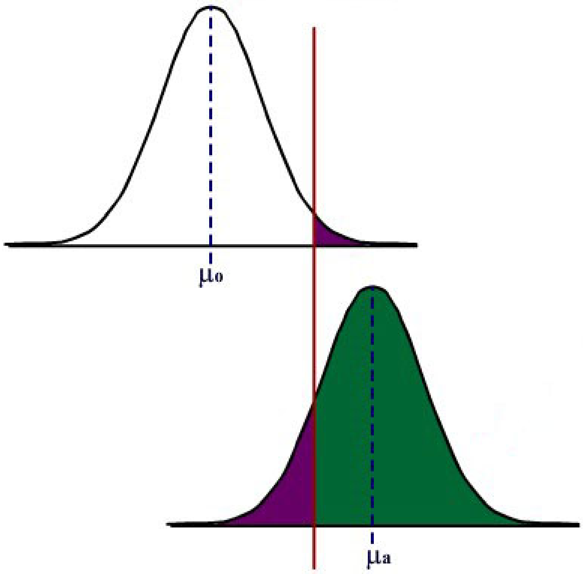
DE ANZA COLLEGE – DEPARTMENT OF MATHEMATICS

Inferential Statistics and Hypothesis Testing

A Holistic Approach

Maurice A. Geraghty

9/1/2011



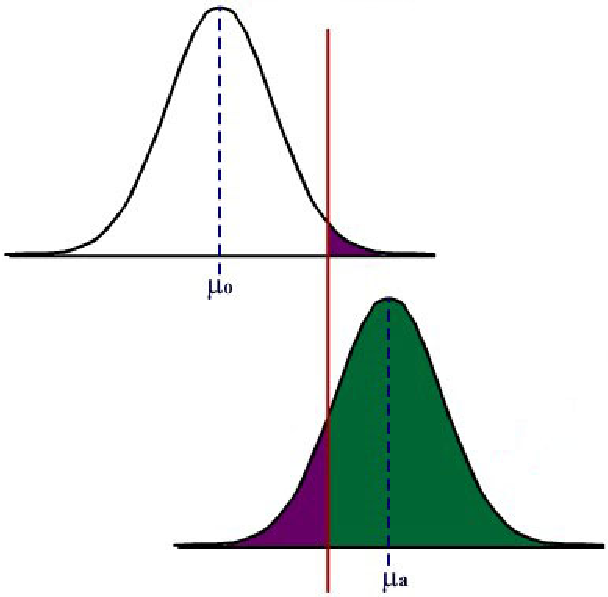
Supplementary material for an introductory lower division course in Probability and Statistics

Inference and Hypothesis Testing – A Holistic Approach

Supplementary Material for an Introductory Lower Division Course in Probability and Statistics

Maurice A. Geraghty, De Anza College

September 1, 2011



1. Introduction – a Classroom Story and an Inspiration……………………………………………..Page 02
2. The Six Blind Men and the Elephant…………………………………………………………………Page 05
3. Two News Stories of Research……………………………………………………………………….Page 06

4. Review and Central Limit Theorem…………………………………………………………………Page 08
5. Point Estimation and Confidence Intervals…………………………………………………………Page 13
6. One Population Hypothesis Testing…………………………………………………………………Page 20
7. Two Population Inference………………………………………………………………………………Page 41
8. One Factor Analysis of Variance (ANOVA)…………………………………………………………..Page 51
9. Glossary of Statistical Terms used in Inference………………………………………………….Page 56

10. Flash Animations……………………………………………………………………………………………..Page 62
11. PowerPoint Slides……………………………………………………………………………………………Page 63
12. Notes and Sources……………………………………………………………………………………………Page 64

1. 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.

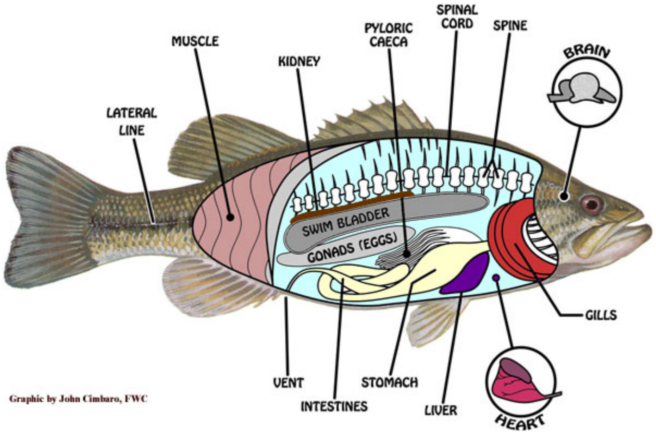
These students, (to their credit) continued to work hard; they wanted to understand the material, not simply pass the class. Since these students had excellent math skills, I went deeper into the explanation of Type II error and the statistical power function. Although they could compute power and sample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



Graphic by John Cimbaro, FWC

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.

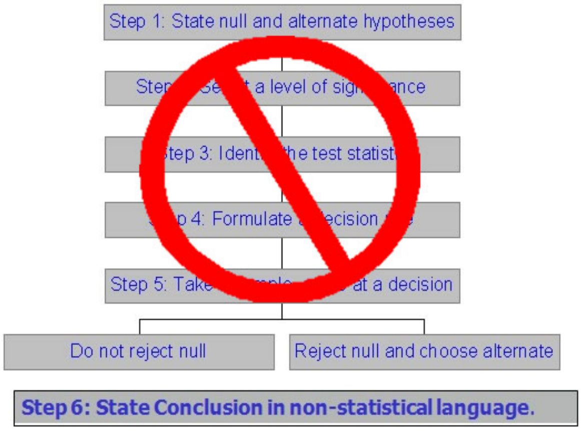
Although science has traditionally been reluctant to embrace, often hostile to including holistic philosophy in the scientific method, there have been many who now support a multicultural or multi-philosophical approach. In his book Holism and Reductionism in Biology and Ecology[4], Looijen claims that "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.

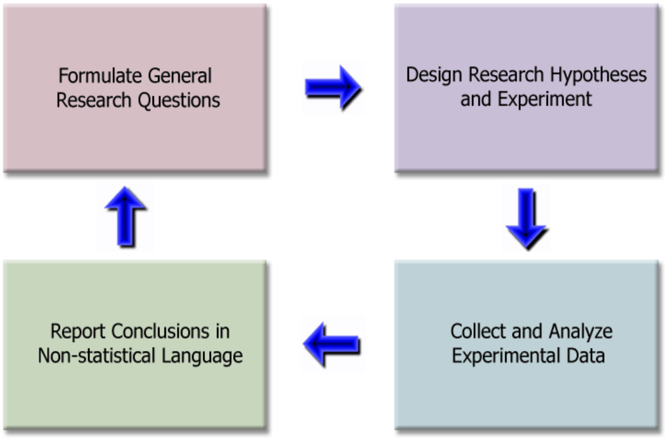
I realize that the problem of my high-achieving students being unable to comprehend hypothesis testing could be cultural – these were international students who may have been schooled under a more holistic philosophy. The Introductory Statistics curriculum and most texts give an incomplete explanation 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.


Step 1: State null and alternate hypotheses
Step Se t a level of sig ance
tep 3: Ident he test statist
p 4: Formulate ecision
Step 5: Take at a decision
Do not reject null
Reject null and choose alternate
Step 6: State Conclusion in non-statistical language.


Formulate General Research Questions
Design Research Hypotheses and Experiment
Collect and Analyze Experimental Data
Report Conclusions in Non-statistical Language

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.

2. The Six 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

3. Two News Stories of Research

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.

4. Review and Central Limit Theorem

4.1 Empirical Rule

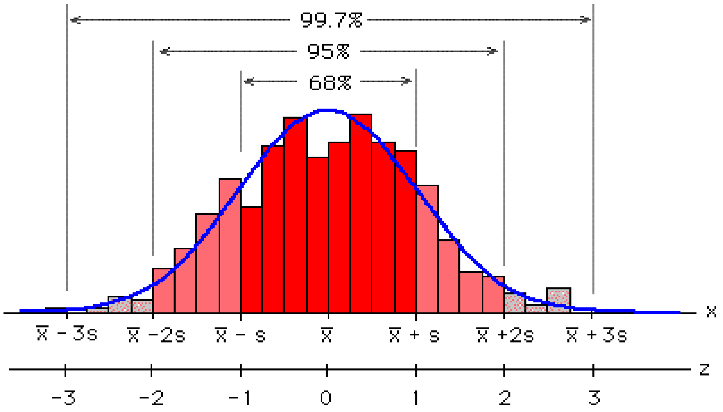
A student asked me about the distribution of exam scores after she saw her score of 87 out of 100. I told her the distribution of test scores were approximately bell-shaped with a mean score of 75 and a standard deviation of 10. Most people would have an intuitive grasp of the mean score as being the "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.

4.2 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: $Z = \dfrac{X_i - \overline{X}}{s}$

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

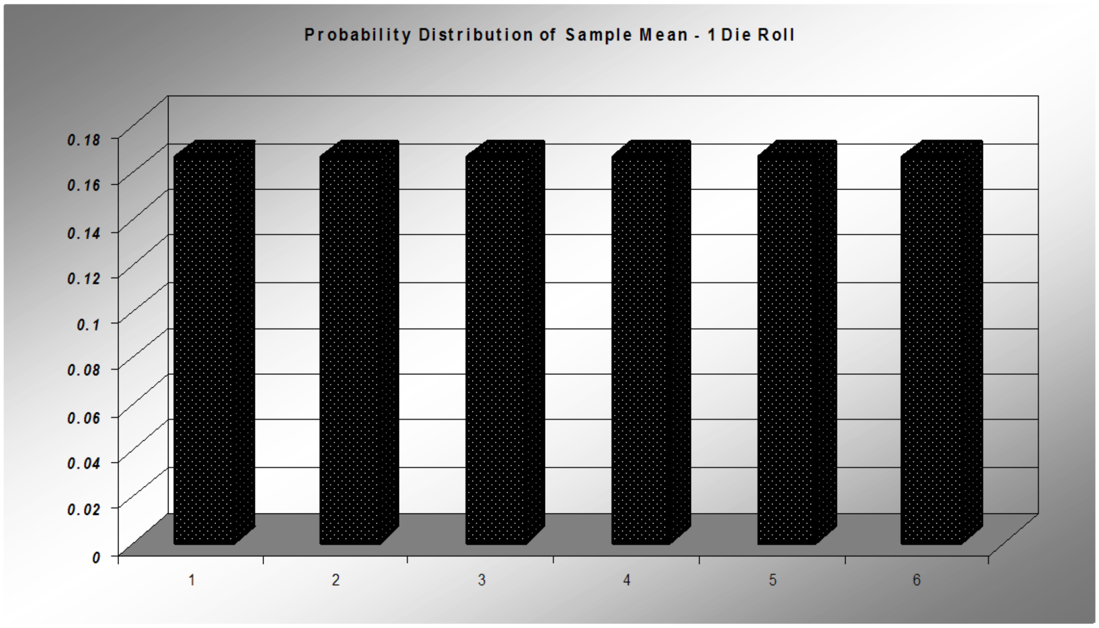
4.3 The Sample Mean as a Random Variable – Central Limit Theorem

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.

