Math 10 (Sec 28) – Statistics – Winter 2016 Syllabus

Instructor: Maurice (Mo) Geraghty
Email: geraghtymo@fhda.edu
Website: http://nebula2.deanza.edu/~mo

Office Location/Phone: S-49A (408) 864-5383
Office Hours: M 12:30-1:20 Tu 6:20-7:00
W 11:30-12:20 Th 11:30-1:00 (in LCW110)

Required Materials: Textbook – *Collaborative Statistics* by Illowsky/Dean **(online or printed copy)**
Textbook – *Inferential Statistics and Hypothesis Testing* by Geraghty **(online only)**

Calculator – Scientific Calculator is sufficient. **Cell phone calculators are not allowed on exams.**

Access to a computer outside of class; we will be using the computer lab and Minitab . Also, you will need an e-mail address and access to the Internet. Course topics, homework, exam information, handouts, data sets, and other information will be posted on the website.

Grading: Grading will be based on the following criteria. **Grades are not negotiable.**

* * * * * * * * * *Grading Scale (points) * * * * * * * * *		
460 - 446 = A+	445 - 427 = A	426 - 414 = A-
413 - 400= B+	399 - 381 = B	380 - 368 = B-
367 - 345 = C+	344 - 322 = C	321 - 299 = D+
298 - 276 = D	0 - 275 = F	

Grading Criteria	
Exams:	200 pts
Final:	100 pts
Labs:	110 pts
Homework:	50 pts

Homework: Completed Homework must be turned in by the due date, but should be completely daily. Homework assignments may also be posted on the website. **There is no credit for late homework**.

Exams: There will be two exams during the quarter. Your final exam score will replace your lowest scoring exam if it improves your grade. **There are no make-up exams.**

Final Exam: A comprehensive exam will be given on the final exam date.

Computer Lab: Lab classes will be held in the math computer lab: S44. You will use Mintab and other statistical software in analyzing data, learning statistical models and working on the class material Computer labs can be done in groups of no more than four people for a common grade and be turned in by email on the due date. **There is no credit for late labs received after midnight on the due date.**

Adding/Dropping: If you choose not to complete the course, it is your responsibility to officially drop or withdraw from the course by the deadline date. **I will not sign late drop or withdrawal forms.**

Attendance: It is expected that you attend both the lecture and labs. Attendance means arriving on time and staying the entire scheduled period.

Changes: Information in this syllabus may be changed during the quarter, but you will be informed in advance.

Other Information: All students are expected to understand the college policy on cheating as outlined in the student handbook. **Plagiarism (submitting another's work as your own) will result in an immediate failure for the course for your entire group.**

Cell phones and and other electronic devices need to be turned off or silenced. Please arrive on time and stay the entire period.

Read the **Frequently Asked Questions** on the website for other policies and procedures. Student Learning Outcomes (SLO's) are also posted on the class website.

If you feel that you may need an accommodation based on the impact of a disability, you should contact me privately to discuss your specific needs. Also, please contact Disability Support Services (864-8753) or Educational Diagnostic Center (864-8839) for information or questions about eligibility, services and accommodations for physical (DSS), psychological (DSS) or learning (EDC) disabilities.

Tentative Schedule - Math 10 - Sec 28
Winter Quarter - 2016
`

	Monday	Tuesday	Wednesday	Thursday	Friday
Jan	4 **Part 1**	5	6 **Part 1** **HW 0**	7 **Lab 1 Due**	8
Jan	11 **Part 1/2**	12	13 **Part 2** **HW 1**	14 **Lab 2 Due**	15 **Drop Deadline (Jan 17)**
Jan	18 **Holiday**	19	20 **Part 3** **HW 2**	21 **Lab 3 Due**	22
Jan	25 **Part 3/4**	26	27 **Part 4/Review** **HW 3**	28 **Lab 4 Due**	29
Feb	1 **Exam 1** **Part 4/5**	2	3 **Part 5** **HW 4**	4 **Lab 5 Due**	5
Feb	8 **Part 5/6**	9	10 **Part 6** **HW 5**	11 **Lab 6 Due**	12 **Holiday**
Feb	15 **Part 6**	16	17 **Part 6**	18 **Lab 7 Due**	19
Feb	22 **Holiday**	23	24 **Part 7** **HW 6**	25 **Lab 8 Due**	26 **Withdraw Deadline**
Feb/Mar	29 **Part 7**	1	2 **Review/Part 8**	3 **Lab 9 Due**	4
Mar	7 **Exam 2** **Part 8** **HW 7**	8	9 **Part 8**	10 **Lab 10 Due**	11
Mar	14 **Part 8/9**	15	16 **Part 9** **HW 8** **Review**	17 **Lab 11 Due**	18
Mar	21	22	23 **Final Exam 4:00-6:00** **HW 9**	24	25

Slides	Topic	Illowsky/Dean	Geraghty
Part 1	Descriptive Statistics	1 (all), 2 (all), 6.3, 12.4, 12.6, 12.7	Sec 4 - outliers
Part 2	Probability	3 (all)	
Part 3	Discrete Random Variables	4 (omit 4.7)	
Part 4	Continuous Random Vars/Central Limit Theorem	5 (all), 6 (all), 7 (omit 7.3)	Sec 4
Part 5	Confidence Intervals	8 (all)	Sec 5
Part 6	1 Population Hypothesis Testing/Power	9 (all), 11.6	Sec 6
Part 7	2 Population Hypothesis Testing	10 (omit 10.4), 13.5	Sec 7
Part 8	Chi-Square tests/ANOVA	11 (all), 13 (all)	Sec 8
Part 9	Regression	12	

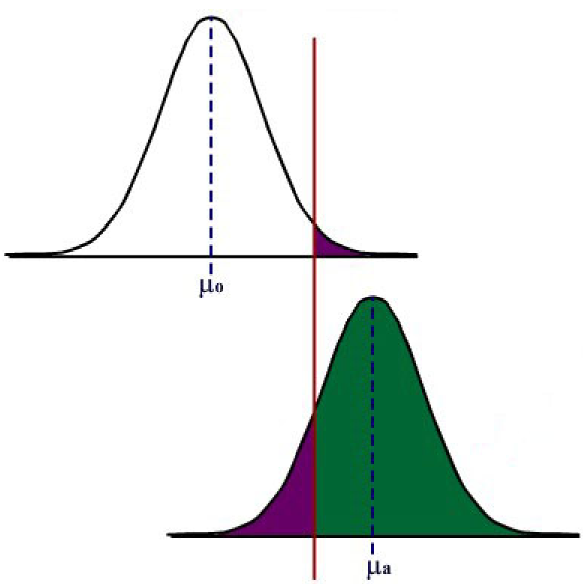
DE ANZA COLLEGE – DEPARTMENT OF MATHEMATICS

Inferential Statistics and Hypothesis Testing

A Holistic Approach

Maurice A. Geraghty

1/1/2016

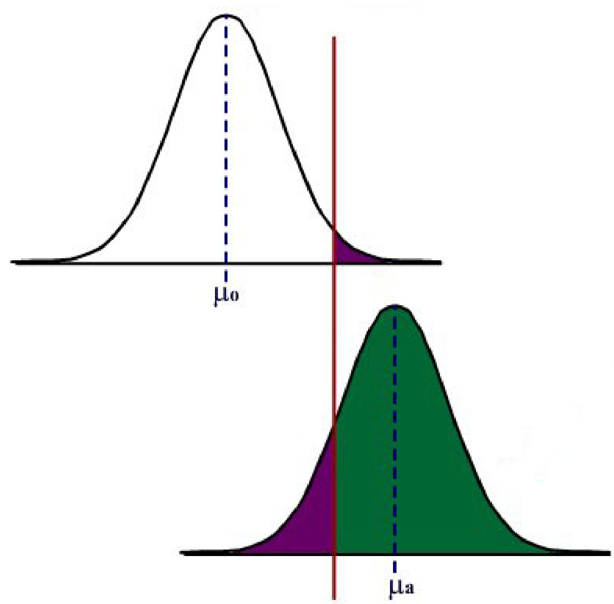


Supplementary material for an introductory lower division course in Probability and Statistics

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
originally published September 1, 2009
revised January 1, 2016



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.

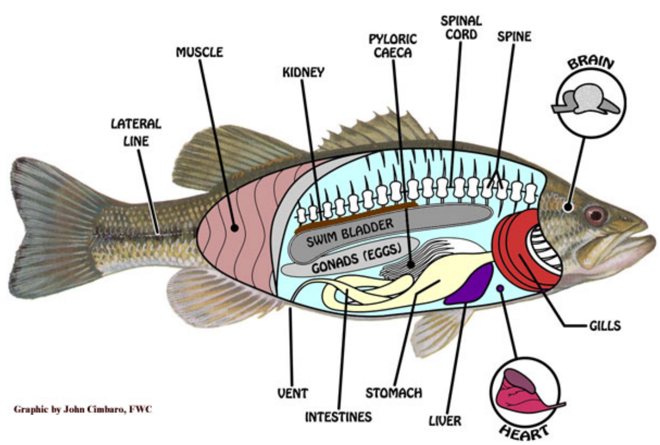
These students, (to their credit) continued to work hard; they wanted to understand the material, not simply pass the class. Since these students had excellent math skills, I went deeper into the explanation of Type II error and the statistical power function. Although they could compute power and sample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



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

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

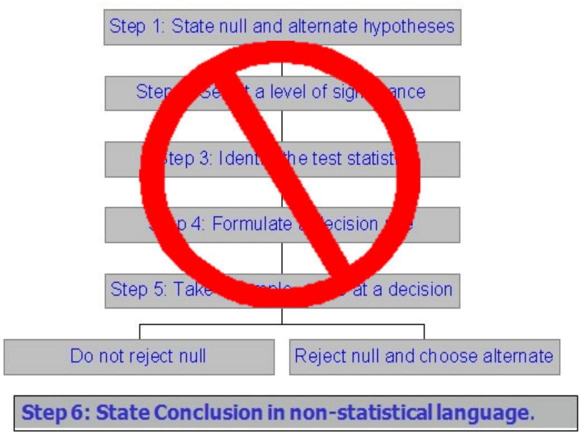
Although science has traditionally been reluctant to embrace, often hostile to including holistic philosophy in the scientific method, there have been many who now support a multicultural or multi-philosophical approach. In his book Holism and Reductionism in Biology and Ecology[4], Looijen claims that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

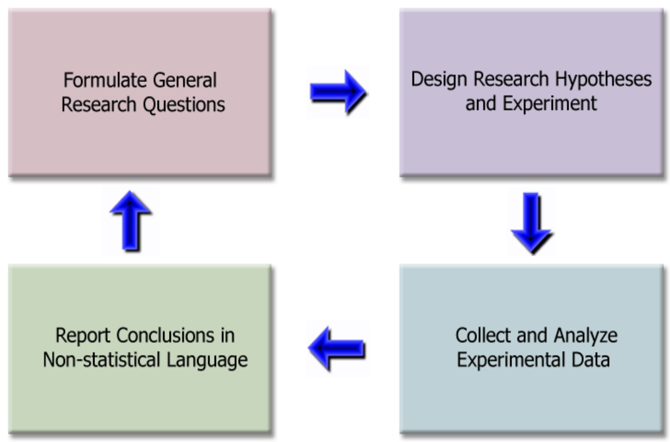
I realize that the problem of my high-achieving students being unable to comprehend hypothesis testing could be cultural – these were international students who may have been schooled under a more holistic philosophy. The Introductory Statistics curriculum and most texts give an incomplete explanation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

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

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

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





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

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

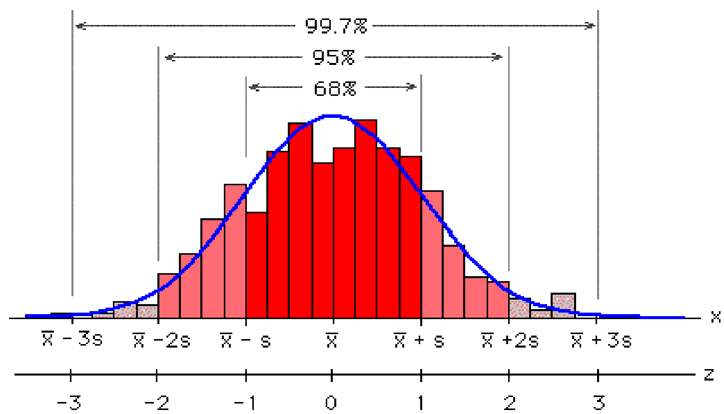
A student asked me about the distribution of exam scores after she saw her score of 87 out of 100. I told her the distribution of test scores were approximately bell-shaped with a mean score of 75 and a standard deviation of 10. Most people would have an intuitive grasp of the mean score as being the "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: $$Z = \frac{X_i - \overline{X}}{s}$$

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

Interpreting Z-score for Several Students

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

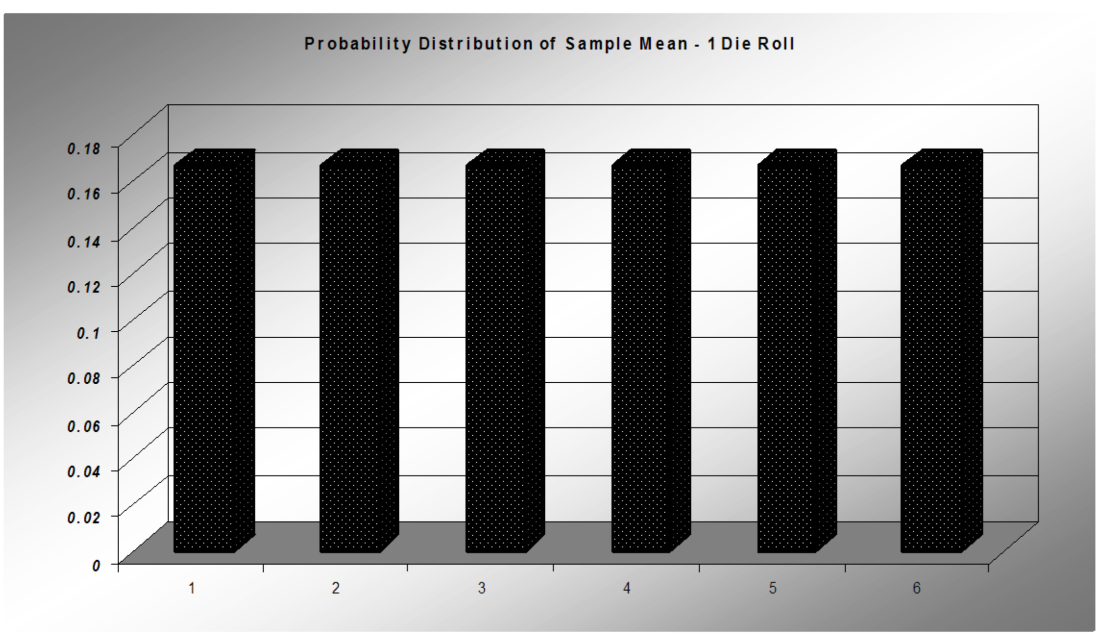
4.3 The Sample Mean as a Random Variable – Central Limit Theorem

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.

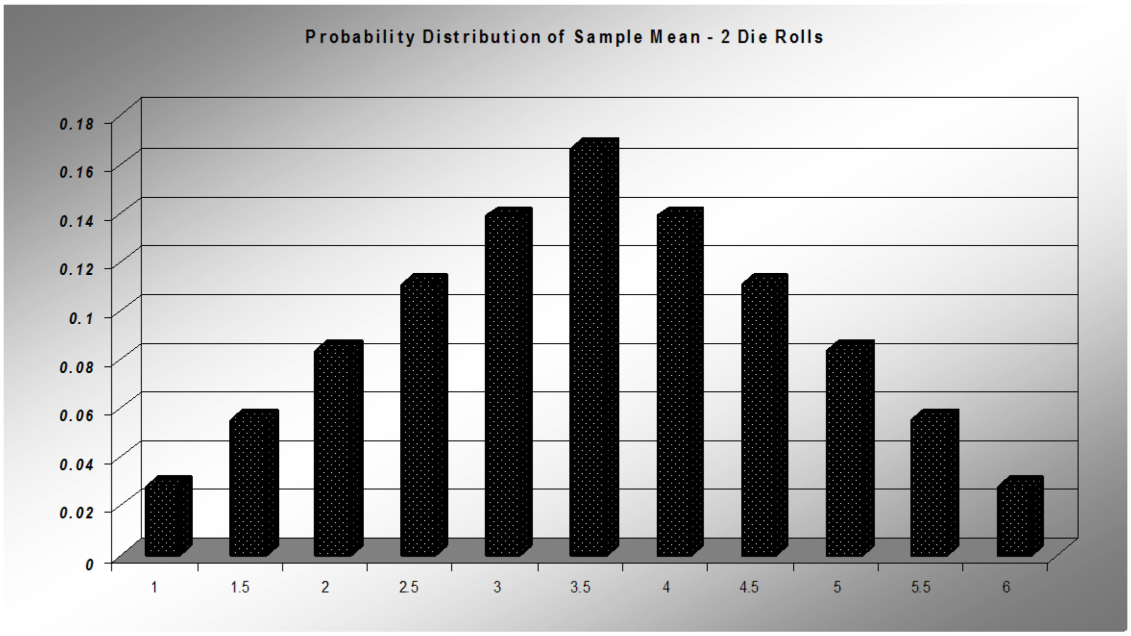
We start with a Random Sample X_1, X_2, …, X_n where each of the random variables X_i has the same probability distribution and are mutually independent of each other. The sample mean is a function of these random variables (add them up and divide by the sample size), so $\bar{X}$ is a random variable.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

