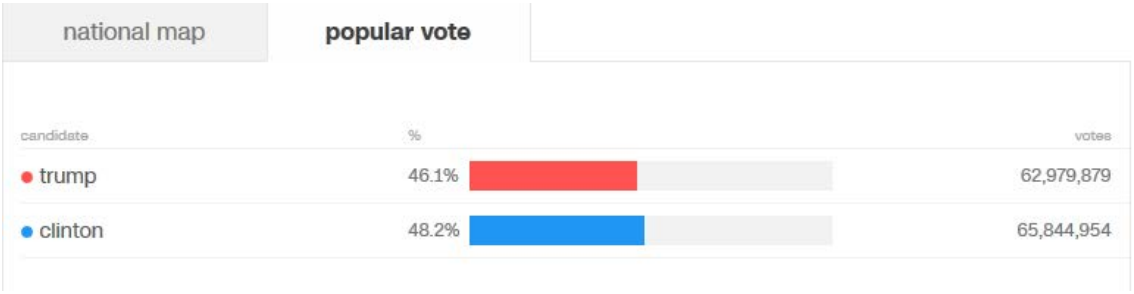
Conditional Probabilities can be determined from the contingency table by using the inside values of the table divided by the outside total value of the conditional event.

- Probability a driver was DUI **given** the driver had an accident = P(D|A) = 70/100 = 0.70
- Probability a DUI driver had an accident = P(A|D) = 70/200 = 0.35

Creating a two-table from reported probabilities

We can create a hypothetical two-way table from reported cross tabulated probabilities, such as the CNN exit poll for the 2016 presidential election:





Step 1: Choose a convenient total number. (This is called the **radix** of the table)

GENDER			
VOTED FOR	Female	Male	Total
Trump			
Clinton			
Other			
Total			10000

Radix chosen = 10000 random voters

Step 2: Determine the outside values of the table by multiplying the radix times the marginal probabilities for gender.

GENDER			
VOTED FOR	Female	Male	Total
Trump			
Clinton			
Other			
Total	5300	4700	10000

Total Female = (0.53)(10000) = 5300
Total Male = (0.47)(10000) = 4700

Step 3: Determine the inside values of the table by multiplying the appropriate gender total times the conditional probabilities from the exit polls

GENDER			
VOTED FOR	Female	Male	Total
Trump	2173	2444	
Clinton	2862	1927	
Other	265	329	
Total	5300	4700	10000

Trump Female = (0.41)(5300) = 2173
Clinton Female = (0.54)(5300) = 2862
Other Female = (0.05)(5300) = 265
Trump Male = (0.52)(4700) = 2444
Clinton Male = (0.41)(4700) = 1927
Other Male = (0.057)(4700) = 329

Step 4: Add each row to get the row totals.

GENDER			
VOTED FOR	Female	Male	Total
Trump	2173	2444	4617
Clinton	2862	1927	4789
Other	265	329	594
Total	5300	4700	10000

Trump = 2173 + 2444 = 4617
Clinton = 2862 + 1927 = 4789
Other = 265 + 329 = 594

From the last column, we can now get the marginal probabilities (which are slightly off from the actual vote due to rounding in the exit polls): Donald Trump received 46%, Hillary Clinton received 48% and other candidates received 6% of the total vote.

4.8 Multiplicative Rule and Tree Diagrams

Earlier, we learned about the additive rule for finding the joint probability of the Union of two events. There are corresponding multiplicative rule to find the probability of the Intersection of two events. Using algebra, thia rule can calculated directly from the Rules for Conditional Probability.

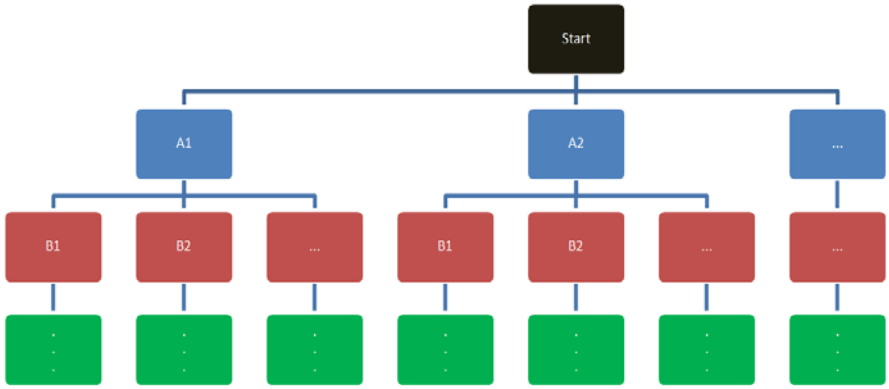
Multiplicative Rule of Probability

$$P(A \text{ and } B) = P(A) \times P(B \mid A)$$

$$P(A \text{ and } B) = P(B) \times P(A \mid B)$$

One useful way to express the Multiplicative Rule is by creating a **tree diagram**, a simple way to express a sequence of events.

The first level of branches connecting to the start are marginal probabilities, and all lower levels of branches are conditional probabilities. To find the probability of getting to the end of any last branch, multiply the probabilities all branches that connect back to Start.

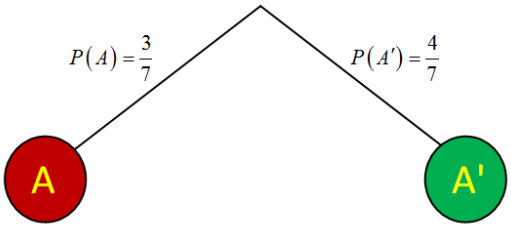


Example:

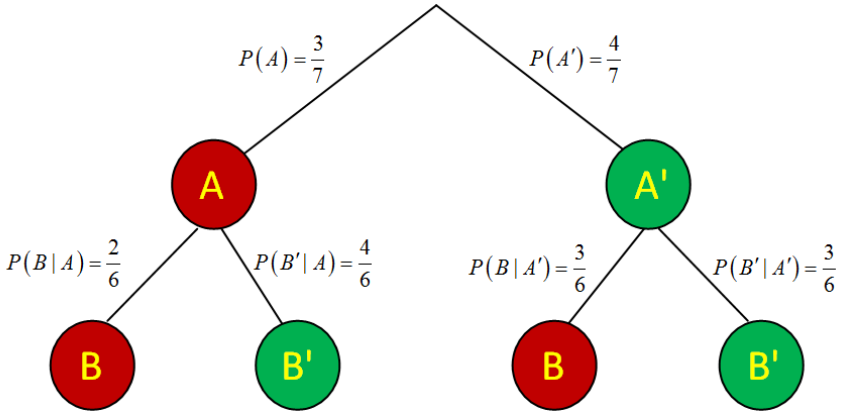
A box contains 4 green balls and 3 red balls. Two balls are drawn, one at a time, without replacement. Make a tree diagram and find the probability of choosing two red balls.

Let A be the event red on the first Draw and B be the event Red on second draw. Then in this example A' would be the event green (not red) on the first and B' would be the event green on the second draw.

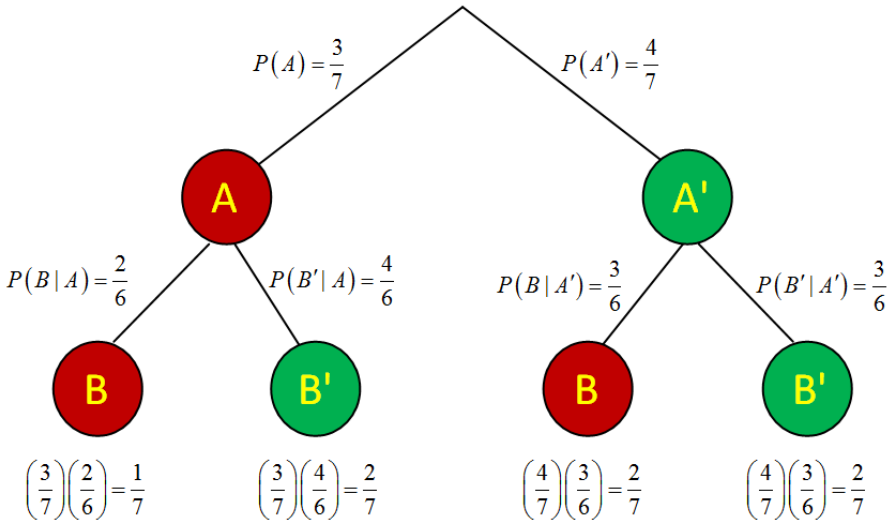
First, make a tree of the first draw and assign probabilities based on the number of balls in the box; 3 out of 7 are red and 4 out of 7 are green.



Next, conduct the second draw, assuming the ball chosen on the first draw is gone. For example, if the first draw was red, the chance of getting another red is 2 out of 6, since there are 2 remaining reds and 4 remaining greens. However, if the first draw was green, the chance of getting red is 3 out of 6.



Finally, use the multiplicative rule and multiply down the branch to get all joint probabilities. If you have constructed the tree diagram correctly, all of these probabilities must add to 1.



The probability of getting 2 red balls is 1/7 or approximately 0.143

Example:

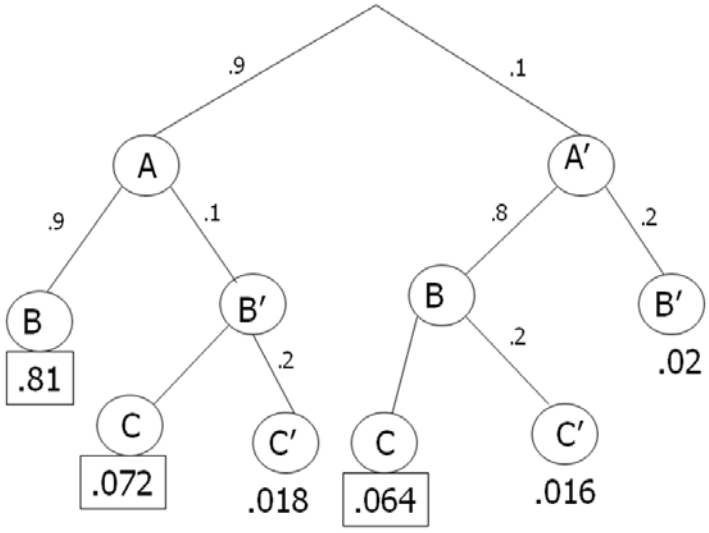
A Circuit has three linear switches. If at least two of the switches function, the Circuit will succeed. Each switch has a 10% failure rate if all are operating, and a 20% failure rate if one switch has already failed. Construct a tree diagram and find the probability the circuit will succeed.

Event A = first switch succeeds
Event A' = first switch fails

Event B = second switch succeeds
Event B' = second switch fails

Event C = third switch succeeds
Event C' = third switch fails

P(2 or more successes) = 0.81 + 0.072 + 0.064
= 0.946



The switch has a 94.6% chance of succeeding. Notice that we did not need a tie-breaking third branch for the cases of the first 2 switches succeeding, or the first 2 switches failing.

4.9 Independence

Two events are considered **independent** if the probability of one event occurring is not changed by knowing if the other event occurred or not. Events that are not independent are called dependent.

Here are examples of independent (unrelated) events:

- A fair coin flip comes up heads; the coin is flipped again and comes up heads.
- A student is unable to attend a math class at De Anza College; it rains today in New York City.
- A house in San Francisco starts on fire; on the same day a house in Dallas starts on fire.
- A patient is diagnosed with cancer; on the same day a patient is diagnosed with pneumonia

In these independent events, the probability of the second event occurring is not affected whether or not the first event occurs.

Examples of dependent (related) events

- A student gets an A on the first exam; the same student gets an A on the second exam.
- A person has never smoked; the same person gets lung cancer.
- An earthquake destroys a home in San Francisco; on the same day an earthquake destroys a home in Oakland.
- A student majors in computer science; the same student wants to work for Google.

In these dependent events, the probability of the second event occurring is affected whether or not the first event occurs:

- A student who gets an A on exam is more likely to get an A on another exam.
- A non-smoker is less likely to get lung cancer than a smoker.
- An single strong earthquake will affect homes all over the Bay Area.
- A computer science major is more likely to work for a tech company, like Google.

The mathematical definition of independent events mean that the marginal probability of the first event occurring is the same as the conditional probability of the first occurring given the second event occurred. We can then adjust the Multiplicative Rule to get three formulas, any of which can be used to test for independence:

If events A and B are **independent**, then the following statements are all true

$$P(A) = P(A \mid B)$$

$$P(B) = P(B \mid A)$$

$$P(A \text{ and } B) = P(A) \times P(B)$$

The last formula is particularly useful and can easily generalized to finding the joint probability of many independent events from looking at the simple marginal probabilities, making random sampling in statistical research so critical.

Example:

A fair coin is flipped ten times. Find the probability of getting heads on all 10 tosses.
Because the coin tosses are independent, the multiplicative rule requires only marginal probabilities:
$P(\text{all Heads}) = P(H)^{10} = 0.5^{10} = 0.0009766$

Example:

On Monday, there is a 10% chance your history instructor will have a surprise quiz. On the same day, there is a 20% chance that your Math instructor will also have a surprise quiz. No other class you are taking has surprise quizzes. What is the probability that you will have a least one surprise quiz on Monday? Assume that all events are independent.

Let H be the event "Surprise quiz in History" and M be the event "Surprise quiz in Math." Then use both the Additive Rule and the Multiplicative Rule for independent events.

P(H or M) = P(H) + P(M) - P(M and H)
P(H) = 0.10 P(M) = 0.20
P(H and M) = P(H) x P(M) = 0.10 x 0.20 = 0.02
P(H or M) = 0.10 + 0.20 - 0.02 =0.28

There is a 28% chance that there will be at least one surprise quiz on Monday.

Example:

1000 drivers were asked if they were involved in accident in the last year. They were also asked if during this time, they were DUI, driving under the influence of alcohol or drugs. Are the events "Driver was DUI" and "Driver was involved in an accident" independent or dependent events?

	Accident	No Accident	Total
DUI	70	130	200
Non- DUI	30	770	800
Total	100	900	1000

Let A be the event drive had an accident and D be the event driver was DUI. We can use any of the rules for independence answer this question. Let's show all three possible methods here, but in practice choose the most convenient formula given the provided data.

Use Formula 1
P(A) = 100/1000 = 0.10
P(A|D) = 70/200 =0.35
P(A) ≠ P(A|D)

Use Formula 2
P(D) = 200/1000 = 0.20
P(D|A) = 70/100 =0.70
P(D) ≠ P(D|A)

Use Formula 3
P(A) = 100/1000 = 0.10
P(D) = 200/1000 = 0.20
P(A and D) = 70/1000 = 0.07
P(A) x P(D) = (0.10)(0.20) = 0.02
P(A and D) ≠ P(A) x P(D)

"Driver was DUI" and "Driver was involved in an accident" are dependent events.

Example:

1000 drivers were asked if they were involved in accident in the last year. They were also asked if during this time, did they drive a domestic car or an import. Are the events "Driver drives a domestic car" and "Driver was involved in an accident" independent or dependent events?

	Accident	No Accident	Total
Domestic Car	60	540	600
Import Car	40	360	400
Total	100	900	1000

Let A be the event drive had an accident and D be the event driver drives a domestic car. Let's again show all three possible methods here, but in practice choose the most convenient formula given the provided data.

Use Formula 1
P(A) = 100/1000 = 0.10
P(A|D) = 60/600 =0.10
P(A) = P(A|D)

Use Formula 2
P(D) = 600/1000 = 0.60
P(D|A) = 60/100 =0.60
P(D) = P(D|A)

Use Formula 3
P(A) = 100/1000 = 0.10
P(D) = 600/1000 = 0.60
P(A and D) = 60/1000 = 0.06
P(A) x P(D) = (0.10)(0.60) = 0.06
P(A and D) = P(A) x P(D)

"Driver was DUI" and "Driver drives a domestic car" are independent events.

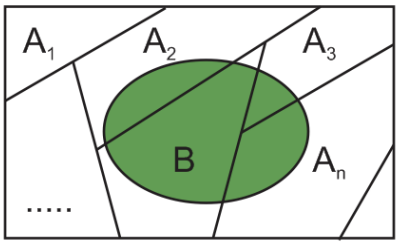
4.10 Changing the Conditionality and Bayesian Statistics

A trucking company is concerned that some of their drivers may be using amphetamine drugs to stay awake, exposing the company to lawsuits. They hire a testing agency to randomly test drivers. The marketing material for this testing agency claims that 99% of drivers who are using amphetamines will have a positive test result, so the company can be assured that any driver who tests positive will almost certainly be using the amphetamines.

This marketing material presented by the testing agency represents faulty reasoning. The 99% represents the probability that a driver tests positive given the driver is using amphetamines, while the claim was that the probability would be near-certain that a driver was using amphetamines given the test was positive. The conditionality has been incorrectly switched because in general: $P(A|B) \neq P(B|A)$.

To switch the conditionality requires several pieces of information and is often explained in statistics books by using Bayes' Theorem: If the sample space is the union of mutually events A_1, A_2, ..., A_n, then

$$P(A_i \mid B) = \frac{P(A_i) \times P(B \mid A_i)}{P(A_1) \times P(B \mid A_1) + P(A_2) \times P(B \mid A_2) + \cdots + P(A_n) \times P(B \mid A_n)}$$



A more straightforward approach to solving this type of problem to use techniques that have already been covered in this section:

- First construct a tree diagram.
- Second, create a Contingency Table using a convenient radix (sample size)
- From the Contingency table it is easy to calculate all conditional probabilities.

Example:

10% of prisoners in a Canadian prison are HIV positive. (This is also known in medical research as the **incidence rate**.) A test will correctly detect HIV 95% of the time, but will incorrectly "detect" HIV in non-infected prisoners 15% of the time (false positive). If a randomly selected prisoner tests positive, find the probability the prisoner is HIV+

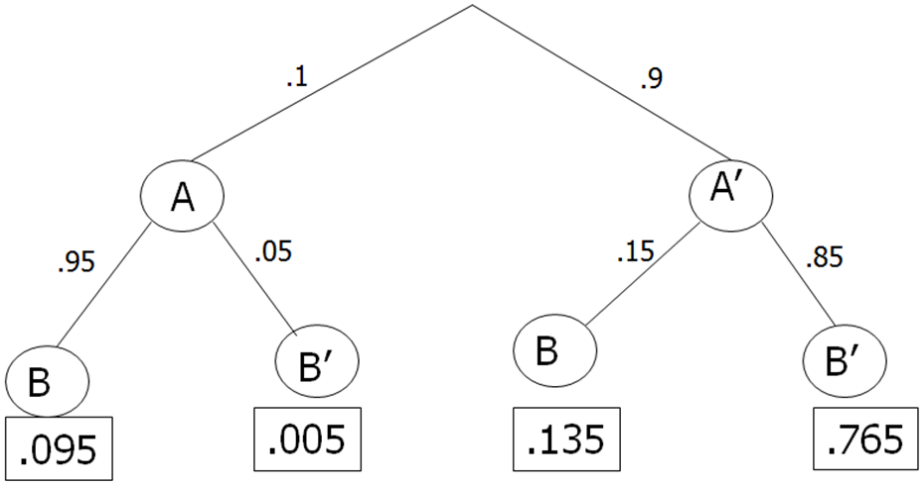
Let A be the event that a prisoner is HIV positive and B the event that a prisoner tests positive. Then A' would be the event that a prisoner is HIV negative and B' would be the event that the prisoner tests negative.

There are four possible outcomes in this probability model:

- True Positive (also known as in medical research as **sensitivity**.) - The prisoner correctly tests positive and is actually HIV positive.
- False Negative - The prisoner incorrectly tests negative and is actually HIV positive.
- False Positive - The prisoner incorrectly tests positive and is actually HIV negative.
- True Negative (also known as in medical research as **specificity**.) - The prisoner correctly tests negative and is actually is HIV negative.

From the information given, first construct a tree diagram.

P(A) = 0.10 P(A') = 1 - 0.10 = 0.90
P(B|A) = 0.95 P(B|A') = 0.15 P(B'|A) = 1 - 0.95 = 0.05 P(B'|A') = 1 - 0.15 = 0.85



Next, construct a contingency table. It is helpful to choose a convenient radix (sample size) like 10000 and multiply by each joint probability from the tree diagram:

- Samples in A and B = (.095)(10000) = 950
- Samples in A and B' = (.005)(10000) = 50
- Samples in A' and B = (.135)(10000) = 1350
- Samples in A' and B' = (.765)(10000) = 7650

	HIV+ A	HIV- A′	Total
Test+ B	950	1350	2300
Test- B′	50	7650	7700
Total	1000	9000	10000

To find the probability that a prisoner who tests positive really is HIV positive, find P(A|B):

$$P(A|B) = \frac{950}{2300} = 0.413$$

So the probability that a prisoner who tests positive really is HIV positive is only 41.3%. This result may seem unusual, but when the incidence rate is lower than the false positive rate, it is more likely that a positive result on a test will be incorrect.

This problem could have also been answered directly, but much less straightforward by using Bayes' Theroem:

$$P(B|A) = \frac{P(A)\times P(B|A)}{P(A)\times P(B|A) + P(A')\times P(B|A')}$$
$$= \frac{(0.10)(0.95)}{(0.10)(0.95)+(0.90)(0.85)} = 0.413$$

5. Discrete Random Variables

Not available yet - refer to slides and other textbooks

6. Continuous Random Variables

Not available yet - refer to slides and other textbooks

7. The Central Limit Theorem

7.1 Empirical Rule

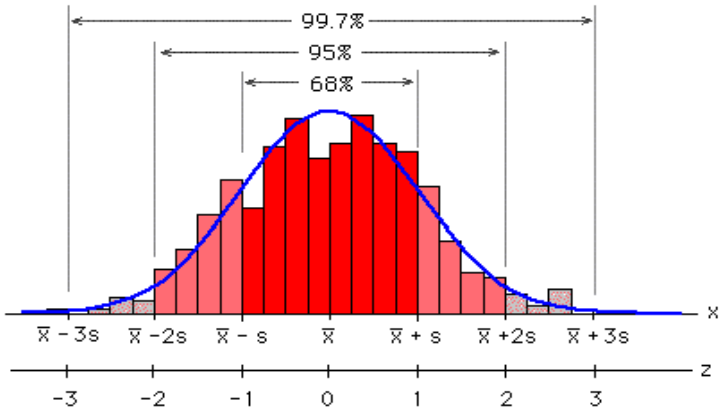
A student asked me about the distribution of exam scores after she saw her score of 87 out of 100. I told her the distribution of test scores were approximately bell-shaped with a mean score of 75 and a standard deviation of 10. Most people would have an intuitive grasp of the mean score as being the "average student's score" and would say this student did better than average. However, having an intuitive grasp of standard deviation is more challenging. The Empirical Rule is a helpful tool in explaining standard deviation.

The standard deviation is a measure of variability or spread from the center of the data as defined by the mean. The empirical rules states that for bell-shaped data:

68% of the data is within 1 standard deviation of the mean.

95% of the data is within 2 standard deviations of the mean.

99.7% of the data is within 3 standard deviations of the mean.



In the example, our interpretation would be:

68% of students scored between 65 and 85.
95% of students scored between 55 and 95.
99.7% of students scored between 45 and 105.

The student who scored an 87 would be in the upper 16% of the class, more than one standard deviation above the mean score.

7.2 The Z-score

Related to the Empirical Rule is the Z-score which measures how many standard deviations a particular data point is above or below the mean. Unusual observations would have a Z-score over 2 or under -2. Extreme observations would have Z-scores over 3 or under -3 and should be investigated as potential outliers.

Formula for Z-score: $$Z = \frac{X_i - \overline{X}}{s}$$

The student who received an 87 on the exam would have a Z-score of 1.2, meaning her score was well above average, but not highly unusual.

Interpreting Z-score for Several Students

Test Score	Z-score	Interpretation
87	+1.2	well above average
71	-0.4	slightly below average
99	+2.4	unusually above average
39	-3.6	extremely below average

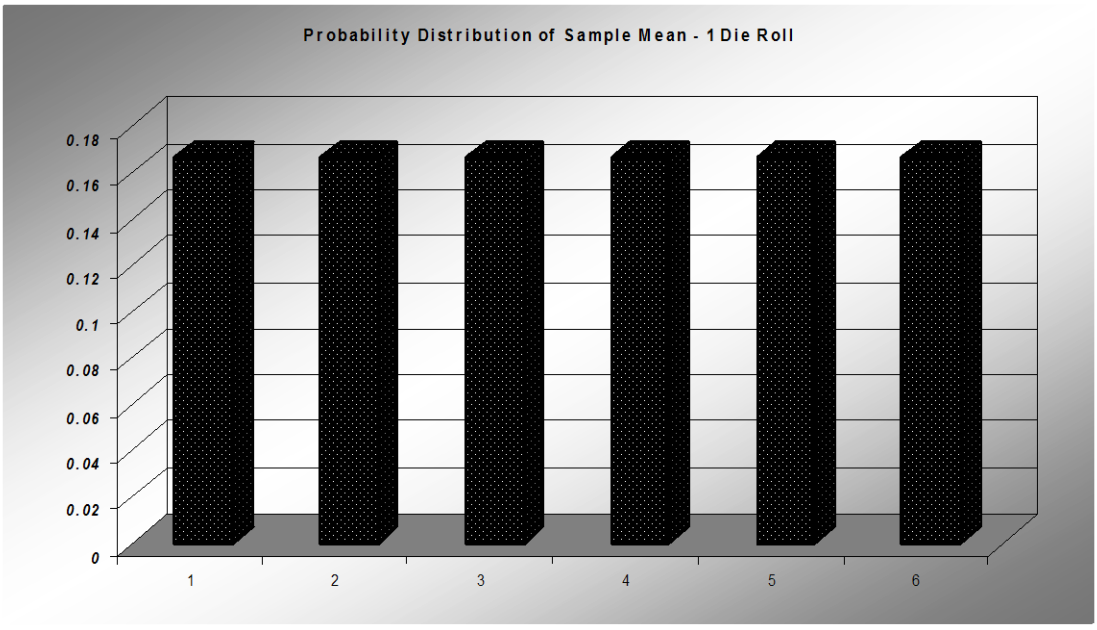
7.3 The Sample Mean as a Random Variable – Central Limit Theorem

In the section on descriptive statistics, we studied the sample mean, $\bar{X}$, as measure of central tendency. Now we want to consider $\bar{X}$ as a Random Variable.