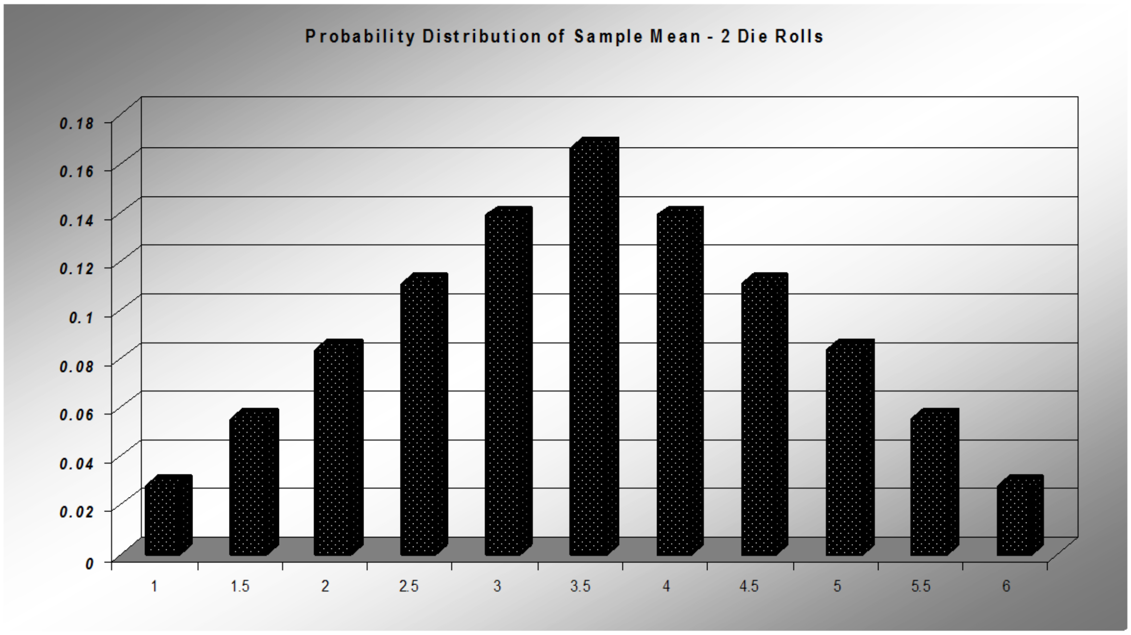
We start with a Random Sample X_1, X_2, …, X_n where each of the random variables X_i has the same probability distribution and are mutually independent of each other. The sample mean is a function of these random variables (add them up and divide by the sample size), so $\bar{X}$ is a random variable.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

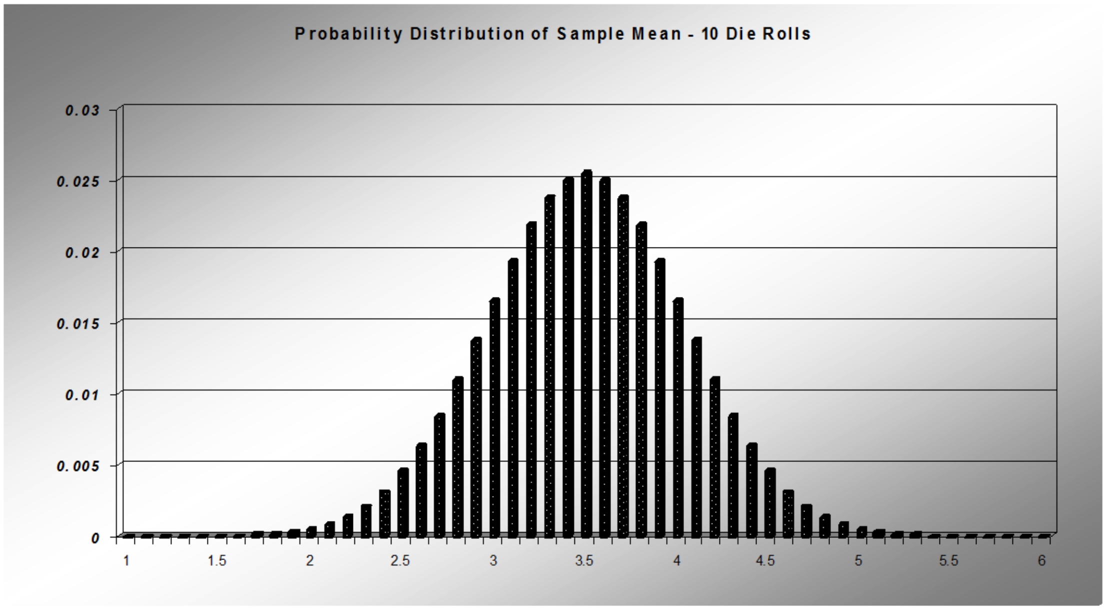
For the case n=1, the distribution of the sample mean is the same as the distribution of the random variable. Since each die has the same chance of being chosen, the distribution is rectangular shaped centered at 3.5:



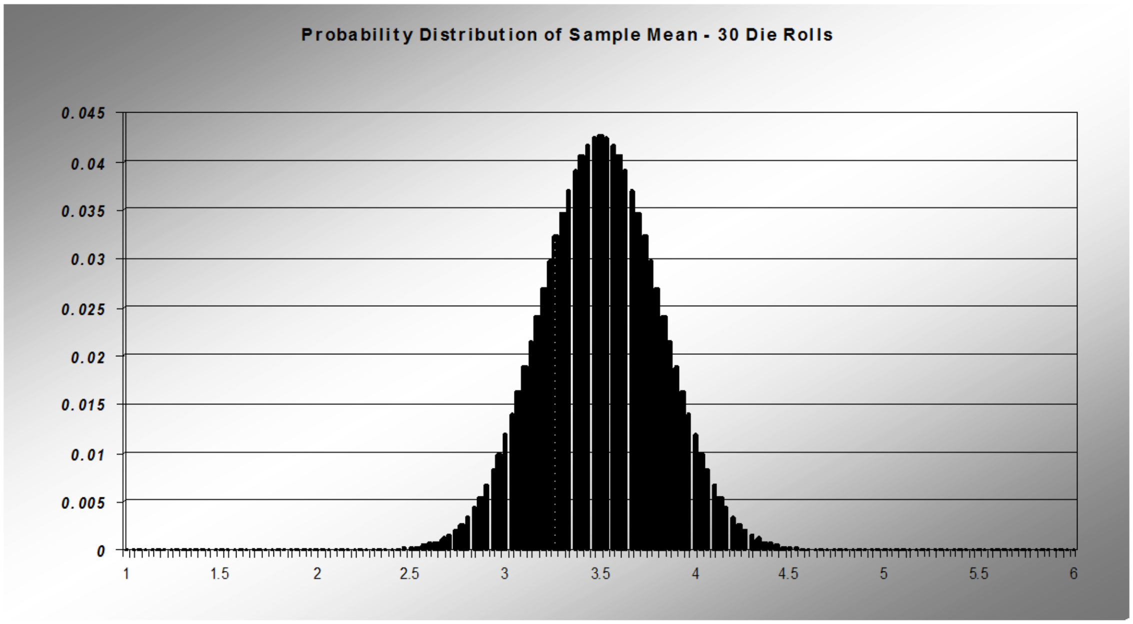
For the case n=2, the distribution of the sample mean starts to take on a triangular shape since some values are more likely to be rolled than others. For example, there six ways to roll a total of 7 and get a sample mean of 3.5, but only one way to roll a total of 2 and get a sample mean of 1. Notice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

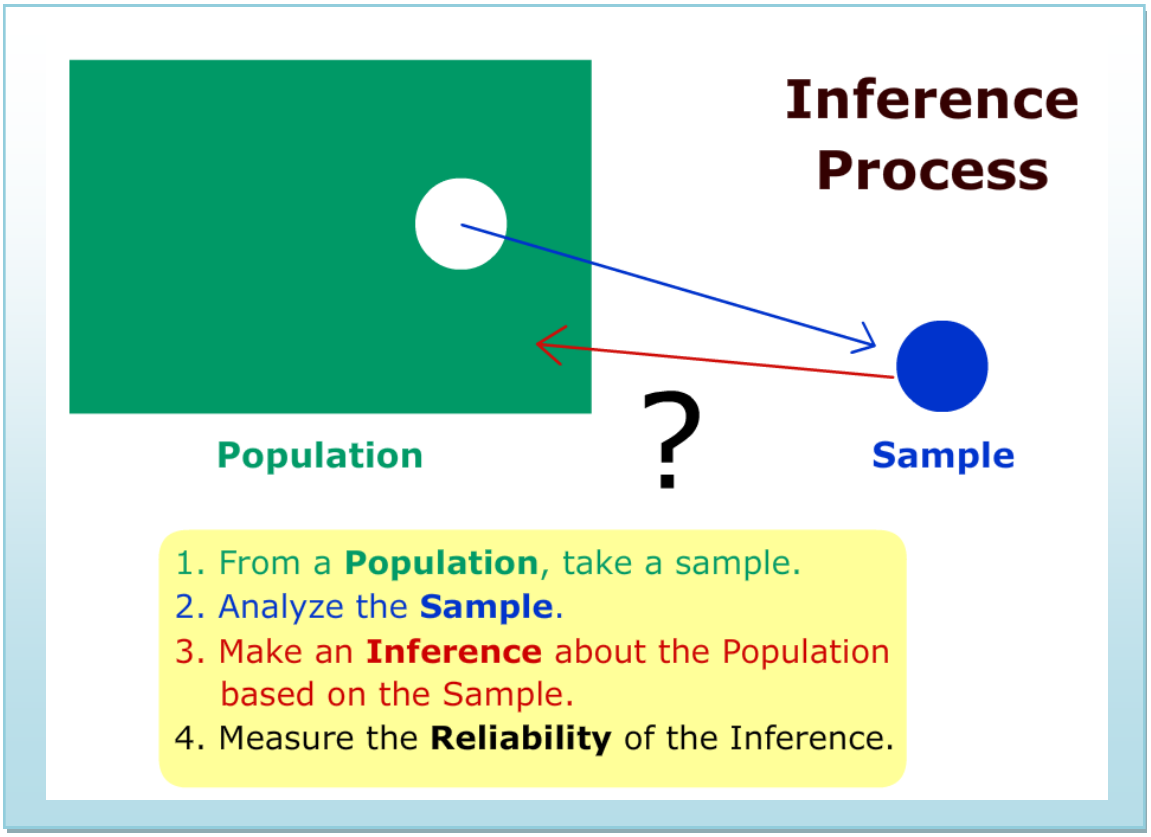
$$P(X > 70) = P\left(Z > \frac{(70 - 69.2)}{2.9} \right) = P(Z > 0.28) = 0.3897$$

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

5. Point Estimation and Confidence Intervals

5.1 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.