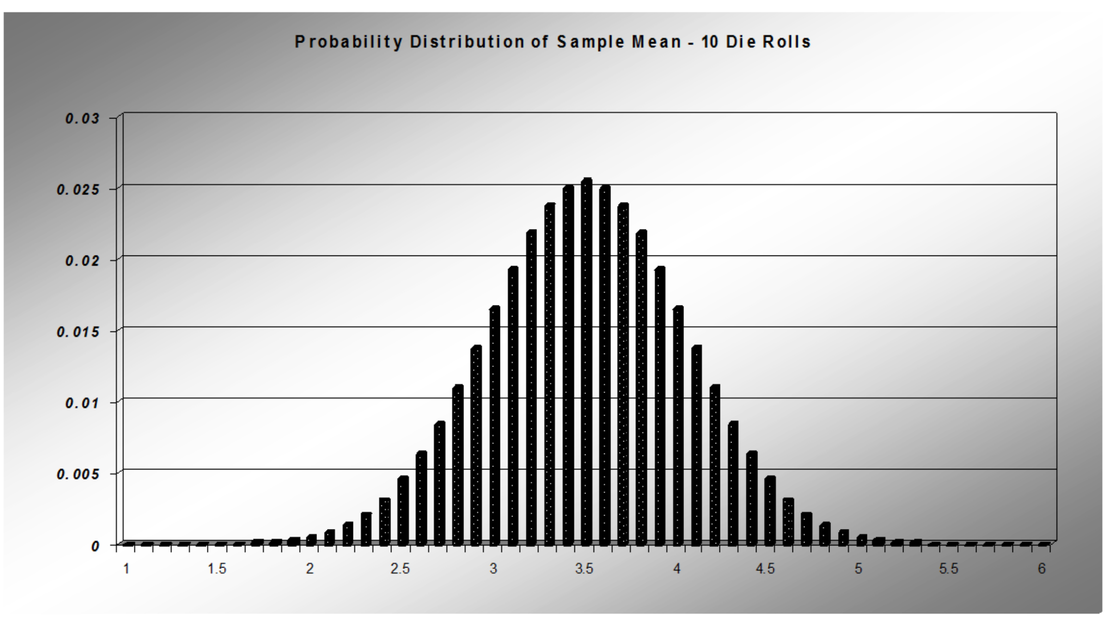
For the case n=1, the distribution of the sample mean is the same as the distribution of the random variable. Since each die has the same chance of being chosen, the distribution is rectangular shaped centered at 3.5:



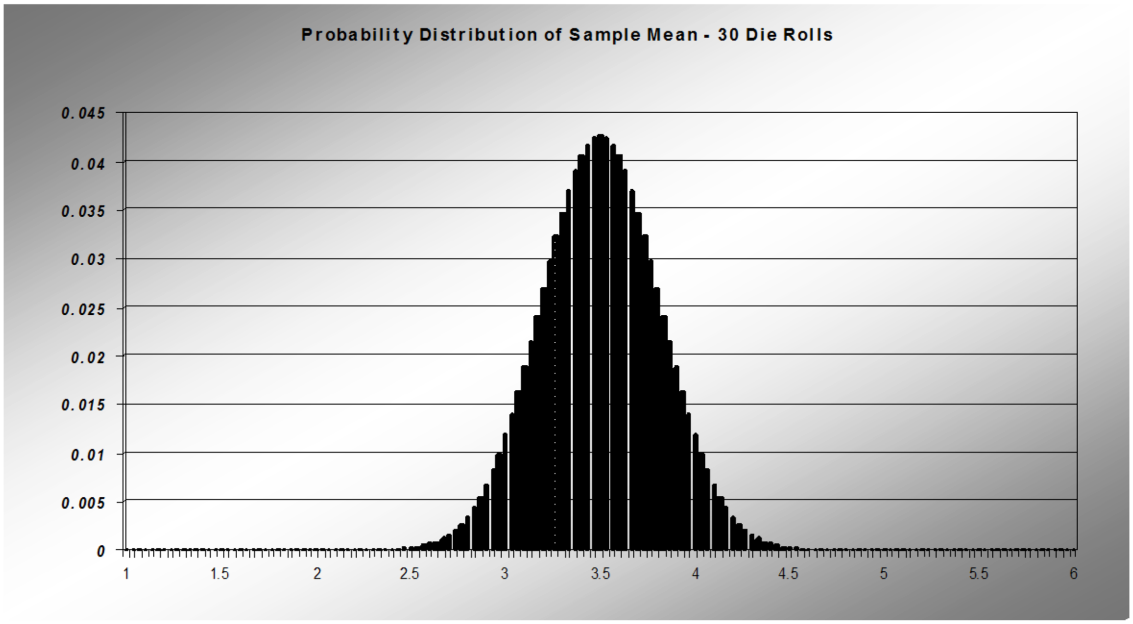
For the case n=2, the distribution of the sample mean starts to take on a triangular shape since some values are more likely to be rolled than others. For example, there six ways to roll a total of 7 and get a sample mean of 3.5, but only one way to roll a total of 2 and get a sample mean of 1. Notice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

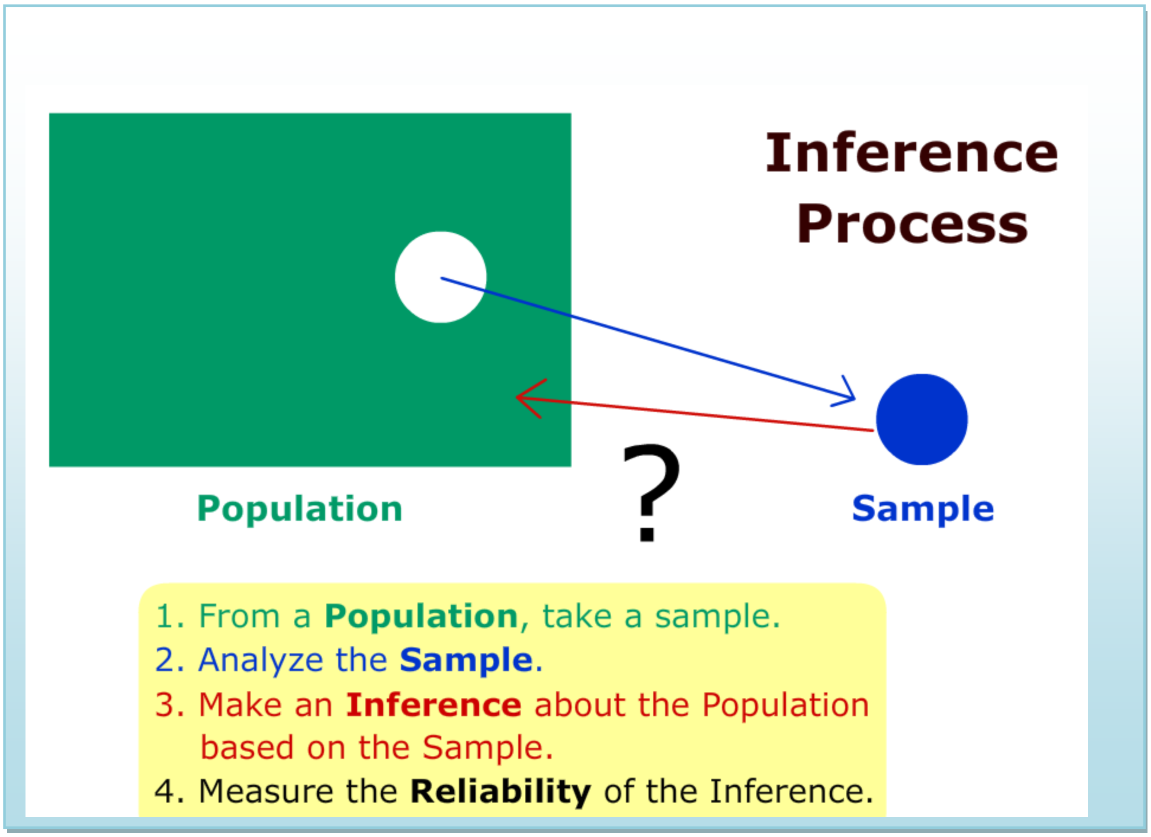
$$P(X > 70) = P\left(Z > \frac{(70 - 69.2)}{2.9} \right) = P(Z > 0.28) = 0.3897$$

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.



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

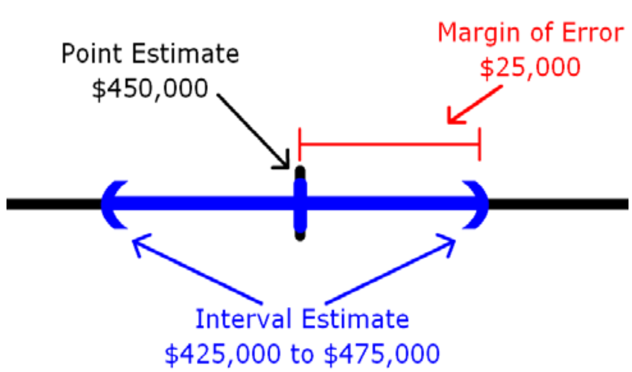
In the example above, Lupe's realtor estimated the population mean of similar homes in Lupe's neighborhood by using the sample mean of $450,000 from the adjusted price of the sampled homes.

Interval Estimation

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

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

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



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.

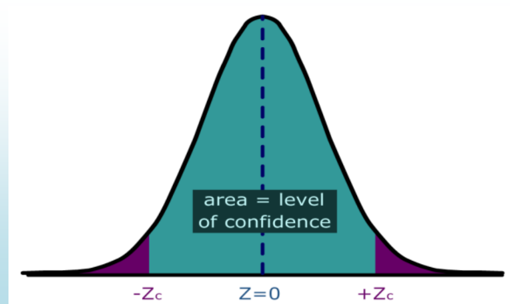
5.3.1 Confidence Interval for Population Mean

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?

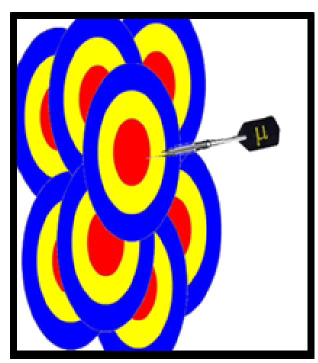
$$24 \pm \frac{1.96 \cdot 4}{\sqrt{49}} = 24 \pm 1.12 = (22.88, 25.12) \text{ hours per week}$$

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



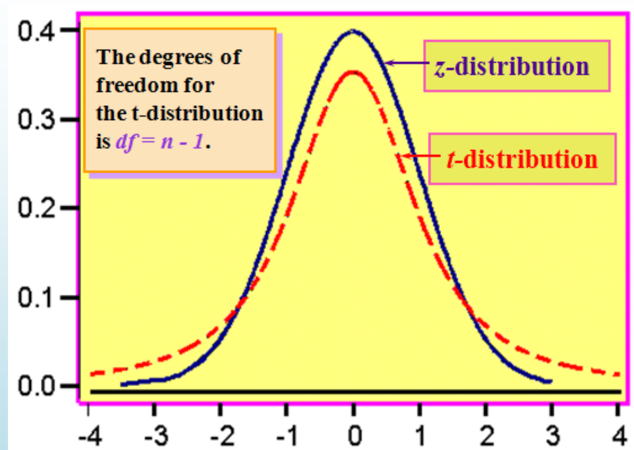
- The confidence interval is constructed from random variables calculated from sample data and attempts to predict an unknown but fixed population parameter with a certain level of confidence.
- Increasing the level of confidence will always increase the margin of error.
- It is impossible to construct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

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

Example

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

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

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

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

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 $\widehat{p}$ 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.

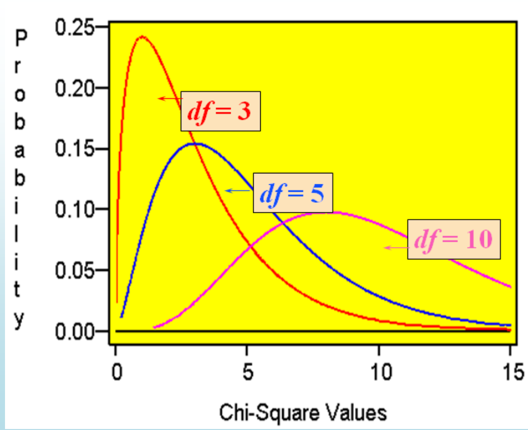
To create a confidence interval for an estimate of standard deviation, we need to introduce a new distribution, called the Chi-square (χ^2) distribution.

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.



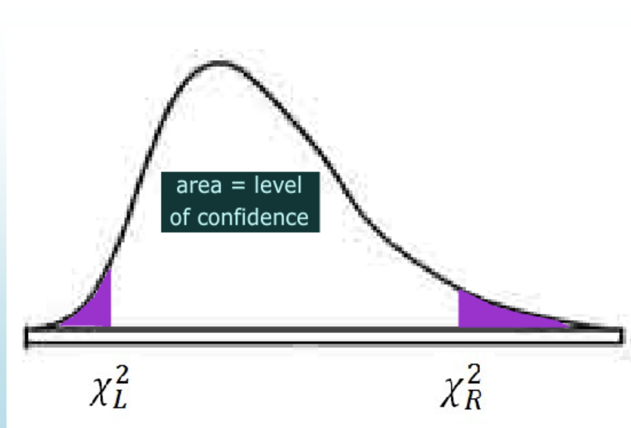
5.3.5 Confidence Interval for Population Variance and Standard Deviation

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 $\chi_R^2 = 32.8523$.

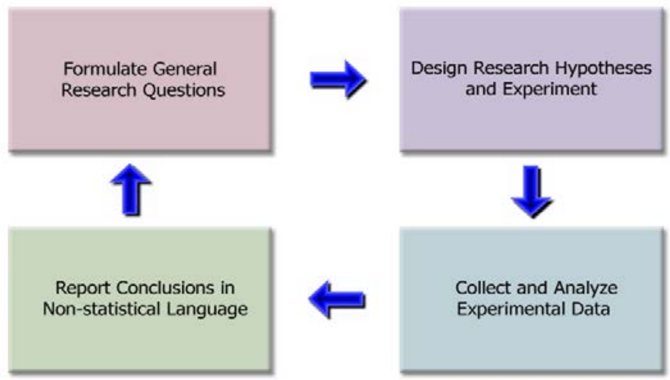
Formula for confidence interval for σ is: $\left(\sqrt{\frac{(19)5^2}{32.8523}}, \sqrt{\frac{(19)5^2}{8.90655}}\right) = (3.8, 7.3)$

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.

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.

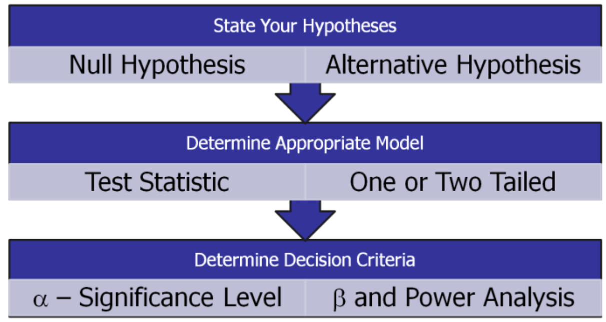
6.2 Formulate General Research Questions

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

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

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

Examples of Null and Alternative Hypotheses:

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

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.

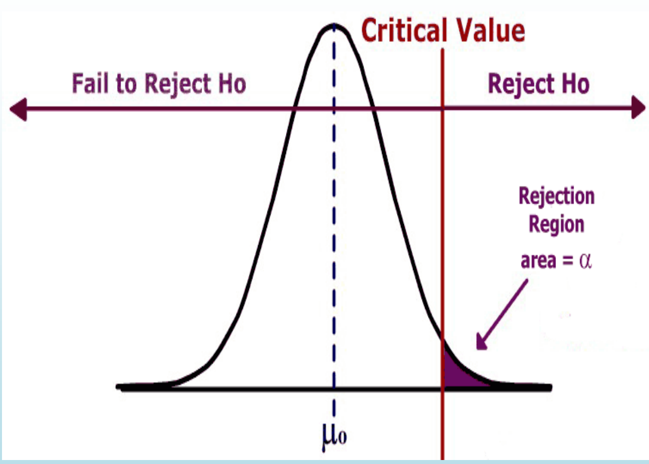
In designing an experiment, we set a maximum probability of making Type I error. This probability is called the **level of significance** or **significance level** of the test and designated by the Greek letter α.

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.

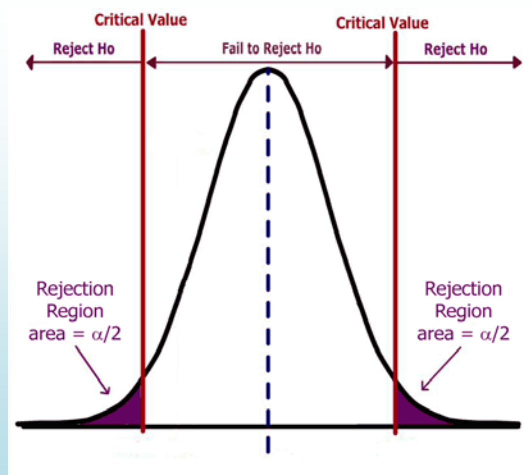
6.3.4 Critical Value and Rejection Region

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.

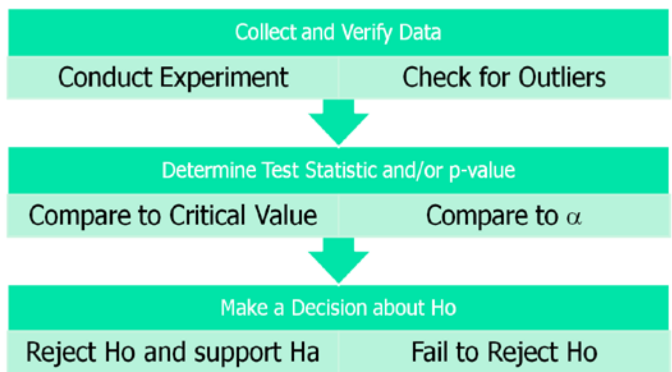
In a two tailed-test, the significance level is split into two parts since there are two rejection regions. In hypothesis testing where the statistical model is symmetrical (eg: the Standard Normal Z or Student's t distribution) these two regions would be equal. There is a relationship between a confidence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

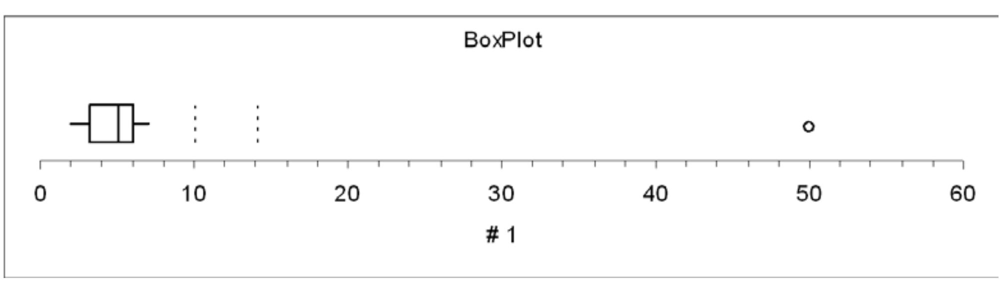
2 2 3 4 5 5 6 6 7 50

	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.