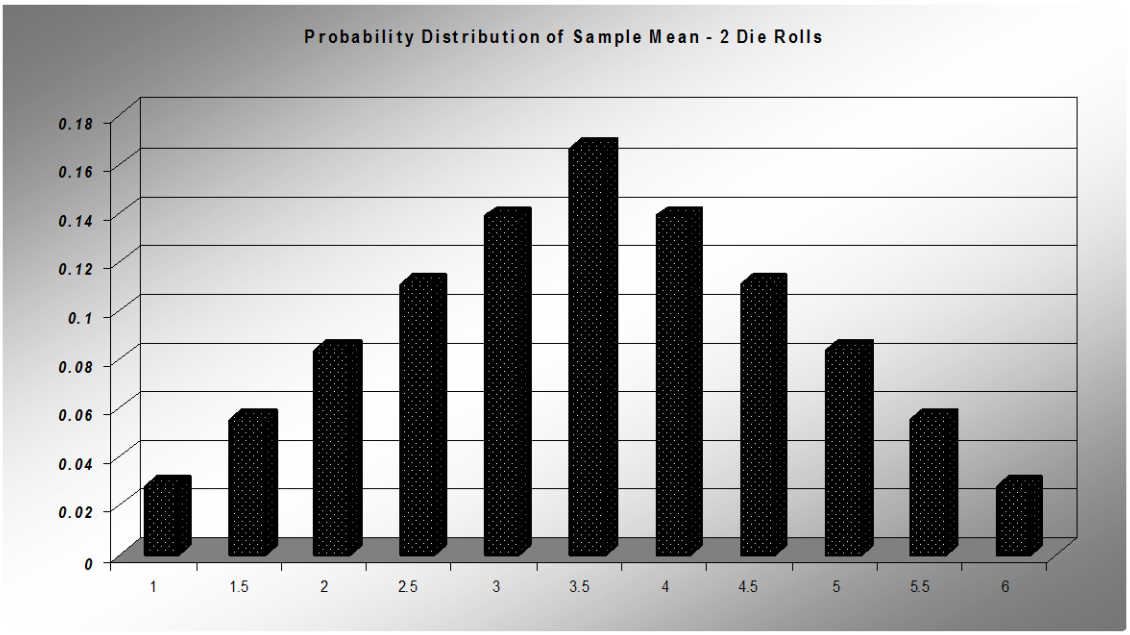
We start with a Random Sample X_1, X_2, …, X_n where each of the random variables X_i has the same probability distribution and are mutually independent of each other. The sample mean is a function of these random variables (add them up and divide by the sample size), so $\bar{X}$ is a random variable. So what is the Probability Distribution Function (PDF) of $\bar{X}$?

To answer this question, conduct the following experiment. We will roll samples of n dice, determine the mean roll, and create a PDF for different values of n.

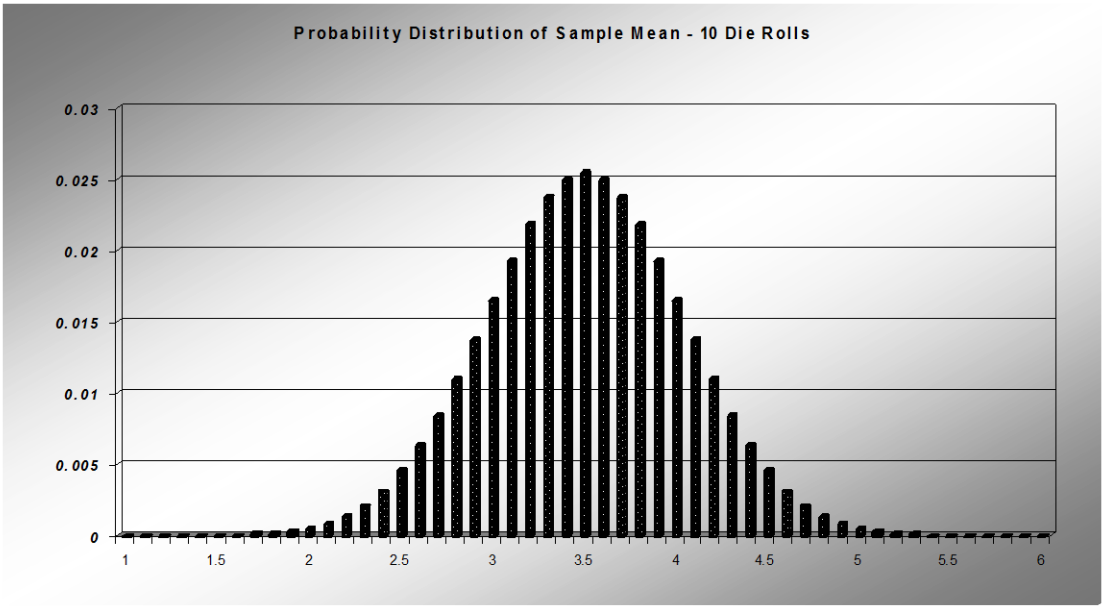
For the case n=1, the distribution of the sample mean is the same as the distribution of the random variable. Since each die has the same chance of being chosen, the distribution is rectangular shaped centered at 3.5:



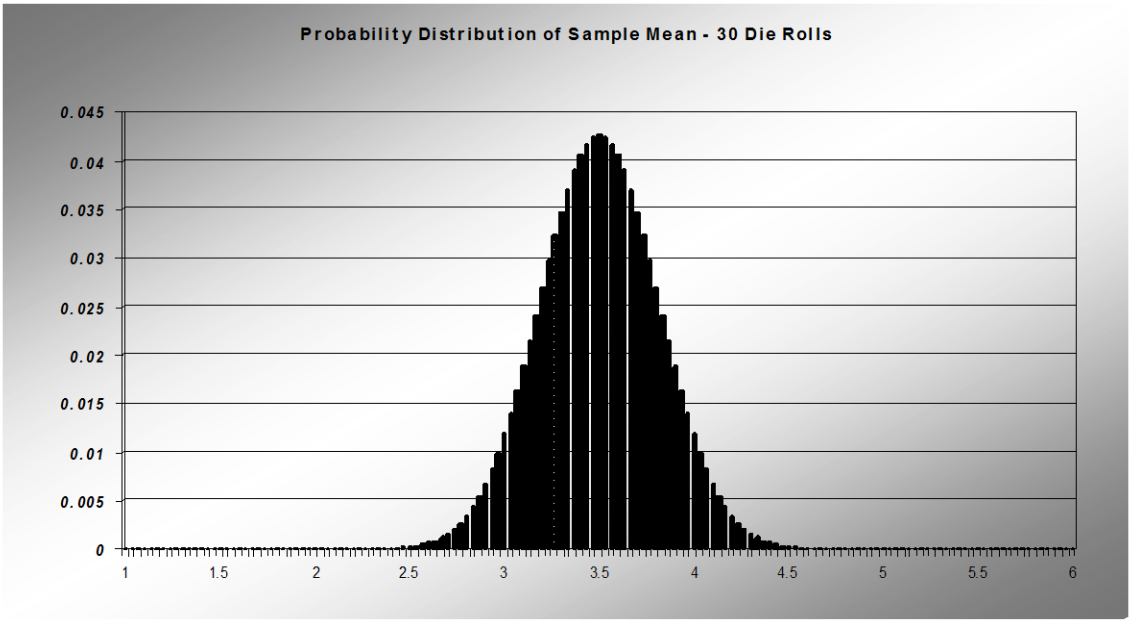
For the case n=2, the distribution of the sample mean starts to take on a triangular shape since some values are more likely to be rolled than others. For example, there six ways to roll a total of 7 and get a sample mean of 3.5, but only one way to roll a total of 2 and get a sample mean of 1. Notice the PDF is still centered at 3.5.



For the case n=10, the PDF of the sample mean now takes on a familiar bell shape that looks like a Normal Distribution. The center is still at 3.5 and the values are now more tightly clustered around the mean, implying that the standard deviation has decreased.



Finally, for the case n=30, the PDF continues to look like the Normal Distribution centered around the same mean of 3.5, but more tightly clustered than the prior example:



This die-rolling example demonstrates the Central Limit Theorem's three important observations about the PDF of $\bar{X}$ compared to the PDF of the original random variable.

1. The mean stays the same.
2. The standard deviation gets smaller.
3. As the sample size increase, the PDF of $\bar{X}$ is approximately Normal.

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample from a population that has a mean μ and a standard deviation σ, and n is sufficiently large then:

1. $\mu_{\bar{X}} = \mu$
2. $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
3. The Distribution of $\bar{X}$ is approximately Normal.

Combining all of the above into a single formula: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

where Z represents the Standard Normal Distribution.

This powerful result allows us to use the sample mean $\bar{X}$ as an estimator of the population mean μ. In fact, most inferential statistics practiced today would not be possible without the Central Limit Theorem.

Example:

The mean height of American men (ages 20-29) is $\mu = 69.2$ inches. If a random sample of 60 men in this age group is selected, what is the probability the mean height for the sample is greater than 70 inches? Assume $\sigma = 2.9$".

Due to the Central Limit Theorem, we know the distribution of the Sample will have approximately a Normal Distribution:

$$P(\overline{X} > 70) = P\left(Z > \frac{(70 - 69.2)}{2.9/\sqrt{60}}\right) = P(Z > 2.14) = 0.0162$$

Compare this to the much larger probability that one male chosen will be over 70 inches tall:

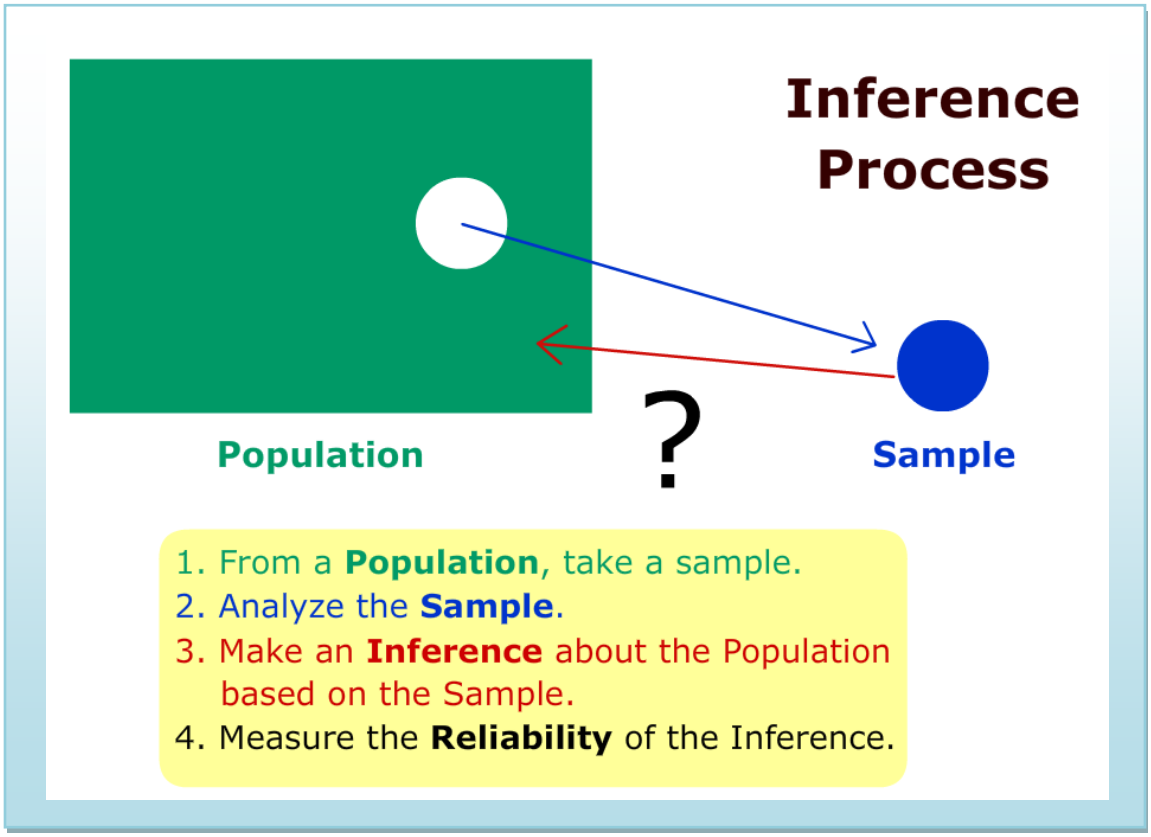
$$P(X > 70) = P\left(Z > \frac{(70 - 69.2)}{2.9}\right) = P(Z > 0.28) = 0.3897$$

This example demonstrates how the sample mean will cluster towards the population mean as the sample size increases.

8. Point Estimation and Confidence Intervals

8.1 Inferential Statistics

The reason we conduct statistical research is to obtain an understanding about phenomena in a population. For example, we may want to know if a potential drug is effective in treating a disease. Since it is not feasible or ethical to distribute an experimental drug to the entire population, we instead must study a small subset of the population called a sample. We then analyze the sample and make an inference about the population based on the sample. Using probability theory and the Central Limit Theorem, we can then measure the reliability of the inference.



Example: Lupe is trying to sell her house and needs to determine the market value of the home. The **population** in this example would be all the homes that are similar to hers in the neighborhood.

Lupe's realtor chooses for the **sample** nine recent homes in this neighborhood that sold in the last six months. The realtor then adjusts some of the sales prices to account for differences between Lupe's home and the sold homes.

Sampled Homes Adjusted Sales Price		
$420,000	$440,000	$470,000
$430,000	$450,000	$470,000
$430,000	$460,000	$480,000

Next the realtor takes the mean of the adjusted sample and recommends to Lupe a market value for Lupe's home of $450,000. The realtor has made an **inference** about the mean value of the population.

To measure the **reliability** of the inference, the realtor should look at factors like: the sample size being small, values of homes may have changed in the last six months, or that Lupe's home is not exactly like the sampled homes.

8.2 Point Estimation

The example above is an example of **Estimation**, a branch of Inferential Statistics where sample statistics are used to estimate the values of a population parameter. Lupe's realtor was trying to estimate the population mean (μ) based on the sample mean ($\bar{X}$).

	Sample Statistics		Population Parameters
Mean	$\bar{X}$	$\longrightarrow$	μ
Standard Deviation	s	$\longrightarrow$	σ
Proportion	$\hat{p}$	$\longrightarrow$	p

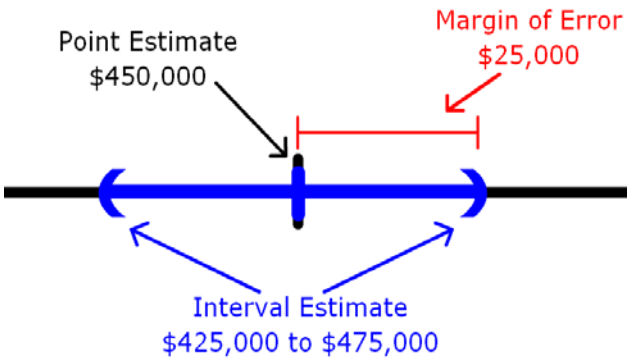
In the example above, Lupe's realtor estimated the population mean of similar homes in Lupe's neighborhood by using the sample mean of $450,000 from the adjusted price of the sampled homes.

Interval Estimation

A point estimate is our "best" estimate of a population parameter, but will most likely not exactly equal the parameter. Instead, we will choose a range of values called an **Interval Estimate** that is likely to include the value of the population parameter.

If the Interval Estimate is symmetric, the distance from the Point Estimator to either endpoint of the Interval Estimate is called the **Margin of Error.**

In the example above, Lupe's realtor could instead say the true population mean is probably between $425,000 and $475,000, allowing a $25,000 Margin of Error from the original estimate of $450,000. This Interval estimate could also be reported as $450,000 $\pm$ $25,000.



8.3 Confidence Intervals

Using probability and the Central Limit Theorem, we can design an Interval Estimate called a **Confidence Interval** that has a known probability (**Level of Confidence**) of capturing the true population parameter.

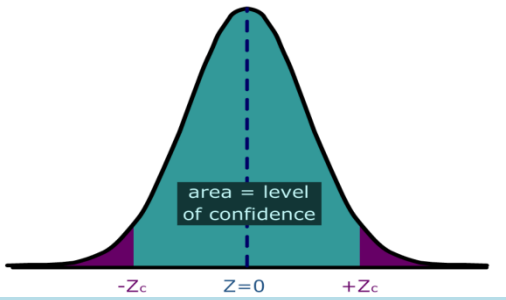
8.3.1 Confidence Interval for Population Mean

To find a confidence interval for the population mean (μ) when the population standard deviation (σ) is known, and n is sufficiently large, we can use the Standard Normal Distribution probability distribution function to calculate the critical values for the Level of Confidence:

Constructing Confidence Intervals for μ

c=Level of Confidence	Z_c=Critical Value
90%	1.645
95%	1.960
99%	2.578

$$\bar{X} \pm Z_c \cdot \frac{\sigma}{\sqrt{n}}$$



Example: The Dean wants to estimate the mean number of hours worked per week by students. A sample of 49 students showed a mean of 24 hours with a standard deviation of 4 hours. The point estimate is 24 hours (sample mean). What is the 95% confidence interval for the average number of hours worked per week by the students?

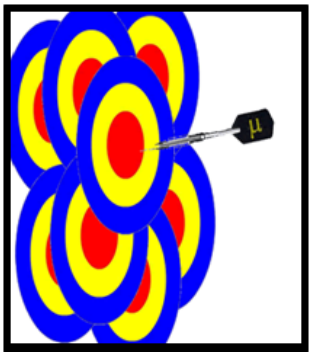
$$24 \pm \frac{1.96 \cdot 4}{\sqrt{49}} = 24 \pm 1.12 = (22.88, 25.12) \text{ hours per week}$$

The margin of error for the confidence interval is 1.12 hours. We can say with 95% confidence that mean number of hours worked by students is between 22.88 and 25.12 hours per week.

If the level of confidence is increased, then the margin of error will also increase. For example, if we increase the level of confidence to 99% for the above example, then:

$$24 \pm \frac{2.578 \cdot 4}{\sqrt{49}} = 24 \pm 1.47 = (22.53, 25.47) \text{ hours per week}$$

Some important points about Confidence Intervals



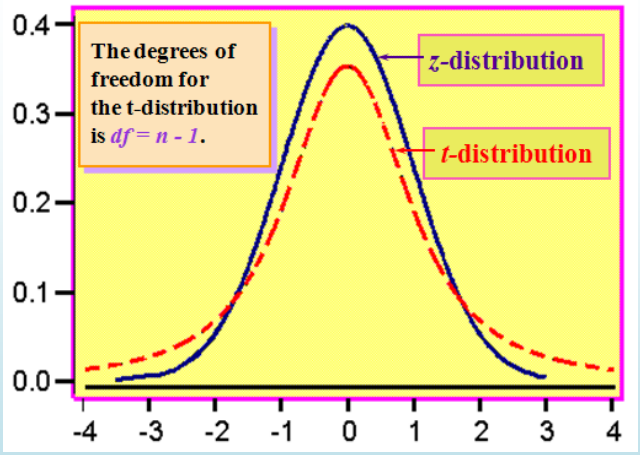
- The confidence interval is constructed from random variables calculated from sample data and attempts to predict an unknown but fixed population parameter with a certain level of confidence.
- Increasing the level of confidence will always increase the margin of error.
- It is impossible to construct a 100% Confidence Interval without taking a census of the entire population.
- Think of the population mean like a dart that always goes to the same spot, and the confidence interval as a moving target that tries to "catch the dart." A 95% confidence interval would be like a target that has a 95% chance of catching the dart.

8.3.2 Confidence Interval for Population Mean using Sample Standard Deviation – Student's t Distribution

The formula for the confidence interval for the mean requires the knowledge of the population standard deviation (σ). In most real-life problems, we do not know this value for the same reasons we do not know the population mean. This problem was solved by the Irish statistician William Sealy Gosset, an employee at Guiness Brewing. Gosset, however, was prohibited by Guiness in using his own name in publishing scientific papers. He published under the name "A Student", and therefore the distribution he discovered was named "Student's t-distribution"[9].

Characteristics of Student's t Distribution

- It is continuous, bell-shaped, and symmetrical about zero like the z distribution.
- There is a **family** of *t*-distributions sharing a mean of zero but having different standard deviations based on **degrees of freedom**.
- The t-distribution is more spread out and flatter at the center than the Z-distribution, but approaches the Z-distribution as the sample size gets larger.



Confidence Interval for μ

$$\bar{X} \pm t_c \frac{s}{\sqrt{n}}$$ **with degrees of freedom = n - 1**

Example

Last year Sally belonged to an Health Maintenance Organization (HMO) that had a population average rating of 62 (on a scale from 0-100, with '100' being best); this was based on records accumulated about the HMO over a long period of time. This year Sally switched to a new HMO. To assess the population mean rating of the new HMO, 20 members of this HMO are polled and they give it an average rating of 65 with a standard deviation of 10. Find and interpret a 95% confidence interval for population average rating of the new HMO.

The t distribution will have 20-1 =19 degrees of freedom. Using table or technology, the critical value for the 95% confidence interval will be t_c=2.093

$$65 \pm \frac{2.093 \cdot 10}{\sqrt{20}} = 65 \pm 4.68 = (60.32, 69.68) \text{ HMO rating}$$

With 95% confidence we can say that the rating of Sally's new HMO is between 60.32 and 69.68. Since the quantity 62 is in the confidence interval, we cannot say with 95% certainty that the new HMO is either better or worse than the previous HMO.

8.3.3 Confidence Interval for Population Proportion

Recall from the section on random variables the binomial distribution where p represented the proportion of successes in the population. The binomial model was analogous to coin-flipping, or yes/no question polling. In practice, we want to use sample statistics to estimate the population proportion (p).

The sample proportion ($\hat{p}$) is the proportion of successes in the sample of size n and is the point estimator for p. Under the Central Limit Theorem, if $np > 5$ and $n(1-p) > 5$, the distribution of the sample proportion $\hat{p}$ will have an approximately Normal Distribution.

Normal Distribution for $\hat{p}$ if Central Limit Theorem conditions are met.

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Using this information we can construct a confidence interval for p, the population proportion:

Confidence interval for p: $\hat{p} \pm Z\sqrt{\frac{p(1-p)}{n}} \approx \hat{p} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Example

200 California drivers were randomly sampled and it was discovered that 25 of these drivers were illegally talking on the cell phone without the use of a hands-free device. Find the point estimator for the proportion of drivers who are using their cell phones illegally and construct a 99% confidence interval.

The point estimator for p is $\hat{p} = \frac{25}{200} = .125$ or 12.5%.

A 99% confidence interval for p is:

$0.125 \pm 2.576\sqrt{\frac{.125(1-.125)}{200}} = .125 \pm .060$

The margin of error for this poll is 6% and we can say with 99% confidence that true percentage of drivers who are using their cell phones illegally is between 6.5% and 18.5%

8.3.4 Point Estimator for Population Standard Deviation

We often want to study the variability, volatility or consistency of a population. For example, two investments both have expected earnings of 6% per year, but one investment is much riskier, having higher ups and downs. To estimate variation or volatility of a data set, we will use the sample standard deviation (s) as a point estimator of the population standard deviation (σ).

Example

Investments A and B are both known to have a rate of return of 6% per year. Over the last 24 months, Investment A has sample standard deviation of 3% per month, while for Investment B, the sample standard deviation is 5% per month. We would say that Investment B is more volatile and riskier than Investment A due to the higher estimate of the standard deviation.

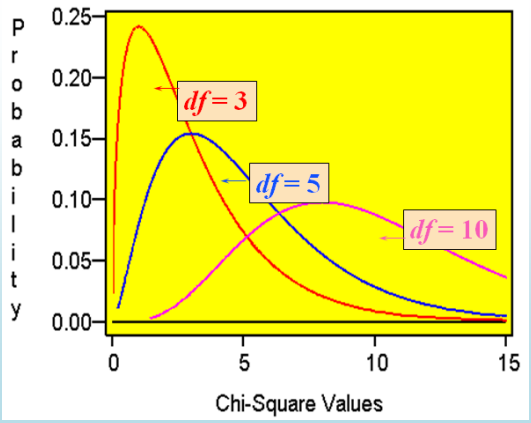
To create a confidence interval for an estimate of standard deviation, we need to introduce a new distribution, called the Chi-square (χ^2) distribution.

The Chi-square (χ^2) Distribution

The Chi-square distribution is a family of distributions related to the Normal Distribution as it represents a sum of independent squared standard Normal Random Variables. Like the Student's t distribution, the degrees of freedom will be n-1 and determine the shape of the distribution. Also, since the Chi-square represents squared data, the inference will be about the variance rather than the standard deviation.

Characteristics of Chi-square (χ^2) Distribution

- It is positively skewed
- It is non-negative
- It is based on degrees of freedom (n-1)
- When the degrees of freedom change, a new distribution is created
- $\frac{(n-1)s^2}{\sigma^2}$ will have Chi-square distribution.