Inference Process

1. From a **Population**, take a sample.
2. Analyze the **Sample**.
3. Make an **Inference** about the Population based on the Sample.
4. 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.

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

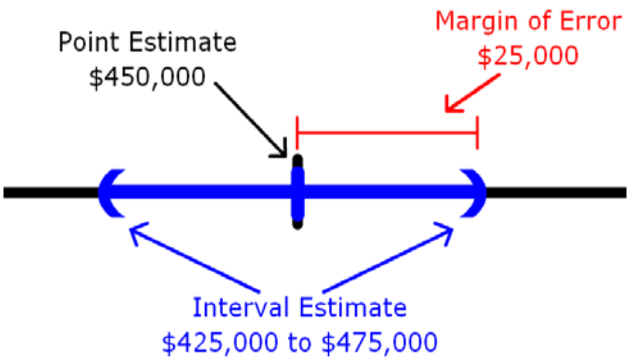
In the example above, Lupe's realtor estimated the population mean of similar homes in Lupe's neighborhood by using the sample mean of $450,000 from the adjusted price of the sampled homes.

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 ± $25,000.



5.3 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.

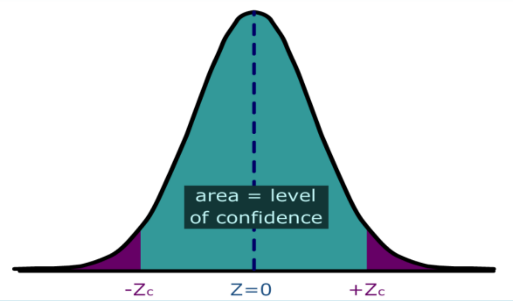
5.3.1 Confidence Interval for Population Mean

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?

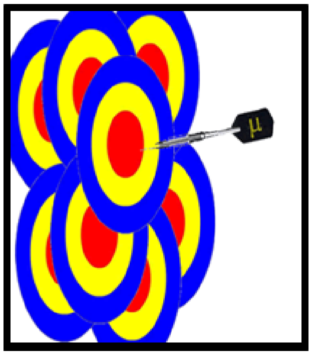
$$24 \pm \frac{1.96 \cdot 4}{\sqrt{49}} = 24 \pm 1.12 = (22.88, 25.12) \text{ hours per week}$$

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



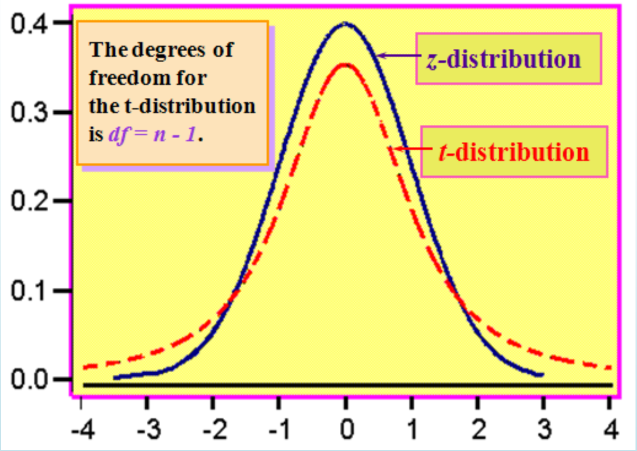
- The confidence interval is constructed from random variables calculated from sample data and attempts to predict an unknown but fixed population parameter with a certain level of confidence.
- Increasing the level of confidence will always increase the margin of error.
- It is impossible to construct 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.

5.3.2 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"[8].

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

$$\boldsymbol{\overline{X} \pm t_c \frac{s}{\sqrt{n}}}$$ **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.

5.3.3 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%

5.3.4 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.

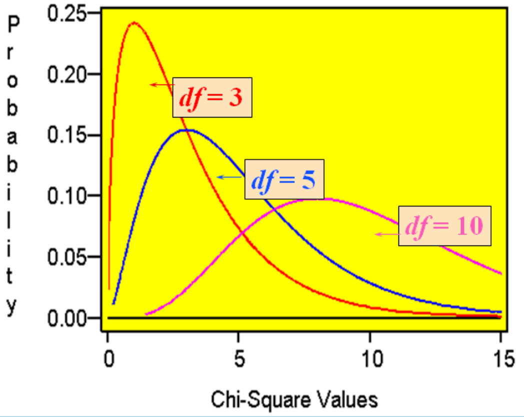
To create a confidence interval for an estimate of standard deviation, we need to introduce a new distribution, called the Chi-square (χ^2) distribution.

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.



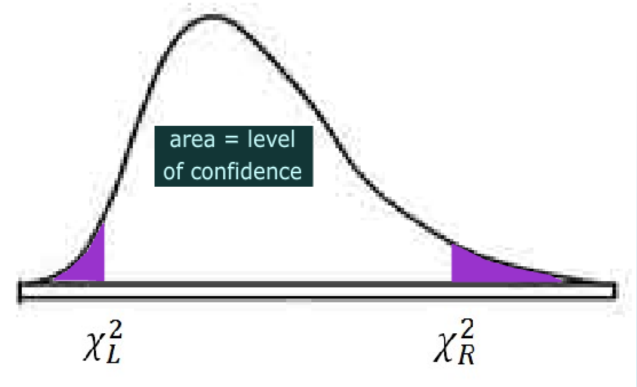
5.3.5 Confidence Interval for Population Variance and Standard Deviation

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: $\left(\sqrt{\frac{(19)5^2}{32.8523}}, \sqrt{\frac{(19)5^2}{8.90655}}\right) = (3.8, 7.3)$

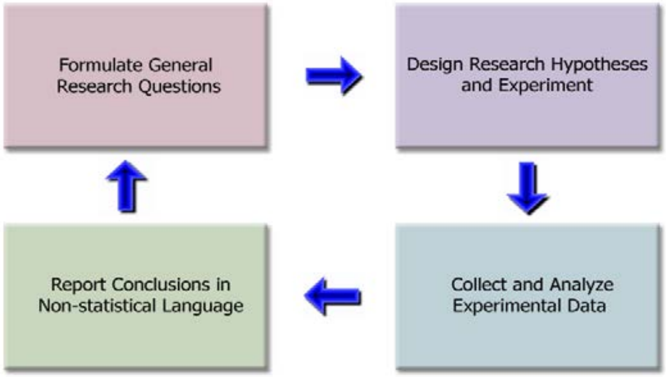
One can say with 95% confidence that the standard deviation for this mutual fund is between 3.8% and 7.3% per month.

6. 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.

6.1 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.


Formulate General Research Questions
Design Research Hypotheses and Experiment
Collect and Analyze Experimental Data
Report Conclusions in Non-statistical Language

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.

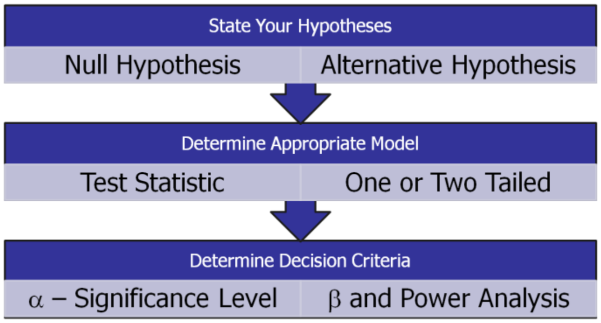
6.2 Formulate General Research Questions

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.

6.3 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.

6.3.1 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:

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

6.3.2 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.

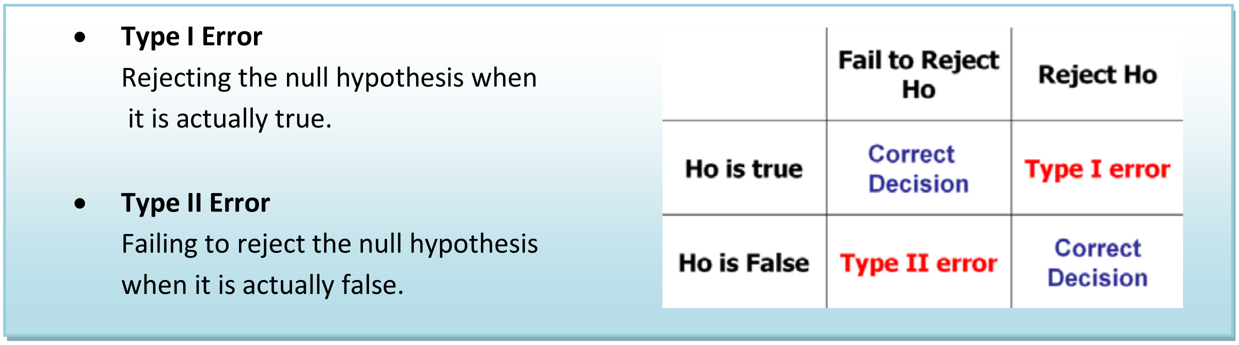
Once the model is chosen, we can then determine a **test statistic**, a value derived from the data that is used to decide whether to **reject** or **fail to reject** the Null Hypothesis.

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	$Z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_0)}{n}}}$
Variance vs. Hypothesized Value	$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

6.3.3 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.