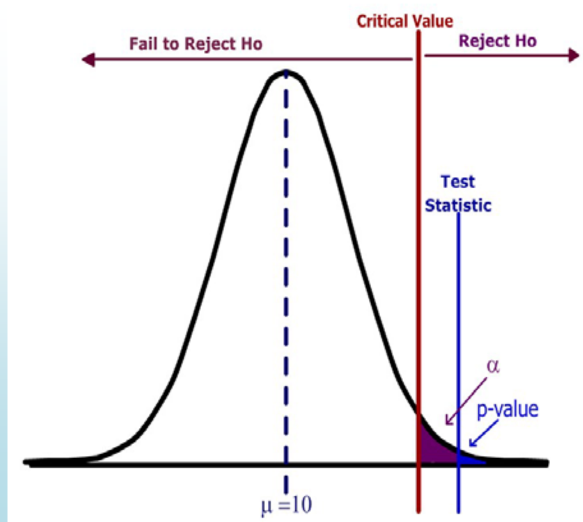
1. If the test statistic lies in the rejection region, reject Ho. (critical value method)
2. If the p-value < α, reject Ho. (p-value method)

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_0 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

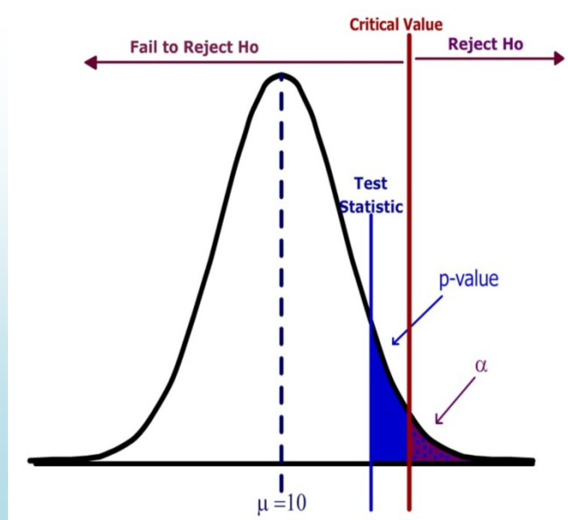
- Ho: $\mu = 10$
  Ha: $\mu > 10$
- Design: Critical value is determined by significance level α.
- Data Analysis: p-value is determined by test statistic
- Test statistic falls in rejection region.
- p-value (blue) < α (purple)
- Reject Ho.
- Strong statement: Data supports the Alternative Hypothesis.



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_0.

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_0.

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

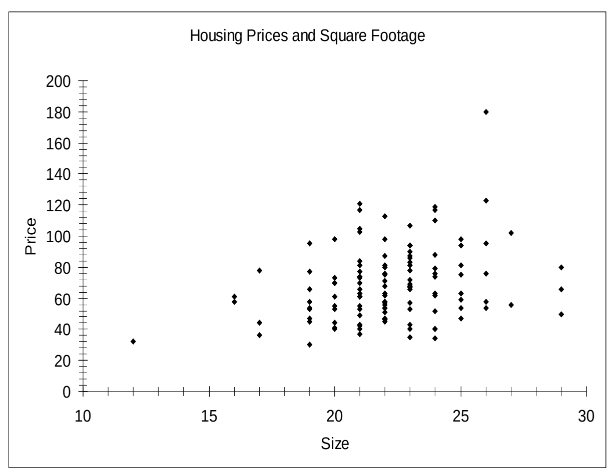
Conclusion if Ho is rejected: The mean number of text messages sent by students exceeds 1000.

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**.

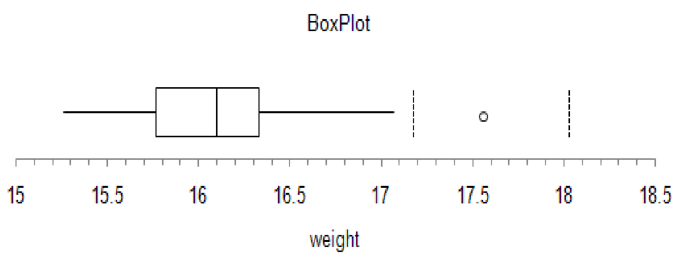
The quality control statistician has been given the authority to sample 36 bottles of soy sauce and knows from past testing that the population standard deviation is 0.5 ounces. The model will be a **test of population mean vs. hypothesized valu**e of 16 oz. A two-tailed test is selected since the company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

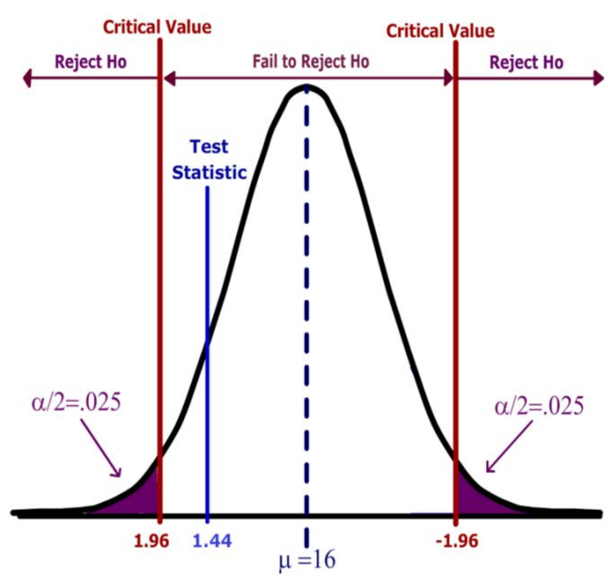


Next, the sample mean and the test statistic are calculated.

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

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.

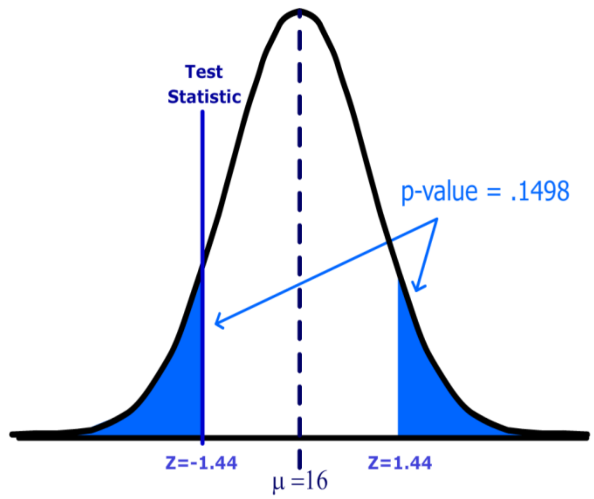


Alternatively (and preferably) the statistician would use the p-value method of decision rule. The p-value for a two-tailed test must include all values (positive and negative) more extreme than the Test Statistic, so in this example we find the probability that $Z < -1.44$ or $Z > 1.44$ (the area shaded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

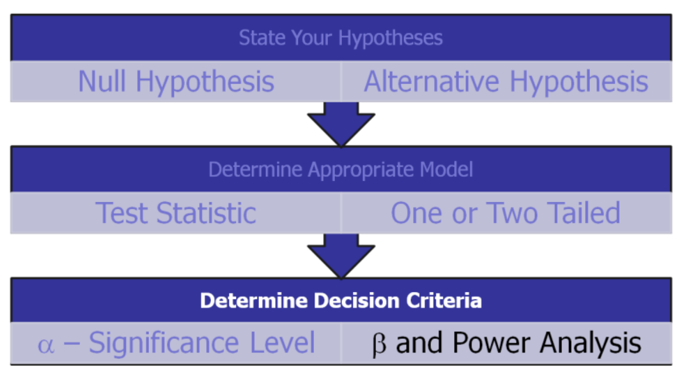
> "There is insufficient evidence to conclude that the machine that fills 16 ounce soy sauce bottles is operating improperly. This conclusion is based on 36 measurements taken during a single hour's production run. I recommend continued monitoring of the machine during different employee shifts to account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



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 (μ_a)** that is "**practically different**" from the **mean under the Null Hypothesis (μ_o).** This practical difference is called the **effect size**.

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

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

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

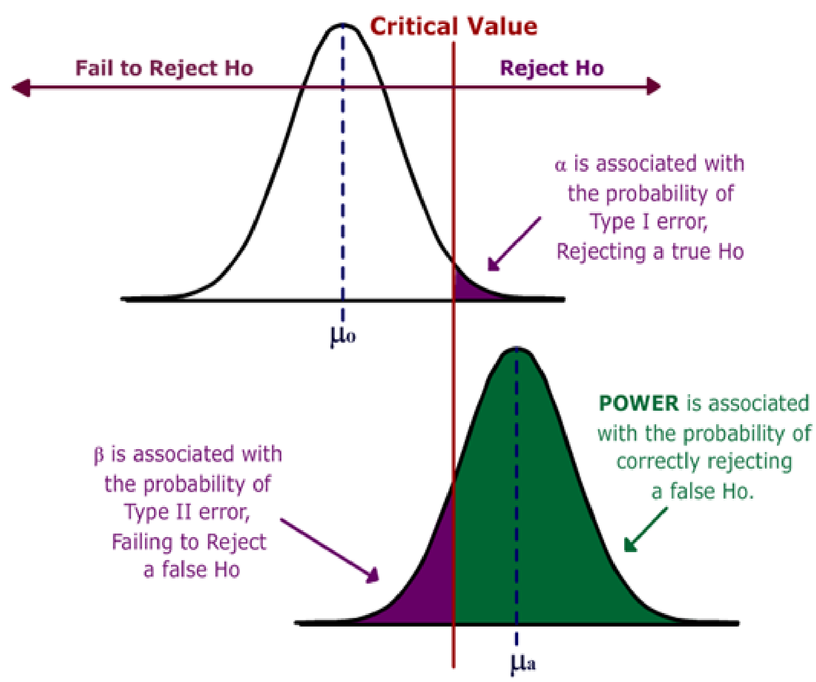
Suppose we are conducting a one-tailed test of the population mean:

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.

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

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



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

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

Example

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

First, find the critical value of the test.

Reject Ho when Z < -1.645

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

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

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

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

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

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

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

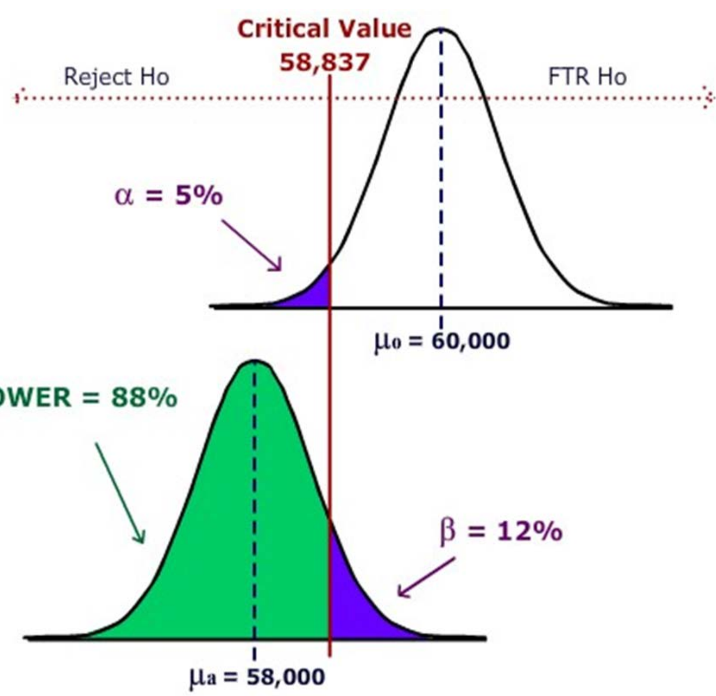
Power Calculation Values

Input Values

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

Calculated Values

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



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:

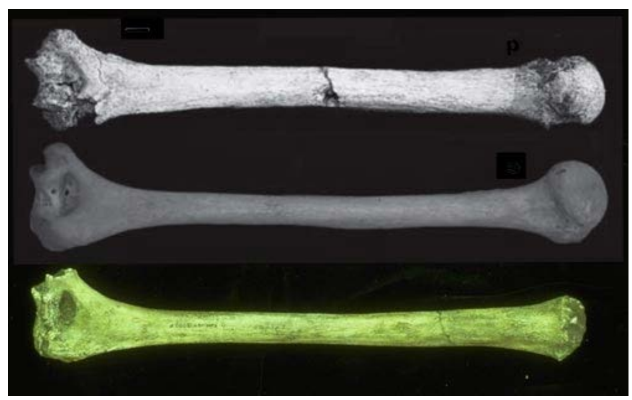
- Test of population mean vs. hypothesized value, population standard deviation unknown.
- Test of population proportion vs. hypothesized value.
- Test of population standard deviation (or variance) vs. hypothesized value.

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.

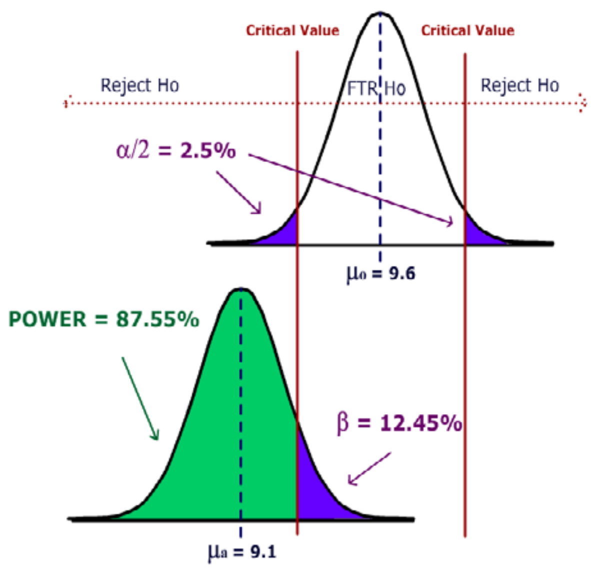
2. Determine the power of this test if the bones actually came from Species B (assume a standard deviation of 0.7)

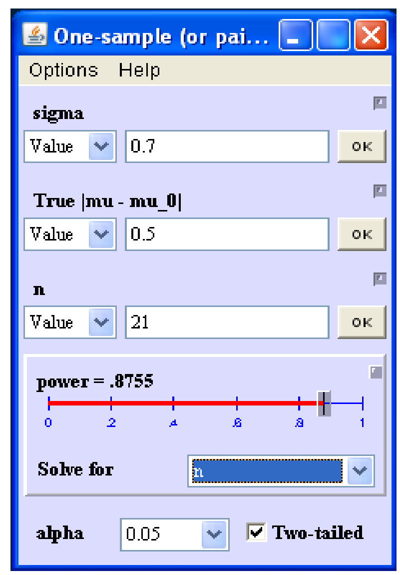
Information needed for Power Calculation

- μ_o = 9.6 (Species A)
- μ_a = 9.1 (Species B)
- Effect Size =| mo - ma | = 0.5
- s = 0.7 (given)
- α = .05
- n = 21 (sample size)
- Two tailed test

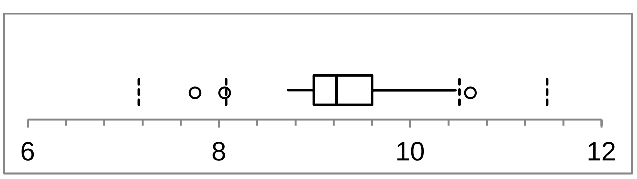
Results using Online Power Calculator[12]

- Power =.8755
- β = 1 - Power = .1245
- If humerus bones are from Species B, test has an 87.55% chance of correctly rejecting Ho and a maximum Type II error of 12.55%





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



Hypothesis Test: Mean vs. Hypothesized Value

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

From MegaStat[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

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

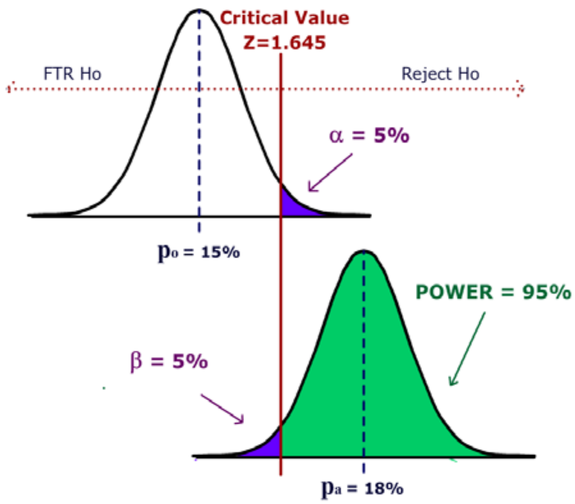
Test Statistic: $Z = \frac{\hat{p}-p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$ **Requirement for Normality Assumption:** $np(1-p) > 5$

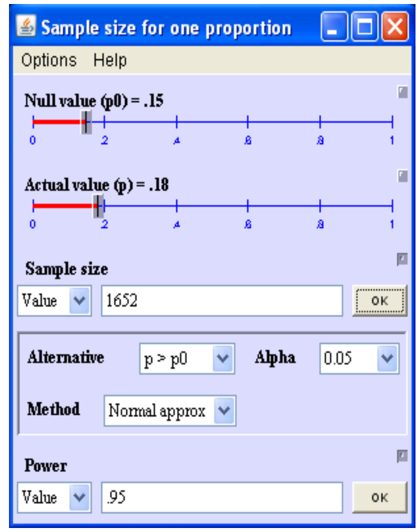
Example

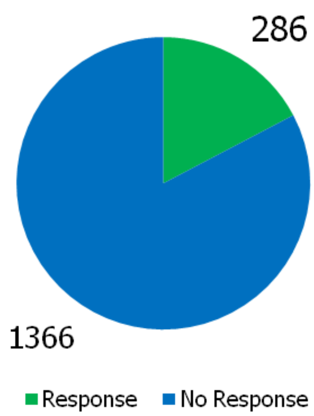
In the past, 15% of the mail order solicitations for a certain charity resulted in a financial contribution. A new solicitation letter has been drafted and will be sent to a random sample of potential donors. A hypothesis test will be run to determine if the new letter is more effective. Determine the 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







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

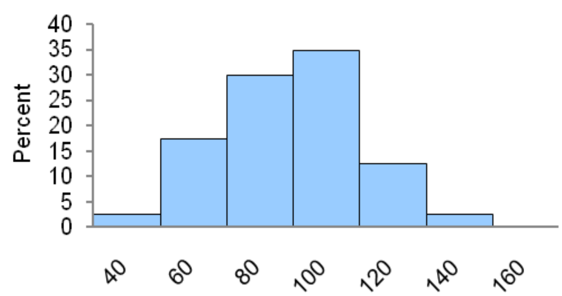
$\boldsymbol{\sigma^2}$ = population variance $\boldsymbol{\sigma_o^2}$ = population variance under Ho

$\boldsymbol{s^2}$ = 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.