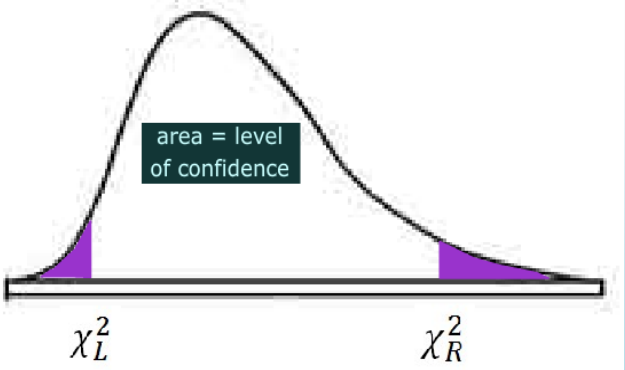
8.3.5 Confidence Interval for Population Variance and Standard Deviation

Since the Chi-square represents **squared data,** we can construct confidence intervals for the population variance (σ^2), and take the square root of the endpoints to get a confidence interval for the population standard deviation. Due to the skewness of the Chi-square distribution the resulting confidence interval will not be centered at the point estimator, so the margin of error form used in the prior confidence intervals doesn't make sense here.

Confidence Interval for population variance (σ^2)

- Confidence is **NO**T symmetric since chi-square distribution is not symmetric.
- Take square root of both endpoints to get confidence interval the population standard deviation (σ).

$$\left(\frac{(n-1)s^2}{\chi_R^2}, \frac{(n-1)s^2}{\chi_L^2}\right)$$



Example

In performance measurement of investments, standard deviation is a measure of volatility or risk. Twenty monthly returns from a mutual fund show an average monthly return of 1% and a sample standard deviation of 5%. Find a 95% confidence interval for the monthly standard deviation of the mutual fund.

The Chi-square distribution will have 20-1 =19 degrees of freedom. Using technology, the two critical values are $\chi_L^2 = 9.90655$ and $\chi_R^2 = 32.8523$.

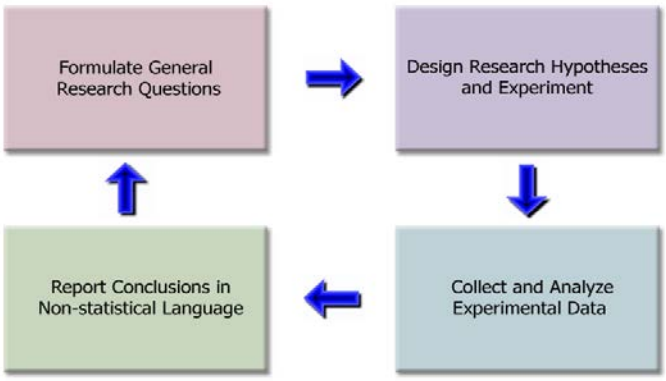
Formula for confidence interval for σ is: $\left(\sqrt{\frac{(19)5^2}{32.8523}}, \sqrt{\frac{(19)5^2}{8.90655}}\right) = (3.8, 7.3)$

One can say with 95% confidence that the standard deviation for this mutual fund is between 3.8% and 7.3% per month.

9. One Population Hypothesis Testing

In the prior section we used statistical inference to make an estimate of a population parameter and measure the reliability of the estimate through a confidence interval. In this section, we will explore in detail the use of statistical inference in testing a claim about a population parameter, which is the heart of the scientific method used in research.

9.1 Procedures of Hypotheses Testing and the Scientific Method



The actual conducting of a hypothesis test is only a small part of the scientific method. After formulating a general question, the scientific method consists of: the designing of an experiment, the collecting of data through observation and experimentation, the testing of hypotheses, and the reporting of overall conclusions. The conclusions themselves lead to other research ideas making this process a continuous flow of adding to the body of knowledge about the phenomena being studied.

Others may choose a more formalized and detailed set of procedures, but the general concepts of inspiration, design, experimentation, and conclusion allow one to see the whole process.

9.2 Formulate General Research Questions

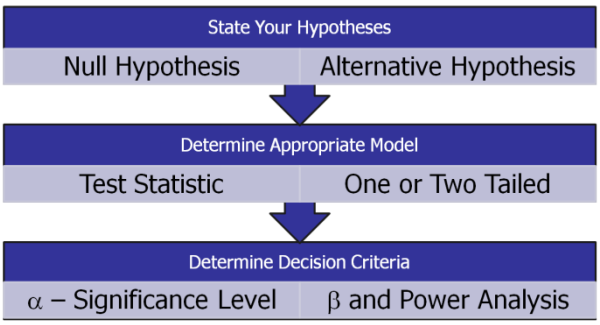
Most general questions start with an inspiration or an idea about a topic or phenomenon of interest. Some examples of general questions:

- (Health Care) Would a public single payer health care system be more effective than the current private insurance system?
- (Labor) What is the effect of undocumented immigration and outsourcing of jobs on the current unemployment rate.
- (Economy) Is the federal economic stimulus package effective in lessening the impact of the recession?
- (Education) Are colleges too expensive for students today?

It is important to not be so specific in choosing these general questions. Based on available or potentially available data, we can decide later what specific research hypotheses will be formulated and tested to address the general question. During the data collection and testing process other ideas may come up and we may choose to redefine the general question. However, we always want to have an overriding purpose for our research.

9.3 Design Research Hypotheses and Experiment

After developing a general question and having some sense of the data that is available or to be collected, it is time to design and an experiment and set of hypotheses.



9.3.1 Hypotheses and Hypothesis Testing

For purposes of testing, we need to design **hypotheses** that are statements about population parameters. Some examples of hypotheses:

- At least 20% of juvenile offenders are caught and sentenced to prison.
- The mean monthly income for college graduates is $5000.
- The mean standardized test score for schools in Cupertino is the same as the mean scores for Los Altos.
- The lung cancer rates in California are lower than the rates in Texas.
- The standard deviation of the New York Stock Exchange today is greater than 10 percentage points per year.

These same hypotheses could be written in symbolic notation:

- $p > 0.20$
- $\mu > 5000$
- $\mu_1 = \mu_2$
- $p_1 < p_2$
- $\sigma > 10$

Hypothesis Testing is a procedure, based on sample evidence and probability theory, used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected. This hypothesis that is tested is called the **Null Hypothesis** designated by the symbol H_0. If the Null Hypothesis is unreasonable and needs to be rejected, then the research supports an **Alternative Hypothesis** designated by the symbol H_a.

Null Hypothesis (H_0): A statement about the value of a population parameter that is assumed to be true for the purpose of testing.

Alternative Hypothesis (H_a): A statement about the value of a population parameter that is assumed to be true if the Null Hypothesis is rejected during testing.

From these definitions it is clear that the Alternative Hypothesis will necessarily contradict the Null Hypothesis; both cannot be true at the same time. Some other important points about hypotheses:

- Hypotheses must be statements about population parameters, never about sample statistics.
- In most hypotheses tests, equality ($=, \leq, \geq$) will be associated with the Null Hypothesis while non-equality ($\neq, <, >$) will be associated with the Alternative Hypothesis.
- It is the Null Hypothesis that is always tested in attempt to "disprove" it and support the Alternative Hypothesis. This process is analogous in concept to a "proof by contradiction" in Mathematics or Logic, but supporting a hypothesis with a level of confidence is not the same as an absolute mathematical proof.

Examples of Null and Alternative Hypotheses:

- $H_o: p \leq 0.20$ $H_a: p > 0.20$
- $H_o: \mu \leq 5000$ $H_a: \mu > 5000$
- $H_o: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$
- $H_o: p_1 \geq p_2$ $H_a: p_1 < p_2$
- $H_o: \sigma \leq 10$ $H_a: \sigma > 10$

9.3.2 Statistical Model and Test Statistic

To test a hypothesis we need to use a **statistical model** that describes the behavior for data and the type of population parameter being tested. Because of the Central Limit Theorem, many statistical models are from the Normal Family, most importantly the Z, t, χ^2, and F distributions. Other models that are used when the Central Limit Theorem is not appropriate are called non-parametric Models and will not be discussed here.

Each chosen model has requirements of the data called **model assumptions** that should be checked for appropriateness. For example, many models require the sample mean has approximately a Normal Distribution, which may not be true for some smaller or heavily skewed data sets.

Once the model is chosen, we can then determine a **test statistic**, a value derived from the data that is used to decide whether to **reject** or **fail to reject** the Null Hypothesis.

Some Examples of Statistical Models and Test Statistics

Statistical Model	Test Statistic
Mean vs. Hypothesized Value	$t = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$
Proportion vs. Hypothesized Value	$Z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$
Variance vs. Hypothesized Value	$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

9.3.3 Errors in Decision Making

Whenever we make a decision or support a position, there is always a chance we make the wrong choice. The hypothesis testing process requires us to either to reject the Null Hypothesis and support the Alternative Hypothesis or fail to reject the Null Hypothesis. This creates the possibility of two types of error:

- **Type I Error**
  Rejecting the null hypothesis when it is actually true.

- **Type II Error**
  Failing to reject the null hypothesis when it is actually false.

	Fail to Reject Ho	Reject Ho
Ho is true	Correct Decision	Type I error
Ho is False	Type II error	Correct Decision

In designing hypothesis tests, we need to carefully consider the probability of making either one of these errors.

Example:

Recall the two news stories discussed earlier in Section 3. In the first story, a drug company marketed a suppository that was later found to be ineffective (and often dangerous) in treatment. Before marketing the drug, the company determined that the drug was effective in treatment, which means the company rejected a Null Hypothesis that the suppository had no effect on the disease. This is an example of Type I error.

In the second story, research was abandoned when the testing showed Interferon was ineffective in treating a lung disease. The company in this case failed to reject a Null Hypothesis that the drug was ineffective. What if the drug really was effective? Did the company make Type II error? Possibly, but since the drug was never marketed, we have no way of knowing the truth.

These stories highlight the problem of statistical research: errors can be analyzed using probability models, but there is often no way of indentifying specific errors. For example, there are unknown innocent people in prison right now because a jury made Type I error in wrongfully convicting defendants. We must be open to the possibility of modification or rejection of currently accepted theories when new data is discovered.

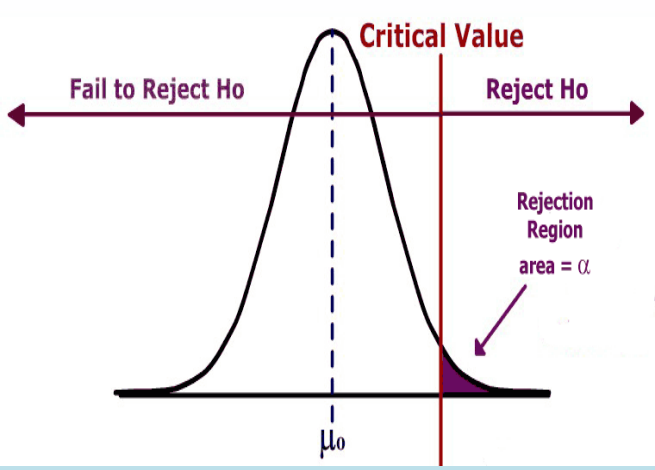
In designing an experiment, we set a maximum probability of making Type I error. This probability is called the **level of significance** or **significance level** of the test and designated by the Greek letter α.

The analysis of Type II error is more problematic as there many possible values that would satisfy the Alternative Hypothesis. For a specific value of the Alternative Hypothesis, the design probability of making Type II error is called **Beta (β)** which will be analyzed in detail later in this section.

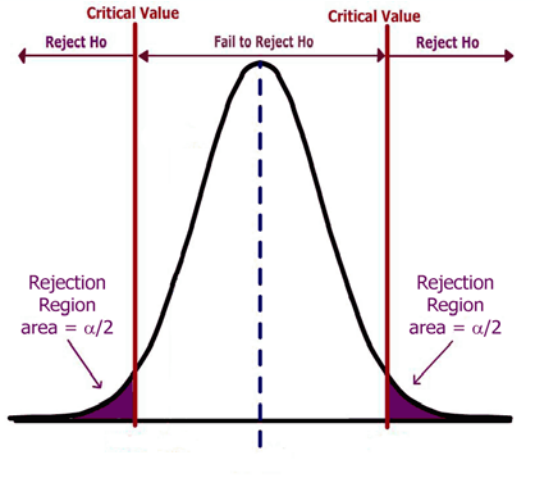
9.3.4 Critical Value and Rejection Region

Once the significance level of the test is chosen, it is then possible to find region(s) of the probability distribution function of the test statistic that would allow the Null Hypothesis to be rejected. This is called the **Rejection Region** and the boundry between the Rejection Region and the "Fail to Reject" is called the **Critical Value**.

There can be more than one critical value and rejection region. What matters is that the total area of the rejection region equals the significance level α.



One-tailed Hypothesis Test



Two-tailed Hypothesis Test

9.3.5 One and Two tailed Tests

A test is one-tailed when the Alternative Hypothesis, H_a , states a direction, such as:

H_0: The mean income of females is less than or equal to the mean income of male.
H_a : The mean income of females is greater than males.

Since equality is usually part of the Null Hypothesis, it is the Alternative Hypothesis which determines which tail to test.

A test is two-tailed when no direction is specified in the alternate hypothesis H_a , such as:

H_0 : The mean income of females is equal to the mean income of males.
H_a : The mean income of females is not equal to the mean income of the males.

In a two tailed-test, the significance level is split into two parts since there are two rejection regions. In hypothesis testing where the statistical model is symmetrical (eg: the Standard Normal Z or Student's t distribution) these two regions would be equal. There is a relationship between a confidence interval and a two-tailed test: If the level of confidence for a confidence interval is equal to 1-α, where α is the significance level of the two-tailed test, the critical values would be the same.

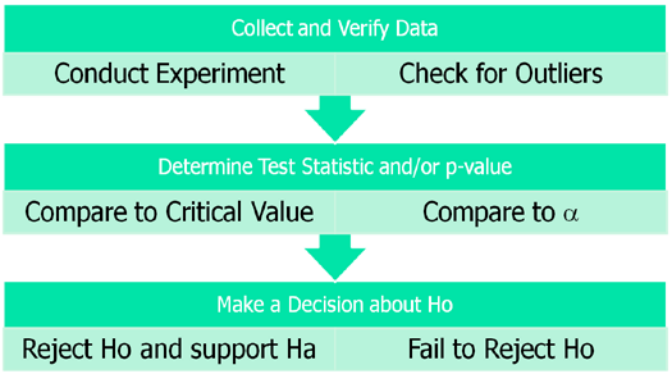
Here are some examples for testing the mean μ against a hypothesized value μ_0:

> H_a: $\mu > \mu_0$ means test the upper tail and is also called a right-tailed test.
> H_a: $\mu < \mu_0$ means test the lower tail and is also called a left-tailed test.
> H_a: $\mu \neq \mu_0$ means test both tails.

Deciding when to conduct a one or two-tailed test is often controversial and many authorities even go so far as to say that only two-tailed tests should be conducted. Ultimately, the decision depends on the wording of the problem. If we want to show that a new diet reduces weight, we would conduct a lower tailed test since we don't care if the diet causes weight gain. If instead, we wanted to determine if mean crime rate in California was different from the mean crime rate in the United States, we would run a two-tailed test, since different means greater than or less than.

9.4 Collect and Analyze Experimental Data

After designing the experiment, the next procedure would be to actually collect and verify the data. For the purposes of statistical analysis, we will assume that all sampling is either random, or uses an alternative technique that adequately simulates a random sample.



9.4.1 Data Verification

After collecting the data but before running the test, we need to verify the data. First, get a picture of the data by making a graph (histogram, dot plot, box plot, etc.) Check for skewness, shape and any potential outliers in the data.

9.4.2 Working with Outliers

An outlier is data point that is far removed from the other entries in the data set. Outliers could be caused by:

- Mistakes made in recording data
- Data that don't belong in population
- True rare events

The first two cases are simple to deal with as we can correct errors or remove data that that does not belong in the population. The third case is more problematic as extreme outliers will increase the standard deviation dramatically and heavily skew the data.

In *The Black Swan*, Nicholas Taleb argues that some populations with extreme outliers should not be analyzed with traditional confidence intervals and hypothesis testing.[10] He defines a Black Swan to be an

unpredictable extreme outlier that causes dramatic effects on the population. A recent example of a Black Swan was the catastrophic drop in the value of unregulated Credit Default Swap (CDS) real estate insurance investments which caused the near collapse of international banking system in 2008. The traditional statistical analysis that measured the risk of the CDS investments did not take into account the consequence of a rapid increase in the number of foreclosures of homes. In this case, statistics that measure investment performance and risk were useless and created a false sense of security for large banks and insurance companies.

Example

Here are the quarterly home sales for 10 realtors

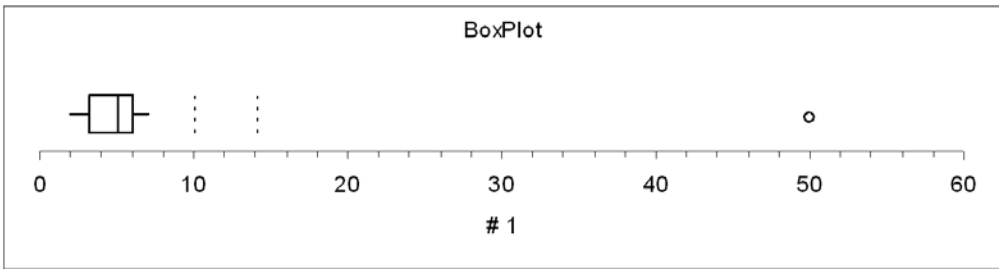
2 2 3 4 5 5 6 6 7 50

	With outlier	Without Outlier
Mean	9.00	4.44
Median	5.00	5.00
Standard Deviation	14.51	1.81
Interquartile Range	3.00	3.50

In this example, the number 50 is an outlier. When calculating summary statistics, we can see that the mean and standard deviation are dramatically affected by the outlier, while the median and the interquartile range (which are based on the ranking of the data) are hardly changed. One solution when dealing with a population with extreme outliers is to use inferential statistics using the ranks of the data, also called non-parametric statistics.

Using Box Plot to find outliers

- The "box" is the region between the 1st and 3rd quartiles.
- Possible outliers are more than 1.5 IQR's from the box (inner fence)
- Probable outliers are more than 3 IQR's from the box (outer fence)
- In the box plot below of the realtor example, the dotted lines represent the "fences" that are 1.5 and 3 IQR's from the box. See how the data point 50 is well outside the outer fence and therefore an almost certain outlier.



9.4.3 The Logic of Hypothesis Testing

After the data is verified, we want to conduct the hypothesis test and come up with a decision, whether or not to reject the Null Hypothesis. The decision process is similar to a "proof by contradiction" used in mathematics:

- We assume H_o is true before observing data and design H_a to be the complement of H_o.
- Observe the data (evidence). How unusual are these data under H_o?
- If the data are too unusual, we have "proven" H_o is false: Reject H_o and support H_a (strong statement).
- If the data are not too unusual, we fail to reject H_o. This "proves" nothing and we say data are inconclusive. (weak statement) .
- We can never "prove" H_o , only "disprove" it.
- "Prove" in statistics means support with (1-α)100% certainty. (example: if α=.05, then we are at least 95% confident in our decision to reject H_o.

9.4.4 Decision Rule – Two methods, Same Decision

Earlier we introduced the idea of a **test statistic** which is a value calculated from the data under the appropriate Statistical Model from the data that can be compared to the **critical value** of the Hypothesis test. If the test statistic falls in the **rejection region** of the statistical model, we reject the Null Hypothesis.

Recall that the critical value was determined by design based on the chosen **level of significance α.** The more preferred method of making decisions is to calculate the probability of getting a result as extreme as the value of the test statistic. This probability is called the **p-value**, and can be compared directly to the significance level.

- **p-value:** the probability, assuming that the null hypothesis is true, of getting a value of the test statistic at least as extreme as the computed value for the test.
- If the p-value is smaller than the significance level α, H_0 is rejected.
- If the p-value is larger than the significance level α, H_0 is not rejected.

Comparing p-value to α

Both the p-value and α are probabilities of getting results as extreme as the data assuming H_o is true.

The p-value is determined by the data is related to the actual probability of making Type I error (Rejecting a True Null Hypothesis). The smaller the p-value, the smaller the chance of making Type I error and therefore, the more likely we are to reject the Null Hypothesis.

The significance level α is determined by design and is the maximum probability we are willing to accept of rejecting a true H_0.

Two Decision Rules lead to the same decision.

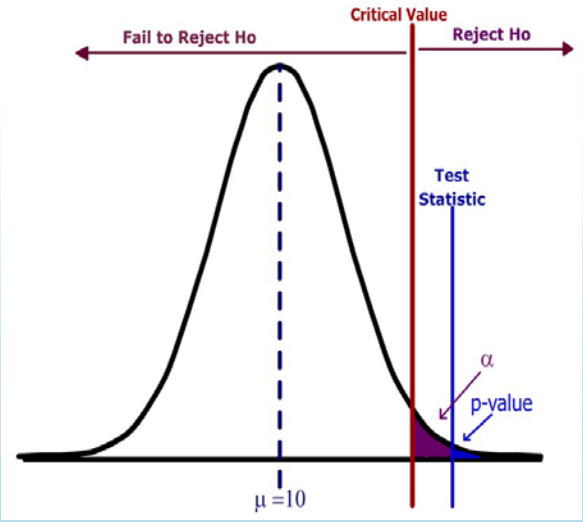
1. If the test statistic lies in the rejection region, reject Ho. (critical value method)
2. If the p-value < α, reject Ho. (p-value method)

This p-value method of comparison is preferred to the critical value method because the rule is the same for all statistical models: Reject H_o if p-value < α.

Let's see why these two rules are equivalent by analyzing a test of mean vs. hypothesized value.

Decision is Reject Ho

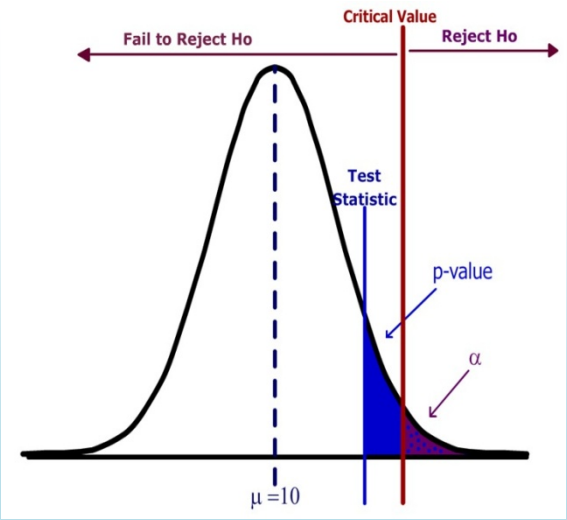
- Ho: $\mu = 10$
  Ha: $\mu > 10$
- Design: Critical value is determined by significance level α.
- Data Analysis: p-value is determined by test statistic
- Test statistic falls in rejection region.
- p-value (blue) < α (purple)
- Reject Ho.
- Strong statement: Data supports the Alternative Hypothesis.



In this example, the test statistic lies in the rejection region (the area to the right of the critical value). The p-value (the area to the right of the test statistic) is less than the significance level (the area to the right of the critical value). The decision is Reject H_o.

Decision is Fail to Reject Ho

- Ho: $\mu = 10$
  Ha: $\mu > 10$
- Design: critical value is determined by significance level α.
- Data Analysis: p-value is determined by test statistic
- Test statistic does not fall in the rejection region.
- p-value (blue) > α (purple)
- Fail to Reject Ho.
- Weak statement: Data is inconclusive and does not support the Alternative Hypothesis.



In this example, the Test Statistic does not lie in the Rejection Region. The p-value (the area to the right of the test statistic) is greater than the significance level (the area to the right of the critical value). The decision is Fail to Reject H_o.

9.5 Report Conclusions in Non-statistical Language

The hypothesis test has been conducted and we have reached a decision. We must now communicate these conclusions so they are complete, accurate, and understood by the targeted audience. How a conclusion is written is open to subjective analysis, but here are a few suggestions:

9.5.1 Be consistent with the results of the Hypothesis Test.

Rejecting Ho requires a **strong statement** in support of Ha, while failing to reject Ho does NOT support Ho, but requires a **weak statement** of insufficient evidence to support Ha.

Example: A researcher wants to support the claim that, on average, students send more than 1000 text messages per month and the research hypotheses are Ho: μ=1000 vs. Ha: μ>1000

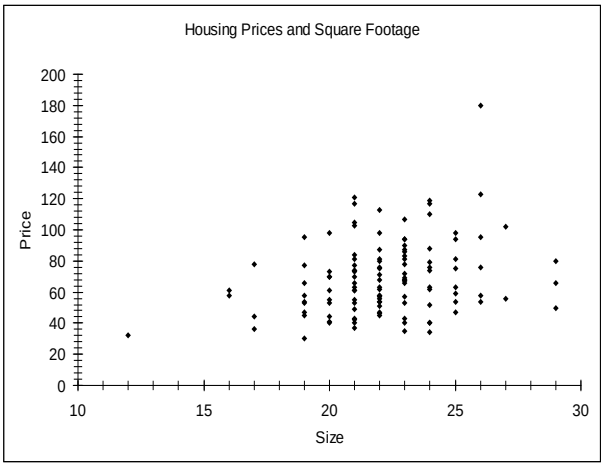
Conclusion if Ho is rejected: The mean number of text messages sent by students exceeds 1000.