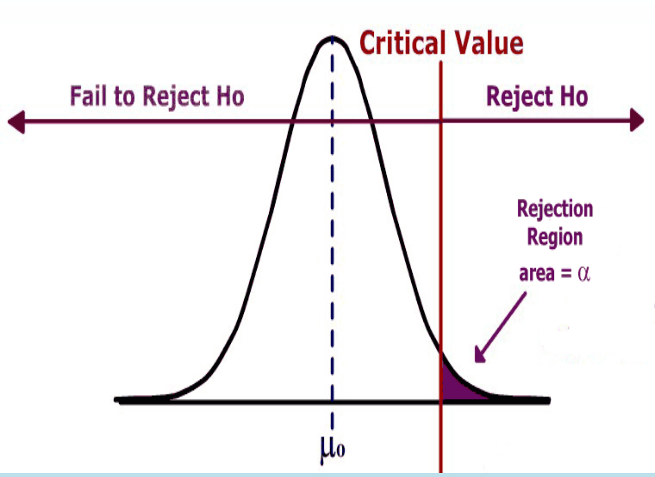
In designing an experiment, we set a maximum probability of making Type I error. This probability is called the **level of significance** or **significance level** of the test and designated by the Greek letter α.

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.

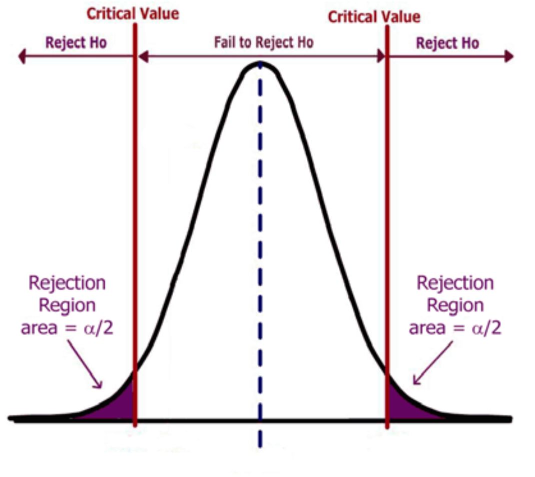
6.3.4 Critical Value and Rejection Region

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

6.3.5 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.

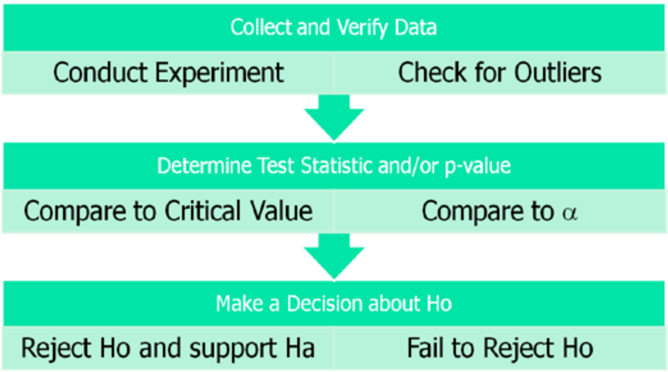
Here are some examples for testing the mean μ against a hypothesized value μ_0:

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.

6.4 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.



6.4.1 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.

6.4.2 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.[9] 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

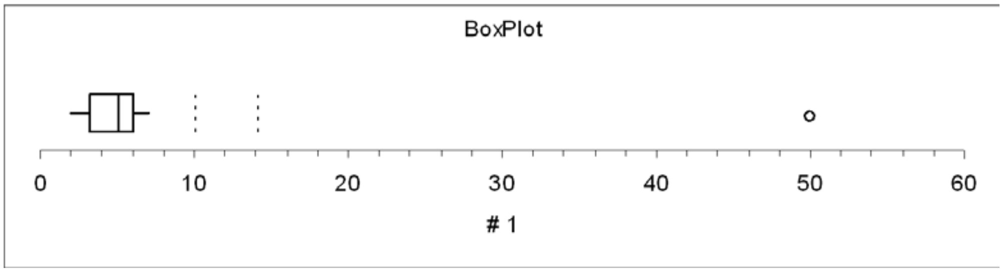
2 2 3 4 5 5 6 6 7 50

	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.



6.4.3 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.

6.4.4 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 Ho 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.

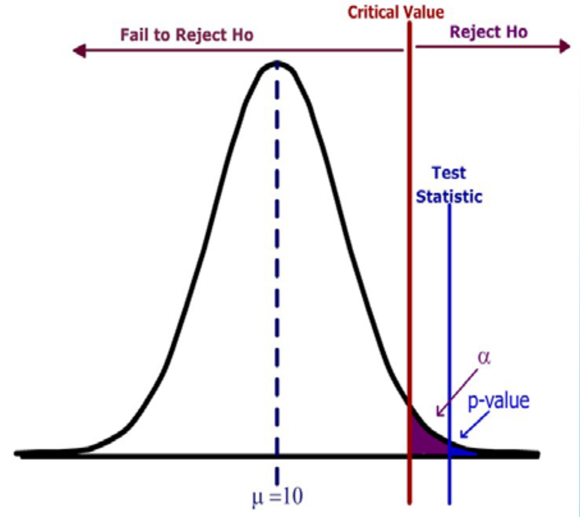
1. If the test statistic lies in the rejection region, reject Ho. (critical value method)
2. If the p-value < α, reject Ho. (p-value method)

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

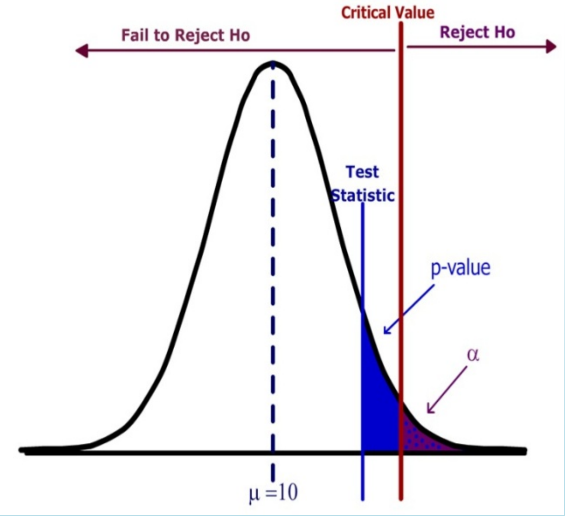
- Ho: $\mu = 10$
  Ha: $\mu > 10$
- Design: Critical value is determined by significance level α.
- Data Analysis: p-value is determined by test statistic
- Test statistic falls in rejection region.
- p-value (blue) < α (purple)
- Reject Ho.
- Strong statement: Data supports the Alternative Hypothesis.



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.

6.5 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:

6.5.1 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

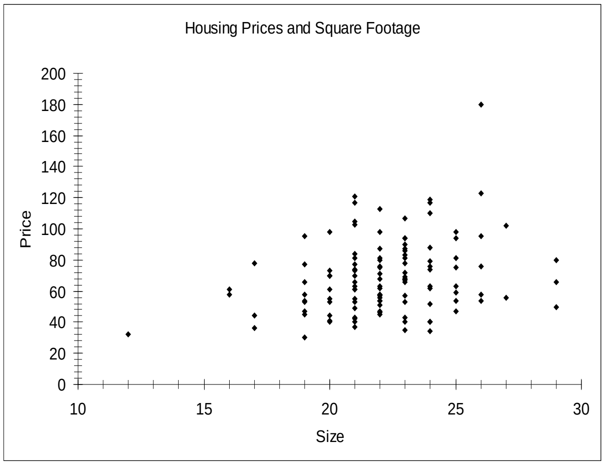
Conclusion if Ho is rejected: The mean number of text messages sent by students exceeds 1000.

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.

6.5.2 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.

6.5.3 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."[10] 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.

6.5.4 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.[11]

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.

6.5.5 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.

6.6 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**.

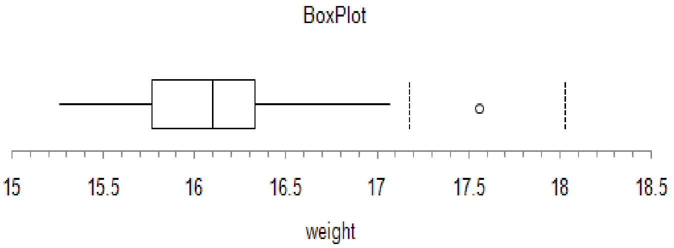
The quality control statistician has been given the authority to sample 36 bottles of soy sauce and knows from past testing that the population standard deviation is 0.5 ounces. The model will be a **test of population mean vs. hypothesized valu**e of 16 oz. A two-tailed test is selected since the company 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$_0$: 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.

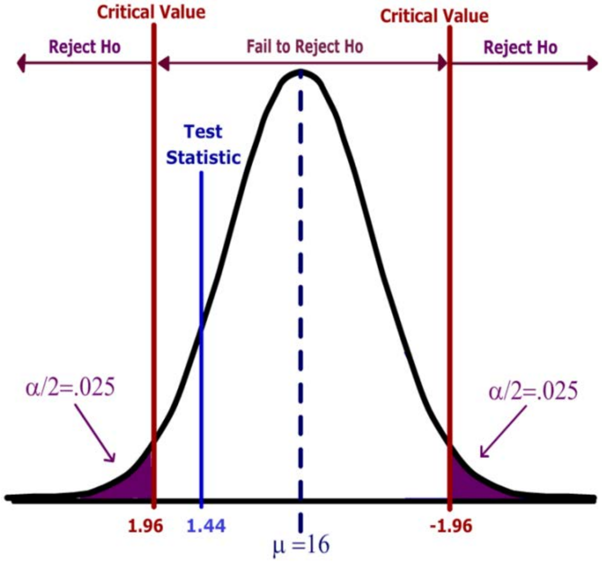


Next, the sample mean and the test statistic are calculated.

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

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$_0$ when Z >1.96 or Z<-1.96.

Based on this result, the decision is **fail to reject H$_0$** since the test statistic does not fall in the rejection region.