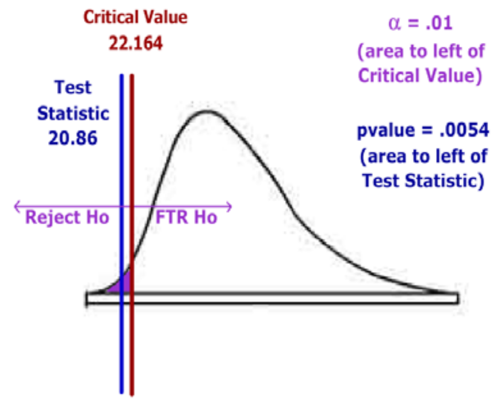
Hypotheses In terms of the population variance:

- 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 8^{th} grade test scores is less than 30. The 40 test scores were the results of the recently administered exam to the 8^{th} 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

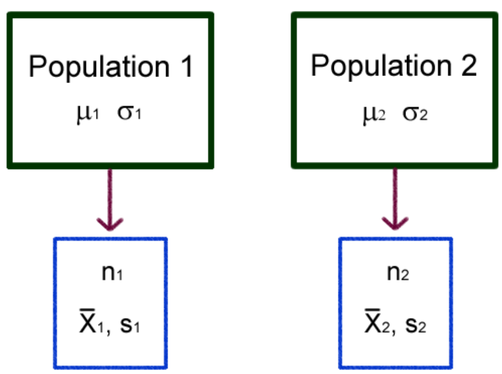
In this section we consider expanding the concepts from the prior section to design and conduct hypothesis testing with two samples. Although the logic of hypothesis testing will remain the same, care must be taken to choose the correct model. We will first consider comparing two population means.

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

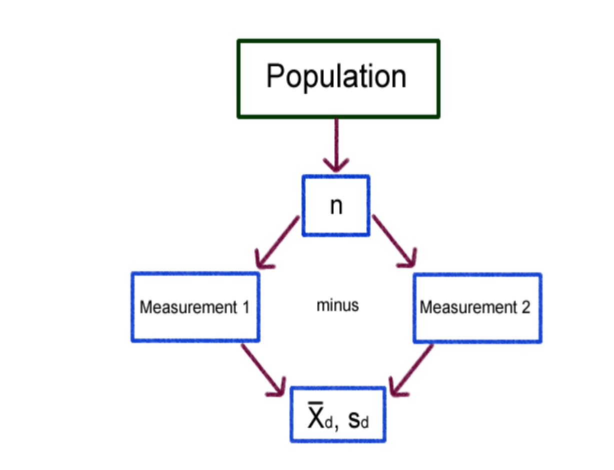
- n_1 **is the sample size from population 1.**
- n_2 **is the sample size from population 2.**
- $\bar{X}_1$ **is the sample mean from population 1.**
- $\bar{X}_2$ **is the sample mean from population 2.**
- s_1 **is the sample standard deviation from population 1.**
- s_2 **is the sample standard 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 $\boldsymbol{Ho}: \boldsymbol{\mu_{1=}\mu_2}$. 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 $\boldsymbol{Ho}: \boldsymbol{\mu_{1-}\mu_2 = 0}$ and use the difference of sample means $\bar{X}_1 - \bar{X}_2$ as a point estimator for the difference in population means $\mu_1 - \mu_2$.

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

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

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

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 $\boldsymbol{Ho}: \boldsymbol{\mu_{1=}\mu_2}$ 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%.

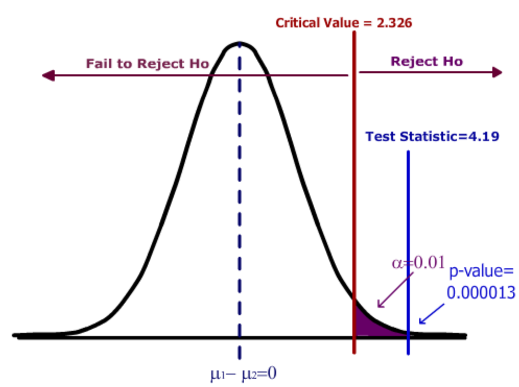
The decision rule under the critical value method would be to reject the Null Hypothesis when the value of the test statistic is in the rejection region. In other words, reject Ho when Z > 2.326. The decision under the p-value method is to reject Ho if the p-value is < α.

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

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

Degrees of freedom = $n_1 + n_2 - 2$

Example

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

Example - Design

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

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

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

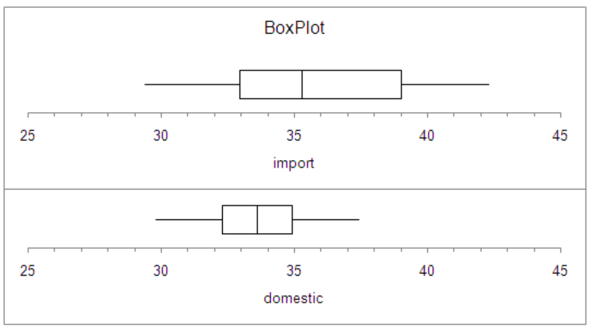
Type I error would be to reject the Null Hypothesis and claim imports has a higher mean MPG, when they do not have higher MPG. The test will be run at a level of significance (α) of 5%.

The degrees of freedom for this test is 25, so the decision rule under the critical value method would be to reject 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

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

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

Example

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

Example - Design

Again we will let domestic cars be population 2.

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

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

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

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

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

Example - Data/Results

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

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

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

16	df
2.17000	difference (import - domestic)
1.24606	standard error of difference
0	hypothesized difference
1.74	t
.0504	p-value (one-tailed, upper)

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

Example - Conclusion

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

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

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: $\mu_d\neq0$**

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

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

Example - Data/Results

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

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

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

Hypothesis Test: Paired Observations

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

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

Example – Conclusion

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

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

7.4 Independent sampling – comparing two population variances or standard deviations

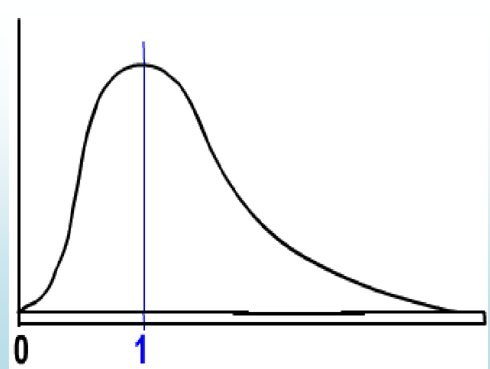
Sometimes we want to test if two populations have the same spread or variation, as measured by variance or standard deviation. This may be a test on its own or a way of checking assumptions when deciding between two different models (e.g.: pooled variance t-test vs. unequal variance t-test). We will 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
$H_o: \sigma_1 \geq \sigma_2$ $H_a: \sigma_1 < \sigma_2$	$F = \frac{s_2^2}{s_1^2}$ *use α table*
$H_o: \sigma_1 \leq \sigma_2$ $H_a: \sigma_1 > \sigma_2$	$F = \frac{s_1^2}{s_2^2}$ *use α table*
$H_o: \sigma_1 = \sigma_2$ $H_a: \sigma_1 \neq \sigma_2$	$F = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$ *use α / 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: $\sigma_1 \leq \sigma_2$ (Software stocks do not have more variation)**
Ha: $\sigma_1 > \sigma_2$ (Software stocks do have more variation)

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

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

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

Example - Data/Results

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

Example – Conclusion

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

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.

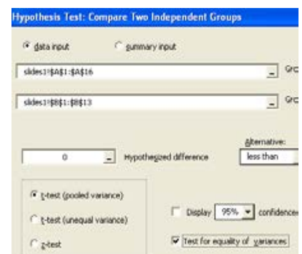
Example - Design

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

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

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

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



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

Example - Data/Results

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

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

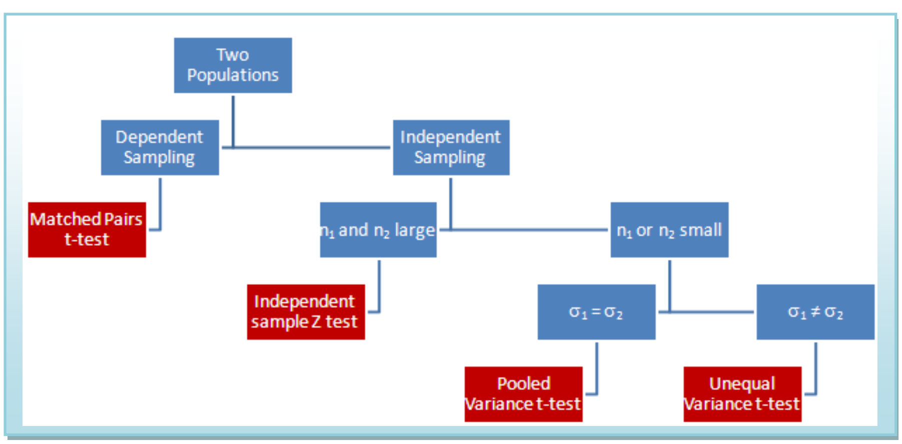
Also p-value = 0.0438 < 0.10 which also makes the result significant.

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. Chi-square Tests and Analysis of Variance (ANOVA)

Often we want to conduct tests claims about the characteristics of qualitative or categorical non-numeric data. In Section 6, we covered a test of one population proportion. In reality, this was a test of a categorical variable with 2 choices (success, failure). Now in this section, we will expand our study of hypothesis tests involving categorical data to include categorical random variables with more than two choices using a goodness-of-fit test. In addition, we will compare two categorical variables for independence. Both of these models will use a Chi-square test statistic, by looking at deviations between the observed values and expected values of the data.

8.1.1 Chi-Square Goodness-of-Fit test

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

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

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

Chi-square Goodness of Fit Test

Model Assumptions

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

Test Statistic

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

k = number of categories

n = sample size

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

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

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

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

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

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

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

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

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

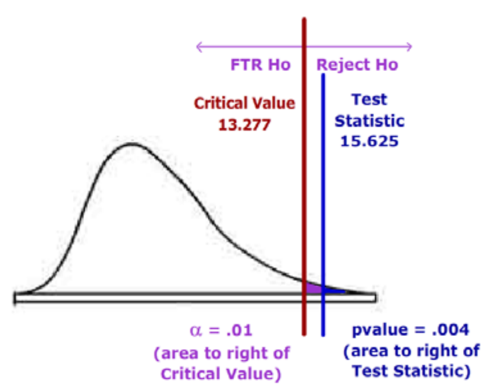
Important Assumption: The Expected Value of Each Category needs to be greater than or equal to 5. In this example, $E_i = np_i = (400)(.20) = 80 \geq 5$ for each category, so the model is appropriate.

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

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

Results:

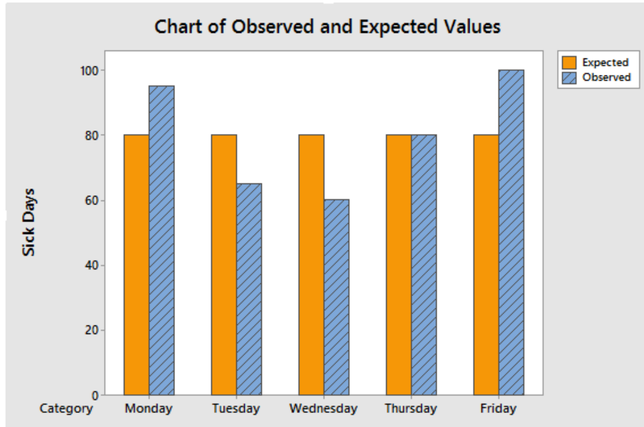
Day of Week	Observed Frequency O_i	Expected proportion p_i	Expected Frequency E_i	$\frac{(O_i - E_i)^2}{E_i}$
Monday	95	.20	80	2.8125
Tuesday	65	.20	80	2.8125
Wednesday	60	.20	80	5.0000
Thursday	80	.20	80	0.0000



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

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

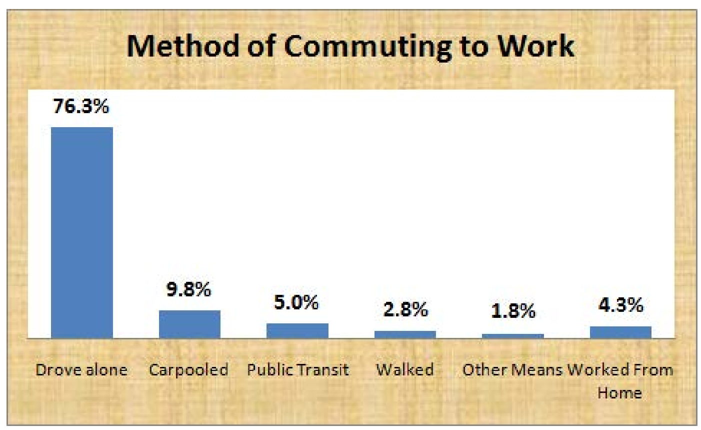
Conclusion: There is a difference in the proportion of employees who call in sick due to the day of the week. Employees are more likely to call in sick on days close to the weekend.



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

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

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



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

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

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

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

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

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

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

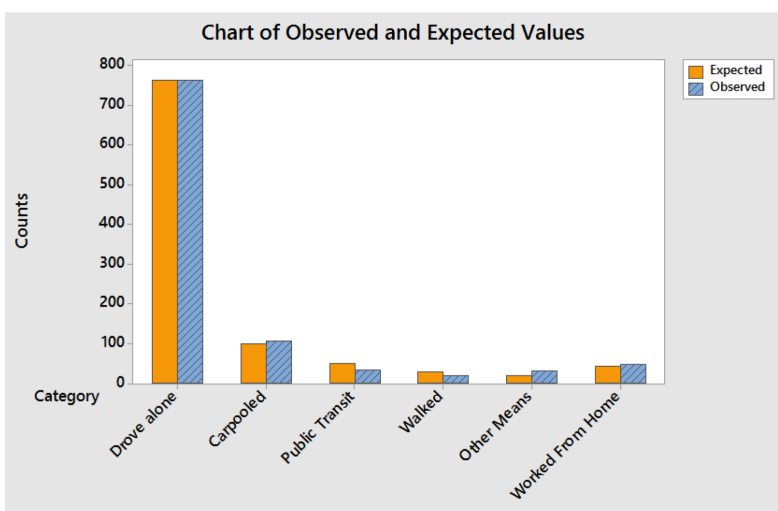
Decision Rule (Critical Value Method): Reject Ho if $\chi^2 > 11.071$ ($\alpha = .05$, 5 df)

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

Results:

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

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



Conclusion: Workers in Santa Clara County do not have the same frequencies of method of commuting as the entire United States.

8.2.1 Chi-Square Test of Independence

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

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

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

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

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

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

8.2.2 Two-way tables

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

8.2.3 Chi-square test for Independence

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

Chi-square Test for Independence

Model Assumptions

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

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

Test Statistic

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

r = number of row categories

c = number of column categories

n = sample size

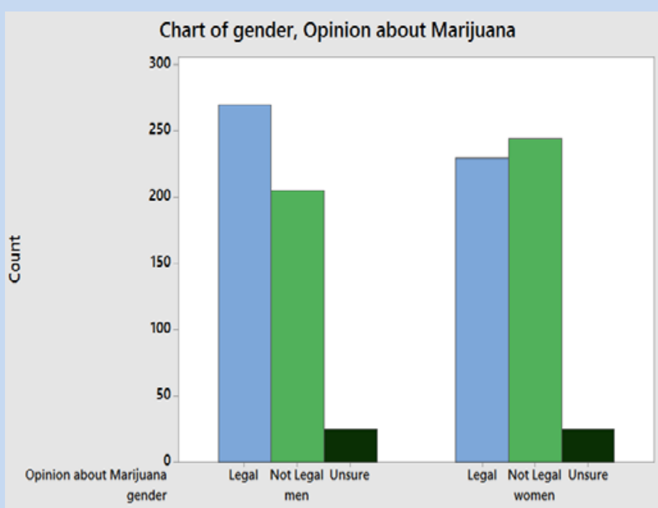
Research Hypotheses: **Ho:** Gender and Opinion about legalization of marijuana for recreational use are independent events.

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

Statistical Model: Chi-square Test of Independence

Results

```
Rows: Opinion about Marijuana
Columns: gender

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

                men  women   All

Legal           270    230   500
                250    250
              1.600  1.600

Not Legal       205    245   450
                225    225
              1.778  1.778

Unsure           25     25    50
                 25     25
              0.000  0.000

All             500    500  1000
```



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

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

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

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

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

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

8.3 One Factor Analysis of Variance (ANOVA)

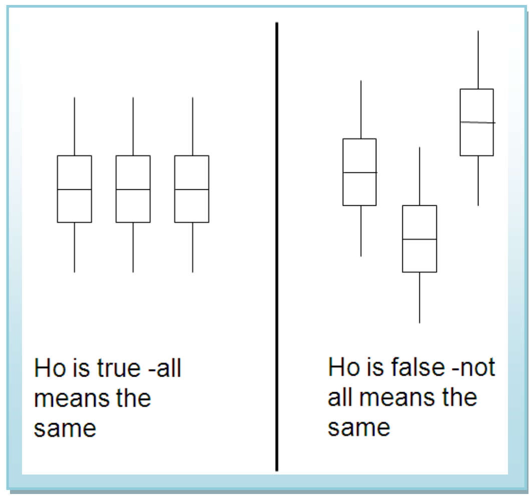
In the Section 7 we used statistical inference to compare two population means under variety of models. These models can be expanded to compare more than two populations using a technique called Analysis of Variance, or ANOVA for short. There are many ANOVA models, but we limit our study to one of them,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

8.3.1 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 $\boldsymbol{Ho}: \boldsymbol{\mu_1} = \boldsymbol{\mu_2} = \cdots = \boldsymbol{\mu_k}$. 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.

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

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



8.4 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

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

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

8.5 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.

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

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

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

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

Example

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

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

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

Example - Design

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

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

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

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

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

Example - Data/Results

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

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

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

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

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

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

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

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

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

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

Example – Conclusion

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

8.6 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.

8.6.1 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.