Conclusion if Ho is not rejected: There is insufficient evidence to support the claim that the mean number of text messages sent by students exceeds 1000.

9.5.2 Use language that is clearly understood in the context of the problem.

Do not use technical language or jargon, but instead refer back to the language of the original general question or research hypotheses. Saying less is better than saying more.

Example: A test supported the Alternative Hypothesis that housing prices and size of homes in square feet were positively correlated. Compare these two conclusions and decide which is clearer:



- Conclusion 1: By rejecting the Null Hypothesis we are inferring that the Alterative Hypothesis is supported and that there exists a significant correlation between the independent and dependent variables in the original problem comparing home prices to square footage.
- Conclusion 2: Homes with more square footage generally have higher prices.

9.5.3 Limit the inference to the population that was sampled.

Care must be taken to describe the population being sampled and understand that the any claim is limited to this sampled population. If a survey was taken of a subgroup of a population, then the inference applies only to the subgroup.

For example, studies by pharmaceutical companies will only test adult patients, making it difficult to determine effective dosage and side effects for children. "In the absence of data, doctors use their medical judgment to decide on a particular drug and dose for children. 'Some doctors stay away from drugs, which could deny needed treatment,' Blumer says. 'Generally, we take our best guess based on what's been done before.' The antibiotic chloramphenicol was widely used in adults to treat infections resistant to penicillin. But many newborn babies died after receiving the drug because their immature livers couldn't break down the antibiotic."[11] We can see in this example that applying inference of the drug testing results on adults to the un-sampled children led to tragic results.

9.5.4 Report sampling methods that could question the integrity of the random sample assumption.

In practice it is nearly impossible to choose a random sample, and scientific sampling techniques that attempt to simulate a random sample need to be checked for bias caused by under-sampling.

Telephone polling was found to under-sample young people during the 2008 presidential campaign because of the increase in cell phone only households. Since young people were more likely to favor Obama, this caused bias in the polling numbers. Additionally, caller ID has dramatically reduced the percentage of successful connections with people being surveyed. The pollster Jay Leve of SurveyUSA said telephone polling was "doomed" and said his company was already developing new methods for polling.[12]

Sampling that didn't occur over the weekend may exclude many full time workers while self-selected and unverified polls (like ratemyprofessors.com) could contain immeasurable bias.

9.5.5 Conclusions should address the potential or necessity of further research, sending the process back to the first procedure.

Answers often lead to new questions. If changes are recommended in a researcher's conclusion, then further research is usually needed to analyze the impact and effectiveness of the implemented changes. There may have been limitations in the original research project (such as funding resources, sampling techniques, unavailability of data) that warrants more a comprehensive study.

For example, a math department modifies its curriculum based on a performance statistics for an experimental course. The department would want to do further study of student outcomes to assess the effectiveness of the new program.

9.6 Test of Mean vs. Hypothesized Value – A Complete Example

A food company has a policy that the stated contents of a product match the actual results. A **General Question** might be "Does the stated net weight of a food product match the actual weight?" The quality control statistician decides to test the 16 ounce bottle of Soy Sauce and must now **design the experiment**.

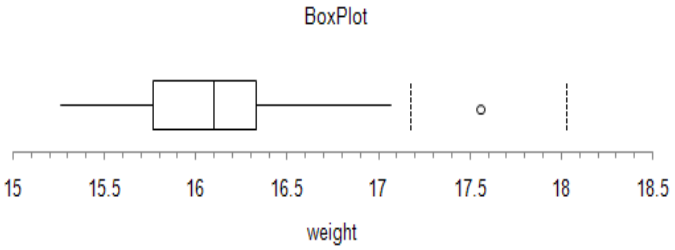
The quality control statistician has been given the authority to sample 36 bottles of soy sauce and knows from past testing that the population standard deviation is 0.5 ounces. The model will be a **test of population mean vs. hypothesized valu**e of 16 oz. A two-tailed test is selected since the company is concerned about both overfilling and underfilling the bottles as the stated policy is the stated weight match the actual weight of the product.

Research Hypotheses: **H_o: μ=16 (The filling machine is operating properly)**
H_a: μ ≠16 (The filling machine is not operating properly)

Since the population standard deviation is known the **test statistic** will be $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. This model is appropriate since the sample size assures the distribution of the sample mean is approximately Normal from the Central Limit Theorem.

Type I error would be to reject the Null Hypothesis and say the machine is not running properly when in fact it was operating properly. Since the company does not want to needlessly stop production and recalibrate the machine, the statistician chooses to limit the probability of Type I error by setting the **level of significance (α)** to 5%.

The statistician now **conducts the experiment** and samples 36 bottles in the last hour and determines from a box plot of the data that there is one unusual observation of 17.56 ounces. The value is rechecked and kept in the data set.

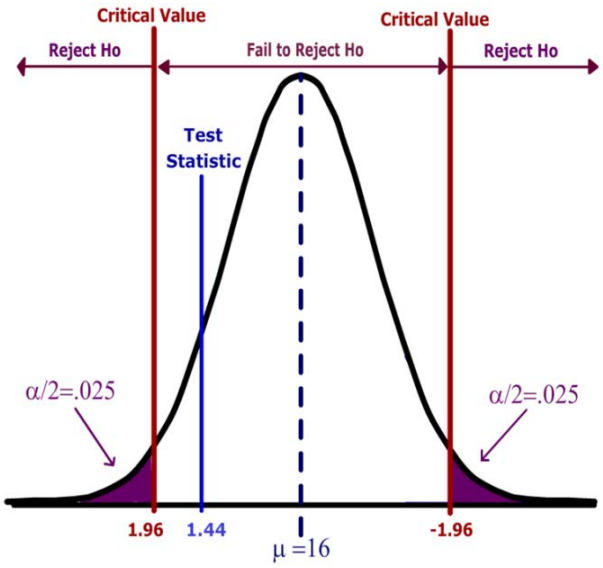


Next, the sample mean and the test statistic are calculated.

$$\bar{X} = 16.12 \text{ ounces} \qquad Z = \frac{16.12-16}{0.5/\sqrt{36}} = 1.44$$

The **decision rule** under the critical value method would be to reject the Null Hypothesis when the value of the test statistic is in the rejection region. In other words, reject H_o when Z >1.96 or Z<-1.96.

Based on this result, the decision is **fail to reject H_o** since the test statistic does not fall in the rejection region.

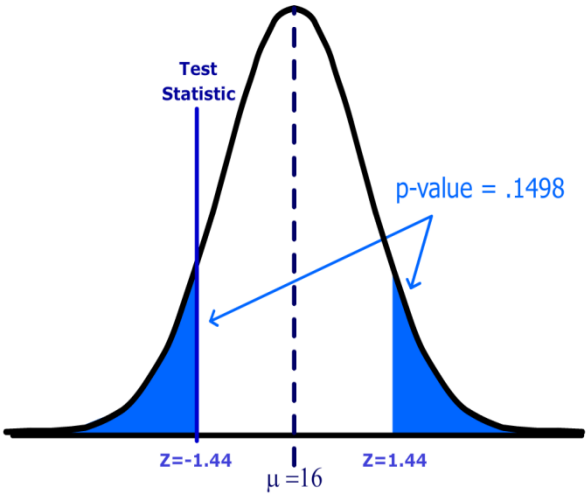


Alternatively (and preferably) the statistician would use the p-value method of decision rule. The p-value for a two-tailed test must include all values (positive and negative) more extreme than the Test Statistic, so in this example we find the probability that Z < -1.44 or Z > 1.44 (the area shaded blue).

Using a calculator, computer software or a Standard Normal table, **the p-value=0.1498.** Since the p-value is greater than α, the decision again is **fail to reject Ho.**

Finally the statistician must **report the conclusions** and make a recommendation to the company's management:

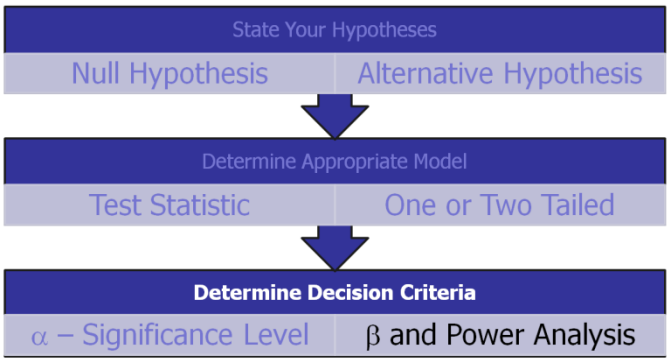
> "There is insufficient evidence to conclude that the machine that fills 16 ounce soy sauce bottles is operating improperly. This conclusion is based on 36 measurements taken during a single hour's production run. I recommend continued monitoring of the machine during different employee shifts to account for the possibility of potential human error."



The statistician makes the weak statement and is not stating that the machine is running properly, only that there is not enough evidence to state machine is running improperly. The statistician also reporting concerns about the sampling of only one shift of employees (restricting the inference to the sampled population) and recommends repeating the experiment over several shifts.

9.7 Type II Error and Statistical Power

In the prior example, the statistician failed to reject the Null Hypothesis because the probability of making Type I error (rejecting a true Null Hypothesis) exceeded the significance level of 5%. However, the statistician could have made Type II error if the machine is really operating improperly. One of the important and often overlooked tasks is to analyze the probability of making Type II error (β). Usually statisticians look at statistical power which is the complement of β.



Beta (β): The probability of failing to reject the null hypothesis when it is actually false.

Power (or Statistical Power): The probability of rejecting the null hypothesis when it is actually false.

Both beta and power are calculated for specific possible values of the Alternative Hypothesis.

	Fail to Reject Ho	Reject Ho
Ho is true	$1-\alpha$	α Type I error
Ho is False	β Type II error	$1-\beta$ Power

If a hypothesis test has low power, then it would difficult to reject H_o, even if H_o were false; the research would be a waste of time and money. However, analyzing power is difficult in that there are many values of the population parameter that support H_a. For example, in the soy sauce bottling example, the Alternative Hypothesis was that the mean was not 16 ounces. This means the machine could be filling the bottles with a mean of 16.0001 ounces, making H_a technically true. So when analyzing power and Type II error we need to choose a value for the **population mean under the Alternative Hypothesis (μ_a)** that is "**practically different**" from the **mean under the Null Hypothesis (μ_o).** This practical difference is called the **effect size**.

μ_o: The value of the population mean under the Null Hypothesis

μ_a: The value of the population mean under the Alternative Hypothesis

Effect Size: The "practical difference" between μ_o and μ_a = $|\mu_o - \mu_a|$

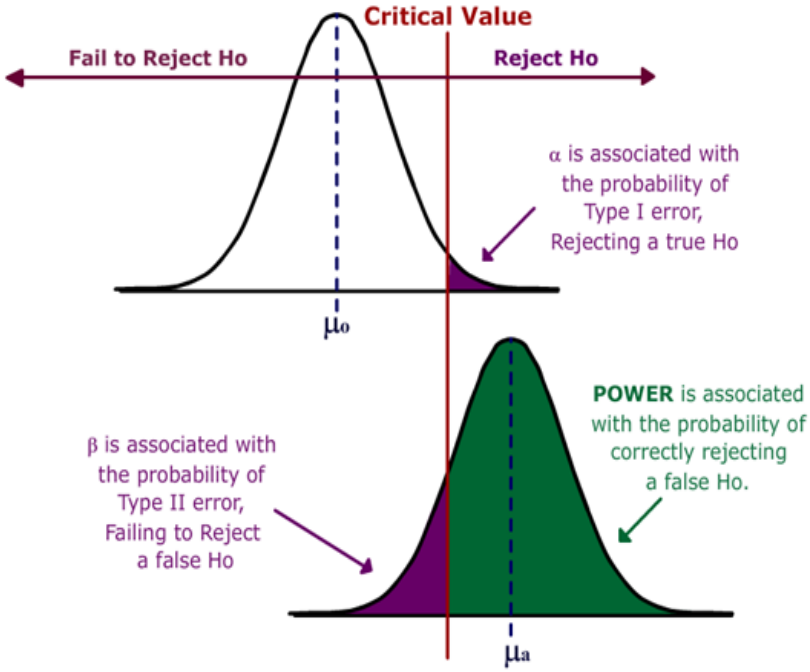
Suppose we are conducting a one-tailed test of the population mean:

$H_o: \mu = \mu_0$ $H_a: \mu > \mu_0$

Consider the two graphs shown to the right. The top graph is the distribution of the sample mean under the Null Hypothesis that we covered in an earlier section. The area to the right of the critical value is the rejection region.

We now add the bottom graph which represents the distribution of the sample mean under the Alternative Hypothesis for the specific value μ_a.

We can now measure the Power of the test (the area in green) and beta (the area in purple) on the lower graph.



There are several methods of increasing Power, but they all have trade-offs:

Ways to increase power	Trade off
Increase sample size	Increased cost or unavailability of data
Increase significance level (α)	More likely to Reject a True Ho (Type I error)
Choose a value of μ_a further from μ_o	Result may be less meaningful
Redefine population to lower standard deviation	Result may be too limited to have value
Do as a one-tail rather than a two-tail test	May produce a biased result

Example

Bus brake pads are claimed to last on average at least 60,000 miles and the company wants to test this claim. The bus company considers a "practical" value for purposes of bus safety to be that the pads last at least 58,000 miles. If the standard deviation is 5,000 and the sample size is 50, find the power of the test when the mean is really 58,000 miles. (Assume α = .05)

First, find the critical value of the test.

Reject Ho when Z < -1.645

Next, find the value of $\overline{X}$ that corresponds to the critical value.

$$\overline{X} = \mu_o + \frac{Z\sigma}{\sqrt{n}} = 60000 - (1.645)(5000)/\sqrt{50} = 58837$$

Ho is rejected when $\overline{X}$ < 58837

Finally, find the probability of rejecting Ho if Ha is true.

$$P(\overline{X} < 58837) = P\left(Z < \frac{(58837 - \mu_a)}{\sigma/\sqrt{n}}\right)$$

$$= P\left(Z < \frac{(58837 - 58000)}{5000/\sqrt{50}}\right) = P(Z < 1.18) = .8810$$

Therefore, this test has 88% power and β would be 12%

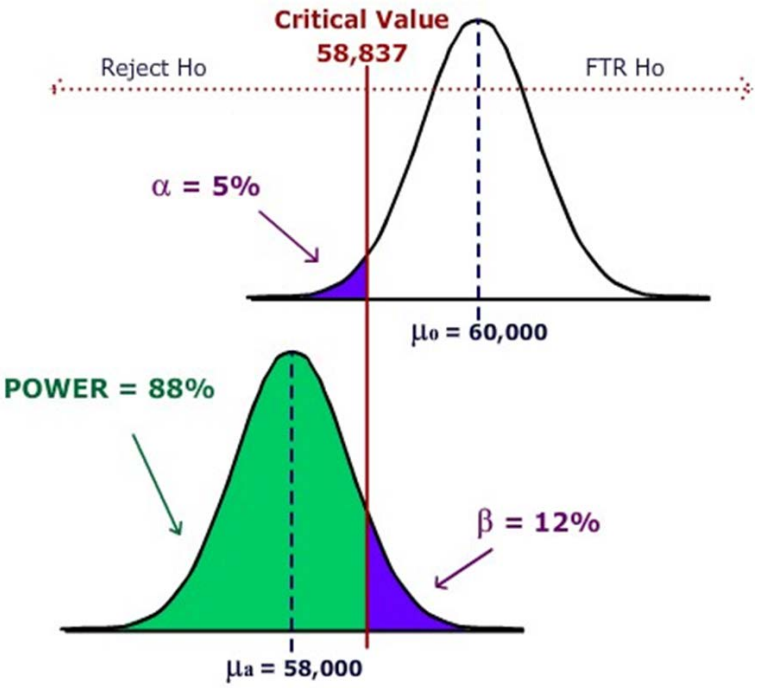
Power Calculation Values

Input Values

$\mu_o = 60{,}000$ miles
$\mu_a = 58{,}000$ miles
$\alpha = 0.05$
$n = 50$
$\sigma = 5000$ miles

Calculated Values

Effect Size = 2000 miles
Critical Value $= 58{,}837$ miles
$\beta = 0.1190$ or about 12%
Power $= 0.8810$ or about 88%



9.8 New Models for One Population Inference, Similar Procedures

The procedures outlined for the test of population mean vs. hypothesized value with known population standard deviation will apply to other models as well. All that really changes is the test statistic.

Examples of some other one population models:

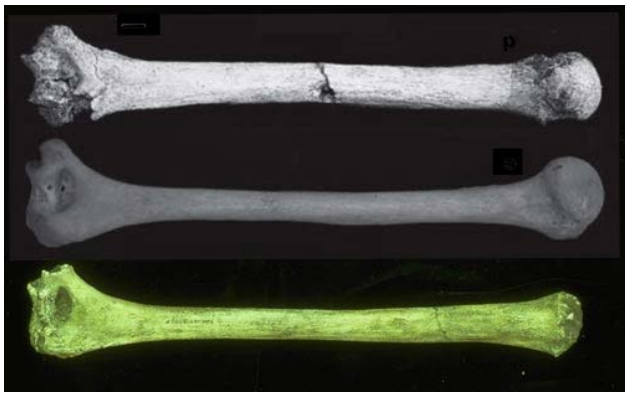
- Test of population mean vs. hypothesized value, population standard deviation unknown.
- Test of population proportion vs. hypothesized value.
- Test of population standard deviation (or variance) vs. hypothesized value.

9.8.1 Test of population mean with unknown population standard deviation

The test statistic for the one sample case changes to a Student's t distribution with degrees of freedom equal to n-1: $t = \frac{\bar{X}-\mu_o}{s/\sqrt{n}}$

The shape of the t distribution is similar to the Z, except the tails are fatter, so the logic of the decision rule is the same as the Z test statistic.

Example



Humerus bones from the same species have approximately the same length-to-width ratios. When fossils of humerus bones are discovered, archaeologists can determine the species by examining this ratio. It is known that Species A has a mean ratio of 9.6. A similar Species B has a mean ratio of 9.1 and is often confused with Species A. 21 humerus bones were unearthed in an area that was originally thought to be inhabited Species A. (Assume all unearthed bones are from the same species.)

1. Design a hypotheses where the alternative claim would be the humerus bones were not from Species A.

   Research Hypotheses

   H_o: $\mu = 9.6$ (The humerus bones are from Species A)

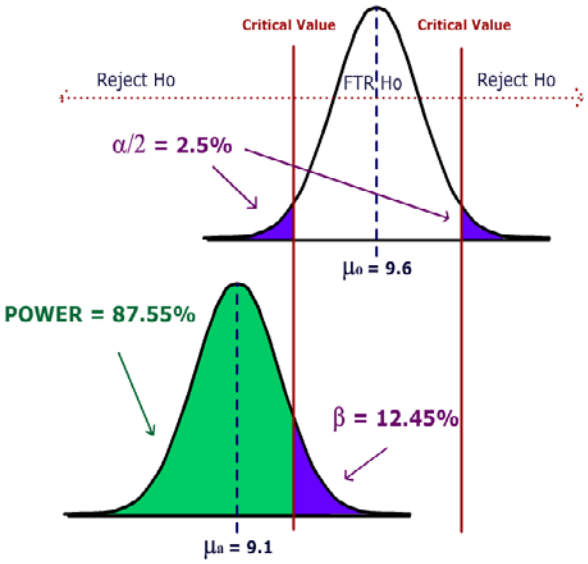
   H_a: $\mu \neq 9.6$ (The humerus bones are not from Species A)

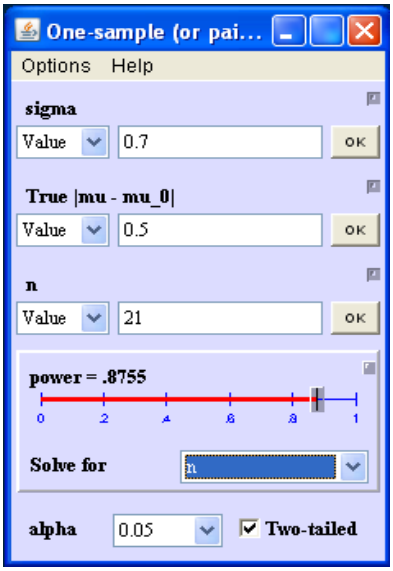
   Significance level: $\alpha = .05$

   Test Statistic (Model): t-test of mean vs. hypothesized value, unknown standard deviation
   Model Assumptions: we may need to check the data for extreme skewness as the distribution of the sample mean is assumed to be approximately the Normal Distribution.

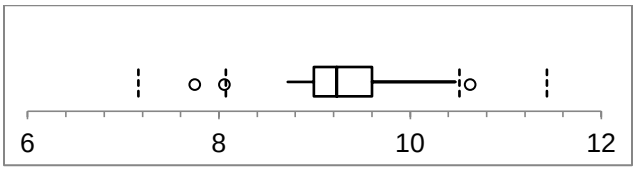
2. Determine the power of this test if the bones actually came from Species B (assume a standard deviation of 0.7)

Information needed for Power Calculation	**Results using Online Power Calculator**[13]
• μ_o = 9.6 (Species A) • μ_a = 9.1 (Species B) • Effect Size =| mo - ma | = 0.5 • s = 0.7 (given) • α = .05 • n = 21 (sample size) • Two tailed test	• Power =.8755 • β = 1 - Power = .1245 • If humerus bones are from Species B, test has an 87.55% chance of correctly rejecting Ho and a maximum Type II error of 12.55%





3. Conduct the test using at a 5% significance level and state overall conclusions.



Hypothesis Test: Mean vs. Hypothesized Value

9.60000	hypothesized value
9.26190	mean Data
0.66700	std. dev.
0.14555	std. error
21	n
20	df
-2.32	t
.0308	p-value (two-tailed)

From MegaStat[14], p-value = .0308 and α =.05.
Since p-value < α , Ho is **rejected** and we support Ha.

Conclusion: The evidence supports the claim (p-value<.05) that the humerus bones are not from Species A. The small sample size limited the power of the test, which prevented us from making a more definitive conclusion. Recommend testing to see if bones are from Species B or other unknown species. We are assuming since the bones were unearthed in the same location, they came from the same species.

9.8.2 Test of population proportion vs. hypothesized value.

When our data is categorical and there are only two possible choices (for example a yes/no question on a poll), we may want to make a claim about a proportion or a percentage of the population (p) being compared to a particular value (p_o). We will then use the sample proportion $(\hat{p})$to test the claim.

Test of proportion vs. hypothesized value

$\boldsymbol{p}$ = population proportion $\quad$ $\boldsymbol{p_o}$ = population proportion under Ho

$\boldsymbol{\hat{p}}$ = sample proportion $\quad$ $\boldsymbol{p_a}$ = population proportion under Ha

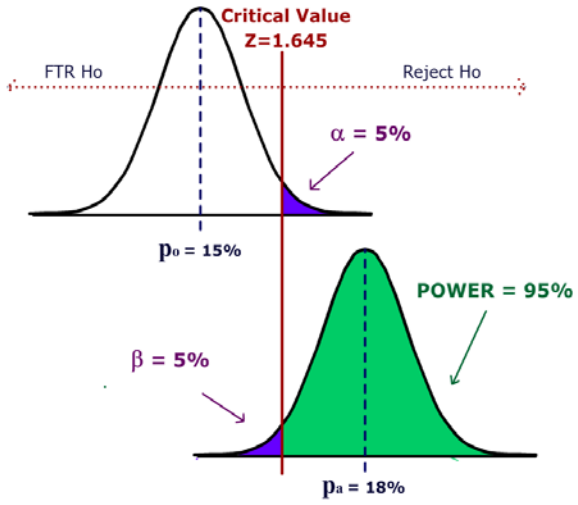
Test Statistic: $Z = \dfrac{\hat{p} - p_o}{\sqrt{\dfrac{p_o(1-p_o)}{n}}}$ **Requirement for Normality Assumption:** $np(1-p) > 5$

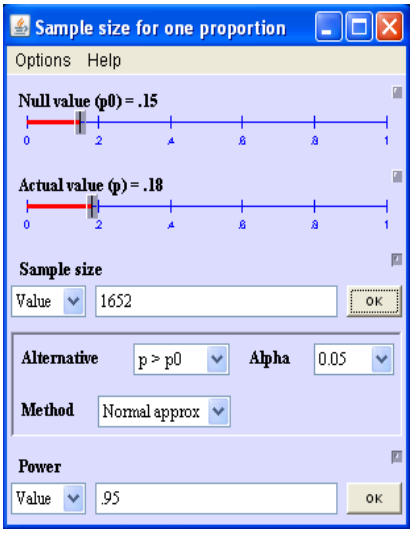
Example

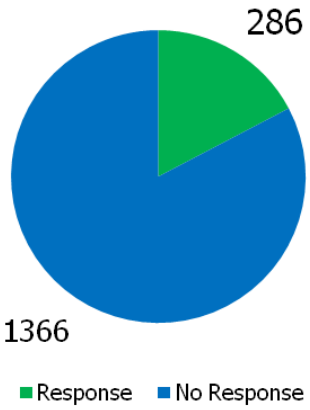
In the past, 15% of the mail order solicitations for a certain charity resulted in a financial contribution. A new solicitation letter has been drafted and will be sent to a random sample of potential donors. A hypothesis test will be run to determine if the new letter is more effective. Determine the sample so that (1) the test will be run at the 5% significance level and (2) If the letter has an 18% success rate, (an effect size of 3%), the power of the test will be 95%. After determining the sample size, conduct the test.

- Ho: p ≤ 0.15 (The new letter is not more effective.)
- Ha: p > 0.15 (The new letter is more effective.)
- Test Statistic – Z-test of proportion vs. hypothesized value.