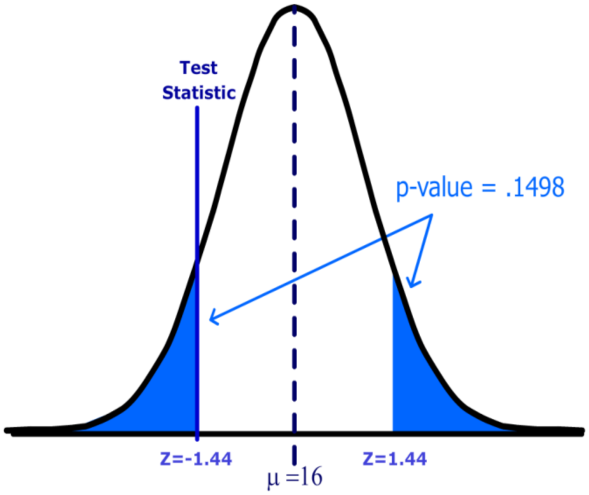


Alternatively (and preferably) the statistician would use the p-value method of decision rule. The p-value for a two-tailed test must include all values (positive and negative) more extreme than the Test Statistic, so in this example we find the probability that Z < -1.44 or Z > 1.44 (the area shaded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

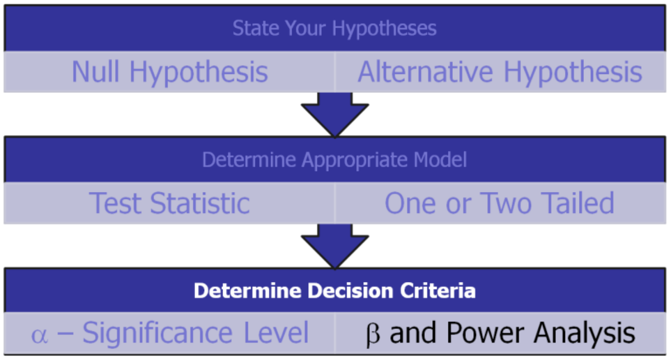
> "There is insufficient evidence to conclude that the machine that fills 16 ounce soy sauce bottles is operating improperly. This conclusion is based on 36 measurements taken during a single hour's production run. I recommend continued monitoring of the machine during different employee shifts to account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

6.7 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.



State Your Hypotheses

Null Hypothesis | Alternative Hypothesis

Determine Appropriate Model

Test Statistic | One or Two Tailed

Determine Decision Criteria

α – Significance Level | β and Power Analysis

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 Ho, even if Ho 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 Ha. 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 Ha technically true. So when analyzing power and Type II error we need to choose a value for the **population mean under the Alternative Hypothesis (μa)** that is "**practically different**" from the **mean under the Null Hypothesis (μo).** This practical difference is called the **effect size**.

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

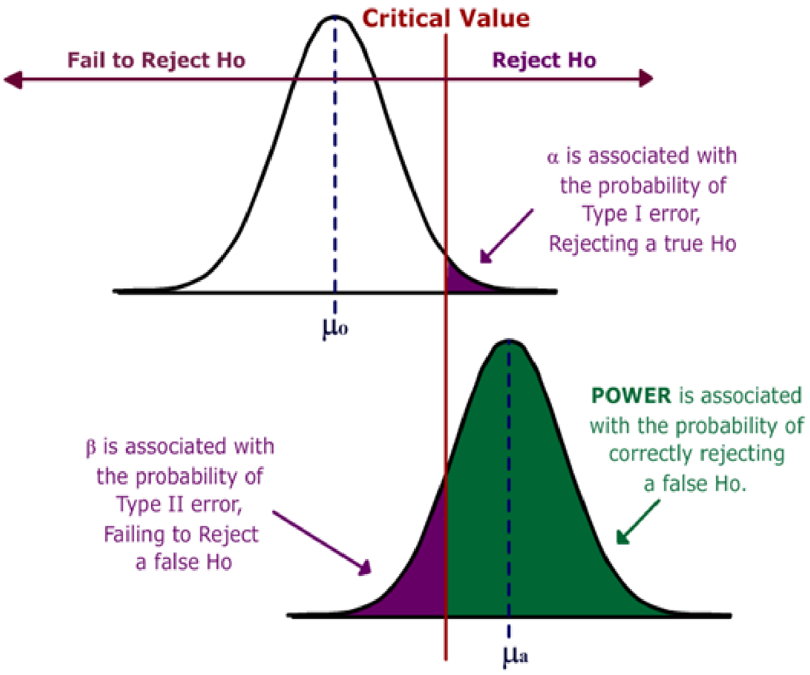
μa: The value of the population mean under the Alternative Hypothesis

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

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

Ho: $\mu = \mu_0$ Ha: $\mu > \mu_0$

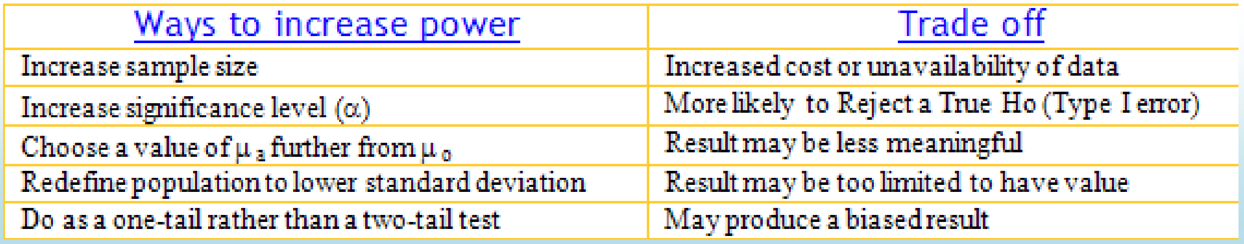
Consider the two graphs shown to the right. The top graph is the distribution of the sample mean under the Null Hypothesis that we covered in an earlier section. The area to the right of the critical value is the rejection region.


Critical Value
Fail to Reject Ho
Reject Ho
α is associated with the probability of Type I error, Rejecting a true Ho
μo
POWER is associated with the probability of correctly rejecting a false Ho.
β is associated with the probability of Type II error, Failing to Reject a false Ho
μa

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

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%

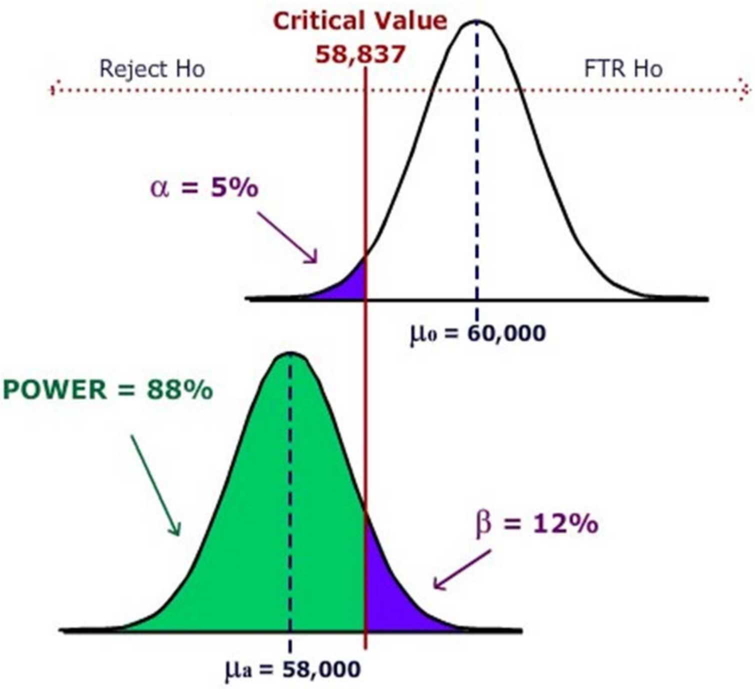
Power Calculation Values

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%



6.8 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:

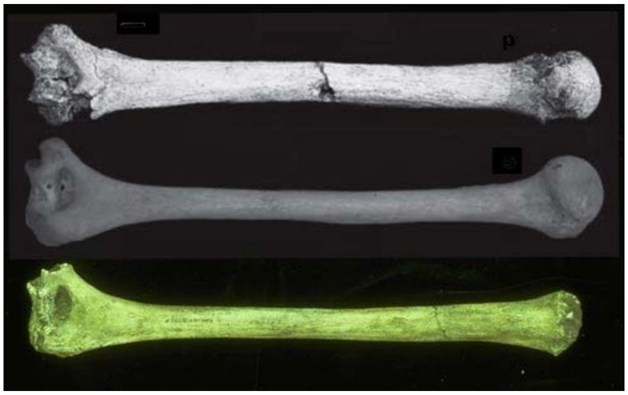
- Test of population mean vs. hypothesized value, population standard deviation unknown.
- Test of population proportion vs. hypothesized value.
- Test of population standard deviation (or variance) vs. hypothesized value.

6.8.1 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.

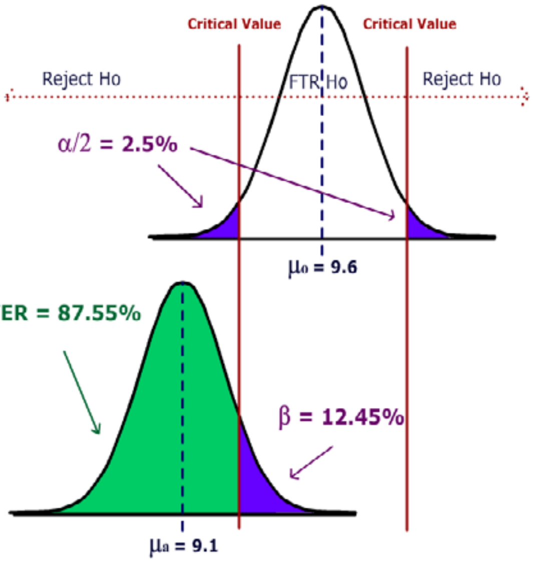
   Research Hypotheses
   H_o: μ = 9.6 (The humerus bones are from Species A)
   H_a: $\mu \neq$ 9.6 (The humerus bones are not from Species A)

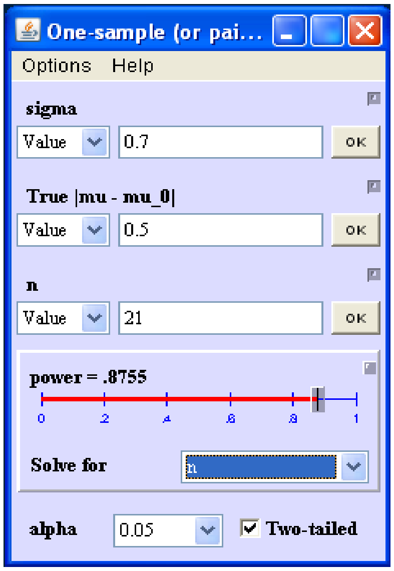
   Significance level: α =.05

   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[12]
- μo = 9.6 (Species A) - μa = 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%





sigma Value 0.7

True |mu - mu_0| Value 0.5

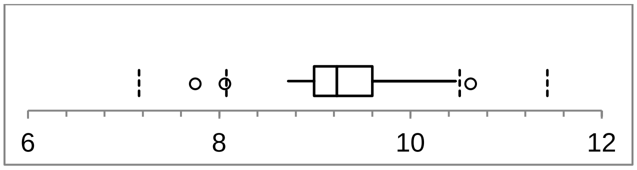
n Value 21

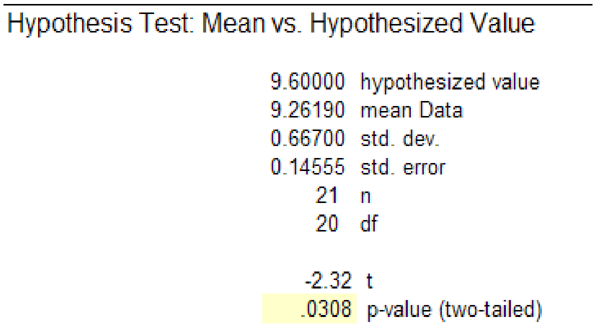
power = .8755

Solve for n

alpha 0.05 ☑ Two-tailed

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[13], 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.