8.6.2 The Tukey HSD test

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

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

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

q = value from studentized range table

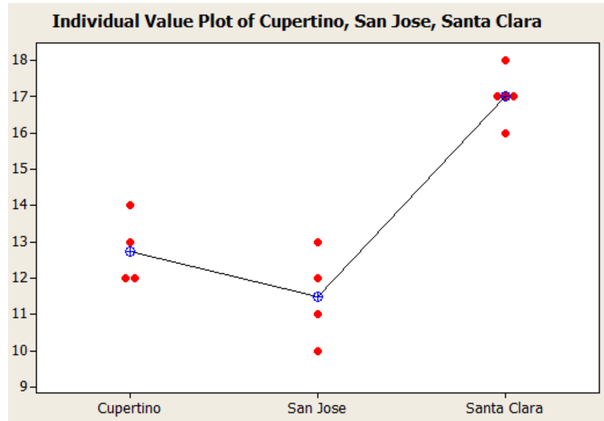
MSE = Mean Square Error from ANOVA table

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

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

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

Example



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

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

Example - Design

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

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

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

Example - Data/Results/Conclusion

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

```
Grouping Information Using Tukey Method

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

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

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

8.7 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.

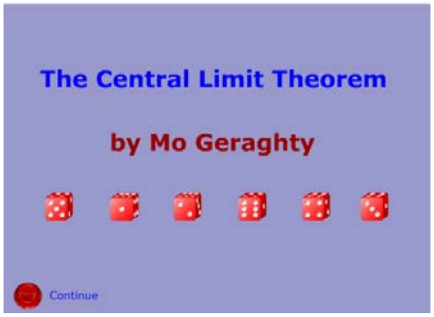
Type II Error
Failing to reject the Null Hypothesis when it is actually false.

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.

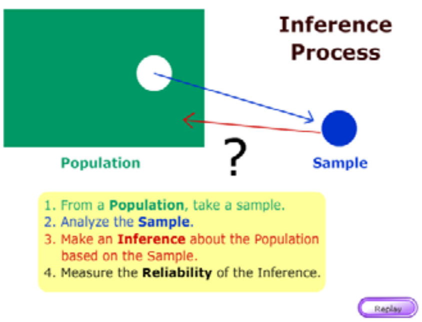
Z-score
A measure of relative standing that shows the distance in standard deviations a particular data point is above or below the mean.

10. Flash Animations

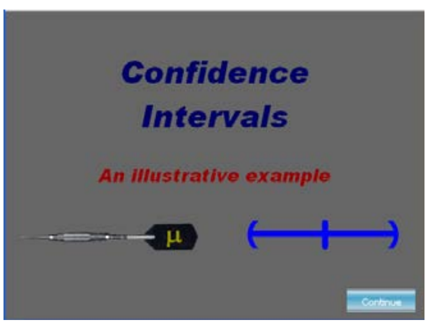
I have designed four interactive Flash animations that will provide the student with deeper insight of the major concepts of inference and hypothesis testing. These animations are on my website http://nebula2.deanza.edu/~mo/ .



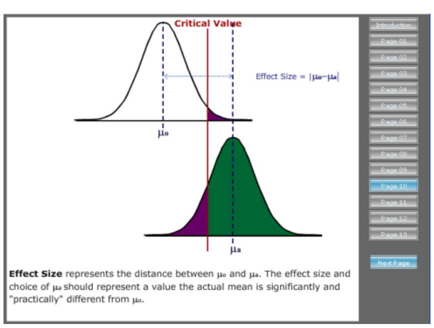
Central Limit Theorem (Section 4.3)
Using die rolling with progressively increasing sample sizes, this animation shows the three main properties of the Central Limit Theorem.



Inference Process (Section 5.1)
This animation walks a student through the logic of the statistical inference and is presented just before confidence intervals and hypothesis testing.



Confidence Intervals (Section 5.3.1)
This animation compares hypothesis testing to an unusual method of playing darts and compares it to a practical example from the 2008 presidential election.



Statistical Power in Hypothesis Testing (Section 6.7)
This animation explains power, Type I and Type II error conceptually, and demonstrates the effect of changing model assumptions.

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

Math 10

Part 1

Data and Descriptive Statistics

1

Introduction

- Green Sheet – Homework 0
- Projects
- Computer Lab – S44
  - Minitab
- Website
  - http://nebula2.deanza.edu/~mo
- Tutor Lab - S43
  - Drop in or assigned tutors – get form from lab.
  - Group Tutoring
- Other Questions

2

Descriptive Statistics

- Organizing, summarizing and displaying data
  - Graphs
  - Charts
  - Measure of Center
  - Measures of Spread
  - Measures of Relative Standing

3

Problem Solving

- The Role of Probability
- Modeling
- Simulation
- Verification

4

Inferential Statistics

- Population – the set of all measurements of interest to the sample collector
- Sample – a subset of measurements selected from the population
- Inference – A conclusion about the population based on the sample
- Reliability – Measure the strength of the Inference

5

Raw Data – Apple

Monthly Adjusted Stock Price: 12/1999 to 12/2014

113.91	66.32	79.01	47.15	28.51	15.38	16.51	8.48	4.36	1.4	1.2	1.49
118.93	67.77	78.16	47.78	27.04	22.93	16.39	9.27	4.54	1.53	1.58	1.23
107.53	62.53	79	45.9	25.5	21.5	13.5	10.22	3.54	1.43	1.64	1.46
100.32	54.8	81.11	43.64	25.07	22.65	12.57	9.73	2.62	1.29	1.6	1.01
102.06	62.15	73.38	42.09	22.76	25.53	11.45	9.17	2.33	1.21	1.47	1.12
94.72	60.78	61.75	40.72	22.1	23.53	11.6	7.79	2.19	0.96	1.67	1.32
92.07	60.77	54.79	38.39	19.27	19.41	11.48	7.25	2.2	0.96	1.48	1.74
89.59	60.6	51.71	32.89	18.37	16.91	12.4	6.34	1.9	1.02	1.44	4.12
83.06	62.17	54.76	34.8	17.02	18.31	10.97	5.77	1.74	0.97	1.19	3.44
75.55	72.63	51.59	34.03	14.22	26.8	10.41	4.98	1.83	0.97	1.05	3.54
74.07	79.88	52.06	34.75	12.08	24.65	9.18	5.38	1.62	1.05	1.25	2.84
70.04	80.88	52.83	35.32	12.19	25.7	9.19	4.88	1.53	1.09	1.27	4.2
78.49	90.63	45.41	31.79	11.55	20.76	7.75	5.64	1.45	0.98	1.57	4.59
77.8	90.38	47.06	27.68	12.54	18.73	8.09	6.07	1.41	1	1.35	3.88
72.71	82.63	47.37	25.98	14.56	17.82	9.52	5.2	1.55	1.03	1.72	3.51

6

Apple – Adjusted Stock Price 15 Years

7

Crime Rate

- In the last 18 years, has violent crime:
  - Increased?
  - Stayed about the Same?
  - Decreased?

8

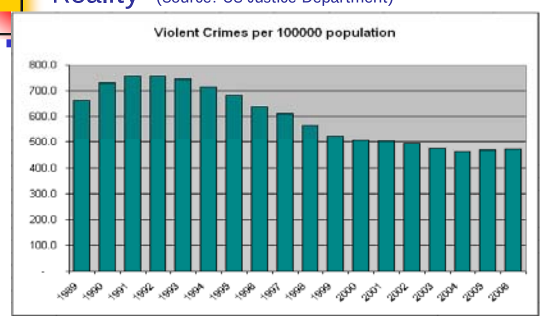
Perception – Gallup Poll

"Is there more crime in the U.S. than there was a year ago, or less?"

	More	Less	Same (vol.)	Unsure
	%	%	%	%
10/13-16/05	67	21	9	3
10/11-14/04	53	28	14	5
10/03	60	25	11	4
10/02	62	21	11	6
10/01	41	43	10	6
8-9/00	47	41	7	5
10/98	52	35	8	5
8/97	64	25	6	5
7/96	71	15	8	6
10/93	87	4	5	4
3/92	89	3	4	4
9/90	84	3	7	6
6/89	84	5	5	6

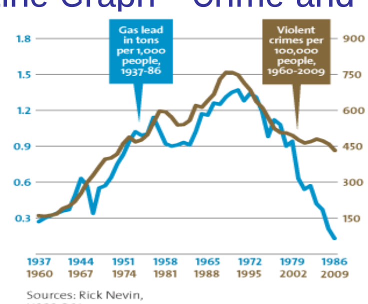
9

Reality (Source: US Justice Department)



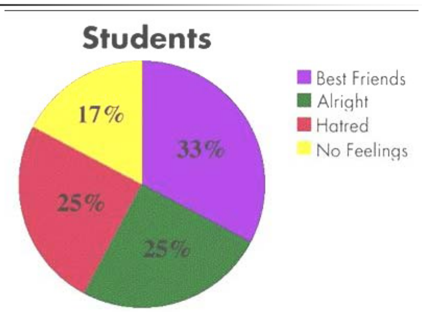
10

Line Graph - Crime and Lead



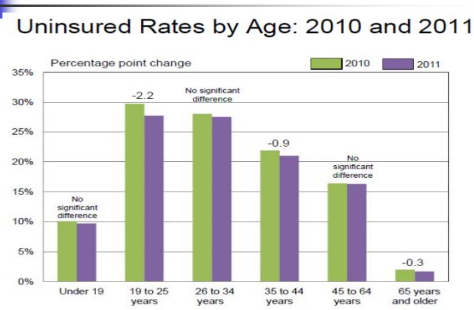
11

Pie Chart - What do you think of your College roommate?



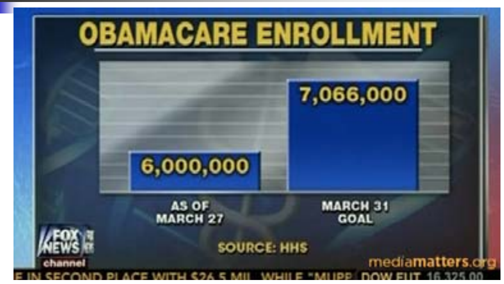
12

Bar Chart - Health Care



13

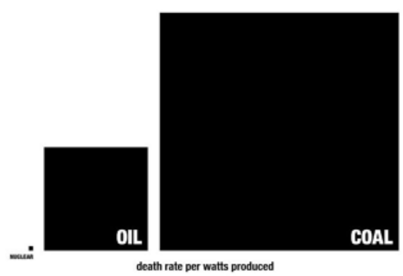
Distorting the truth with Statistics



14

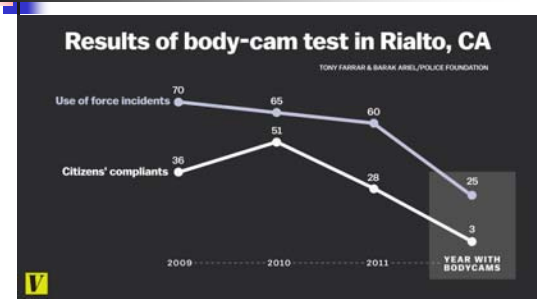
Nuclear, Oil and Coal Energy Deaths per terawatt-hour produced

source: thebigfuture.com



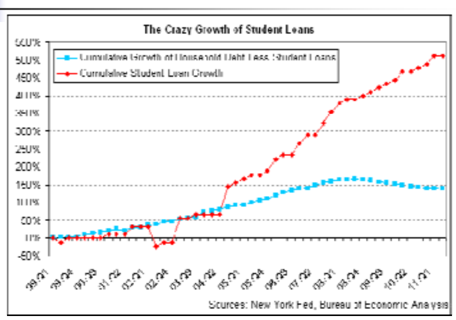
15

Should Police wear Body Cameras?



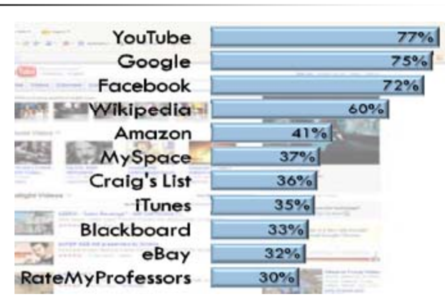
16

Increase in Debt since 1999



17

Most Popular Websites for College Students in 2007



18

Decline of MySpace

19

RATE MY PROFESSORS

De Anza College

Professor's Name	Department	Total Ratings	Overall Quality	Ease	Hot?
	Mandarin	3	4.3	2.0	
	Mandarin	8	1.6	1.6	
	Marketing	1	5.0	5.0	
	Mathematics	66	4.7	4.0	
	Mathematics	73	1.4	1.7	
	Mathematics	15	2.7	2.6	
	Mathematics	41	1.6	2.1	

20

Types of Data

- Qualitative
  - Non-numeric
  - Always categorical
- Quantitative
  - Numeric
  - Categorical numbers are actually qualitative
  - Continuous or discrete

21

Levels of Measurement

- **Nominal:** Names or labels only
  - Example: What city do you live in?
- **Ordinal:** Data can be ranked, but no quantifiable difference.
  - Example: Ratings Excellent, Good, Fair, Poor
- **Interval:** Data can be ranked with quantifiable differences, but no true zero.
  - Example: Temperature
- **Ratio:** Data can be ranked with quantifiable differences and there is a true zero.
  - Example: Age

22

Examples of Data

- Distance from De Anza College
- Number of Grandparents still alive
- Eye Color
- Amount you spend on food each week.
- Number of Facebook "Friends"
- Zip Code
- City you live in.
- Year of Birth
- How to prepare Steak? (rare, medium, well-done)
- Do you own an SUV?

23

Data Collection

- **Personal** – individual interviews
- **Phone** – voice and automated
- **Impersonal Survey** – Internet or Mail
- **Direct Observation** – measurements

- **Scientific Studies** – control for lurking variables
- **Observational Studies** – difficult to establish a cause and effect relationship.

24

Sampling

- Random Sampling
  - Each member of the population has the same chance of being sampled.
- Systematic Sampling
  - The sample is selected by taking every k_{th} member of the population.
- Stratified Sampling
  - The population is broken into more homogenous subgroups (strata) and a random sample is taken from each strata.
- Cluster Sampling
  - Divide population into smaller clusters, randomly select some clusters and sample each member of the selected clusters.
- Convenience Sampling
  - Self selected and non-scientific methods which are prone to extreme bias.

25

Graphical Methods

- Stem and Leaf Chart
- Grouped data
- Pie Chart
- Histogram
- Ogive
- Grouping data
- Example

26

Daily Minutes spent on the Internet by 30 students

102	104	85	67	101
71	116	107	99	82
103	97	105	103	95
105	99	86	87	100
109	108	118	87	125
124	112	122	78	92

27

Stem and Leaf Graph