Information needed for Sample Size Calculation	**Results using online Power Calculator and Megastat**
• p_o = 0.15 (current letter) • p_a = 0.18 (potential new letter) • Effect Size =| pa - po | = 0.03 • Desired Power = 0.95 • α = .05 • One tailed test	• Sample size = 1652 • The charity sent out 1652 new solicitation letters to potential donors and ran the test, receiving 286 positive responses. • p-value for test = 0.0042







Hypothesis test for proportion vs hypothesized value

Observed	*Hypothesized*	
0.1731	0.15	p (as decimal)
286/1652	248/1652	p (as fraction)
286.	247.8	X
1652	1652	n
	0.0088	std. error
	2.63	z
	.0042	p-value (one-tailed, upper)

Since p-value < α, reject H_0 and support H_a. Since the p-value is actually less than 0.01, we would go further and say that the data supports rejecting H_0 for $\alpha = .01$.

Conclusion: The evidence supports the claim that the new letter is more effective. The 1652 test letters were selected as a random sample from the charity's mailing list. All letters were sent at the same time period. The letters needed to be sent in a specific time period, so we were not able to control for seasonal or economic factors. We recommend testing both solicitation methods over the entire year to eliminate seasonal effects and to create a control group.

9.8.3 Test of population standard deviation (or variance) vs. hypothesized value.

We often want to make a claim about the variability, volatility or consistency of a population random variable. Hypothesized values for population variance σ^2 or standard deviation s are tested with the Chi-square (χ^2) distribution.

Examples of Hypotheses:

- Ho: $\sigma = 10$ Ha: $\sigma \neq 10$
- Ho: $\sigma^2 = 100$ Ha: $\sigma^2 > 100$

The sample variance s^2 is used in calculating the Chi-square Test Statistic.

Test of variance vs. hypothesized value

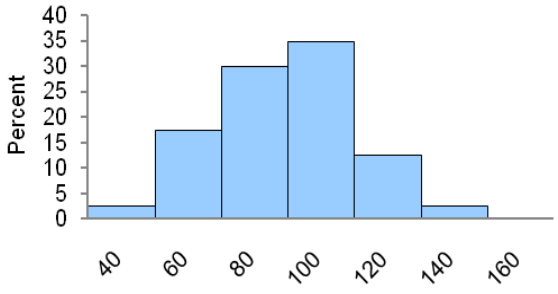
$\boldsymbol{\sigma^2}$ = population variance $\boldsymbol{\sigma_o^2}$ = population variance under Ho

$\boldsymbol{s^2}$ = sample variance

Test Statistic: $\chi^2 = \frac{(n-1)s^2}{\sigma_o^2}$ $\boldsymbol{n-1}$ = degrees of freedom

Example

A state school administrator claims that the standard deviation of test scores for 8th grade students who took a life-science assessment test is less than 30, meaning the results for the class show consistency. An auditor wants to support that claim by analyzing 41 students recent test scores. The test will be run at 1% significance level.



Design:

Research Hypotheses:

- Ho: Standard deviation for test scores equals 30.
- Ha: Standard deviation for test scores is less than 30.

57	75	86	92	101	108	110	120	155
63	77	88	96	102	108	111	122	
66	78	88	96	107	109	115	135	
68	81	92	98	107	109	115	137	
72	82	92	99	107	110	118	139	

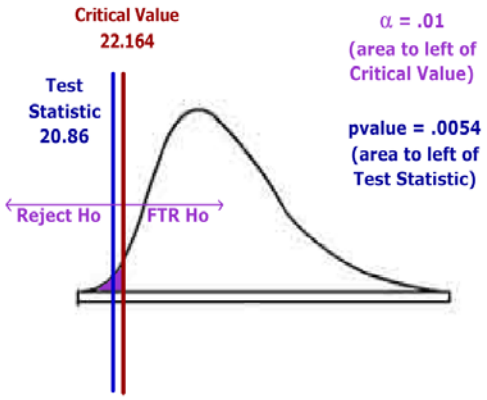
Hypotheses In terms of the population variance:

- Ho: $\sigma^2 = 900$
- Ha: $\sigma^2 < 900$

Results:

Chi-square Variance Test

900.000	hypothesized variance
469.426	observed variance of Data
41	n
40	df
20.86	chi-square
.0054	p-value (one-tailed, lower)



Decision: Reject Ho

Conclusion:

The evidence supports the claim (p-value<.01) that the standard deviation for 8th grade test scores is less than 30. The 40 test scores were the results of the recently administered exam to the 8th grade students. Since the exams were for the current class only, there is no assurance that future classes will achieve similar results. Further research would be to compare results to other schools that administered the same exam and to continue to analyze future class exams to see if the claim is holding true.

10.Two Population Inference

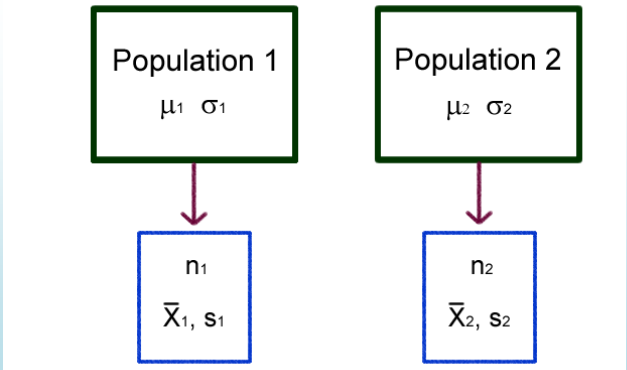
In this section we consider expanding the concepts from the prior section to design and conduct hypothesis testing with two samples. Although the logic of hypothesis testing will remain the same, care must be taken to choose the correct model. We will first consider comparing two population means.

10.1 Independent vs. dependent sampling

In designing a two population test of means, first determine whether the experiment involves data that is collected by independent or dependent sampling.

10.1.1 Independent sampling

The data is collected by two simple random samples from separate and unrelated populations. This data will then be used to compare the two population means. This is typical of an experimental or **treatment** population versus a **control** population.



INDEPENDENT SAMPLING

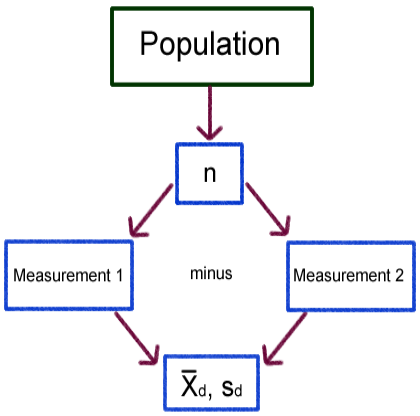
- n_1 **is the sample size from population 1.**
- n_2 **is the sample size from population 2.**
- $\overline{X}_1$ **is the sample mean from population 1.**
- $\overline{X}_2$ **is the sample mean from population 2.**
- s_1 **is the sample standard deviation from population 1.**
- s_2 **is the sample standard deviation from population 2.**

Example

A community college mathematics department wants to know if an experimental algebra course has higher success rates when compared to a traditional course. The mean grade points for 80 students in the experimental course (treatment) is compared to the mean grade points for 100 students in the traditional course (control).

10.1.2 Dependent sampling

The data consists of a single population and two measurements. A simple random sample is taken from the population and pairs of measurement are collected. This is also called related sampling or matched pair design. Dependent sampling actually reduces to a one population model of differences.



DEPENDENT SAMPLING

- n **is the sample size from the population, the number of pairs**
- $\bar{X}_d$ **is the sample mean of the differences of each pair.**
- s_d **is the sample standard deviation of the differences of each pair.**

Example

An instructor of a statistics course wants to know if student scores are different on the second midterm compared to the first exam. The first and second midterm scores for 35 students is taken and the mean difference in scores is determined.

10.2 Independent sampling models

We will first consider the case when we want to compare the population means of two populations using independent sampling.

10.2.1 Distribution of the difference of two sample means

Suppose we wanted to test the hypothesis $\boldsymbol{Ho: \mu_{1=}\mu_2}$. We have point estimators for both μ_1 and μ_2, namely $\bar{X}_1$ and $\bar{X}_2$, which have approximately Normal Distributions under the Central Limit Theorem, but it would useful to combine them both into a single estimator. Fortunately it is known that if two random variables have a Normal Distribution, then so does the sum and difference. Therefore we can restate the hypothesis as $\boldsymbol{Ho: \mu_{1-}\mu_2 = 0}$ and use the difference of sample means $\bar{X}_1 - \bar{X}_2$ as a point estimator for the difference in population means $\mu_1 - \mu_2$.

Distribution of $\bar{X}_1 - \bar{X}_2$ under the Central Limit Theorem

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_{1-}\mu_2 \qquad \sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$Z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_{1-}\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$ if n_1 and n_2 are sufficiently large.

10.2.2 Comparing two means, independent sampling: Model when population variances known

When the population variances are known, the test statistic for the Hypothesis $\boldsymbol{Ho}$: $\boldsymbol{\mu_1=\mu_2}$ can be tested with Normal distribution Z test statistic shown above. Also, if both sample size n_1 and n_2 exceed 30, this model can also be used.

Example

Are larger homes more likely to have pools? The square footage (size) data for single family homes in California was separated into two populations: Homes with pools and homes without pools. We have data from 130 homes with pools and 95 homes without pools.

Example - Design

Research Hypotheses: **Ho: $\mu_1 \leq \mu_2$ (Homes with pools do not have more mean square footage)**
Ha: $\mu_1 > \mu_2$ (Homes with pools do have more mean square footage)

Since both sample sizes are over 30, the model will be a **Large sample Z test comparing two population means with independent sampling**. This model is appropriate since the sample sizes assures the distribution of the sample mean is approximately Normal from the Central Limit Theorem. A one-tailed test is selected since we want to support the claim that homes with pools are larger. The test statistic will be $= \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$.

Type I error would be to reject the Null Hypothesis and claim home with pools are larger, when they are not larger. It was decided to limit this error by setting the level of significance (α) to 1%.

The decision rule under the critical value method would be to reject the Null Hypothesis when the value of the test statistic is in the rejection region. In other words, reject Ho when Z > 2.326. The decision under the p-value method is to reject Ho if the p-value is < α.

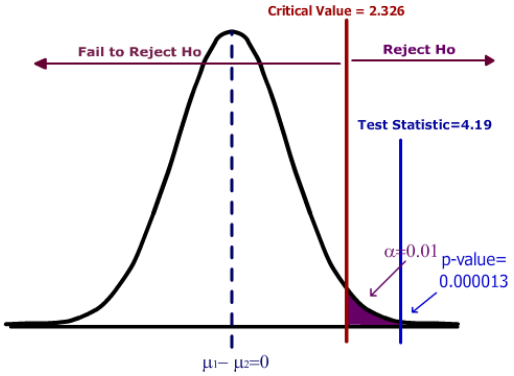
Example - Data/Results

Hypothesis Test: Independent Groups (z-test)

SqFt Pool	SqFt no Pool	
26.25	23.04	mean
6.93	4.55	std. dev.
130	95	n

3.212 difference (SqFt Pool - SqFt no Pool)
0.766 standard error of difference
0 hypothesized difference

4.19 z
1.37E-05 p-value (one-tailed, upper)



Since the test statistic (Z = 4.19) is greater than the critical value (2.326), Ho is rejected. Also the p-value (0.000013) is less than α (0.01), the decision is Reject Ho.

Example - Conclusion

The researcher makes the strong statement that homes with pools have a significantly higher mean square footage than home without pools.

10.2.3 Model when population variances unknown, but assumed to be equal

In the case when the population standard deviations are unknown, it seems logical to simply replace the population standard deviations for each population with the sample standard deviations and use a t-distribution as we did for the one population case. However, this is not so simple when the sample size for either group is under 30.

We will consider two models. This first model (which we prefer to use since it has higher power) assumes the population variances are equal and is called the **pooled variance t-test**. In this model we combine or "pool" the two sample standard deviations into a single estimate called the pooled standard deviation, s_p . If the central limit theorem is working, we then can substitute s_p for s_1 and s_2 get a t-distribution with $n_1 + n_2 - 2$ degrees of freedom:

Pooled variance t-test to compare the means for two independent populations

Model Assumptions

- Independent Sampling
- $\bar{X}_1 - \bar{X}_2$ approximately Normal
- $\sigma_1^2 = \sigma_2^2$

Test Statistic

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad s_p = \sqrt{\frac{(n_1-1)s_1^2 - (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

Degrees of freedom = $n_1 + n_2 - 2$

Example

A recent EPA study compared the highway fuel economy of domestic and imported passenger cars. A sample of 15 domestic cars revealed a mean of 33.7 MPG (mile per gallon) with a standard deviation of 2.4 mpg. A sample of 12 imported cars revealed a mean of 35.7 mpg with a standard deviation of 3.9. At the .05 significance level can the EPA conclude that the MPG is higher on the imported cars?

Example - Design

It is best to associate the subscript 2 with the control group, in this case we will let domestic cars be population 2.

Research Hypotheses: **Ho: $\mu_1 \leq \mu_2$ (Imported compact cars do not have a higher mean MPG)**

Ha: $\mu_1 > \mu_2$ (Imported compact cars have a higher mean MPG)

We will assume the population variances are equal $\sigma_1^2 = \sigma_2^2$, so the model will be a **Pooled variance t-test**. This model is appropriate if the distribution of the differences of sample means is approximately Normal from the Central Limit Theorem. A one-tailed test is selected based on Ha.

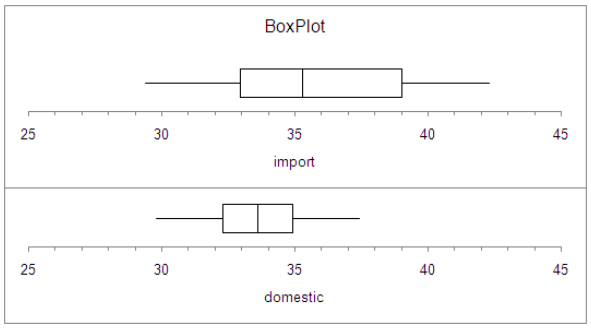
Type I error would be to reject the Null Hypothesis and claim imports has a higher mean MPG, when they do not have higher MPG. The test will be run at a level of significance (α) of 5%.

The degrees of freedom for this test is 25, so the decision rule under the critical value method would be to reject H_0 when t > 1.708. The decision under the p-value method is to reject Ho if the p-value is < α.

Example - Data/Results

$$s_p = \sqrt{\frac{(12-1)3.86^2 - (12-1)2.16^2}{15+12-2}} = 3.03 \qquad t = \frac{(35.76-33.59)-0}{3.03\sqrt{\frac{1}{12}+\frac{1}{15}}} = 1.85$$

Since 1.85 > 1.708, the decision would be to Reject Ho. Also the p-value is calculated to be .0381 which again shows that the result is significant at the 5% level.



import	domestic	
35.76	33.59	mean
3.86	2.16	std. dev.
12	15	n

25	df
2.17000	difference (import - domestic)
9.16856	pooled variance
3.02796	pooled std. dev.
1.17273	standard error of difference
0	hypothesized difference
1.85	t
.0381	p-value (one-tailed, upper)

Example - Conclusion

Imported compact cars have a significantly higher mean MPG rating when compared to domestic cars.

10.2.4 Model when population variances unknown, but assumed to be unequal

In the prior example, we assumed the population variances were equal. However, when looking at the box plot of the data or the sample standard deviations, it appears that the import cars have more variability MPG than domestic cars, which would violate the assumption of equal variances required for the Pooled Variance t-test.

Fortunately, there is an alternative model that has been developed for when population variances are unequal, called the Behrens-Fisher model [15], or the **unequal variances t-test**.

Unequal variance t-test to compare the means for two independent populations

Model Assumptions

- Independent Sampling
- $\bar{X}_1 - \bar{X}_2$ approximately Normal
- $\sigma_1^2 \neq \sigma_2^2$

Test Statistic

$$t' = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \qquad df = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left[\frac{(s_1^2/n_1)^2}{(n_1-1)}+\frac{(s_2^2/n_2)^2}{(n_2-1)}\right]}$$

The degrees of freedom will be less then or equal to $n_1 + n_2 - 2$, so this test will usually have less power than the pooled variance t-test.

Example

We will repeat the prior example to see if we can support the claim that imported compact cars have higher mean MPG when compared to domestic compact cars. This time we will assume that the population variances are not equal.

Example - Design

Again we will let domestic cars be population 2.

Research Hypotheses: **Ho: $\mu_1 \leq \mu_2$ (Imported compact cars do not have a higher mean MPG)**

Ha: $\mu_1 > \mu_2$ (Imported compact cars have a higher mean MPG)

We will assume the population variances are unequal $\sigma_1^2 \neq \sigma_2^2$, so the model will be an **unequal variance t-test**. This model is appropriate if the distribution of the differences of sample means is approximately Normal from the Central Limit Theorem. A one-tailed test is selected based on Ha.

Type I error would be to reject the Null Hypothesis and claim imports has a higher mean MPG, when they do not have higher MPG. The test will be run at a level of significance (α) of 5%.

The degrees of freedom for this test is 16 (see calculation below), so the decision rule under the critical value method would be to reject Ho when t > 1.746. The decision under the p-value method is to reject Ho if the p-value is < α.

Example - Data/Results

$$df = \frac{\left(\frac{2.16^2}{15} + \frac{3.86^2}{12}\right)^2}{\left[\frac{\left(2.16^2/15\right)^2}{(15-1)} + \frac{\left(3.86^2/12\right)^2}{(12-1)}\right]} = 16$$

$$t = \frac{(35.76 - 33.59) - 0}{\sqrt{\frac{2.16^2}{15} + \frac{3.86^2}{12}}} = 1.74$$

import	domestic	
35.76	33.59	mean
3.86	2.16	std. dev.
12	15	n

16 df
2.17000 difference (import - domestic)
1.24606 standard error of difference
0 hypothesized difference

1.74 t
.0504 p-value (one-tailed, upper)

Since 1.74 <1.708, the decision would be Fail to Reject Ho. Also the p-value is calculated to be .0504 which again shows that the result is not significant (barely) at the 5% level.

Example - Conclusion

Insufficient evidence to claim imported compact cars have a significantly higher mean MPG rating when compared to domestic cars.

You can see the lower power of this test when compared to the pooled variance t-test example where Ho was rejected. We always prefer to run the test with higher power when appropriate.

10.3 Dependent sampling – matched pairs t-test

The independent models shown above compared samples that were not related. However, it is often advantageous to have related samples that are paired up – Two measurements from a single population. The model we will consider here is called the **matched pairs t-test** also known as the paired difference t-test. The advantage of this design is that we can eliminate variability due to other factors not being studied, increasing the power of the design.

In this model we take the difference of each pair and create a new population of differences, so if effect, the hypothesis test is a one population test of mean that we already covered in the prior section.

Matched pairs t-test to compare the means for two dependent populations

Model Assumptions

- Dependent Sampling
- $X_d = X_1 - X_2$
- $\bar{X}_d = \bar{X}_1 - \bar{X}_2$ approximately Normal

Test Statistic

$$t = \frac{\bar{X}_d - \mu_d}{s_d/\sqrt{n}} \qquad df = n - 1$$

Example

An independent testing agency is comparing the daily rental cost for renting a compact car from Hertz and Avis. A random sample of 15 cities is obtained and the following rental information obtained.

At the .05 significance level can the testing agency conclude that there is a difference in the rental charged?

City	Hertz	Avis
Atlanta	42	40
Baltimore	51	47
Boston	46	42
Chicago	56	52
Cleveland	45	43
Denver	48	48
Dallas	56	54
Honolulu	37	32
Los Angeles	51	48
Kansas City	45	48
Miami	41	39
New York	44	42
San Francisco	48	45
Seattle	46	50
Washington DC	44	43

Notice in this example that cities are the single population being sampled and two measurements (Hertz and Avis) are being taken from each city. Using the matched pair design, we can eliminate the variability due to cities being differently priced (Honolulu is cheap because you can't drive very far on Oahu!)

Example - Design

Research Hypotheses: **H_0: $\mu_1=\mu_2$ (Hertz and Avis have the same mean price for compact cars.)**
H_a: $\mu_1\neq\mu_2$ (Hertz and Avis do not have the same mean price for compact cars.)

Model will be matched pair t-test and these hypotheses can be restated as: **H_0: $\mu_d=0$ H_a: $\mu_d\neq0$**

The test will be run at a level of significance (α) of 5%.

Model is two tailed matched pairs t-test with 14 degrees of freedom. Reject Ho if t < -2.145 or t >2.145.

Example - Data/Results

We take the difference for each pair and find the sample mean and standard deviation.

$\overline{X}_d = 1.80$

$s_d = 2.513$

$n = 15$

$$t = \frac{1.80 - 0}{2.513/\sqrt{15}} = 2.77$$

Reject Ho under either the critical value or p-value method.

Hypothesis Test: Paired Observations

0.000	hypothesized value
46.667	mean Hertz
44.867	mean Avis
1.800	mean difference (Hertz - Avis)
2.513	std. dev.
0.649	std. error
15	n
14	df
2.77	t
.0149	p-value (two-tailed)

City	Hertz	Avis	Difference
Atlanta	42	40	2
Baltimore	51	47	4
Boston	46	42	4
Chicago	56	52	4
Cleveland	45	43	2
Denver	48	48	0
Dallas	56	54	2
Honolulu	37	32	5
Los Angeles	51	48	3
Kansas City	45	48	-3
Miami	41	39	2
New York	44	42	2
San Francisco	48	45	3
Seattle	46	50	-4
Washington DC	44	43	1

Example – Conclusion

There is a difference in mean price for compact cars between Hertz and Avis. Avis has lower mean prices.

The advantage of the matched pair design is clear in this example. The sample standard deviation for the Hertz prices is $5.23 and for Avis it is $5.62. Much of this variability is due to the cities, and the matched pairs design dramatically reduces the standard deviation to $2.51, meaning the matched pairs t-test has significantly more power in this example.

10.4 Independent sampling – comparing two population variances or standard deviations

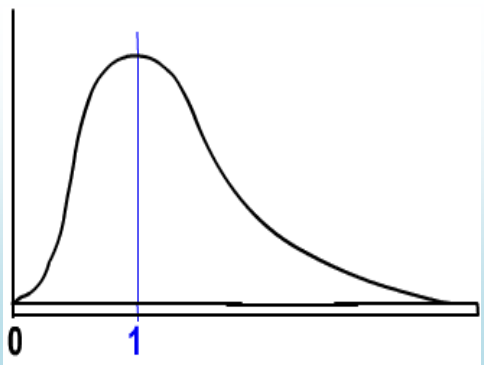
Sometimes we want to test if two populations have the same spread or variation, as measured by variance or standard deviation. This may be a test on its own or a way of checking assumptions when deciding between two different models (e.g.: pooled variance t-test vs. unequal variance t-test). We will now explore testing for a difference in variance between two independent samples.

10.4.1 F distribution

The F distribution is a family of distributions related to the Normal Distribution. There are two different degrees of freedom, usually represented as numerator (df_{num}) and denominator (df_{den}). Also, since the F represents squared data, the inference will be about the variance rather than the standard deviation.

Characteristics of F Distribution

- It is positively skewed
- It is non-negative
- There are 2 different degrees of freedom (df_{num}, df_{den})
- When the degrees of freedom change, a new distribution is created
- The expected value is 1.



10.4.2 F test for equality of variances

Suppose we wanted to test the Null Hypothesis that two population standard deviations are equal, $Ho: \sigma_1 = \sigma_2$. This is equivalent to testing that the population variances are equal: $\sigma_1^2 = \sigma_2^2$. We will now instead write these as an equivalent ratio: $Ho: \frac{\sigma_1^2}{\sigma_2^2} = 1$ or $Ho: \frac{\sigma_2^2}{\sigma_1^2} = 1$. This is the logic behind the F test; If two population variances are equal, then the ratio of sample variances from each population will have F distribution. F will always be an upper tailed test in practice, so the larger variance goes in the numerator. The test statistics are summarized in the table.

Hypotheses	Test Statistic
$H_o: \sigma_1 \geq \sigma_2$ $H_a: \sigma_1 < \sigma_2$	$F = \frac{s_2^2}{s_1^2}$ *use* α *table*
$H_o: \sigma_1 \leq \sigma_2$ $H_a: \sigma_1 > \sigma_2$	$F = \frac{s_1^2}{s_2^2}$ *use* α *table*
$H_o: \sigma_1 = \sigma_2$ $H_a: \sigma_1 \neq \sigma_2$	$F = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$ *use* $\alpha/2$ *table*

10.4.3 Example - Stand alone test

A stockbroker at brokerage firm, reported that the mean rate of return on a sample of 10 software stocks (population 1)was 12.6 percent with a standard deviation of 4.9 percent. The mean rate of return on a sample of 8 utility stocks (population 2) was 10.9 percent with a standard deviation of 3.5 percent. At the .05 significance level, can the broker conclude that there is more variation in the software stocks?

Example - Design

Research Hypotheses: **Ho: $\sigma_1 \leq \sigma_2$ (Software stocks do not have more variation)**
Ha: $\sigma_1 > \sigma_2$ (Software stocks do have more variation)

Model will be F test for variances and the test statistic from the table will be F= $\frac{s_1^2}{s_2^2}$. The degrees of freedom for numerator will be n_1-1=9 and the degrees of freedom for denominator will be n_2-1=7.

The test will be run at a level of significance (α) of 5%.

Critical Value for F with df_{num}=9 and df_{den}=7 is 3.68. Reject Ho if F >3.68.

Example - Data/Results

$F = {4.9^2}/{3.5^2} = 1.96$, which is less than critical value, so Fail to Reject Ho.