6.8.2 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

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

$\boldsymbol{\hat{p}}$ = sample proportion

$\boldsymbol{p_a}$ = population proportion under Ha

Test Statistic: $Z = \frac{\hat{p}-p_O}{\sqrt{\frac{p_O(1-p_O)}{n}}}$

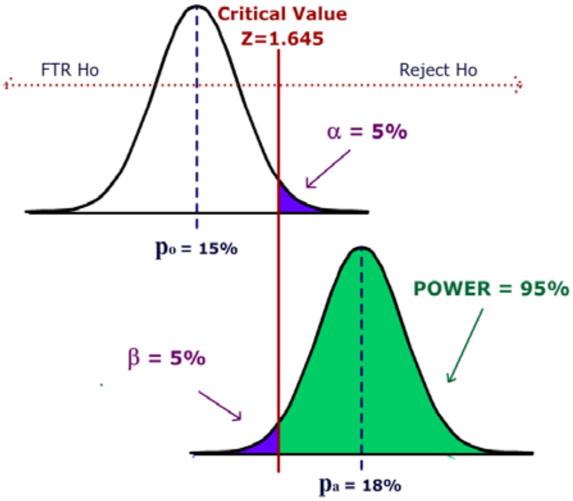
Requirement for Normality Assumption: $np(1-p) > 5$

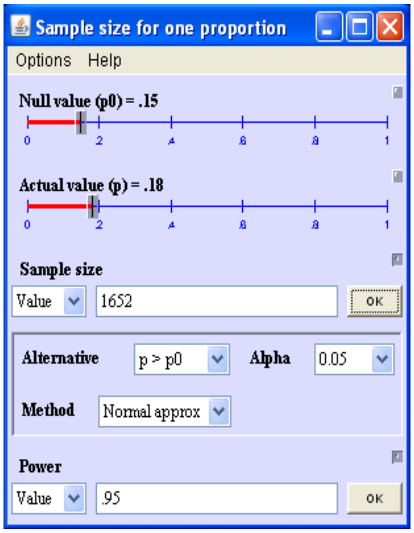
Example

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





Sample size for one proportion

Options Help

Null value (p0) = .15

Actual value (p) = .18

Sample size

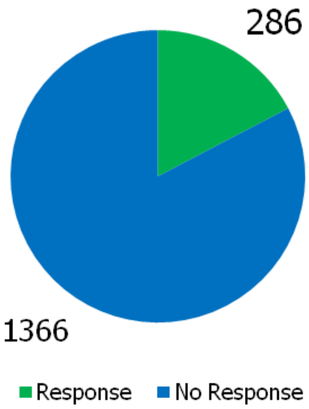
Value 1652 OK

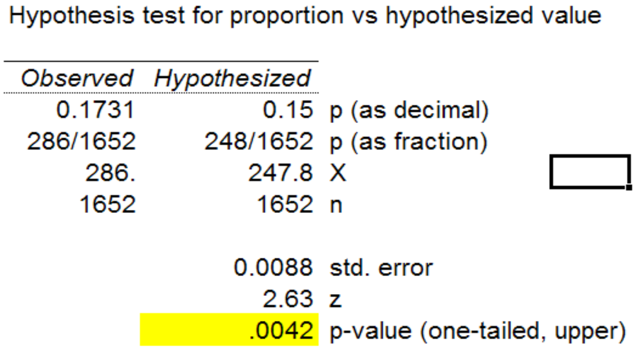
Alternative p > p0 Alpha 0.05

Method Normal approx

Power

Value .95 OK





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 Ho and support Ha. Since the p-value is actually less than 0.01, we would go further and say that the data supports rejecting Ho for α = .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.

6.8.3 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

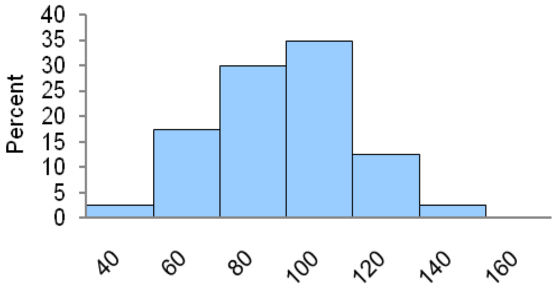
$\boldsymbol{\sigma^2}$ = population variance $\boldsymbol{\sigma_o^2}$ = population variance under Ho

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

Test Statistic: $\chi^2 = \dfrac{(n-1)s^2}{\sigma_o^2}$ $\boldsymbol{n-1}$ = 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:

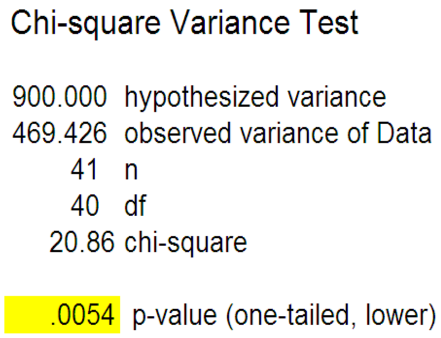
57	75	86	92	101	108	110	120	155
63	77	88	96	102	108	111	122	
66	78	88	96	107	109	115	135	
68	81	92	98	107	109	115	137	
72	82	92	99	107	110	118	139	

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

Hypotheses In terms of the population variance:

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

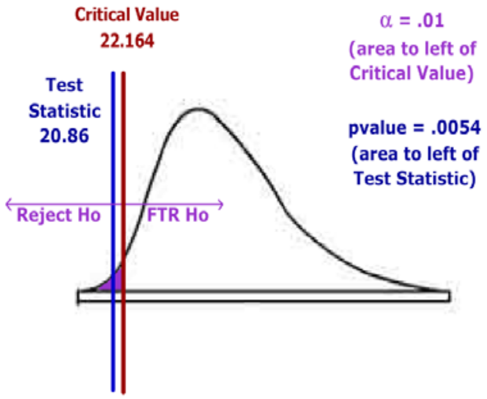
Results:



```
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.

7. Two Population Inference

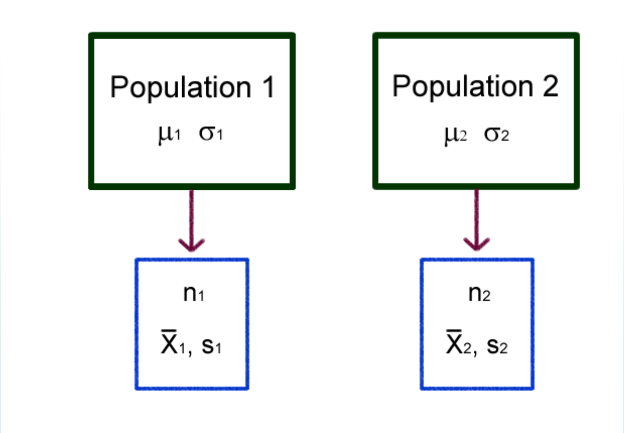
In this section we consider expanding the concepts from the prior section to design and conduct hypothesis testing with two samples. Although the logic of hypothesis testing will remain the same, care must be taken to choose the correct model. We will first consider comparing two population means.

7.1 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.

7.1.1 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

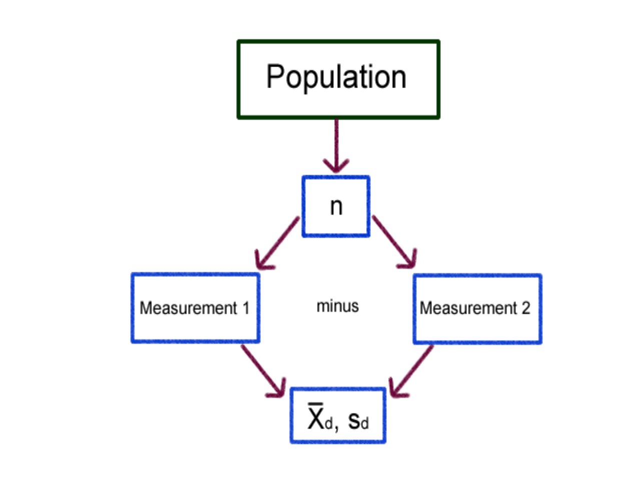
- n_1 **is the sample size from population 1.**
- n_2 **is the sample size from population 2.**
- $\bar{X}_1$ **is the sample mean from population 1.**
- $\bar{X}_2$ **is the sample mean from population 2.**
- s_1 **is the sample standard deviation from population 1.**
- s_2 **is the sample standard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

7.1.2 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**
- $\overline{X}_d$ **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.

7.2 Independent sampling models

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

7.2.1 Distribution of the difference of two sample means

Suppose we wanted to test the hypothesis $\boldsymbol{Ho\colon \mu_{1=}\mu_2.}$ 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 $\boldsymbol{Ho\colon \mu_{1-}\mu_2 = 0}$ 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 $\boldsymbol{\bar{X}_1 - \bar{X}_2}$ under the Central Limit Theorem

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

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

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

When the population variances are known, the test statistic for the Hypothesis $\boldsymbol{Ho}\colon \boldsymbol{\mu_{1=}\mu_2}$ 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

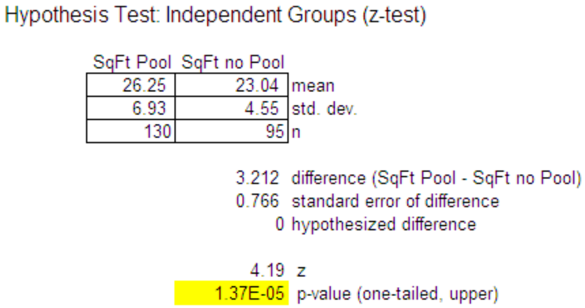
Research Hypotheses: **Ho: $\mu_1 \le \mu_2$ (Homes with pools do not have more mean square footage)**
Ha: $\mu_1 > \mu_2$ (Homes with pools do have more mean square footage)

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

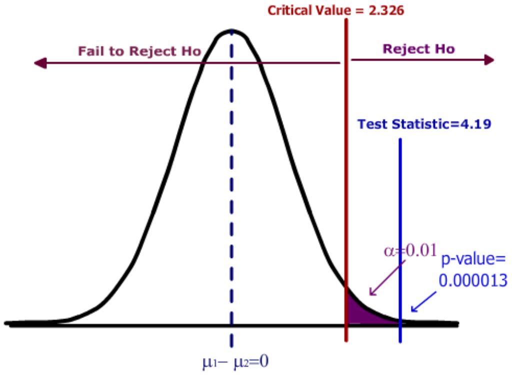


Hypothesis Test: Independent Groups (z-test)

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.

7.2.3 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: $\mu_1 \le \mu_2$ (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.

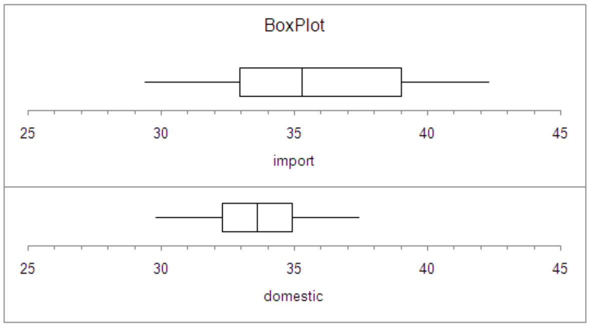
Type I error would be to reject the Null Hypothesis and claim imports has a higher mean MPG, when they do not have higher MPG. The test will be run at a level of significance (α) of 5%.

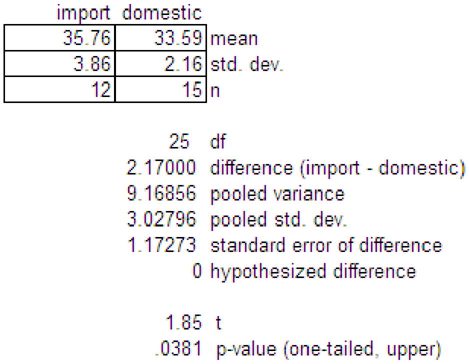
The degrees of freedom for this test is 25, so the decision rule under the critical value method would be to reject Ho when t > 1.708. The decision under the p-value method is to reject Ho if the p-value is < α.

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.

7.2.4 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 [14], 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

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

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: $\mu_1 \le \mu_2$ (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.

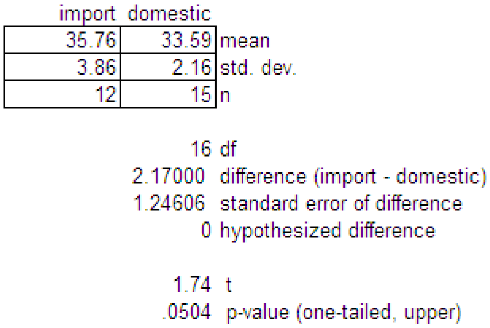
Type I error would be to reject the Null Hypothesis and claim imports has a higher mean MPG, when they do not have higher MPG. The test will be run at a level of significance (α) of 5%.

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.

7.3 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

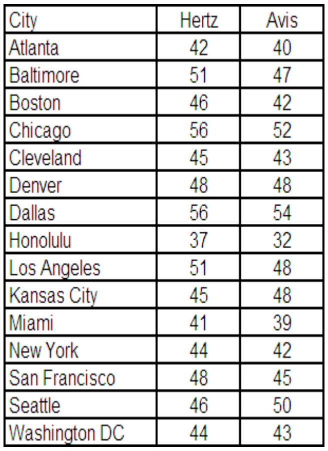
Test Statistic

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

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: **Ho: $\mu_1=\mu_2$ (Hertz and Avis have the same mean price for compact cars.)**
Ha: $\mu_1\neq\mu_2$ (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: **Ho: $\mu_d=0$ Ha: $\mu_d\neq 0$**

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

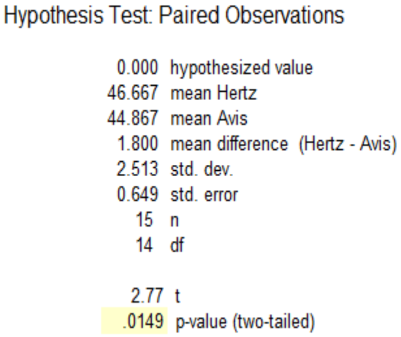
Model is two tailed matched pairs t-test with 14 degrees of freedom. Reject Ho if t < -2.145 or t >2.145.

Example - Data/Results

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

$$\overline{X}_d = 1.80$$
$$s_d = 2.513$$
$$n = 15$$
$$t = \frac{1.80 - 0}{2.513/\sqrt{15}} = 2.77$$

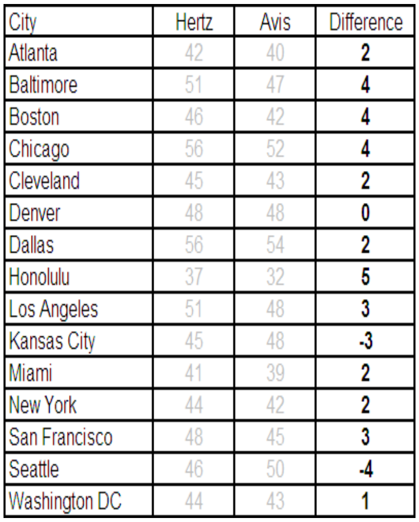
Reject Ho under either the critical value or p-value method.



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.

7.4 Independent sampling – comparing two population variances or standard deviations

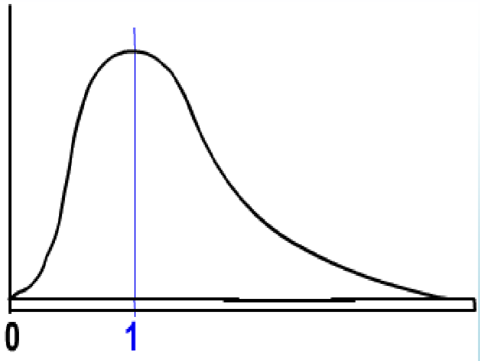
Sometimes we want to test if two populations have the same spread or variation, as measured by variance or standard deviation. This may be a test on its own or a way of checking assumptions when deciding between two different models (e.g.: pooled variance t-test vs. unequal variance t-test). We will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

7.4.1 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.

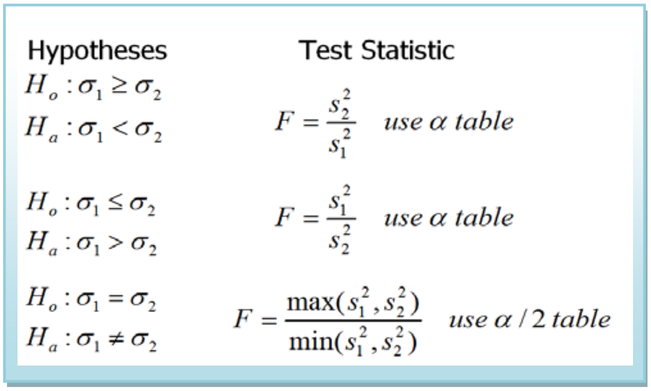
Characteristics of F Distribution

- 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.



7.4.2 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
$H_o: \sigma_1 \ge \sigma_2$ $H_a: \sigma_1 < \sigma_2$	$F = \frac{s_2^2}{s_1^2}$ *use α table*
$H_o: \sigma_1 \le \sigma_2$ $H_a: \sigma_1 > \sigma_2$	$F = \frac{s_1^2}{s_2^2}$ *use α table*
$H_o: \sigma_1 = \sigma_2$ $H_a: \sigma_1 \ne \sigma_2$	$F = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$ *use $\alpha/2$ table*

7.4.3 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: $\sigma_1 \le \sigma_2$ (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.

7.4.4 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

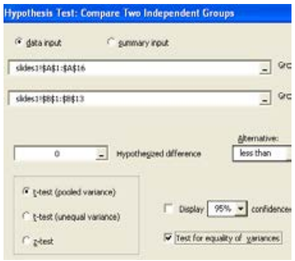
Research Hypotheses: **Ho: σ_1=σ_2 (choose the pooled variance t-test to compare means)**
Ha: σ_1≠σ_2 (choose the unequal variance t-test to compare means)

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.



Hypothesis Test: Compare Two Independent Groups

● data input ○ summary input

slides1!$A$1:$A$16

slides1!$B$1:$B$13

0 Hypothesized difference

Alternative: less than

● t-test (pooled variance)
○ t-test (unequal variance)
○ z-test

□ Display 95% confidence

☑ Test for equality of variances

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.

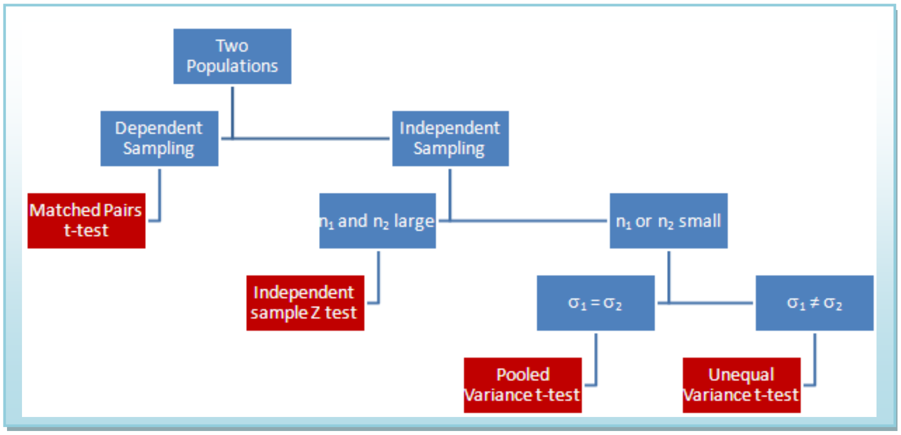
```
F-test for equality of variance
 14.894  variance: import
  4.654  variance: domestic
   3.20  F
  .0438  p-value
```

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.



8. One Factor Analysis of Variance (ANOVA)

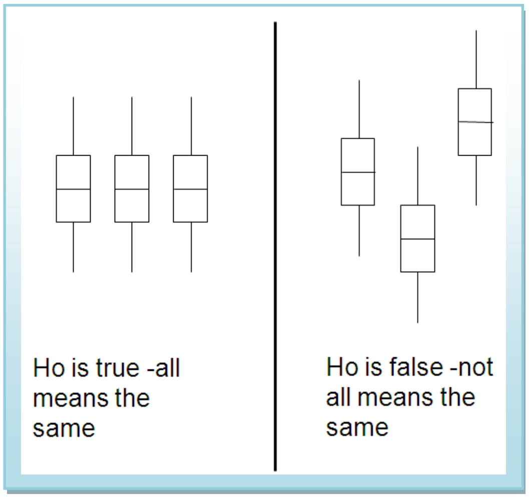
In the prior section we used statistical inference to compare two population means under variety of models. These models can be expanded to compare more than two populations using a technique called Analysis of Variance, or ANOVA for short. There are many ANOVA models, but we limit our study to one of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

8.1 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 $\boldsymbol{Ho}\colon \boldsymbol{\mu_1} = \boldsymbol{\mu_2} = \cdots = \boldsymbol{\mu_k}$. 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.

8.2 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 $Ho\colon \mu_1 = \mu_2 = \mu_3$ 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.



8.3 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

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

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

8.4 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.

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

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: μ_1=μ_2 =μ_3 (Mean sales same at all restaurants)**
Ha: At least μ_i is different (Means sales not the same at all restaurants)

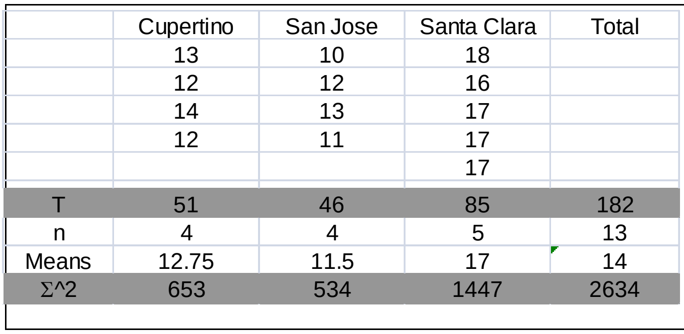
We will assume the population variances are equal $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, so the model will be **One Factor ANOVA**. This model is appropriate if the distribution of the sample means is approximately Normal from the Central Limit Theorem.

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 Megastat.

Example - Data/Results

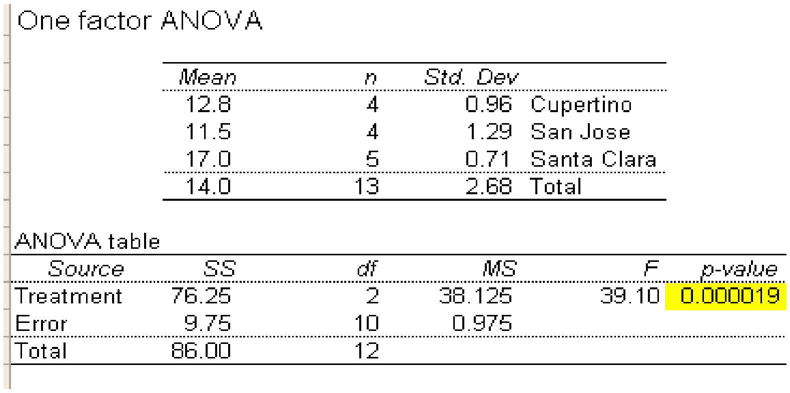


	Cupertino	San Jose	Santa Clara	Total
	13	10	18	
	12	12	16	
	14	13	17	
	12	11	17	
			17	
T	51	46	85	182
n	4	4	5	13
Means	12.75	11.5	17	14
Σ^2	653	534	1447	2634

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

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

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