```
 6 7
 7 18
 8 25677
 9 25799
10 01233455789
11 268
12 245
```

28

Back-to-back Example

- Passenger loading times for two airlines

11, 14, 16, 17, 19, 21, 22, 23, 24, 24, 24, 26, 31, 32, 38, 39	8, 11, 13, 14, 15, 16, 16, 18, 19, 19, 21, 21, 22, 24, 26, 31

29

Back to Back Example

	0	
	0	8
1 4	1	1 3 4
6 7 9	1	5 6 6 8 9 9
1 2 3 4 4 4	2	1 1 2 4
6	2	6
1 2	3	1
8 9	3	

30

Grouping Data

- Choose the number of groups
  - between 5 and 10 is best
- Interval Width = (Range+1)/(Number of Groups)
  - Round **up** to a convenient value
- Start with lowest value and create the groups.
- Example – for 5 categories
  Interval Width = (58+1)/5 = 12 (rounded up)

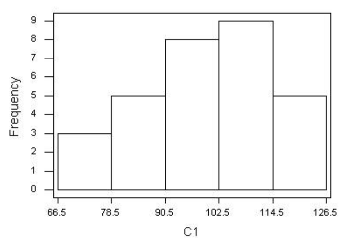
31

Grouping Data

Class Interval	Frequency	Relative Frequency	Cumulative Relative Frequency
66.5-78.5	3	.100	.100
78.5-90.5	5	.167	.267
90.5-102.5	8	.266	.533
102.5-114.5	9	.300	.833
114.5-126.5	5	.167	1.000
Total	30	1.000	

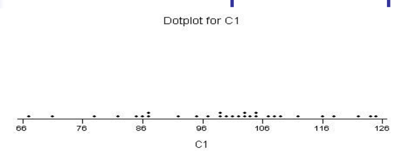
32

Histogram – Graph of Frequency or Relative Frequency



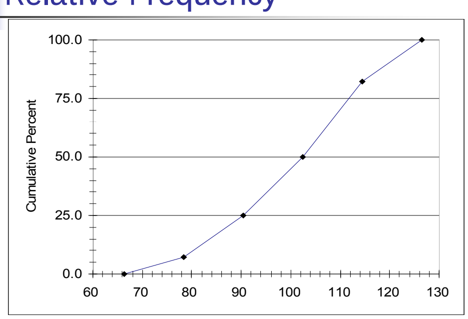
33

Dot Plot – Graph of Frequency



34

Ogive – Graph of Cumulative Relative Frequency



35

Measures of Central Tendency

- Mean
  - Arithmetic Average $\bar{X} = \frac{\sum X_i}{n}$
- Median
  - "Middle" Value after ranking data
  - Not affected by "outliers"
- Mode
  - Most Occurring Value
  - Useful for non-numeric data

36

Example

2 2 5 9 12

Circle the Average

a) 2

b) 5

c) 6

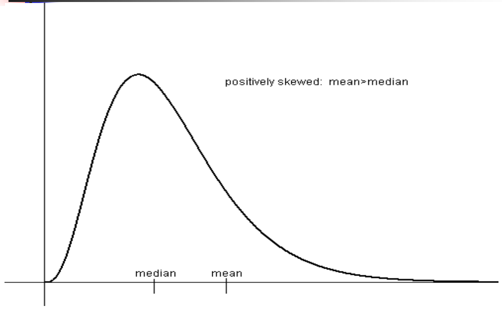
37

Example – 5 Recent Home Sales

- $500,000
- $600,000
- $600,000
- $700,000
- $2,600,000

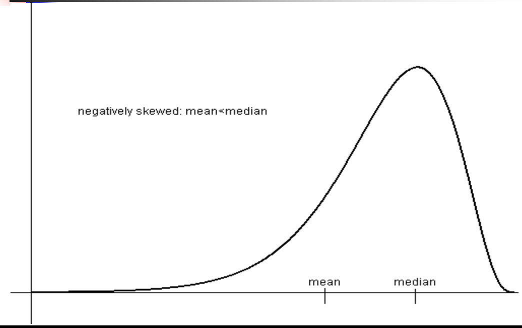
38

Positively Skewed Data Set Mean > Median



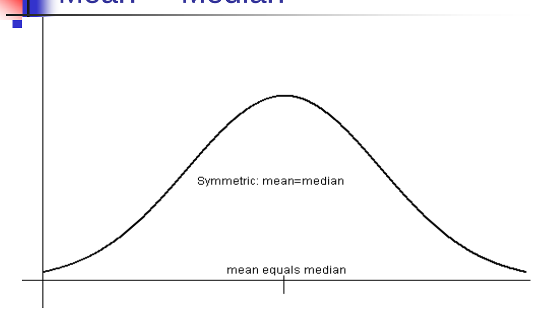
39

Negatively Skewed Data Set Mean < Median



40

Symmetric Data Set Mean = Median



41

Measures of Variability

- Range
- Variance
- Standard Deviation
- Interquartile Range (percentiles)

42

Range

Max(Xi) –Min(Xi)

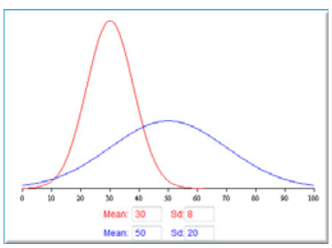
125 – 67 = 58

Sample Variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$s^2 = \frac{\sum x_i^2 - (\sum x_i)^2 / n}{n-1}$$

Sample Standard Deviation



$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Variance and Standard Deviation

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
2	-4	16
2	-4	16
5	-1	1
9	3	9
12	6	36
30	**0**	**78**

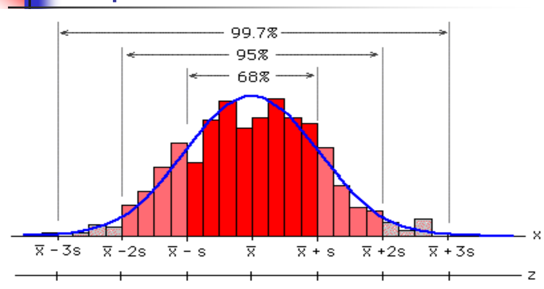
$$s^2 = \frac{78}{4} = 19.5$$

$$s = \sqrt{19.5} \approx 4.42$$

Interpreting the Standard Deviation

- Chebyshev's Rule
  - At least 100 x (1-(1/k)^2)% of any data set must be within k standard deviations of the mean.
- Empirical Rule (68-95-99 rule)
  - Bell shaped data
  - 68% within 1 standard deviation of mean
  - 95% within 2 standard deviations of mean
  - 99.7% within 3 standard deviations of mean

Empirical Rule



Measures of Relative Standing

- Z-score
- Percentile
- Quartiles
- Box Plots

49

Z-score

- The number of Standard Deviations from the Mean
- Z>0, X_i is greater than mean
- Z<0, X_i is less than mean

$$Z = \frac{X_i - \bar{X}}{s}$$

50

Percentile Rank

Formula for ungrouped data

- The location is (n+1)p (interpolated or rounded)
- n= sample size
- p = percentile

51

Quartiles

- 25^{th} percentile is 1^{st} quartile
- 50^{th} percentile is median
- 75^{th} percentile is 3^{rd} quartile
- 75^{th} percentile – 25^{th} percentile is called the Interquartile Range which represents the "middle 50%"

52

IQR example

n+1=31

.25 x 31 = 7.75   location 8 = **87** ← 1^{st} Quartile

.75 x 31 = 23.25   location 23 = **108** ← 3rd Quartile

Interquartile Range (IQR) =108 – 87 = **21**

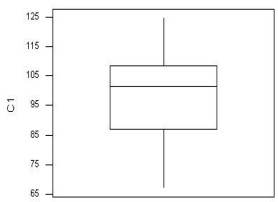
53

Box Plots

- A box plot is a graphical display, based on quartiles, that helps to picture a set of data.
- Five pieces of data are needed to construct a box plot:
  - Minimum Value
  - First Quartile
  - Median
  - Third Quartile
  - Maximum Value.

54

Box Plot



55

Outliers

- An outlier is data point that is far removed from the other entries in the data set.
- Outliers could be
  - Mistakes made in recording data
  - Data that don't belong in population
  - True rare events

56

Outliers have a dramatic effect on some statistics

- Example quarterly home sales for 10 realtors:

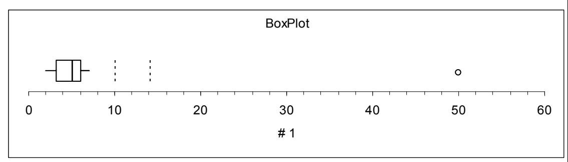
2 2 3 4 5 5 6 6 7 50

	with outlier	without outlier
Mean	9.00	4.44
Median	5.00	5.00
Std Dev	14.51	1.81
IQR	3.00	3.50

57

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, 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.



58

Using Z-score to detect outliers

- Calculate the mean and standard deviation without the suspected outlier.
- Calculate the Z-score of the suspected outlier.
- If the Z-score is more than 3 or less than -3, that data point is a probable outlier.

$$Z = \frac{50 - 4.4}{1.81} = 25.2$$

59

Outliers – what to do

- Remove or not remove, there is no clear answer.
- For some populations, outliers don't dramatically change the overall statistical analysis. Example: the tallest person in the world will not dramatically change the mean height of 10000 people.
- However, for some populations, a single outlier will have a dramatic effect on statistical analysis (called "**Black Swan**" by Nicholas Taleb) and inferential statistics may be invalid in analyzing these populations. Example: the richest person in the world will dramatically change the mean wealth of 10000 people.

60

Bivariate Data

- Ordered numeric pairs (X,Y)
- Both values are numeric
- Paired by a common characteristic
- Graph as Scatterplot

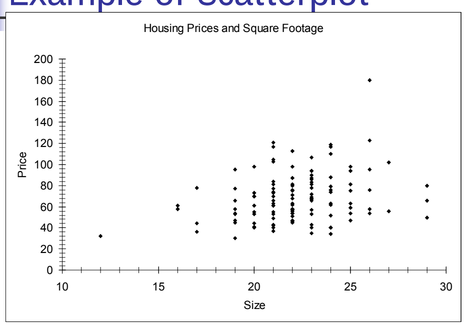
61

Example of Bivariate Data

- Housing Data
  - X = Square Footage
  - Y = Price

62

Example of Scatterplot



63

Another Example

64

Correlation Analysis

- Correlation Analysis: A group of statistical techniques used to measure the strength of the relationship (correlation) between two variables.
- Scatter Diagram: A chart that portrays the relationship between the two variables of interest.
- Dependent Variable: The variable that is being predicted or estimated. “Effect”
- Independent Variable: The variable that provides the basis for estimation. It is the predictor variable. “Cause?” (Maybe!)

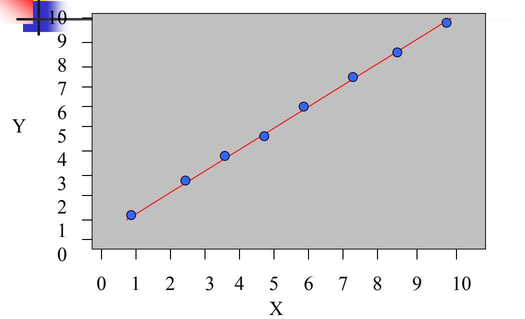
65

The Coefficient of Correlation, r

- The Coefficient of Correlation (r) is a measure of the **strength** of the relationship between two variables.
  - It requires interval or ratio-scaled data (variables).
  - It can range from -1 to 1.
  - Values of -1 or 1 indicate perfect and strong correlation.
  - Values close to 0 indicate weak correlation.
  - Negative values indicate an inverse relationship and positive values indicate a direct relationship.

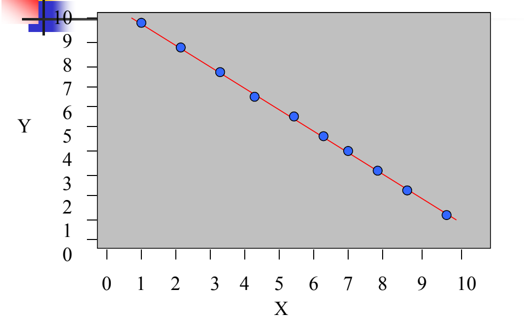
66

Perfect Positive Correlation



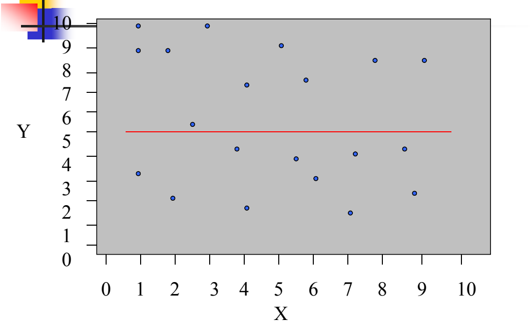
67

Perfect Negative Correlation



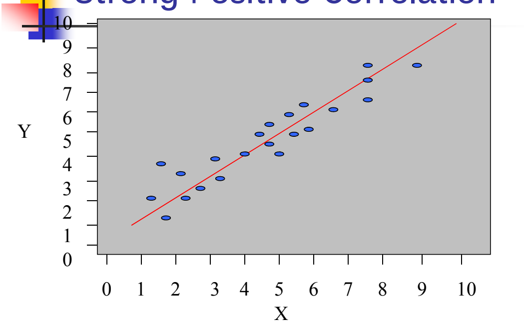
68

Zero Correlation



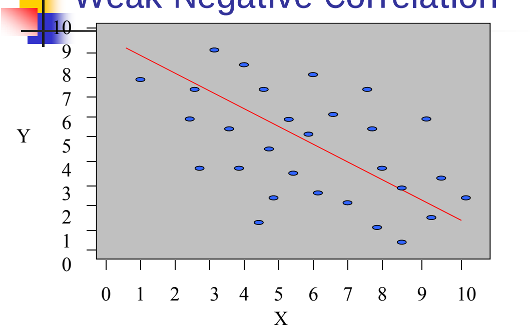
69

Strong Positive Correlation



70

Weak Negative Correlation



71

Causation

- Correlation does not necessarily imply causation.
- There are 4 possibilities if X and Y are correlated:
  1. X causes Y
  2. Y causes X
  3. X and Y are caused by something else.
  4. Confounding - The effect of X and Y are hopelessly mixed up with other variables.

72

Causation - Examples

- City with more police per capita have more crime per capita.
- As Ice cream sales go up, shark attacks go up.
- People with a cold who take a cough medicine feel better after some rest.

73

Formula for correlation coefficient r

$$r = \frac{SSXY}{\sqrt{SSX \cdot SSY}}$$

$$SSX = \Sigma X^2 - \tfrac{1}{n}(\Sigma X)^2$$

$$SSY = \Sigma Y^2 - \tfrac{1}{n}(\Sigma Y)^2$$

$$SSXY = \Sigma XY - \tfrac{1}{n}(\Sigma X \cdot \Sigma Y)$$

74

Example

- X = Average Annual Rainfall (Inches)
- Y = Average Sale of Sunglasses/1000
- Make a Scatter Diagram
- Find the correlation coefficient

X	10	15	20	30	40
Y	40	35	25	25	15

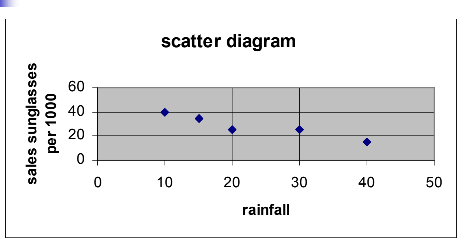
75

Example *continued*

- Make a Scatter Diagram
- Find the correlation coefficient

76

Example *continued*



77

Example *continued*

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

- SSX = 3225 - 115^2/5 = 580
- SSY = 4300 - 140^2/5 = 380
- SSXY= 2775 - (115)(140)/5 = -445

78

Example *continued*

$$r = \frac{SSXY}{\sqrt{SSX \cdot SSY}}$$

$$r = \frac{-445}{\sqrt{580 \cdot 330}} = -0.9479$$

- Strong negative correlation

79

Math 10

Part 2

Probability

1

Probability

- Classical probability
  - Based on mathematical formulas
- Empirical probability
  - Based on the relative frequencies of historical data.
- Subjective probability
  - "one-shot" educated guess.

2

Examples of Probability

- What is the probability of rolling a four on a 6-sided die?
- What percentage of De Anza students live in Cupertino?
- What is the chance that the Golden State Warriors will repeat as NBA champions in 2016?

3

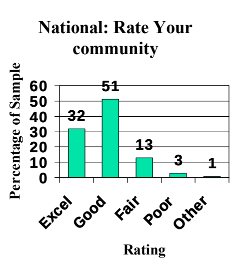
Classical Probability

- Event
  - A result of an experiment
- Outcome
  - A result of the experiment that cannot be broken down into smaller events
- Sample Space
  - The set of all possible outcomes
- Probability Event Occurs
  - # of elements in Event / # Elements in Sample Space
- Example – flip two coins, find the probability of exactly 1 head.
  - {HH, HT, TH, TT}
  - P(1 head) = 2/4 = .5

4

Empirical Probability

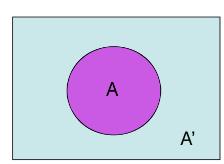
- Historical Data
- Relative Frequencies
- Example: What is the chance someone rates their community as good or better?
- **0.51 + 0.32 = 0.83**



5

Rule of Complement

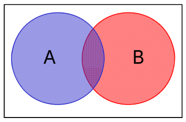
- Complement of an event
- The event does not occur
- A' is the complement of A
- P(A) + P(A') =1
- P(A) = 1 – P(A')



6

Additive Rule

- The UNION of two events A and B is that either A or B occur (or both). (All colored parts)
- The INTERSECTION of two events A and B is that both A and B will occur. (Purple Part only)
- Additive Rule:
  P(A or B) = P(A) + P(B) – P(A and B)



7

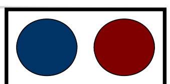
Example

- In a group of students, 40% are taking Math, 20% are taking History.
- 10% of students are taking both Math and History.
- Find the Probability of a Student taking either Math or History or both.
- P(M or H) = 40% + 20% - 10% = 50%

8

Mutually Exclusive

- Mutually Exclusive
- Both cannot occur
- If A and B are mutually exclusive, then
  - P(A or B) = P(A) + P(B)
- Example roll a die
  - A: Roll 2 or less B: Roll 5 or more
  - P(A)=2/6 P(B)=2/6
  - P(A or B) = P(A) + P(B) = 4/6



9

Conditional Probability

- The probability of an event occuring GIVEN another event has already occurred.
- P(A|B) = P(A and B) / P(B)
- Example: Of all cell phone users in the US, 15% have a smart phone with AT&T. 25% of all cell phone users use AT&T. Given a selected cell phone user has AT&T, find the probability the user also has a smart phone.
- A=AT&T subscriber B=Smart Phone User
- P(A and B) = .15 P(A)=.25
- P(B|A) = .15/.25 = .60

10

Contingency Tables

- Two data items can be displayed in a contingency table.
- Example: auto accident during year and DUI of driver.

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

11

Contingency Tables

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

Given the Driver is DUI, find the Probability of an Accident.

A=Accident D=DUI

P(A and D) = .07 P(D) = .2

P(A|D) = .07/.2 = .35

12

Multiplicative Rule

- P(A and B) = P(A) x P(B|A)
- P(A and B) = P(B) x P(A|B)
- Example: A box contains 4 green balls and 3 red balls. Two balls are drawn. Find the probability of choosing two red balls.
- A=Red Ball on 1st draw   B=Red Ball on 2nd Draw
- P(A)=3/7   P(B|A)=2/6
- P(A and B) = (3/7)(2/6) = 1/7

13

Independence

- If A is not dependent on B, then they are INDEPENDENT events, and the following statements are true:
  - P(A|B)=P(A)
  - P(B|A)=P(B)
  - P(A and B) = P(A) x P(B)

14

Example

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

A: Accident   D:DUI Driver

P(A) = .10   P(A|D) = .35 (70/200)

Therefore A and D are **DEPENDENT** events as P(A) < P(A|D)

15

Example

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

A: Accident   U:US Car

P(A) = .10   P(A|U) = .10 (60/600)

Therefore A and U are **INDEPENDENT** events as P(A) = P(A|U)

Also P(A and U) = P(A)xP(U) = (.1)(.6) = .06

16

Random Sample

- A random sample is where each member of the population has an equally likely chance of being chosen, and each member of the sample is INDEPENDENT of all other sampled data.

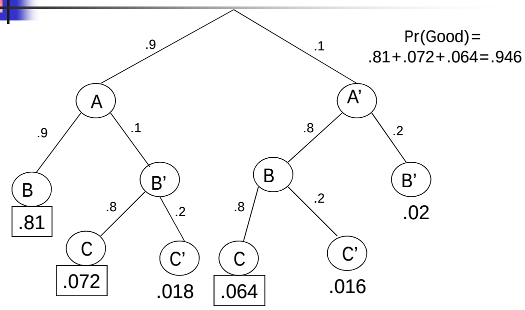
17

Tree Diagram method

- Alternative Method of showing probability
- Example: Flip Three Coins
- Example: A Circuit has three switches. If at least two of the switches function, the Circuit will succeed. Each switch has a 10% failure rate if all are operating, and a 20% failure rate if one switch has already failed. Find the probability the circuit will succeed.

18

Circuit Problem



19

Switching the Conditionality

- Often there are questions where you desire to change the conditionality from one variable to the other variable
- First construct a tree diagram.
- Second, move the information to a Contingency Table
- From the Contingency table it is easy to calculate all conditional probabilities.

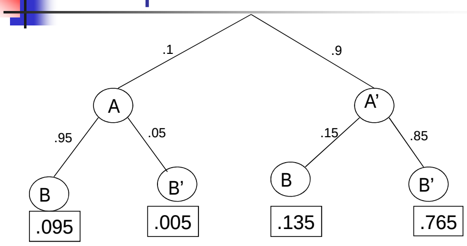
20

Example

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

21

Example



A=Prisoner is HIV+
B=Test is Positive for HIV

22

Example

	HIV+ A	HIV- A'	Total
Test+ B	.095	.135	.230
Test- B'	.005	.765	.770
Total	.100	.900	1.000

$$P(A|B) = \frac{.095}{.230} \approx .413$$

23

Math 10

Part 3
Discrete Random Variables

1

Random Variable

- The value of the variable depends on an experiment, observation or measurement.
- The result is not known in advance.
- For the purposes of this class, the variable will be numeric.

2

Random Variables

- Discrete – Data that you Count
  - Defects on an assembly line
  - Reported Sick days
  - RM 7.0 earthquakes on San Andreas Fault
- Continuous – Data that you Measure
  - Temperature
  - Height
  - Time

3

Discrete Random Variable

- List Sample Space
- Assign probabilities P(x) to each event x
- Use "relative frequencies"
- Must follow two rules
  - $P(x) \geq 0$
  - $\Sigma P(x) = 1$
- P(x) is called a Probability Distribution Function or pdf for short.

4

Probability Distribution Example

- Students are asked 4 questions and the number of correct answers are determined.
- Assign probabilities to each event.

x	P(x)
0	.1
1	.1
2	.2
3	.4
4	

5

Probability Distribution Example

- Students are asked 4 questions and the number of correct answers are determined.
- Assign probabilities to each event.

x	P(x)
0	.1
1	.1
2	.2
3	.4
4	.2

6

Mean and Variance of Discrete Random Variables

- Population mean μ, is the expected value of x

  $\mu = \Sigma[(x) P(x)]$