Example – Conclusion

There is insufficient evidence to claim more variation in the software stock.

10.4.4 Example - Testing model assumptions

When comparing two means from independent samples, you have a choice between the more powerful pooled variance t-test (assumption is $\sigma_1^2 = \sigma_2^2$) or the weaker unequal variance t-test (assumption is $\sigma_1^2 \neq \sigma_2^2$). We can now design a hypothesis test to help us choose the appropriate model. Let us revisit the example of comparing the mpg for import and domestic compact cars. Consider this example a "test before the main test" to help choose the correct model for comparing means.

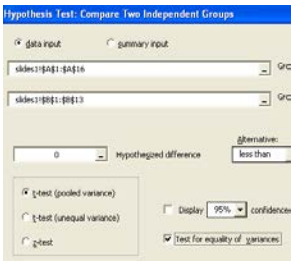
Example - Design

Research Hypotheses: **Ho: $\sigma_1=\sigma_2$ (choose the pooled variance t-test to compare means)**
Ha: $\sigma_1\neq\sigma_2$ (choose the unequal variance t-test to compare means)

Model will be F test for variances and the test statistic from the table will be F= $\frac{s_1^2}{s_2^2}$ (s_1 is larger). The degrees of freedom for numerator will be n_1-1=11 and the degrees of freedom for denominator will be n_2-1=14.

The test will be run at a level of significance (α) of 10%, but use the α=.05 table for a two-tailed test.

Critical Value for F with df_{num}=11 and df_{den}=14 is 2.57. Reject Ho if F >2.57.



We will also run this test the p-value way in Megastat.

Example - Data/Results

$F = {}^{14.894}/_{4.654} = 3.20$, which is more than critical value, Reject Ho.

Also p-value = 0.0438 < 0.10 which also makes the result significant.

F-test for equality of variance

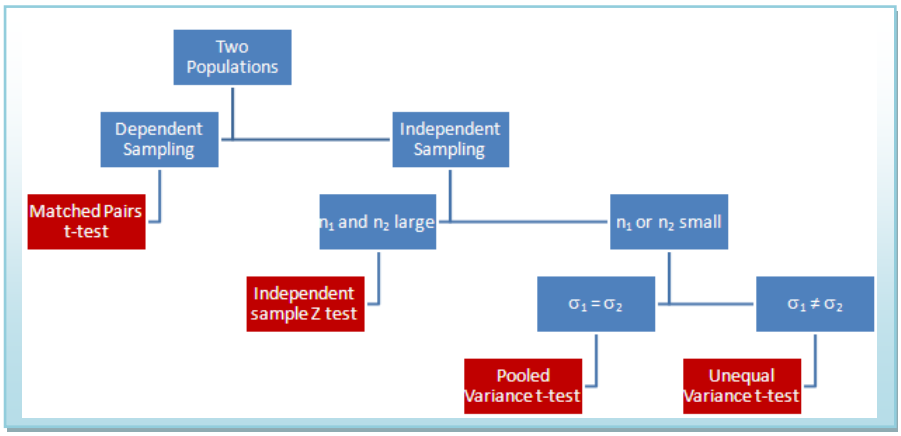
14.894	variance: import
4.654	variance: domestic
3.20	F
.0438	p-value

Example – Conclusion

Do not assume equal variances and run the unequal variance t-test to compare population means

In Summary

This flowchart summarizes which of the four models to choose when comparing two population means. In addition, you can use the F-test for equality of variances to make the decision between the pool variance t-test and the unequal variance t-test.



11. Chi-square Tests for Categorical Data

Often we want to conduct tests claims about the characteristics of qualitative or categorical non-numeric data. In Section 9, we covered a test of one population proportion. In reality, this was a test of a categorical variable with 2 choices (success, failure). Now in this section, we will expand our study of hypothesis tests involving categorical data to include categorical random variables with more than two choices using a goodness-of-fit test. In addition, we will compare two categorical variables for independence. Both of these models will use a Chi-square test statistic, by looking at deviations between the observed values and expected values of the data.

11.1.1 Chi-Square Goodness-of-Fit test

A financial services company had anecdotal evidence that people were calling in sick on Monday and Friday more frequently than Tuesday, Wednesday or Thursday. The speculation was that some employees were using sick days to extend their weekends. A researcher for the company was asked to determine if the data supported a significant difference in absenteeism due to the day of the week.

The categorical variable of interest here is "Day of Week" an employee called in sick (Monday through Friday). This is an example of a **multinomial** random variable, where we will observe a fixed number of trials (the total number of sick days sampled) and at least 2 possible outcomes. (A binomial random variable is a special case of the multinomial random variable where there is exactly 2 possible outcomes and was studied in Section 9 as a Z Test of Proportion.)

The Chi-square goodness-of-fit test is used to test if **observed** data from a categorical variable is consistent with an **expected** assumption about the distribution of that variable.

Chi-square Goodness of Fit Test

Model Assumptions

- $O_i =$ Observed in category i
- $p_i =$ Expected proportion in category i
- $E_i = np_i =$ Expected in category i
- $E_i \geq 5$ for each i

Test Statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad \text{df} = k-1$$

k = number of categories

n = sample size

11.1.2 Chi-Square Goodness-of-Fit test - Example 1 - equal expected frequencies

A researcher for the financial services company collected 400 records of what day of the week employees called in sick to work. Can the researcher conclude that proportion of employees who call in sick is not the same for each day of the week? Design and conduct a hypothesis test at the 1% significance level.

Day of Week	Frequency
Monday	95
Tuesday	65
Wednesday	60
Thursday	80
Friday	100
TOTAL	**400**

Research Hypotheses: **Ho:** There is a no difference in the proportion of employees who call in sick due to the day of the week.

Ha: There is a difference in the proportion of employees who call in sick due to the day of the week.

We can also state the hypotheses in terms of population parameters, p_i for each category. Under the null hypothesis we would expect 20% sick days would occur on each week day.

Research Hypotheses: **Ho:** $p_1 = p_2 = p_3 = p_4 = p_5 = 0.20$

Ha: At least one p_i is different than what was stated in Ho

Statistical Model: Chi-square goodness-of-fit test.

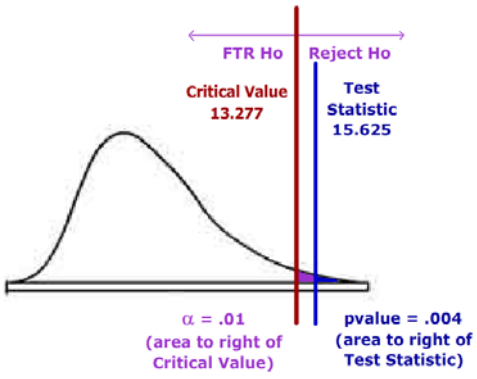
Important Assumption: The Expected Value of Each Category needs to be greater than or equal to 5. In this example, $E_i = np_i = (400)(.20) = 80 \geq 5$ for each category, so the model is appropriate.

Test Statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ df = 5-1=4

Decision Rule (Critical Value Method): Reject Ho if χ^2>13.277 (α=.01, 4df)

Results:

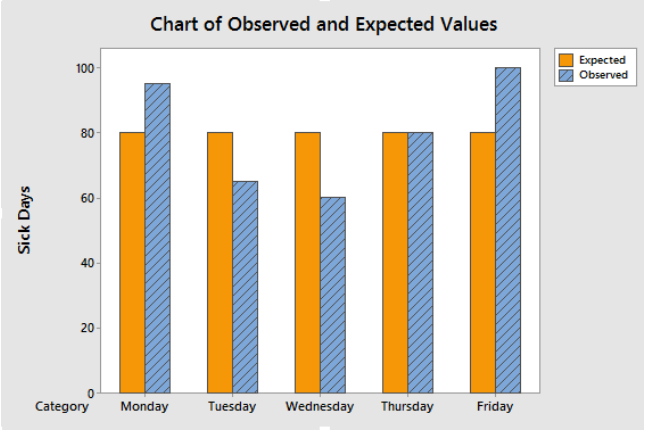
Day of Week	Observed Frequency O_i	Expected proportion p_i	Expected Frequency E_i	$\frac{(O_i - E_i)^2}{E_i}$
Monday	95	.20	80	2.8125
Tuesday	65	.20	80	2.8125
Wednesday	60	.20	80	5.0000
Thursday	80	.20	80	0.0000



Friday	100	.20	80	5.0000
TOTAL	**400**	**1.00**	**400**	**15.625**

Since the Test Statistic is in the Rejection Region, the decision is to **Reject Ho**. Under the p-value method, Ho is also rejected since the **p-value = P(χ^2>15.625) = 0.004** which is less than the Significance Level α of 1%.

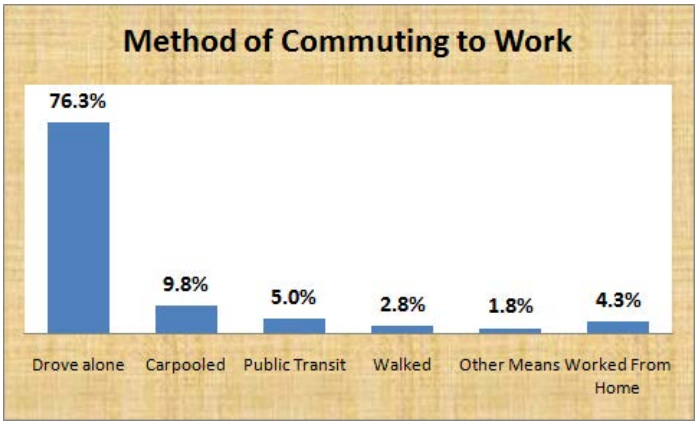
Conclusion: There is a difference in the proportion of employees who call in sick due to the day of the week. Employees are more likely to call in sick on days close to the weekend.



11.1.3 Chi-Square Goodness-of-Fit test - Example 1 - different expected frequencies.

In the prior example, the Null Hypothesis was that all categories had the same proportion, in other words there was no difference in counts due to the choices of a categorical variable. Another set of hypotheses using this same Chi-square goodness-of-fit test can be used to compare current results of an current experiment to prior results. In these tests, it is quite likely that prior proportions were not the same.

In the 2010 United States census, data was collected on how people get to work, their method of commuting. The results are shown in the graph to the right. Suppose you wanted to know if people who live in the San Jose metropolitan area (Santa Clara county) commute with similar proportions as the United States. We will sample 1000 workers from Santa Clara county and conduct a Chi-square goodness-of-fit test. Design and conduct a hypothesis test at the 1% significance level.



Research Hypotheses: **Ho:** Workers in Santa Clara county choose methods of commuting that match the United States averages.

Ha: Workers in Santa Clara county choose methods of commuting that do not match the United States averages.

We can also state the hypotheses in terms of population parameters, p_i for each category. Under the null hypothesis we would expect 20% sick days would occur on each week day.

Research Hypotheses: **Ho:** $p_1 = .763$ $p_2 = .098$ $p_3 = .050$ $p_4 = .028$ $p_5 = .018$ $p_6 = .043$
Ha: At least one p_i is different than what was stated in Ho

Statistical Model: Chi-square goodness-of-fit test.

Important Assumption: The Expected Value of Each Category needs to be greater than or equal to 5. In this example check the **lowest** p_i: $E_5 = np_5 = (1000)(.018) = 18 \geq 5$, so the model is appropriate.

Test Statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ df = 6-1=5

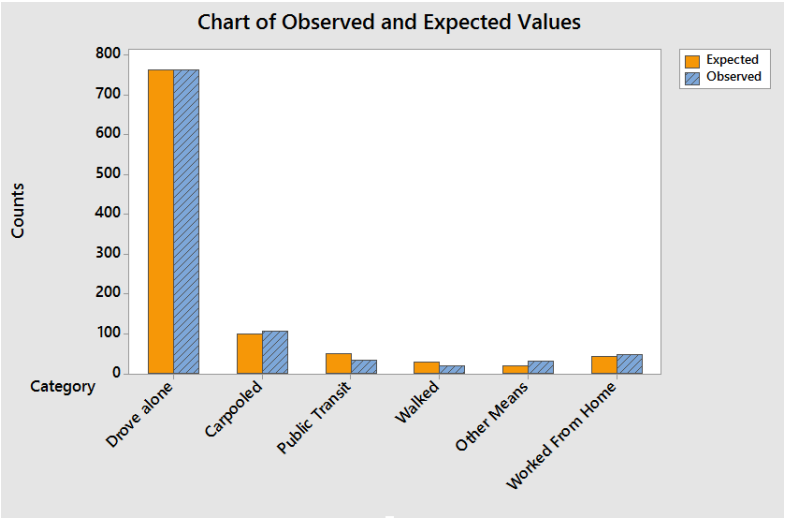
Decision Rule (Critical Value Method): Reject Ho if $\chi^2 > 11.071$ (α=.05, 5 df)

After designing the experiment, we conducted the sample of Santa Clara County, shown in the Observed Frequency Column of the table below. The Expected Proportion and Expected Frequency Columns are calculated using the U.S. 2010 Census.

Results:

Method Of Commuting	**Observed Frequency O_i**	**Expected Proportion p_i**	**Expected Frequency E_i**	$\frac{(O_i - E_i)^2}{E_i}$
Drive Alone	764	0.763	763	0.0013
Carpooled	105	0.098	98	0.5000
Public Transit	34	0.050	50	5.1200
Walked	20	0.028	28	2.2857
Other Means	30	0.018	18	8.0000
Worked from Home	47	0.043	43	0.3721
TOTAL	**1000**	**1.000**	**1000**	**16.2791**

Since the Test Statistic of 16.2791 exceeds the critical value of 11.071, the decision is to **Reject Ho**. Under the p-value method, Ho is also rejected since the **p-value = P(χ^2>16.2791) = 0.006** which is less than the Significance Level α of 5%.



Conclusion: Workers in Santa Clara County do not have the same frequencies of

method of commuting as the entire United States.

11.2.1 Chi-Square Test of Independence

In 2014, Colorado became the first state to legalize the recreational use of marijuana. Other states have joined Colorado, while some have decriminalized or authorized the medical use of marijuana. The question is should marijuana be legalized in all states. Suppose we took a poll of 1000 American adults and asked "Should marijuana be legal or not legal for recreational use" and got the following results:

Marijuana should be	Count	Percent
Legal	500	50%
Not Legal	450	45%
Don't know	50	5%
Total	**1000**	**100%**

The interpretation of this poll is that 50% of adults polled favored the legalization of marijuana for recreational use, while 45% opposed it. The remaining 5% were undecided.

At this time, you might have questions and want to explore this poll in more depth. For example, are younger people more likely to support legalization of marijuana? Do other demographic characteristics such as gender, ethnicity, sexual orientation, religion affect people's opinions about legalization.

Let us explore the possibility of difference of opinion due to gender. Are men more likely (or less likely) to oppose legalization of marijuana compared to women?

In the example above, suppose we have exactly 500 men and 500 women in the survey. What would we expect to see in the data if there was no difference in opinion between men and women?

11.2.2 Two-way tables

Two-way or contingency tables are used to summarize two categorical variables, also known as **bivariate** categorical data. In order to create a two-way table, the researcher must **cross-tabulate** the two responses for each categorical questions.

In the example above, the two categorical variables are gender and opinion on marijuana legalization. Gender has two choices (male or female) while opinion on marijuana legalization has three choices (legal, not legal and unsure).

In the example above, suppose we have exactly 500 men and 500 women in the survey. What would we expect to see in the data if there was no difference in opinion between men and women? We could then simply apply the total percentages to each group.

To create a hypothetical two-way table if there was no difference in opinion between men and women, apply the total percentages for each choice of Opinion to the total number for each choice of Gender.

eg: Men/Legal would 50% of 500 or 250 people.

Marijuana should be	**Men**	**Women**	**Total**
Legal	50%	50%	50%
Not Legal	45%	45%	45%
Unsure	5%	5%	5%
Total	**100%**	**100%**	**100%**

Marijuana should be	**Men**	**Women**	**Total**
Legal	250	250	500
Not Legal	225	225	450
Unsure	25	25	50
Total	**500**	**500**	**1000**

Let's review from probability what independence means. If two events A an B are independent, then the following statements are true:

P(A given B) = P(A)
P(B given A) = P(B)
P(A and B) = P(A)P(B)

You can pick any two events in the table above to verify that Gender and Opinion of Legalization of Marijuana are independent events. For example, compare the events Not **Legal** and **Men.**

P(Not Legal given Men) = 225/500 = 45% same as P(Not Legal) = 45%
P(Men given Not Legal) = 225/450 = 50% same as P(Men) = 50%
P(Not Legal and Men) = 225/1000 = 22.5% same as P(Not Legal)P(Men) = (45%)(50%) = 22.5%

Based on these probability rules we can calculate the expected value of any pair of independent events by using the following formula:

Expected Value = (Row Total)(Column Total)/(Grand Total)

For example, looking at the events **Not Legal and Men:**

Expected Value = (450)(500)/(1000) = 225

What if the events are not independent? Let's review the same survey. What would we expect to see in the data if there was a difference in opinion between men and women? Let's say women were more likely to support legalization. In that case, we would expect the 450 people who supported legalization of marijuana to have a higher number of women (and a smaller number of men) compared to the first table. Note we only change the first six boxes (shaded below), the totals must remain the same.

This is an example of a hypothetical two-way table where women were more likely to support legalization.

Only the six boxes shaded in yellow change from the prior example

Marijuana should be	**Men**	**Women**	**Total**
Legal	40%	60%	50%
Not Legal	55%	35%	45%
Unsure	5%	5%	5%
Total	**100%**	**100%**	**100%**

Marijuana should be	**Men**	**Women**	**Total**
Legal	200	300	500
Not Legal	275	175	450
Unsure	25	25	50
Total	**500**	**500**	**1000**

Now let's see the actual results of this survey and see what is happening:

Actual Poll of 500 men and 500 women adults. Should marijuana be legal for recreational use?

Marijuana should be	**Men**	**Women**	**Total**
Legal	54%	46%	50%
Not Legal	41%	49%	45%
Unsure	5%	5%	5%
Total	**100%**	**100%**	**100%**

Marijuana should be	**Men**	**Women**	**Total**
Legal	270	230	500
Not Legal	205	245	450
Unsure	25	25	50
Total	**500**	**500**	**1000**

In this poll, a higher percentage of men support legalization of marijuana for recreational use compared to women. Question: Is this evidence strong enough to support the claim that gender and opinion about marijuana legalization are not independent events? This question can addressed by conducting a hypothesis test using with the **Chi-square Test for Independence** model.

11.2.3 Chi-square test for Independence

Are Gender and Opinion about legalization of marijuana for recreational use independent events. Conduct a hypothesis test with a significance level of 5%.

Chi-square Test for Independence

Model Assumptions

- O_{ij} = Observed in category ij
- $E_{ij} = np_{ij} = \frac{(ColumnTotal)(RowTotal)}{GrandTotal}$

$E_{ij} \geq 5$ for each ij

Test Statistic

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad df = (r-1)(c-1)$$

r = number of row categories
c = number of column categories
n = sample size

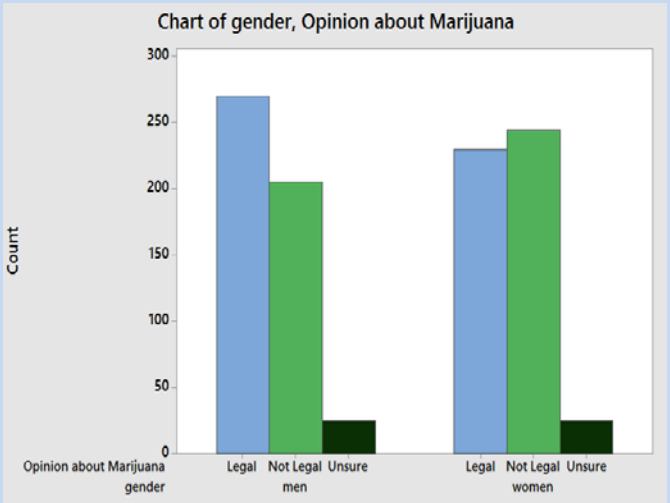
Research Hypotheses: **Ho:** Gender and Opinion about legalization of marijuana for recreational use are independent events.

Ha: Gender and Opinion about legalization of marijuana for recreational use are dependent events.

Statistical Model: Chi-square Test of Independence

Results

```
Rows: Opinion about Marijuana
Columns: gender

1st value = Observed
2nd Value = Expected
3rd Value = Contribution to Chi-square

               men  women   All

Legal          270    230   500
               250    250
             1.600  1.600

Not Legal      205    245   450
               225    225
             1.778  1.778

Unsure          25     25    50
                25     25
             0.000  0.000

All            500    500  1000
```



Important Assumption: The Expected Value of Each Category needs to be greater than or equal to 5. In this example, the lowest expected value is 225 (Men, not legal) so the assumption is met.

Test Statistic: $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}$ df = (3-1)(2-1)=2

Decision Rule (Critical Value Method): Reject Ho if χ^2>5.991 (α=.05, 2df)

$$\chi^2 = 1.600 + 1.600 + 1.778 + 1.778 = 6.756$$

Since the Test Statistic exceeds the critical value, the decision is to **Reject Ho**. Under the p-value method, Ho is also rejected since the **p-value = P(χ^2>6.756) = 0.034** which is less than the Significance Level α of 5%.

Conclusion: Gender and Opinion about legalization of marijuana for recreational use are dependent events. Men are more likely to support legalization of marijuana for recreational use.

12.One Factor Analysis of Variance (ANOVA)

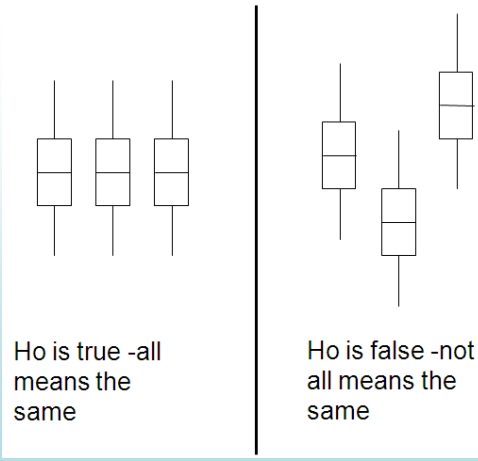
In the Section 7 we used statistical inference to compare two population means under variety of models. These models can be expanded to compare more than two populations using a technique called Analysis of Variance, or ANOVA for short. There are many ANOVA models, but we limit our study to one of them, the One Factor ANOVA model, also known as One Way ANOVA.

12.1 Comparing means from more than two Independent Populations

Suppose we wanted to compare the means of more than two (k) independent populations and want to test the null hypothesis $\boldsymbol{Ho: \mu_1 = \mu_2 = \cdots = \mu_k}$. If we can assume all population variances are equal, we can expand the pooled variance t-test for two populations to one factor ANOVA for k populations.

12.2 The logic of ANOVA - How comparing variances test for a difference in means.

It may seem strange to use a test of "variances" to compare means, but this graph demonstrates the logic of the test. If the null hypothesis $Ho: \mu_1 = \mu_2 = \mu_3$ is true, then each population would have the same distribution and the variance of the combined data would be approximately the same. However, if the Null Hypothesis is false, then the difference between centers would cause the combined data to have an increased variance.



12.3 The One Factor ANOVA model

In ANOVA, we calculate the variance two different ways: The mean square factor (MS_F),also know as mean square between, measures the variability of the means between groups, while the mean square within (MS_E), also know as mean square within, measures the variability within the population. Under the null hypothesis, the ratio of MS_F/MS_E should be close to 1 and has F distribution.

One Factor ANOVA model to compare the means of k independent populations

Model Assumptions

- The populations being sampled are normally distributed.
- The populations have equal standard deviations.
- The samples are randomly selected and are independent.

Test Statistic

$$F = \frac{MS_{Factor}}{MS_{Error}}$$

$$df_{num} = k - 1$$
$$df_{den} = n - k$$

12.4 Understanding the ANOVA table

When running Analysis of Variance, the data is usually organized into a special ANOVA table, especially when using computer software.

Source of Variation	Sum of Squares (SS)	Degrees of freedom (df)	Mean Square (MS)	F
Factor (Between)	SS_{Factor}	k-1	$MS_{Factor}= SS_{Factor}/k-1$	$F= MS_{Factor}/MS_{Error}$
Error (Within)	SS_{Error}	n-k	$MS_{Error}= SS_{Error}/n-k$	
Total	SS_{Total}	n-1		

Sum of Squares: The total variability of the numeric data being compared is broken into the variability between groups (SS_{Factor}) and the variability within groups (SS_{Error}). These formulas are the most tedious part of the calculation. T_c represents the sum of the data in each population and n_c represents the sample size of each population. These formulas represent the numerator of the variance formula.

$$SS_{Total} = \Sigma\left(X^2\right) - \frac{\left(\Sigma X\right)^2}{n} \qquad SS_{Factor} = \Sigma\left(\frac{T_c^2}{n_c}\right) - \frac{\left(\Sigma X\right)^2}{n} \qquad \mathrm{SS}_{\mathrm{Error}} = \mathrm{SS}_{\mathrm{Total}} - \mathrm{SS}_{\mathrm{Factor}}$$

Degrees of freedom: The total degrees of freedom is also partitioned into the Factor and Error components.

Mean Square: This represents calculation of the variance by dividing Sum of Squares by the appropriate degrees of freedom.

F: This is the test statistic for ANOVA: the ratio of two sample variances (mean squares) that are both estimating the same population value has an F distribution. Computer software will then calculate the p-value to be used in testing the Null Hypothesis that all populations have the same mean.