```
One factor ANOVA

  Mean        n     Std. Dev
  12.8        4       0.96  Cupertino
  11.5        4       1.29  San Jose
  17.0        5       0.71  Santa Clara
  14.0       13       2.68  Total

ANOVA table
Source        SS        df          MS            F    p-value
Treatment    76.25       2       38.125        39.10   0.000019
Error         9.75      10        0.975
Total        86.00      12
```

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

Also p-value = 0.000019 < 0.05 which also supports rejecting Ho.

Note that Megastat uses term "Treatment" instead of "Factor" in the ANOVA table.

Example – Conclusion

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

8.5 Post-hoc Analysis – Tukey's Honestly Significant Difference (HSD) Test[15].

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.

8.5.1 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.

8.5.2 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: $HSD = q\sqrt{\dfrac{MSE}{n_c}}$

q = value from studentized range table

MSE = Mean Square Error from ANOVA table

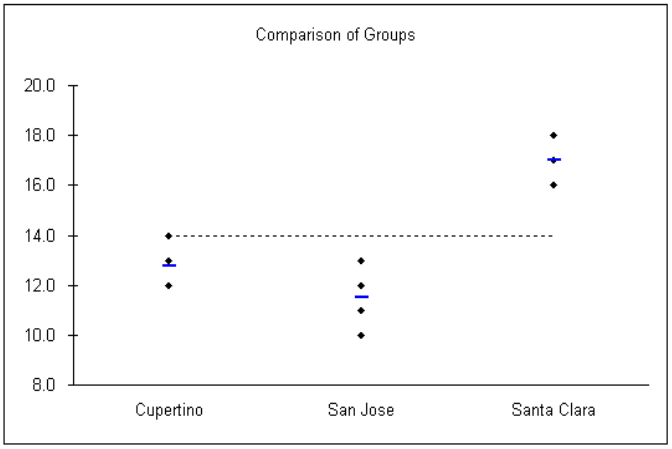
n_c = number of replicates per treatment. An adjustment is made for unbalanced designs.

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.

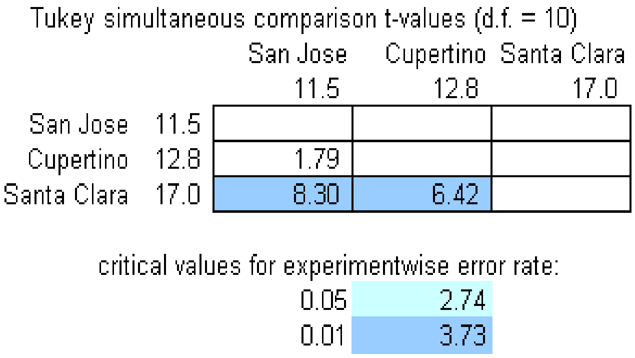
Example - Design

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

These three tests will be conducted with an overall significance level of α = 5%.
The model will be the Tukey HSD test.
The Decision rule will be to reject Ho for each pair where HSD>2.74

Example - Data/Results/Conclusion

Refer to the Megastat output.



Tukey simultaneous comparison t-values (d.f. = 10)

		San Jose	Cupertino	Santa Clara
		11.5	12.8	17.0
San Jose	11.5			
Cupertino	12.8	1.79		
Santa Clara	17.0	8.30	6.42	

critical values for experimentwise error rate:

0.05	2.74
0.01	3.73

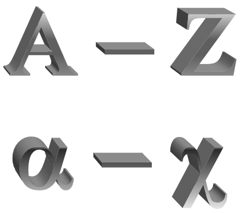
Santa Clara has a significantly higher mean number of tofu pizzas sold compared to both San Jose and Cupertino. There is no significant difference in mean sales between San Jose and Cupertino.

8.6 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.

9. 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.

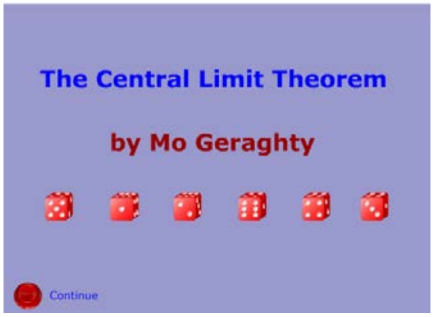
Type II Error
Failing to reject the Null Hypothesis when it is actually false.

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.

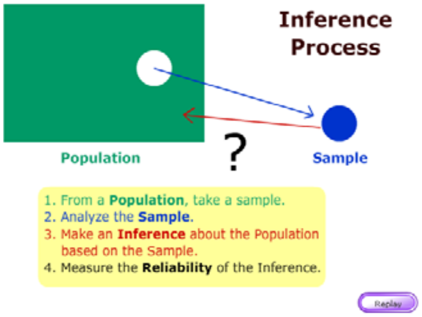
Z-score
A measure of relative standing that shows the distance in standard deviations a particular data point is above or below the mean.

10. 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/ .

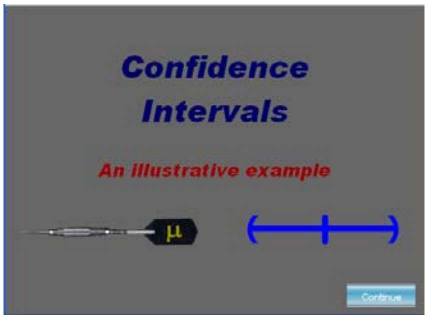


Central Limit Theorem (Section 4.3)
Using die rolling with progressively increasing sample sizes, this animation shows the three main properties of the Central Limit Theorem.

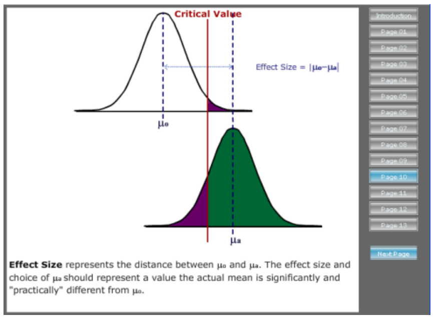


1. From a **Population**, take a sample.
2. Analyze the **Sample**.
3. Make an **Inference** about the Population based on the Sample.
4. Measure the **Reliability** of the Inference.

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.



Effect Size represents the distance between μ_o and μ_a. The effect size and choice of μ_a should represent a value the actual mean is significantly and "practically" different from μ_o.

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.

11. 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

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 (Partially covered in this text)

Section 9:
Correlation and Regression

12. 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] Ronald Walpole & Raymond Meyers & Keying Ye, *Probability and Statistics for Engineers and Scientists*. Pearson Education, 2002, 7th edition.

[9] Taleb, Nicholas, *The Black Swan: The Impact of the Highly Improbable*, Penguin, 2007.

[10] Food and Drug Administration, *FDA Consumer Magazine* , Jan/Feb 2003

[11] 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

[12] Russ Lenth, *Java Applets for Power and Sample Size*, University of Iowa , http://www.stat.uiowa.edu/~rlenth/Power/ , 2009

[13] J. B. Orris, *MegaStat for Excel,* Version 10.1, Butler University, 2007

[14] 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

[15] 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