- Population variance σ, is the expected value of $(x-\mu)^2$

  $\sigma^2 = \Sigma[(x-\mu)^2 P(x)]$

7

Example of Mean and Variance

x	P(x)	xP(x)	$(x-\mu)^2P(x)$
0	0.1	0.0	.625
1	0.1	0.1	.225
2	0.2	0.4	.050
3	0.4	1.2	.100
4	0.2	0.8	.450
Total	**1.0**	**2.5=μ**	**1.450=σ^2**

8

Bernoulli Distribution

- Experiment is one trial
- 2 possible outcomes (Success,Failure)
- p=probability of success
- q=probability of failure
- X=number of successes (1 or 0)
- Also known as Indicator Variable

9

Mean and Variance of Bernoulli

x	P(x)	xP(x)	$(x-\mu)^2P(x)$
0	(1-p)	0.0	$p^2(1-p)$
1	p	p	$p(1-p)^2$
Total	**1.0**	**p=μ**	**p(1-p)=σ^2**

- $\mu = p$
- $\sigma^2 = p(1-p) = pq$

10

Binomial Distribution

- n identical trials
- Two possible outcomes (success/failure)
- Probability of success in a single trial is p
- Trials are mutually independent
- X is the number of successes

- Note: X is a sum of n independent identically distributed Bernoulli distributions

11

Binomial Distribution

- n independent Bernoulli trials
- Mean and Variance of Binomial Distribution is just sample size times mean and variance of Bernoulli Distribution

$$p(x) = {}_nC_x p^x (1-p)^{n-x}$$

$$\mu = E(X) = np$$

$$\sigma^2 = Var(X) = np(1-p)$$

12

Binomial Examples

- The number of defective parts in a fixed sample.
- The number of adults in a sample who support the war in Iraq.
- The number of correct answers if you guess on a multiple choice test.

13

Binomial Example

- 90% of intake valves manufactured are good (not defective). A sample of 10 is selected.
- Find the probability of exactly 8 good valves being chosen.
- Find the probability of 9 or more good valves being chosen.
- Find the probability of 8 or less good valves being chosen.

14

Using Technology

X	p(X)	cumulative probability
0	0.00000	0.00000
1	0.00000	0.00000
2	0.00000	0.00000
3	0.00001	0.00001
4	0.00014	0.00015
5	0.00149	0.00163
6	0.01116	0.01280
7	0.05740	0.07019
8	0.19371	0.26390
9	0.38742	0.65132
10	0.34868	1.00000
	1.00000	

9.000 expected value
0.900 variance
0.949 standard deviation

Use Minitab or Excel to make a table of Binomial Probabilities.

$P(X=8) = .19371$

$P(X\leq 8) = .26390$

$P(X\geq 9) = 1 - P(X\leq 8) = .73610$

15

Poisson Distribution

- Occurrences per time period (rate)
- Rate (μ) is constant
- No limit on occurrences over time period

$$P(x) = \frac{e^{-\mu}\mu^x}{x!}$$

$$\mu = \mu$$

$$\sigma = \sqrt{\mu}$$

16

Examples of Poisson

- Text messages in the next hour
- Earthquakes on a fault
- Customers at a restaurant
- Flaws in sheet metal produced
- Lotto winners

Note: A binomial distribution with a large n and small p is approximately Poisson with $\mu \approx np$.

17

Poisson Example

- Earthquakes of Richter magnitude 3 or greater occur on a certain fault at a rate of twice every year.
- Find the probability of at least one earthquake of RM 3 or greater in the next year.

$$P(X>0) = 1 - P(0)$$

$$= 1 - \frac{e^{-2}2^0}{0!}$$

$$= 1 - e^{-2} \approx .8647$$

18

Poisson Example (cont)

- Earthquakes of Richter magnitude 3 or greater occur on a certain fault at a rate of twice every year.
- Find the probability of exactly 6 earthquakes of RM 3 or greater in the next **2** years.

$$\mu = 2(2) = 4$$

$$P(X = 6) = \frac{e^{-4}4^{6}}{6!} \approx .1042$$

19

Math 10

Part 4 Slides
Continuous Random Variables and the Central Limit Theorem

1

Continuous Distributions

- "Uncountable" Number of possibilities
- Probability of a point makes no sense
- Probability is measured over intervals
- Comparable to Relative Frequency Histogram – Find Area under curve.

2

Discrete vs Continuous

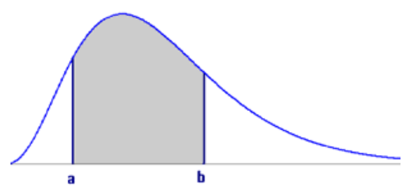
- Countable
- Discrete Points
- p(x) is probability distribution function
- $p(x) \geq 0$
- $\Sigma p(x) = 1$

- Uncountable
- Continuous Intervals
- f(x) is probability density function
- $f(x) \geq 0$
- Total Area under curve =1

3

Continuous Random Variable

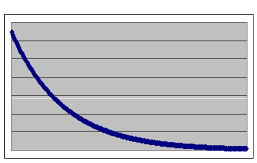
- f(x) is a density function
- P(X<x) is a distribution function.
- P(a<X<b) = area under function between a and b



4

Exponential distribution

- Waiting time
- "Memoryless"
- $f(x) = (1/\mu)e^{-(1/\mu)x}$
- $P(x>a) = e^{-(a/\mu)}$
- $\mu=\mu \quad \sigma^2=\mu^2$
- $P(x>a+b|x>b) = e^{-(a/\mu)}$



5

Examples of Exponential Distributiuon

- Time until...
- a circuit will fail
- the next RM 7 Earthquake
- the next customer calls
- An oil refinery accident
- you buy a winning lotto ticket

6

Relationship between Poisson and Exponential Distributions

- If occurrences follow a Poisson Process with mean = μ, then the waiting time for the next occurrence has Exponential distribution with mean = $1/\mu$.
- Example: If accidents occur at a plant at a constant rate of 3 per month, then the expected waiting time for the next accident is 1/3 month.

7

Exponential Example

The life of a digital display of a calculator has exponential distribution with μ=500hours.

(a) Find the chance the display will last at least 600 hours.

$P(x>600) = e^{-600/500} = e^{-1.2} = .3012$

(b) Assuming it has already lasted 500 hours, find the chance the display will last an additional 600 hours.

$P(x>1100|x>500) = P(x>600) = .3012$

8

Exponential Example

The life of a digital display of a calculator has exponential distribution with μ=500 hours.

(a) Find the median of the distribution

$P(x>med) = e^{-(med)/500} = 0.5$

med = -ln(.5)x500=346.57

p^{th} Percentile = $-\ln(1-p)\mu$

9

Uniform Distribution

- Rectangular distribution
- Example: Random number generator

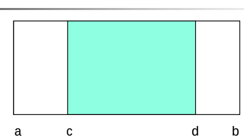
$$f(x) = \frac{1}{b-a} \qquad a \le x \le b$$

$$\mu = E(X) = \frac{b+a}{2}$$

$$\sigma^2 = Var(X) = \frac{(b-a)^2}{12}$$

10

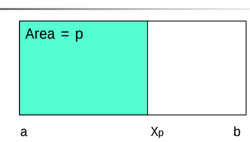
Uniform Distribution - Probability



$$P(c < X < d) = \frac{d-c}{b-a}$$

11

Uniform Distribution - Percentile



Formula to find the pth percentile X_p:

$$X_p = a + p(b-a)$$

12

Uniform Example 1

- Find mean, variance, P(X<3) and 70th percentile for a uniform distribution from 1 to 11.

$$\mu = \frac{1+11}{2} = 6 \quad \sigma^2 = \frac{(11-1)^2}{12} = 8.33$$

$$P(X<3) = \frac{3-1}{11-1} = 0.3$$

$$X_{70} = 1 + 0.7(11-1) = 8$$

13

Uniform Example 2

- A tea lover orders 1000 grams of Tie Guan Yin loose leaf when his supply gets to 50 grams.
- The amount of tea currently in stock follows a uniform random variable.
- Determine this model
- Find the mean and variance
- Find the probability of at least 700 grams in stock.
- Find the 80th percentile

14

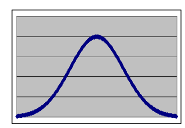
Uniform Example 3

- A bus arrives at a stop every 20 minutes.
  - Find the probability of waiting more than 15 minutes for the bus after arriving randomly at the bus stop.
  - If you have already waited 5 minutes, find the probability of waiting an additional 10 minutes or more. (Hint: recalculate parameters a and b)

15

Normal Distribution

- The normal curve is *bell-shaped*
- The mean, median, and mode of the distribution are equal and located at the peak.
- The normal distribution is symmetrical about its mean. Half the area under the curve is above the peak, and the other half is below it.
- The normal probability distribution is asymptotic - the curve gets closer and closer to the x-axis but never actually touches it.



$$f(x) = \frac{e^{-\frac{1}{2\sigma}(x-\mu)^2}}{\sigma\sqrt{2\pi}}$$

16

The Standard Normal Probability Distribution

- A normal distribution with a mean of 0 and a standard deviation of 1 is called the standard normal distribution.
- Z value: The distance between a selected value, designated x, and the population mean μ, divided by the population standard deviation, σ

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

17

Areas Under the Normal Curve – Empirical Rule

- About 68 percent of the area under the normal curve is within one standard deviation of the mean. $\mu \pm 1\sigma$
- About 95 percent is within two standard deviations of the mean $\mu \pm 2\sigma$
- 99.7 percent is within three standard deviations of the mean. $\mu \pm 3\sigma$

18

EXAMPLE

- The daily water usage per person in a town is normally distributed with a mean of 20 gallons and a standard deviation of 5 gallons.
- About 68% of the daily water usage per person in New Providence lies between what two values?
- $\mu \pm 1\sigma = 20 \pm 1(5)$. That is, about 68% of the daily water usage will lie between 15 and 25 gallons.

19

Normal Distribution – probability problem procedure

- Given: Interval in terms of X
- Convert to Z by $Z = \frac{X-\mu}{\sigma}$
- Look up probability in table.

20

EXAMPLE

- The daily water usage per person in a town is normally distributed with a mean of 20 gallons and a standard deviation of 5 gallons.
- What is the probability that a person from the town selected at random will use less than 18 gallons per day?
- The associated Z value is Z=(18-20)/5=0.
- Thus, P(X<18)=P(Z<-0.40)=.3446

21

EXAMPLE *continued*

- The daily water usage per person in a town is normally distributed with a mean of 20 gallons and a standard deviation of 5 gallons.
- What proportion of the people uses between 18 and 24 gallons?
- The Z value associated with x=18 is Z=-0.40 and with X=24, Z=(24-20)/5=0.80.
- Thus, P(18<X<24)=P(-0.40<Z<0.80)
  =.7881-.3446=**.4435**

22

EXAMPLE *continued*

- The daily water usage per person in a town is normally distributed with a mean of 20 gallons and a standard deviation of 5 gallons.
- What percentage of the population uses more than 26.2 gallons?
- The Z value associated with X=26.2, Z=(26.2-20)/5=1.24.
- Thus P(X>26.2)=P(Z>1.24)
  =1-.8925=**.1075**

23

Normal Distribution – percentile problem procedure

- Given: probability or percentile desired.
- Look up Z value in table that corresponds to probability.
- Convert to X by the formula:

$$X = \mu + Z\sigma$$

24

EXAMPLE

- The daily water usage per person in a town is normally distributed with a mean of 20 gallons and a standard deviation of 5 gallons. A special tax is going to be charged on the top 5% of water users.
- Find the value of daily water usage that generates the special tax
- The Z value associated with 95th percentile =1.645
- X=20 + 5(1.645) = 28.2 gallons per day

25

EXAMPLE

- Professor Kurv has determined that the final averages in his statistics course is normally distributed with a mean of 77.1 and a standard deviation of 11.2.
- He decides to assign his grades for his current course such that the top 15% of the students receive an A.
- What is the lowest average a student can receive to earn an A?
- The top 15% would be the finding the 85th percentile. Find k such that P(X<k)=.85.
- The corresponding Z value is 1.04. Thus we have X=77.1+(1.04)(11.2), or **X=88.75**

26

EXAMPLE

- The amount of tip the servers in an exclusive restaurant receive per shift is normally distributed with a mean of $80 and a standard deviation of $10.
- Shelli feels she has provided poor service if her total tip for the shift is less than $65.
- What percentage of the time will she feel like she provided poor service?
- Let y be the amount of tip. The Z value associated with X=65 is Z= (65-80)/10= -1.5.
- Thus P(X<65)=P(Z<-1.5)=.0668.

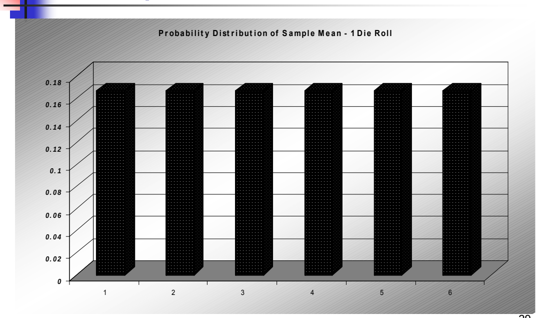
27

Distribution of Sample Mean

- Random Sample: X_1, X_2, X_3, ..., X_n
  - Each X_i is a Random Variable from the same population
  - All X_i's are Mutually Independent
- $\overline{X}$ is a function of Random Variables, so $\overline{X}$ is itself Random Variable.
- In other words, the Sample Mean can change if the values of the Random Sample change.
- What is the Probability Distribution of $\overline{X}$?

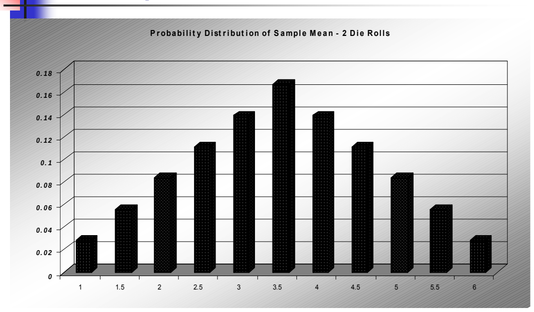
28

Example – Roll 1 Die



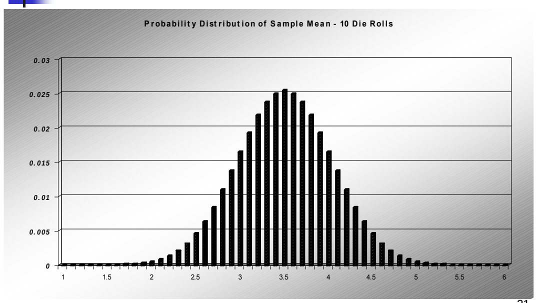
29

Example – Roll 2 Dice



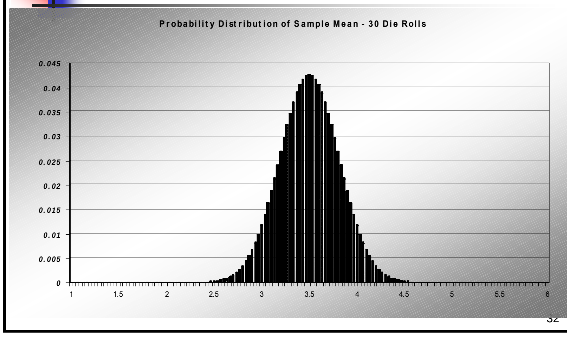
30

Example – Roll 10 Dice



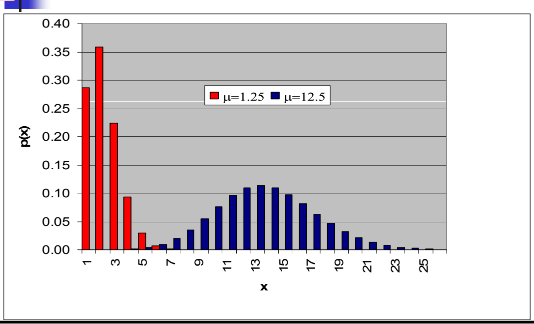
31

Example – Roll 30 Dice



32

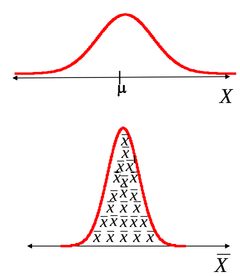
Example - Poisson



Central Limit Theorem – Part 1

- IF a Random Sample of **any size** is taken from a population with a **Normal Distribution** with mean= μ and standard deviation = σ
- THEN the distribution of the sample mean has a Normal Distribution with:

$$\mu_{\overline{X}} = \mu \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

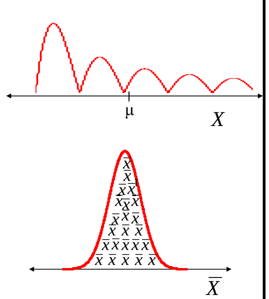


34

Central Limit Theorem – Part 2

- IF a random sample of **sufficiently large size** is taken from a population with **any Distribution** with mean= μ and standard deviation = σ
- THEN the distribution of the sample mean has approximately a Normal Distribution with:

$$\mu_{\overline{X}} = \mu \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$



35

Central Limit Theorem

3 important results for the distribution of $\overline{X}$

- Mean Stays the same

$$\mu_{\overline{X}} = \mu$$

- Standard Deviation Gets Smaller

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- If n is sufficiently large, $\overline{X}$ has a Normal Distribution

36

Example

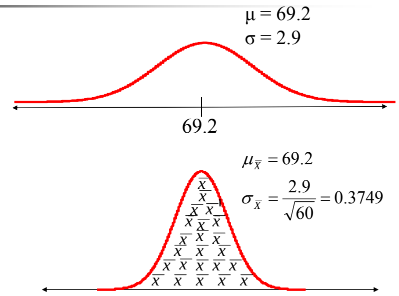
The mean height of American men (ages 20-29) is μ = 69.2 inches. If a random sample of 60 men in this age group is selected, what is the probability the mean height for the sample is greater than 70 inches? Assume σ = 2.9".

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

$$= P(Z > 2.14) = 0.0162$$

37

Example (cont)



38

Example – Central Limit Theorem

The waiting time until receiving a text message follows an exponential distribution with an expected waiting time of 1.5 minutes. Find the probability that the mean waiting time for the 50 text messages exceeds 1.6 minutes.

$\mu = 1.5 \qquad \sigma = 1.5 \qquad n = 40$

Use Normal Distribution (n>30)

$$P(\overline{X} > 1.6) = P\left(Z > \frac{(1.6-1.5)}{1.5/\sqrt{50}}\right) = P(Z > 0.47) = 0.3192$$

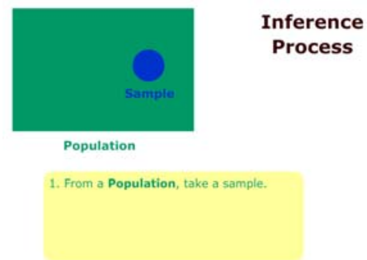
39

Math 10

Part 5 Slides
Confidence Intervals

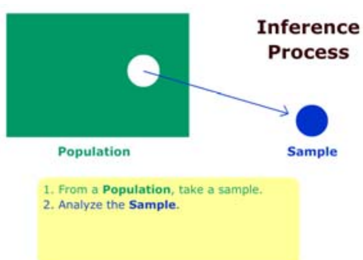
1

Inference Process



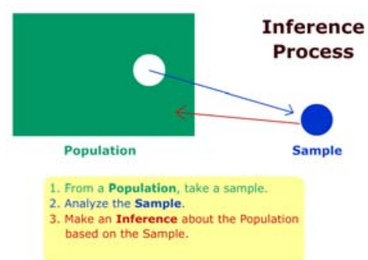
2

Inference Process



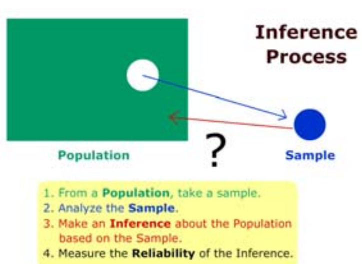
3

Inference Process



4

Inference Process



5

Inferential Statistics

- Population Parameters
  - Mean = μ
  - Standard Deviation = σ
- Sample Statistics
  - Mean = $\overline{X}$
  - Standard Deviation = s

6

Inferential Statistics

- Estimation
  - Using sample data to estimate population parameters.
  - Example: Public opinion polls
- Hypothesis Testing
  - Using sample data to make decisions or claims about population
  - Example: A drug effectively treats a disease

7

Estimation of μ

$\overline{X}$ is an unbiased point estimator of μ

Example: The number of defective items produced by a machine was recorded for five randomly selected hours during a 40-hour work week. The observed number of defectives were 12, 4, 7, 14, and 10. So the sample mean is 9.4.