Example

Party Pizza specializes in meals for students. Hsieh Li, President, recently developed a new tofu pizza.

Before making it a part of the regular menu she decides to test it in several of her restaurants. She would like to know if there is a difference in the mean number of tofu pizzas sold per day at the Cupertino, San Jose, and Santa Clara pizzerias. Data will be collected for five days at each location.

At the .05 significance level can Hsieh Li conclude that there is a difference in the mean number of tofu pizzas sold per day at the three pizzerias?

Example - Design

Research Hypotheses: **Ho: $\mu_1=\mu_2=\mu_3$ (Mean sales same at all restaurants)**
Ha: At least μ_i is different (Means sales not the same at all restaurants)

We will assume the population variances are equal $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, so the model will be **One Factor ANOVA**. This model is appropriate if the distribution of the sample means is approximately Normal from the Central Limit Theorem.

Type I error would be to reject the Null Hypothesis and claim mean sales are different, when they actually are the same. The test will be run at a level of significance (α) of 5%.

The test statistic from the table will be F= $\frac{MSFactor}{MSError}$. The degrees of freedom for numerator will be 3-1=2 and the degrees of freedom for denominator will be 13-1=12. (The total sample size turned out to be only 13, not 15 as planned)

Critical Value for F at α of 5% with df_{num}=2 and df_{den}=12 is 4.10. Reject Ho if F >4.10. We will also run this test using the p-value method with statistical software, such as Minitab.

Example - Data/Results

	Cupertino	San Jose	Santa Clara	Total
	13	10	18	
	12	12	16	
	14	13	17	
	12	11	17	
			17	
T	51	46	85	182
n	4	4	5	13
Means	12.75	11.5	17	14
Σ^2	653	534	1447	2634

$$SS_{Total} = 2634 - \frac{182^2}{13} = 86$$

$$SS_{Factor} = 2624.25 - \frac{182^2}{13} = 76.25$$

$$SS_{Error} = 86 - 76.25 = 9.75$$

$F = {}^{38.125}/_{0.975} = 39.10$, which is more than critical value of 4.10, Reject Ho.

Also from the Minitab output, p-value = 0.000 < 0.05 which also supports rejecting Ho.

One-way ANOVA: Cupertino, San Jose, Santa Clara

```
Source  DF      SS      MS      F      P
Factor   2  76.250  38.125  39.10  0.000
Error   10   9.750   0.975
Total   12  86.000

S = 0.9874   R-Sq = 88.66%   R-Sq(adj) = 86.40%

                                  Individual 95% CIs For Mean Based on
                                  Pooled StDev
Level        N    Mean  StDev  --------+---------+---------+---------+-
Cupertino    4  12.750  0.957        (-----*-----)
San Jose     4  11.500  1.291  (-----*-----)
Santa Clara  5  17.000  0.707                              (-----*-----)
                                  --------+---------+---------+---------+-
                                       12.0      14.0      16.0      18.0
```

Example – Conclusion

There is a difference in the mean number of tofu pizzas sold at the three locations.

12.5 Post-hoc Analysis – Tukey's Honestly Significant Difference (HSD) Test[16].

When the Null Hypothesis is rejected in one factor ANOVA, the conclusion is that not all means are the same. This however leads to an obvious question: Which particular means are different? Seeking further information after the results of a test is called post-hoc analysis.

12.5.1 The problem of multiple tests

One attempt to answer this question is to conduct multiple pairwise independent same t-tests and determine which ones are significant. We would compare μ_1 to μ_2, μ_1 to μ_3, μ_2 to μ_3, μ_1 to μ_4, etc. There is a major flaw in this methodology in that each test would have a significance level of α, so making Type I error would be significantly more than the desired α. Furthermore, these pairwise tests would NOT be mutually independent. There were several statisticians who designed tests that effectively dealt with this problem of determining an "honest" significance level of a set of tests; we will cover the one developed by John Tukey, the Honestly Significant Difference (HSD) test.

12.5.2 The Tukey HSD test

Tests: $H_o : \mu_i = \mu_j \qquad H_a : \mu_i \neq \mu_j$ where the subscripts i and j represent two different populations

Overall significance level of α. This means that **all pairwise tests** can be run at the same time with an overall significance level of α.

Test Statistic: $HSD = q\sqrt{\frac{MSE}{n_c}}$

q = value from studentized range table

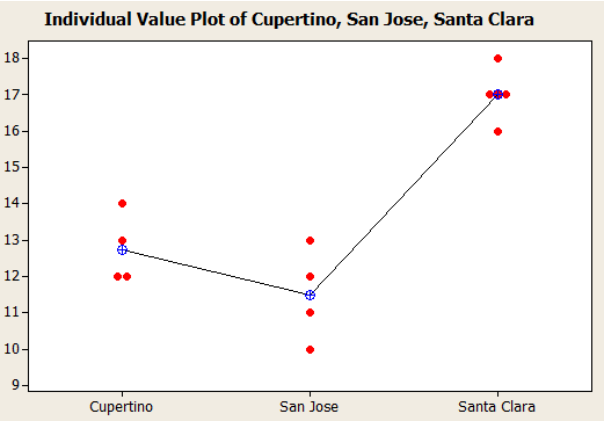
MSE = Mean Square Error from ANOVA table

n_c = number of replicates per treatment. An adjustment is made for unbalanced designs.

Decision: Reject Ho if $\left|\overline{X}_i - \overline{X}_j\right| > HSD$ critical value

Computer software, such as Megastat, will calculate the critical values and test statistics for these series of tests.

Example



Let us return to the Tofu pizza example where we rejected the Null Hypothesis and supported the claim that there was a difference in means among the three restaurants.

In reviewing the graph of the sample means, it appears that Santa Clara has a much higher number of sales than Cupertino and San Jose. There will be three pairwise post-hoc tests to run.

Example - Design

$H_o : \mu_1 = \mu_2 \quad H_a : \mu_1 \neq \mu_2 \qquad H_o : \mu_1 = \mu_3 \quad H_a : \mu_1 \neq \mu_3 \qquad H_o : \mu_2 = \mu_3 \quad H_a : \mu_2 \neq \mu_3$

These three tests will be conducted with an overall significance level of α = 5%.
The model will be the Tukey HSD test.

The Minitab approach for the decision rule will be to reject Ho for each pair that does not share a common group.

Example - Data/Results/Conclusion

Refer to the Minitab output. Santa Clara is in group A while Cupertino and San Jose are in Group B.

```
Grouping Information Using Tukey Method

             N     Mean  Grouping
Santa Clara  5  17.0000  A
Cupertino    4  12.7500    B
San Jose     4  11.5000    B

Means that do not share a letter are significantly different.
```

Santa Clara has a significantly higher mean number of tofu pizzas sold compared to both San Jose and Cupertino. There is no significant difference in mean sales between San Jose and Cupertino.

12.6 Factorial Design – an insight to other ANOVA procedures

A different way of looking at this model is considering a single population with one numeric and one categorical variable being sampled. The numeric variable is called the **response** (tofu pizzas sold) and the categorical variable is the **factor** (location of restaurant). The possible responses to the factor are called the **levels** (Cupertino, San Jose and Sunnyvale). The number of observations per level are called the replicates (n_1=4, n_2=4, n_3=5 in our example). If the replicates are equal, the design is **balanced.** (our example is not balanced).

By thinking of the model in this way, it easy to extend the concept to the multi-factor ANOVA models that are prevalent in the research you will encounter in future studies.

13. Correlation and Linear Regression

Often in statistical research, we want to discover if there is a relationship between two variables. The **explanatory variable** is the "cause" and the **response variable** is the "effect", although a true cause and effect relationship can only be established in a scientific study that controls for all confounding (lurking) variables.

In Chapter 11, we were interested in determining if a person's gender was a valid explanatory variable of the person's opinion about legalization of marijuana for recreational use. In this case, both the explanatory and response variables are categorical and the appropriate model was the Chi-square Test of Independence.

In Chapter 12, we explored if tofu pizza sales (the response variable) were explained by location of the restaurant (the explanatory variable). In this case, the explanatory variable was categorical but the response was numeric. The appropriate model for this example is One Factor Analysis of Variance (ANOVA).

What if we want to determine if a relationship exists when both the explanatory and response variables are both numeric? For example, does annual rainfall in a city help explain sales of sunglasses? This chapter explores and defines the appropriate model for this type of problem.

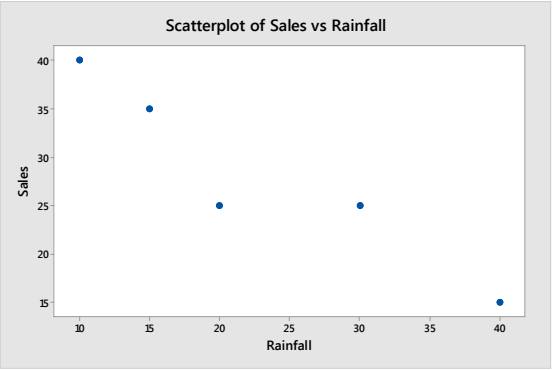
13.1 Bivariate data and scatterplots

In chapter 2, we defined bivariate data as data that has two different numeric variables. In an algebra class, these are also known as ordered pairs. We will let X represent the **independent** (or explanatory) variable and Y represent the **dependent** (or response) variable in this definition. Here is an example of five total pairs where X represents the annual rainfall in inches in a city and Y represents annual sales of sunglasses per 1000 population.

The best way to graph bivariate data is by using a **Scatterplot** where X, the independent variable is the vertical axes and Y, the dependent variable is the horizontal axis.

X=rainfall	Y=sales
10	40
15	35
20	25
30	25
40	15

Here is an example and scatterplot of five total pairs where X represents the annual rainfall in inches in a city and Y represents annual sales of sunglasses per 1000 population.



In the scatterplot for this data, it appears that cities with more rainfall have lower sales. It also appears that this relationship is linear, which can than go forward to explain in a statistical model.

13.2 The Simple Linear Regression Model

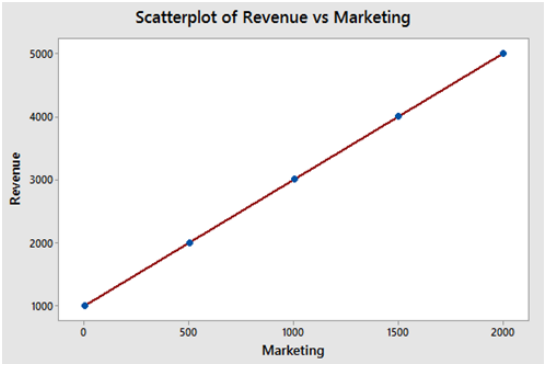
In the scatterplot example shown above, we saw linear correlation between the two dependent variables. We are now going to create a statistical model relating these two variables, but let's start by reviewing a **mathematical linear model** from algebra:

$Y = \beta_0 + \beta_1 X$
Y : *Dependent Variable*
X : *Independent Variable*
β_0 : Y *– intercept*
β_1 : *Slope*

Example: You have a small business producing custom t-shirts. Without marketing, your business has revenue (sales) of $1000 per week. Every dollar you spend marketing will increase revenue by 2 dollars. Let variable X represent amount spent on marketing and let variable Y represent revenue per week. Write a **mathematical model** that relates X to Y.

In this example, we are saying that weekly revenue (Y) depends on marketing expense (X). $1000 of weekly revenue represents the vertical intercept, and $2 of weekly revenue per $1 marketing represents the slope, or rate of change of the model. We can choose some value of X and determine Y and then plot the points on a scatterplot to see this linear relationship.

X=marketing	Y=revenue
$0	$1000
$500	$2000
$1000	$3000
$1500	$4000
$2000	$5000

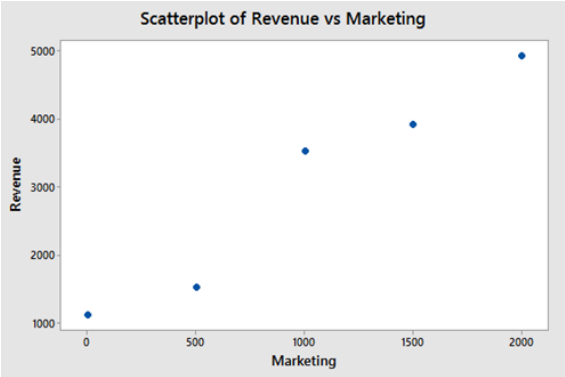


We can then write out the mathematical linear model as an equation:

Linear Model	Example
$Y = \beta_0 + \beta_1 X$	$Y = 1000 + 2X$
Y: *Dependent Variable*	Y: *Revenue*
X: *Independent Variable*	X: *Marketing*
β_0: Y *– intercept*	β_0: $1000
β_1: *Slope*	β_1: $2 *per* $1 *marketing*

We all learned about these linear models in algebra classes, but the real world doesn't generally give such perfect results. In particular, we can choose what to spend on marketing, but the actual revenue will have more uncertainty. For example, the true revenue may look more like this:

X=Marketing	Expected Revenue	Y=Actual Revenue	ε=Residual Error
$0	$1000	$1100	+$100
$500	$2000	$1500	-$500
$1000	$3000	$3500	+$500
$1500	$4000	$3900	-$100
$2000	$5000	$4900	-$100



The difference between the actual revenue and the expected revenue is called the **residual error,** ε If we assume the residual error (represented by ε) is a random variable that follows a normal distribution with μ = 0 and σ a constant for all values of X, we have now created a **statistical model** called a **simple linear regression model**.

Regression Model

$Y = \beta_0 + \beta_1 X + \varepsilon$

Y: *Dependent Variable*

X: *Independent Variable*

β_0: *Y – intercept*

β_1: *Slope*

ε: $Normal(0, \sigma)$

Example

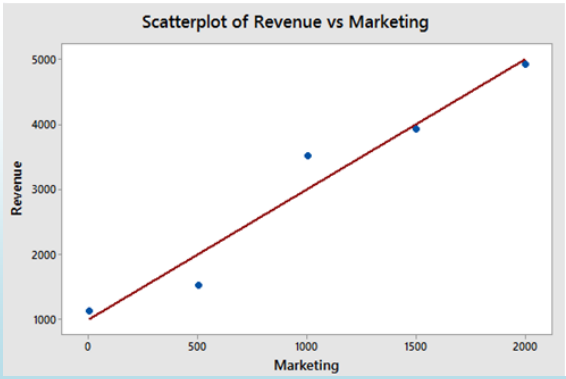
$Y = 1000 + 2X + \varepsilon$

Y: *Revenue*

X: *Marketing*

β_0: $1000

β_1: $2 per $1 marketing



13.3 Estimating the Regression Model with the least square line

We now return to the case where we know the data and can see the linear correlation in a scatterplot, but we do not know the values of the parameters of the underlying model. The three parameters that are unknown to us are the y-intercept β_o, the slope (β_1) and the standard deviation of the residual error (σ):

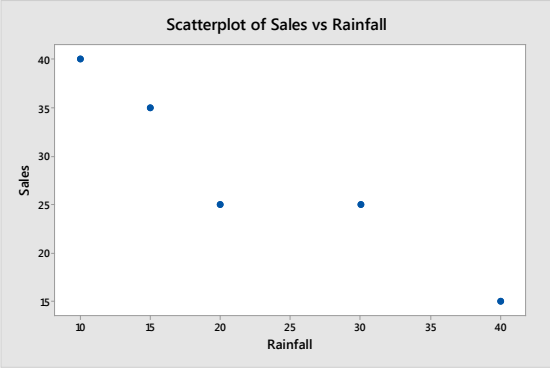
Slope parameter: b_1 will be an estimator for β_1

Y-intercept parameter: b_0 will be an estimator for β_0

Standard deviation: s_e will be an estimator for σ

Regression line: $\hat{Y} = b_0 + b_1 X$

Take the example comparing rainfall to sales of sunglasses where the scatterplot shows a negative correlation. However, there are many lines we could draw. How do we find the line of best fit?



Minimizing Sum of Squared Residual Errors (SSE)

We are going to define the "best line" as the line that minimizes the Sum of Squared Residual Errors (SSE).

Suppose we try to fit this data with a line that goes through the first and last point. We can then calculate the equation of this line using algebra to be : $\hat{Y} = \frac{145}{3} + \frac{5}{6}X$. The SSE for this line is 47.917:

Rainfall	Sales	Predicted Sales	Residual	Squared Residuals
10	40	40	0	0
15	35	35.833	-0.833	0.694
20	25	31.667	-6.667	44.444
30	25	23.333	1.667	2.778
40	15	15	0	0
		Sum of Squared Residuals =		47.917

Although this line is a good fit it, it is not the best line. The slope and intercept for the line that minimizes SSE is be calculated using the least squares principle:

$$SSX = \Sigma X^2 - \frac{1}{n}\left(\Sigma X\right)^2$$

$$SSY = \Sigma Y^2 - \frac{1}{n}\left(\Sigma Y\right)^2$$

$$SSXY = \Sigma XY - \frac{1}{n}\left(\Sigma X \cdot \Sigma y\right)$$

$$b_1 = \frac{SSXY}{SSX}$$

$$b_0 = \overline{Y} - b_1\overline{X}$$

In the Rainfall example where X=Rainfall and Y = Sales of Sunglasses:

	X	Y	X^2	Y^2	XY
	10	40	100	1600	400
	15	35	225	1225	525
	20	25	400	625	500
	30	25	900	625	750
	40	15	1600	225	600
Σ	115	140	3225	4300	2775

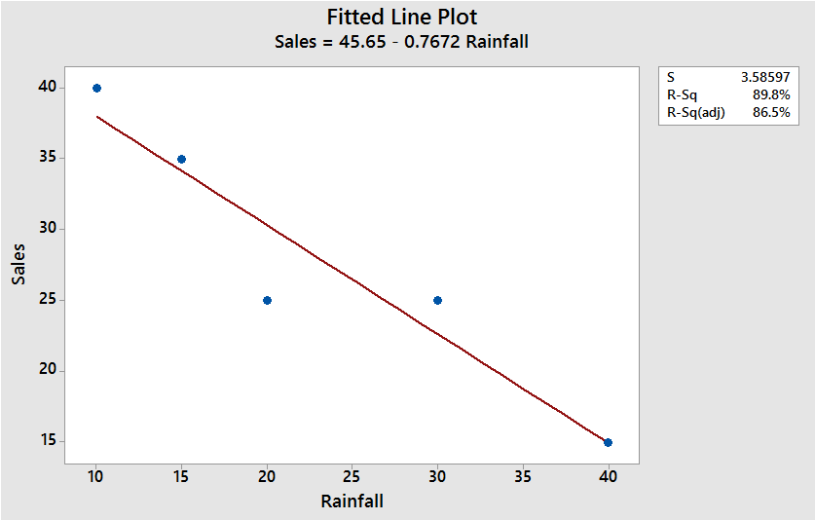
- SSX = 580
- SSY = 380
- SSXY = -445

- b_1 = -.767
- b_0 = 45.647
- $\hat{Y}$ = 45.647 - .767X

The Sum of Squared Residual Errors (SSE) for this line is 38.578, making it the "best line". (Compare to the value above when we picked the line perfectly fitting the two most extreme points).

Rainfall	Sales	Predicted Sales	Residual	Squared Residuals
10	40	37.977	2.023	4.092529
15	35	34.142	0.858	0.736
20	25	30.307	-5.307	28.164
30	25	22.637	2.363	5.584
40	15	14.967	0.033	0.001089
		Sum of Squared Residuals =		38.578

In practice, we will use technology such as Minitab to calculate this line. Here is the example using the Regression Fitted Line Plot option in Minitab which determines and graphs the regression equation. The point (20,25) has the highest residual error, but the overall Sum of Squared Residual Errors (SSE) is minimized.



13.4 Hypothesis test for Simple Linear Regression

We will now describe a hypothesis test to determine if the regression model is meaningful, in other words, does the value of X in any way help predict the expected value of Y?

In simple linear regression this is equivalent to saying "Are X an Y correlated?"

In reviewing the model, $Y = \beta_0 + \beta_1 X + \varepsilon$, as long as the slope ($\beta_1$) has any non-zero value, X will add value in helping predict the expected value of Y. However, if there is no correlation between X and Y, the value of the slope (β_1) will be zero. The model we can use is very similar to One Factor ANOVA.

Simple Linear Regression ANOVA Hypothesis Test

Model Assumptions

- The residual errors are random and are normally distributed.
- The standard deviation of the residual error does not depend on X
- A linear relationship exists between X and Y
- The samples are randomly selected

Test Hypotheses

Ho: X and Y are not correlated
Ha: X and Y are correlated

Ho: β_1 (slope) = 0
Ha: β_1 (slope) ≠ 0

Sum of Squares

$$SS_{Total} = \sum \left(Y - \bar{Y}\right)^2$$

$$SS_{Error} = \sum \left(Y - \hat{Y}\right)^2$$

$$SS_{\text{Regression}} = SS_{Total} - SS_{Error}$$

Test Statistic

$$F = \frac{MS_{Regression}}{MS_{Error}}$$

$$df_{num} = 1$$
$$df_{den} = n - 2$$

The Results of the test can be summarized in a special ANOVA table:

Source of Variation	Sum of Squares (SS)	Degrees of freedom (df)	Mean Square (MS)	F
Factor (due to X)	$SS_{Regression}$	1	$MS_{Factor=}$ $SS_{Factor}/1$	$F= MS_{Factor}/MS_{Error}$
Error (Residual)	SS_{Error}	n-2	$MS_{Error=}$ $SS_{Error}/n-2$	
Total	SS_{Total}	n-1		

Example - Design: Is there a significant correlation between rainfall and sales of sunglasses?

Research Hypotheses:

Ho: Sales and Rainfall are not correlated — Ho: β_1 (slope) = 0
Ha: Sales and Rainfall are correlated — Ha: β_1 (slope) ≠ 0

Type I error would be to reject the Null Hypothesis and claim sales are correlated with sales of sunglasses, when they are not correlated. The test will be run at a level of significance (α) of 5%.

The test statistic from the table will be F= $\frac{MSRegressison}{MSError}$. The degrees of freedom for numerator will be 1 and the degrees of freedom for denominator will be 5-2=3.

Critical Value for F at α of 5% with df_{num}=1 and df_{den}=3 is 10.13. Reject Ho if F >10.13. We will also run this test using the p-value method with statistical software, such as Minitab.

Example - Data/Results

Source	SS	df	MS	F	p-value
Regression	341.422	1	341.422	26.551	0.0142
Error	38.578	3	12.859		
TOTAL	380.000	4			

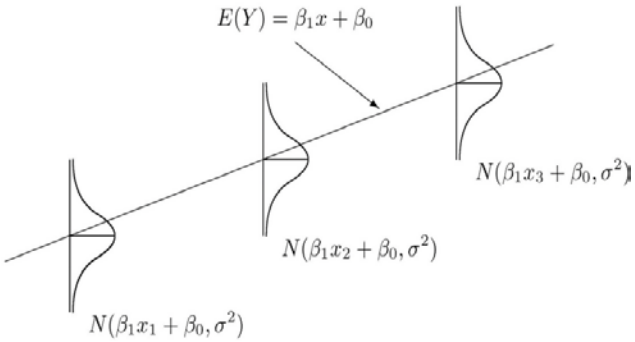
$F = 341.422/12.859 = 26.551$ which is more than critical value of 10.13, so Reject Ho. Also, the p-value = 0.0142 < 0.05 which also supports rejecting Ho.

Example – Conclusion

Sales of Sunglasses and Rainfall are negatively correlated.

13.5 Estimating σ, the standard error of the residuals

The simple linear regression model $\left(Y = \beta_0 + \beta_1 X + \varepsilon\right)$ includes a random variable ε representing the residual which follows a Normal Distribution with an expected value of 0 and a standard deviation σ that is independent of the value of X. The estimate of σ is called the sample standard error of the residuals and is represented by the symbol s_e. We can use the fact that the Mean Square Error (MSE) from the ANOVA table represents the estimated variance of the residuals errors:



$$s_e = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

Example

For the rainfall data, the standard error of the residuals is determined as:

$$s_e = \sqrt{12.859} = 3.586$$

Keep in mind this is the standard deviation of the residual errors and should not be confused with the standard deviation of Y.

13.6 r^2, the Correlation of determination

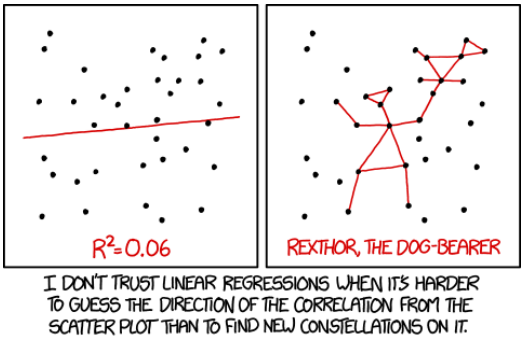
The Regression ANOVA hypothesis test can be used to determine if there is a **significant** correlation between the independent variable (X) and the dependent variable (Y). We now want to investigate the **strength** of correlation.

In the earlier chapter on descriptive statistics, we introduced the correlation coefficient (r), a value between -1 and 1. Values of r close to 0 meant there was little correlation between the variables, while values closer to 1 or -1 represented stronger correlations.

In practice, most statisticians and researchers prefer to use r^2, the coefficient of determination as a measure of strength as it represents the proportion or percentage of the variability of Y that is explained by the variability of X. [17]

$$r^2 = \frac{SS_{regression}}{SS_{Total}} \qquad 0\% \le r^2 \le 100\%$$

r^2 represents the percentage of the variability of Y that is explained by the variability of X.



We can also calculate the correlation coefficient (r) by taking the appropriate square root of r^2, depending on whether the estimate of the slope (b_1) is positive or negative:

$$If\ b_1 > 0, r = \sqrt{r^2}$$
$$If\ b_1 < 0, r = -\sqrt{r^2}$$

Example

For the rainfall data, the coefficient of determination is:

$$r^2 = \frac{341.422}{380} = 89.85\%$$

89.85% of the variability of sales of sunglasses is explained by rainfall.

We can calculate the correlation coefficient (r) by taking the appropriate square root of r^2:

$$r = -\sqrt{.8985} = -0.9479$$

Here we take the negative square root since the slope of the regression line is negative. This shows that there is a strong, negative correlation between sales of sunglasses and rainfall.

13.7 Prediction

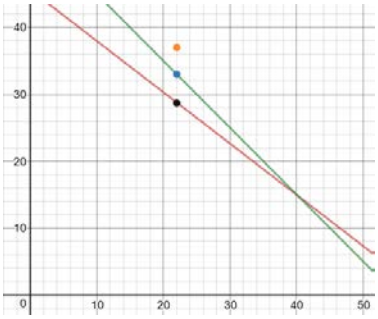
One valuable application of the regression model is to make predictions about the value of the dependent variable if the independent variable is known.

Consider the example about rainfall and sales of sunglasses. Suppose we know a city has 22 inches of rainfall. We can use the regression equation to predict the sales of sunglasses:

$$\hat{Y} = 45.647 - .767X$$
$$\hat{Y}_{22} = 45.647 - .767(22) = 28.7$$

For a city with 22 inches of annual rainfall, the model predicts sales of 28.7 per 1000 population.

To measure the **reliability** of this prediction, we can construct confidence intervals. However, we first have to decide what we are estimating. We could (1) be estimating the **expected** sales for a city with 22 inches of rainfall, or we could (2) be predicting the **actual** sales for a city with 22 inches of rainfall.



In the graph shown, the green line represents $Y = \beta_0 + \beta_1 X + \varepsilon$ the actual regression line which is unknown. The red line represents the least square equation, $\hat{Y} = 45.647 - .767X$, which is derived from the data. The black dot represents our prediction Y_{22}=28.7. The green dot represents the correct population **expected** value of Y_{22} while the yellow dot represents a possible value for the **actual** predicted value of Y_{22}. There is more uncertainty in predicting an actual value of Y_x then the expected value.

The **confidence interval** for the **expected** value of Y for a given value of X is given by:	The **prediction interval** for the **actual** value of Y for a given value of X is given by:
$\hat{Y}_X \pm t \cdot s_e \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{SSX}}$	$\hat{Y}_X \pm t \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{SSX}}$
Degrees of freedom for t =n-2	Degrees of freedom for t =n-2

Example

Find a 95% confidence interval for the expected value of sales for a city with 22 inches of rainfall.

$$28.7 \pm 3.182 \cdot 3.586 \sqrt{\frac{1}{5} + \frac{(22-23)^2}{580}} = 28.7 \pm 5.1 \rightarrow (23.6, 33.8)$$

We are 95% confident that the expected annual sales of sunglasses for a city with 22 inches of annual rainfall is between 23.6 and 33.8 sales per 1000 population.

Find a 95% prediction interval for the actual value of sales for a city with 22 inches of rainfall.

$$28.7 \pm 3.182 \cdot 3.586 \sqrt{1 + \frac{1}{5} + \frac{(22-23)^2}{580}} = 28.7 \pm 12.5 \rightarrow (16.2, 41.2)$$

We are 95% confident that the actual annual sales of sunglasses for a city with 22 inches of annual rainfall is between 16.2 and 41.2 sales per 1000 population.

Extrapolation

When using the model to make predictions, care must be taken to only choose values of X that are in the range of X values of the data. In the rainfall/sales example, the values of X range from 10 to 40 inches of rainfall. Choosing a value of X outside this range is called extrapolation and could lead to invalid results. For example, if we use the model to predict sales for a city with 80 inches of rainfall, we get an impossible negative result for sales:

$$\hat{Y} = 45.647 - .767X$$

$$\hat{Y}_{80} = 45.647 - .767(80) = -15.7$$

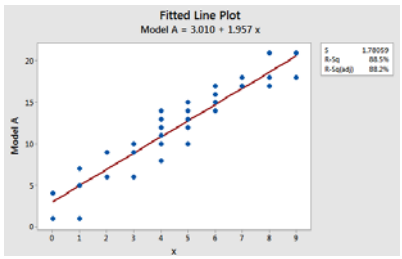
13.9 Residual Analysis

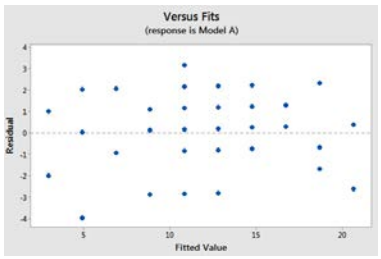
In regression, we assume the model is linear and the residual errors ($Y - \hat{Y}$ for each pair) are random and are normally distributed. We can analyze the residuals to see if these assumptions are valid and if there are any potential outliers. In particular:

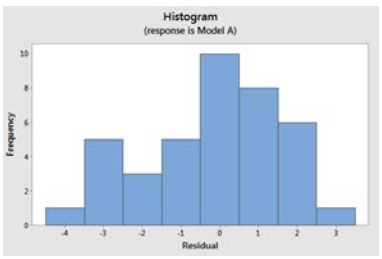
- The residuals should represent a linear model.
- The standard error (standard deviation of the residuals) should not change when the value of X changes.
- The residuals should follow a normal distribution.
- Look for any potential extreme values of X.
- Look for any extreme residual errors.

Example - Model A

Model A is an example of an appropriate linear regression modelmodel. We will make three graphs to test the residual, a scatterplot with the regression line, a plot of the residuals, and a histogram of the residuals.

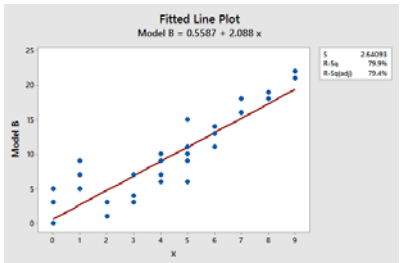


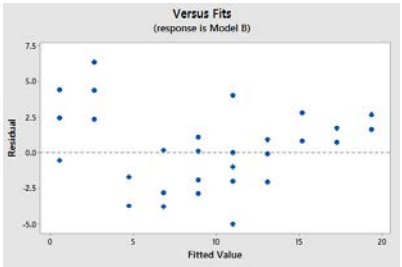


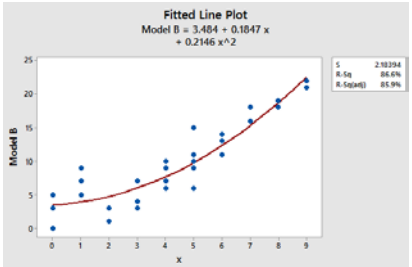


Here we can see the residuals appear to be random, the fit is linear, and the histogram is approximately bell shaped. In addition, there are no extreme outlier values of X or outlier residuals.

Example - Model B

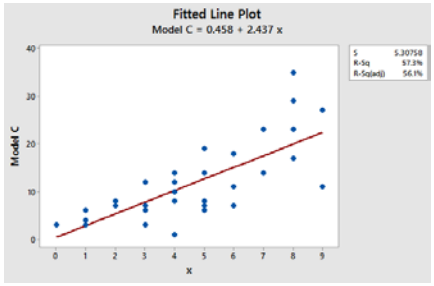


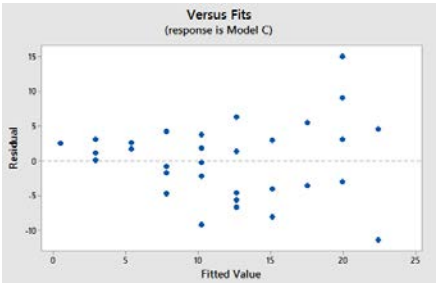




Model B looks like a strong fit, but the residuals are showing a pattern of being positive for low and high values of X and negative for middle values of X. This indicates that the model is not linear and should be fit with a non-linear regression model (for example, the third graph shows a quadratic model).

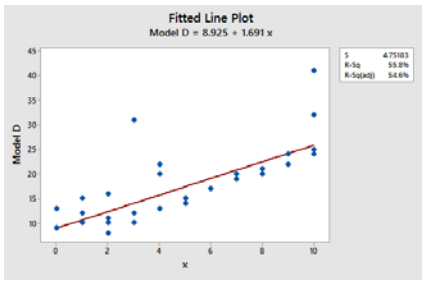
Example - Model C

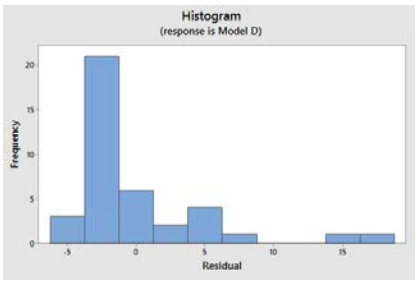




Model C has a linear fit, but the residuals are showing a pattern of being smaller for low values of X and higher for large values of X. This violates the assumption that the standard error should not change when the value of X changes. This phenomena is called **heteroscedasticity**, and requires a data transformation to find a more appropriate model.

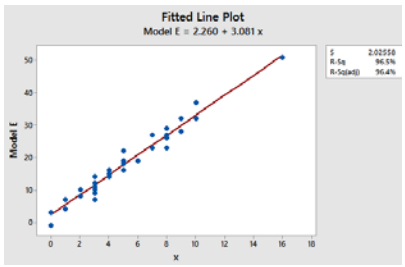
Example - Model D

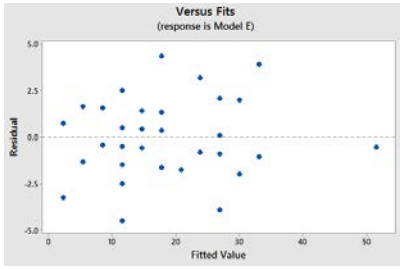


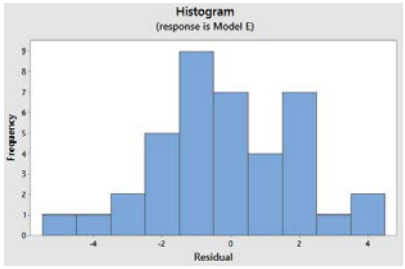


Model D seems to have a linear fit, but the residuals are showing a pattern of being larger when they are positive and smaller when they are negative. This violates the assumption that residuals should follow a normal distribution, as can be seen in the histogram.

Example - Model E

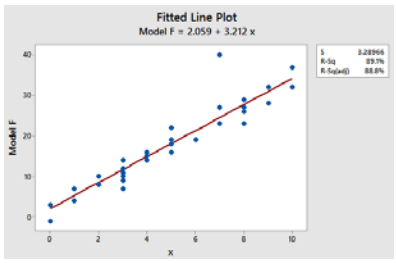


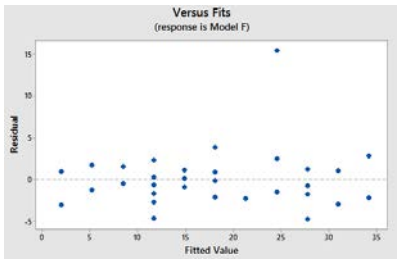


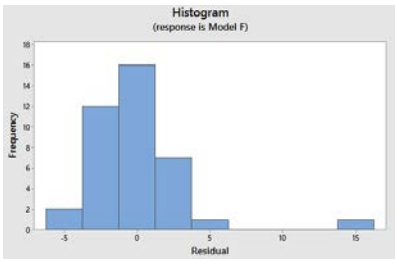


Model E seems to have a linear fit, and the residuals look random and normal. However, the value (16,51) is an extreme outlier value of X and may have an undue influence on the choosing of the regression line.

Example - Model F







Model F seems to have a linear fit, and the residuals look random and normal, except for one outlier at the value (7,40). This outlier is different than the extreme outlier in Model E, but will still have an undue influence on the choosing of the regression line.

14.Glossary of Statistical Terms used in Inference

Alpha (α) – see **Level of Significance**

Alternative Hypothesis (H_a)
A statement about the value of a population parameter that is assumed to be true if the Null Hypothesis is rejected during testing.

Analysis of Variance (ANOVA)
A group of statistical tests used to determine if the mean of a numeric variable (the Response) is affected by one or more categorical variables (Factors).

Beta (β)
The probability, set by design, of failing to reject the Null Hypothesis when it is actually false. Beta is calculated for specific possible values of the Alternative Hypothesis.

Central Limit Theorem
A powerful theorem that allows us to understand the distribution of the sample mean, $\bar{X}$. If X_1, X_2, …, X_n is a random sample from a probability distribution with mean = μ and standard deviation = σ and the sample size is "sufficiently large", then $\bar{X}$ will have a Normal Distribution with the same mean a standard deviation of $\sigma/\sqrt{n}$ (also known as the Standard Error). Because of this theorem, most statistical inference is conducting using a sampling distribution from the Normal Family.

Chi-square Distribution (χ^2)
A family of continuous random variables (based on degrees of freedom) with a probability density function that is from the Normal Family of probability distributions. The Chi-square distribution is non-negative and skewed to the right and has many uses in statistical inference such as inference about a population variance, goodness-of-fit tests and test of independence for categorical data.

Confidence Interval
An Interval estimate that estimates a population parameter from a random sample using a predetermined probability called the level of confidence.

Confidence Level – see **Level of Confidence**

Critical value(s)
The dividing point(s) between the region where the Null Hypothesis is rejected and the region where it is not rejected. The critical value determines the decision rule.

Decision Rule
The procedure that determines what values of the result of an experiment will cause the Null Hypothesis to be rejected. There are two methods that are equivalent decision rules:

1. If the test statistic lies in the Rejection Region, Reject Ho. (Critical Value method)
2. If the p-value < α, Reject Ho. (p-value method)

Dependent Sampling
A method of sampling where 2 or more variables are related to each other (paired or matched). Examples would be the "Before and After" type models using the Matched Pairs t-test.

Effect Size: The "practical difference" between a population parameter under the Null Hypothesis and a selected value of the population parameter under the Alternative Hypothesis.

Empirical Rule (Also known as the 68-95-99.7 Rule)
A rule used to interpret standard deviation for data that is approximately bell-shaped. The rule says about 68% of the data is within one standard deviation of the mean, 95% of the data is within two standard deviations of the mean, and about 99.7% of the data is within three standard deviations of the mean.

Estimation
An inference process that attempts to predict the values of population parameters based on sample statistics.

F Distribution
A family of continuous random variables (based on 2 different degrees of freedom for numerator and denominator) with a probability density function that is from the Normal Family of probability distributions. The F distribution is non-negative and skewed to the right and has many uses in statistical inference such as inference about comparing population variances, ANOVA, and regression.

Factor
In ANOVA, the categorical variable(s) that break the numeric response variable into multiple populations or treatments.

Hypothesis
A statement about the value of a population parameter developed for the purpose of testing.

Hypothesis Testing
A procedure, based on sample evidence and probability theory, used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

Independent Sampling
A method of sampling where 2 or more variables are not related to each other. Examples would be the "Treatment and Control" type models using the independent samples t-test.

Inference – see **Statistical Inference**

Interval Estimate
A range of values based on sample data that used to estimate a population parameter.

Level
In ANOVA, a possible value that a categorical variable factor could be. For example, if the factor was shirt color, levels would be blue, red, yellow, etc.

Level of Confidence
The probability, usually expressed as a percentage, that a Confidence Interval will contain the true population parameter that is being estimated.

Level of Significance (α)
The maximum probability, set by design, of rejecting the Null Hypothesis when it is actually true (maximum probability of making Type I error).

Margin of Error
The distance in a symmetric Confidence Interval between the Point Estimator and an endpoint of the interval. For example a confidence interval for μ may be expressed as $\bar{X} \pm$ Margin of Error.

Model Assumptions

Criteria which must be satisfied to appropriately use a chosen statistical model. For example, a student's t statistic used for testing a population mean vs. a hypothesized value requires random sampling and that the sample mean has an approximately Normal Distribution.

Normal Distribution
Often called the "bell-shaped" curve, the Normal Distribution is a continuous random variable which has Probability Density Function $X = exp[-(x-\mu)^2/2\sigma^2]/\sigma\sqrt{2\pi}$. The special case where $\mu = 0$ and $\sigma = 1$, is called the **Standard Normal Distribution** and designated by Z.

Normal Family of Probability Distributions
The Standard Normal Distribution (Z) plus other Probability Distributions that are functions of independent random variables with Standard Normal Distribution. Examples include the t, the F and the Chi-square distributions.

Null Hypothesis (H_0)
A statement about the value of a population parameter that is assumed to be true for the purpose of testing.

Outlier
A data point that is far removed from the other entries in the data set.

p-value
The probability, assuming that the Null Hypothesis is true, of getting a value of the test statistic at least as extreme as the computed value for the test.

Parameter
A fixed numerical value that describes a characteristic of a population.

Point Estimate
A single sample statistic that is used to estimate a population parameter. For example, $\bar{X}$ is a point estimator for μ.

Population
The set of all possible members, objects or measurements of the phenomena being studied.

Power (or Statistical Power)
The probability, set by design, of rejecting the Null Hypothesis when it is actually false. Power is calculated for specific possible values of the Alternative Hypothesis and is the complement of Beta (β).

Probability Distribution Function (PDF)

A function that assigns a probability to all possible values of a random variable. In the case of a continuous random variable (like the Normal Distribution), the PDF refers to the area to the left of a designated value under a Probability Density Function.

Random Sample
A sample where the values are equally likely to be selected and mutually independent of each other.

Random Variable
A numerical value that is determined by an experiment with a probability distribution function.

Replicate
In ANOVA, the sample size for a specific level of factor. If the replicates are the same for each level, the design is balanced.

Rejection Region
Region(s) of the Statistical Model which contain the values of the Test Statistic where the Null Hypothesis will be rejected. The area of the Rejection Region = α.

Response
In ANOVA, the numeric variable that is being tested under different treatments or populations.

Sample
A subset of the population.

Sample Mean
a) The arithmetic average of a data set.
b) A random variable that has an approximately Normal Distribution if the sample size is sufficiently large.

Significance Level – see **Level of Significance**

Standard Deviation
The square root of the variance and measures the spread of data, distance from the mean. The units of the standard deviation are the same units as the data.

Standard Normal Distribution – see **Normal Distribution**

Statistic
A value that is calculated from sample data only that is used to describe the data. Examples of statistics are the sample mean, sample standard deviation, range, sample median and the interquartile range. Since statistics depend on the sample, they are also random variables.

Statistical Inference
The process of estimating or testing hypotheses of population parameters using statistics from a random sample.

Statistical Model
A mathematical model that describes the behavior of the data being tested.

Student's t distribution (or t distribution)
A family of continuous random variables (based on degrees of freedom) with a probability density function that is from the Normal Family of Probability Distributions. The t distribution is used for statistical inference of the population mean when the population standard deviation is unknown.

Test Statistic
A value, determined from sample information, used to determine whether or not to reject the Null Hypothesis.

Type I Error
Rejecting the Null Hypothesis when it is actually true.

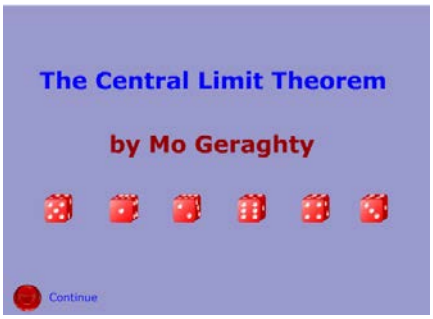
Type II Error
Failing to reject the Null Hypothesis when it is actually false.

Variance
A measure of the mean squared deviation of the data from the mean. The units of the variance are the square of the units of the data.

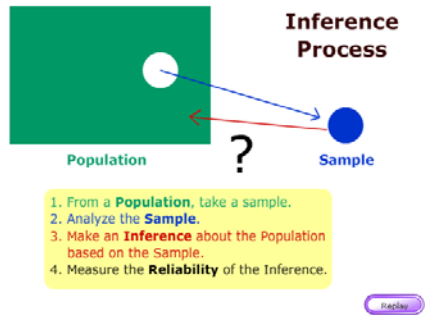
Z-score
A measure of relative standing that shows the distance in standard deviations a particular data point is above or below the mean.

15. Flash Animations

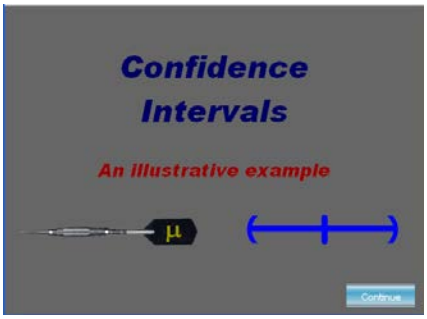
I have designed four interactive Flash animations that will provide the student with deeper insight of the major concepts of inference and hypothesis testing. These animations are on my website http://nebula2.deanza.edu/~mo/ .



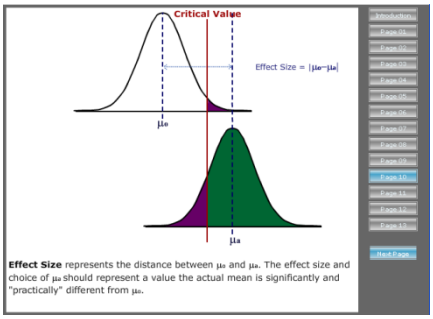
Central Limit Theorem (Section 4.3)
Using die rolling with progressively increasing sample sizes, this animation shows the three main properties of the Central Limit Theorem.



Inference Process (Section 5.1)
This animation walks a student through the logic of the statistical inference and is presented just before confidence intervals and hypothesis testing.



Confidence Intervals (Section 5.3.1)
This animation compares hypothesis testing to an unusual method of playing darts and compares it to a practical example from the 2008 presidential election.



Statistical Power in Hypothesis Testing (Section 6.7)
This animation explains power, Type I and Type II error conceptually, and demonstrates the effect of changing model assumptions.

16. PowerPoint Slides

I have developed PowerPoint Slides that follow the material presented in the course. This material is presented online at as a slideshow as well as note pages that can be downloaded at http://nebula2.deanza.edu/~mo/.

Section 1:
Descriptive Statistics

Section 2:
Probability (Covered in this text)

Section 3:
Discrete Random Variables

Section 4:
Continuous Random Variables and the Central Limit Theorem (Partially covered in this text)

Section 5:
Point Estimation and Confidence Intervals (Covered in this text)

Section 6:
One Population Hypothesis Testing (Covered in this text)

Section 7:
Two Population Inference (Covered in this text)

Section 8:
Chi-square and ANOVA Tests (Covered in this text)

Section 9:
Correlation and Regression (Covered in this text)

17. Notes and Sources

[1] Talk of the Nation, National Public Radio Archives, http://www.npr.org/

[2] John Cimbaro, *Fish Anatomy*, http://www.fws.gov/midwest/lacrossefishhealthcenter/PhotoAlbum.html

[3] Chen Zheng-Long, Chinese Koi Fish, http://www.orientaloutpost.com/proddetail.php?prod=czl-kf135-1

[4] Richard Christian Looijen, *Holism and Reductionism in Biology and Ecology: The Mutual Dependence of Higher and Lower Level Research Programmes,* Springer, 2000

[5] *The Poems of John Godfrey Saxe* (Highgate Edition), Boston: Houghton, Mifflin and Company, 1881

[6] Donna Young, *American Society of Health System Pharmacists*, April 6, 2007, http://www.ashp.org/import/News/HealthSystemPharmacyNews/newsarticle.aspx?id=2517

[7] *The Lancet*, news release, June 29, 2009, http://www.nlm.nih.gov/medlineplus/news/fullstory_86206.html

[8] *CNN*, Election 2016, National President Exit Polls, November 23, 2016, http://www.cnn.com/election/results/exit-polls/national/president

[9] Ronald Walpole & Raymond Meyers & Keying Ye, *Probability and Statistics for Engineers and Scientists*. Pearson Education, 2002, 7th edition.

[10] Taleb, Nicholas, *The Black Swan: The Impact of the Highly Improbable*, Penguin, 2007.

[11] Food and Drug Administration, *FDA Consumer Magazine* , Jan/Feb 2003

[12] Mark Blumenthal, *Is Polling as we Know it Doomed?,* The National Journal Online, http://www.nationaljournal.com/njonline/mp_20090810_1804.php, August 10, 2009

[13] Russ Lenth, *Java Applets for Power and Sample Size*, University of Iowa , http://www.stat.uiowa.edu/~rlenth/Power/ , 2009

[14] J. B. Orris, *MegaStat for Excel,* Version 10.1, Butler University, 2007

[15] Shlomo S. Sawilowsky, *Fermat, Schubert, Einstein, and Behrens-Fisher: The Probable Difference Between Two Means When* $\sigma_1^2 \neq \sigma_2^2$, Journal of Modern Applied Statistical Methods, Vol. 1, No 2, Fall 2002

[16] Lowry, Richard. One Way ANOVA – Independent Samples. Vassar.edu, 2011

Additional reference used but not specifically cited:

Dean Fearn, Elliot Nebenzahl, Maurice Geraghty, *Student Guide for Elementary Business Statistics,* Kendall/Hunt, 2003

[17] Munroe, Randall. http://xkcd.com/1725/ . 2016

Additional reference used but not specifically cited:

Dean Fearn, Elliot Nebenzahl, Maurice Geraghty, *Student Guide for Elementary Business Statistics,* Kendall/Hunt, 2003