Thus a point estimate for μ, the hourly mean number of defectives, is 9.4.

8

Confidence Intervals

- An Interval Estimate states the range within which a population parameter "probably" lies.
- The interval within which a population parameter is expected to occur is called a Confidence Interval.
- The distance from the center of the confidence interval to the endpoint is called the **"Margin of Error"**
- The three confidence intervals that are used extensively are the 90%, 95% and 99%.

9

Confidence Intervals

- A 95% confidence interval means that about 95% of the similarly constructed intervals will contain the parameter being estimated, or 95% of the sample means for a specified sample size will lie within 1.96 standard deviations of the hypothesized population mean.
- For the 99% confidence interval, 99% of the sample means for a specified sample size will lie within 2.58 standard deviations of the hypothesized population mean.
- For the 90% confidence interval, 90% of the sample means for a specified sample size will lie within 1.645 standard deviations of the hypothesized population mean.

10

90%, 95% and 99% Confidence Intervals for μ

- The 90%, 95% and 99% confidence intervals for μ are constructed as follows when $n \geq 30$
- 90% CI for the population mean is given by

$$\overline{X} \pm 1.645 \frac{\sigma}{\sqrt{n}}$$

- 95% CI for the population mean is given by

$$\overline{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

- 99% CI for the population mean is given by

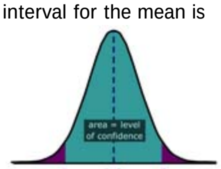
$$\overline{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$$

11

Constructing General Confidence Intervals for μ

- In general, a confidence interval for the mean is computed by:

$$\overline{X} \pm Z \frac{\sigma}{\sqrt{n}}$$



- This can also be thought of as:

Point Estimator ± Margin of Error

12

The nature of Confidence Intervals

- The Population mean μ is fixed.
- The confidence interval is centered around the sample mean which is a Random Variable.
- So the Confidence Interval (Random Variable) is like a target trying hit a fixed dart (μ).

13

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?

14

EXAMPLE *continued*

- Using the 95% CI for the population mean, we have

$$24 \pm 1.96(4/7) = 22.88 \ to \ 25.12$$

- The endpoints of the confidence interval are the confidence limits. The lower confidence limit is 22.88 and the upper confidence limit is 25.12

15

EXAMPLE *continued*

- Using the 99% CI for the population mean, we have

$$24 \pm 2.58(4/7) = 22.53 \ to \ 25.47$$

- Compare to the 95% confidence interval. A higher level of confidence means the confidence interval must be wider.

16

Selecting a Sample Size

- There are 3 factors that determine the size of a sample, none of which has any direct relationship to the size of the population. They are:
  - The degree of confidence selected.
  - The maximum allowable error.
  - The variation of the population.

17

Sample Size for the Mean

- A convenient computational formula for determining n is:

$$n = \left(\frac{Z\sigma}{E} \right)^2$$

- where E is the allowable error (margin of error), Z is the z score associated with the degree of confidence selected, and σ is the sample deviation of the pilot survey.
- σ can be estimated by past data, target sample or range of data.

18

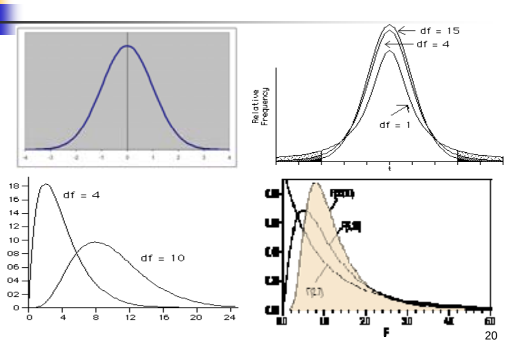
EXAMPLE

- A consumer group would like to estimate the mean monthly electric bill for a single family house in July. Based on similar studies the standard deviation is estimated to be $20.00. A 99% level of confidence is desired, with an accuracy of $5.00. How large a sample is required?

$$n = [(2.58)(20)/5]^2 = 106.5024 \approx 107$$

19

Normal Family of Distributions: Z, t, χ^2, F

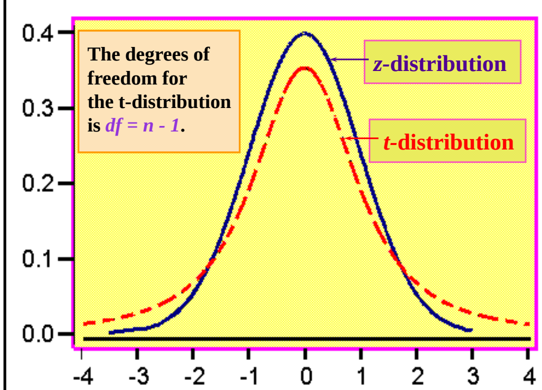


20

Characteristics of Student's *t*-Distribution

- The *t*-distribution has the following properties:
  - 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.

21



Confidence Interval for μ (small sample σ unknown)

Formula uses the t-distribution, a (1-α)100% confidence interval uses the formula shown below:

$$\overline{X} \pm (t_{\alpha/2})\left(\frac{s}{\sqrt{n}}\right) \quad df = n-1$$

23

Example – Confidence Interval

- In a random sample of 13 American adults, the mean waste recycled per person per day was 5.3 pounds and the standard deviation was 2.0 pounds.
- Assume the variable is normally distributed and construct a 95% confidence interval for μ.

24

Example- Confidence Interval

$\alpha/2=.025$

df=13-1=12

t=2.18

$$5.3 \pm 2.18 \frac{2.0}{\sqrt{13}}$$

$$5.3 \pm 1.2 = (4.1, 6.5)$$

25

Confidence Intervals, Population Proportions

- Point estimate for proportion of successes in population is: $\hat{p} = \frac{X}{n}$
- X is the number of successes in a sample of size n.
- Standard deviation of $\hat{p}$ is $\sqrt{\frac{(p)(1-p)}{n}}$
- Confidence Interval for p: $\hat{p} \pm Z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$

26

Population Proportion Example

- In a May 2006 AP/ISPOS Poll, 1000 adults were asked if "Over the next six months, do you expect that increases in the price of gasoline will cause financial hardship for you or your family, or not?"
- 700 of those sampled responded yes!
- Find the **sample proportion** and **margin of error** for this poll. (This means find a 95% confidence interval.)

27

Population Proportion Example

- Sample proportion

$$\hat{p} = \frac{700}{1000} = .70 = 70\%$$

- Margin of Error

$$MOE = 1.96\sqrt{\frac{.70(1-.70)}{1000}} = .028 = 2.8\%$$

28

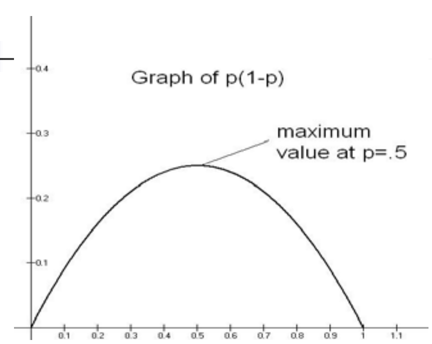
Sample Size for the Proportion

- A convenient computational formula for determining n is:

$$n = \left(p(1-p)\right)\left(\frac{Z}{E}\right)^2$$

- where E is the allowable margin of error, Z is the z-score associated with the degree of confidence selected, and p is the population proportion.
- If p is completely unknown, p can be set equal to ½ which maximizes the value of (p)(1-p) and guarantees the confidence interval will fall within the margin of error.

29



30

Example

- In polling, determine the minimum sample size needed to have a margin of error of 3% when p is unknown.

$$n = (.5)(1-.5)\left(\frac{1.96}{.03}\right)^2 = 1068$$

31

Example

- In polling, determine the minimum sample size needed to have a margin of error of 3% when p is known to be close to 1/4.

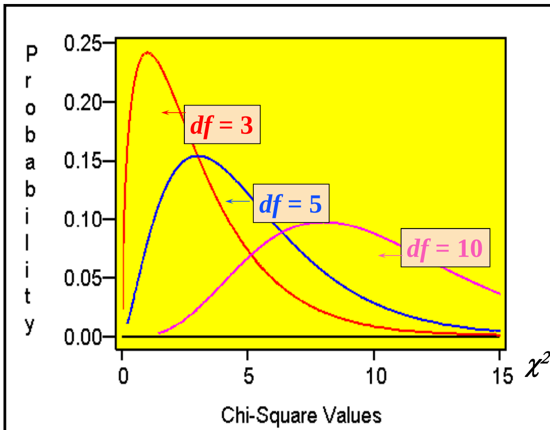
$$n = (.25)(1-.25)\left(\frac{1.96}{.03}\right)^2 = 801$$

32

Characteristics of the Chi-Square Distribution

- The major characteristics of the chi-square distribution are:
  - It is positively skewed
  - It is non-negative
  - It is based on degrees of freedom
  - When the degrees of freedom change, a new distribution is created

33



Inference about Population Variance and Standard Deviation

- s^2 is an unbiased point estimator for σ^2
- s is a point estimator for σ
- Interval estimates and hypothesis testing for both σ^2 and σ require a new distribution – the χ^2 (Chi-square)

35

Distribution of s^2

- $\frac{(n-1)s^2}{\sigma^2}$ has a chi-square distribution

- n-1 is degrees of freedom
- s^2 is sample variance
- σ^2 is population variance

36

Confidence interval for σ^2

- Confidence is **NO**T symmetric since chi-square distribution is not symmetric
- We can construct a (1-α)100% confidence interval for σ^2

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right)$$

- Take square root of both endpoints to get confidence interval for σ, the population standard deviation.

37

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.

38

Example (cont)

- df = n-1 =19
- 95% CI for σ

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

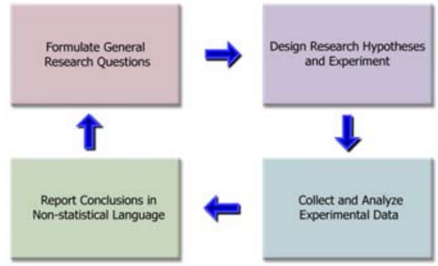
39

Math 10

Part 6
Hypothesis Testing

1

Procedures of Hypotheses Testing



2

Hypotheses Testing – Procedure 1

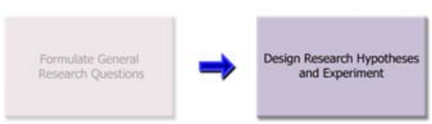
3

General Research Question

- Decide on a topic or phenomena that you want to research.
- Formulate general research questions based on the topic.
- Example:
  - Topic: Health Care Reform
  - Some General Questions:
    - Would a Single Payer Plan be less expensive than Private Insurance?
    - Do HMOs provide the same quality care as PPOs?
    - Would the public support mandated health coverage?

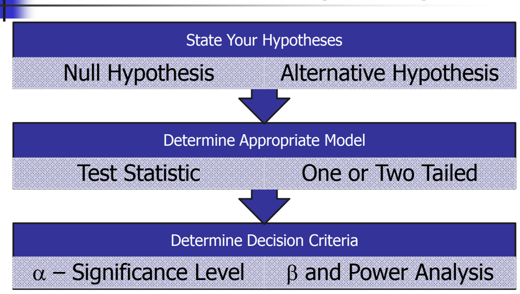
4

Hypotheses Testing – Procedure 2



5

Hypothesis Testing Design



6

What is a Hypothesis?

- Hypothesis: A statement about the value of a population parameter developed for the purpose of testing.
- Examples of hypotheses made about a population parameter are:
  - The mean monthly income for programmers is $9,000.
  - At least twenty percent of all juvenile offenders are caught and sentenced to prison.
  - The standard deviation for an investment portfolio is no more than 10 percent per month.

7

What is Hypothesis 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.

8

Hypothesis Testing Design

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

9

Definitions

- 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.

10

Hypothesis Testing Design

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

11

Definitions

- Statistical Model: A mathematical model that describes the behavior of the data being tested.
- Normal Family = the Standard Normal Distribution (Z) and functions of independent Standard Normal Distributions (eg: t, χ^2, F).
  - Most Statistical Models will be from the Normal Family due to the Central Limit Theorem.
- Model Assumptions: Criteria which must be satisfied to appropriately use a chosen Statistical Model.
- Test statistic: A value, determined from sample information, used to determine whether or not to reject the null hypothesis.

12

Hypothesis Testing Design

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

13

Definitions

- Level of Significance: The probability of rejecting the null hypothesis when it is actually true. (signified by α)
- 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.

14

Outcomes of Hypothesis Testing

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

15

Hypothesis Testing Design

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

16

Definitions

- 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.
- 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

17

One-Tailed Tests of Significance

- A test is one-tailed when the alternate 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 males.
  - H_a : The mean income of females is greater than males.
- Equality is part of H_0
- H_a determines which tail to test
  - H_a: $\mu>\mu_0$ means test upper tail.
  - H_a: $\mu<\mu_0$ means test lower tail.

18

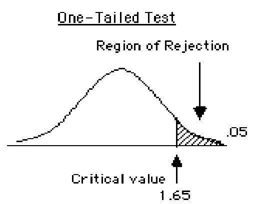
One-tailed test

$$H_0 : \mu \le \mu_0$$

$$H_a : \mu > \mu_0$$

$$\alpha = .05$$

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$



19

Two-Tailed Tests of Significance

- 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.
- Equality is part of H_0
- H_a determines which tail to test
  - H_a: $\mu \neq \mu_0$ means test both tails.

20

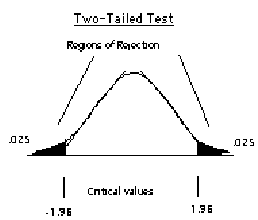
Two-tailed test

$$H_0 : \mu = \mu_0$$

$$H_a : \mu \neq \mu_0$$

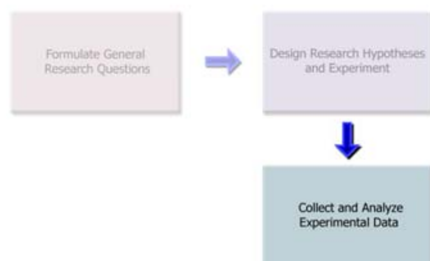
$$\alpha = .05 \quad \alpha/2 = .025$$

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$



21

Hypotheses Testing – Procedure 3



22

Collect and Analyze Experimental Data

Collect and Verify Data

Conduct Experiment | Check for Outliers

Determine Test Statistic and/or p-value

Compare to Critical Value | Compare to α

Make a Decision about Ho

Reject Ho and support Ha | Fail to Reject Ho

23

Collect and Analyze Experimental Data

Collect and Verify Data

Conduct Experiment | Check for Outliers

Determine Test Statistic and/or p-value

Compare to Critical Value | Compare to α

Make a Decision about Ho

Reject Ho and support Ha | Fail to Reject Ho

24

Outliers

- An outlier is data point that is far removed from the other entries in the data set.
- Outliers could be
  - Mistakes made in recording data
  - Data that don't belong in population
  - True rare events

25

Outliers have a dramatic effect on some statistics

- Example quarterly home sales for 10 realtors:

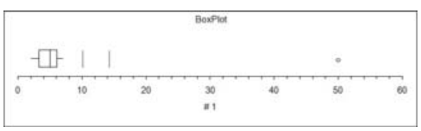
2 2 3 4 5 5 6 6 7 50

	with outlier	without outlier
Mean	9.00	4.44
Median	5.00	5.00
Std Dev	14.51	1.81
IQR	3.00	3.50

26

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, 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.



27

Using Z-score to detect outliers

- Calculate the mean and standard deviation without the suspected outlier.
- Calculate the Z-score of the suspected outlier.
- If the Z-score is more than 3 or less than -3, that data point is a probable outlier.

$$Z = \frac{50 - 4.4}{1.81} = 25.2$$

28

Outliers – what to do

- Remove or not remove, there is no clear answer.
- For some populations, outliers don't dramatically change the overall statistical analysis. Example: the tallest person in the world will not dramatically change the mean height of 10000 people.
- However, for some populations, a single outlier will have a dramatic effect on statistical analysis (called "**Black Swan**" by Nicholas Taleb) and inferential statistics may be invalid in analyzing these populations. Example: the richest person in the world will dramatically change the mean wealth of 10000 people.

29

Collect and Analyze Experimental Data

Collect and Verify Data

Conduct Experiment | Check for Outliers

Determine Test Statistic and/or p-value

Compare to Critical Value | Compare to α

Make a Decision about Ho

Reject Ho and support Ha | Fail to Reject Ho

30

The logic of Hypothesis Testing

- This is a "Proof" by contradiction.
  - 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 go with 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 $(1-\alpha)100\%$ certainty. (example: if α=.05, then we are 95% certain.

31

Test Statistic

- Test Statistic: 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 fall in the Rejection Region, Ho is rejected.
- The Test Statistic will also be used to calculate the p-value as will be defined next.

32

Example - Testing for the Population Mean
Large Sample, Population Standard Deviation Known

- When testing for the population mean from a large sample and the population standard deviation is known, the test statistic is given by:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{\mathrm{n}}}$$

33

p-Value in Hypothesis Testing

- **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, H_0 is rejected.
- If the p-value is larger than the significance level, H_0 is not rejected.

34

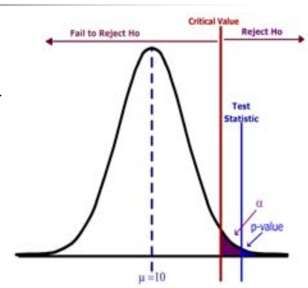
Comparing p-value to α

- Both **p-value** and α are probabilities.
- The **p-value** is determined by the **data**, and is the probability of getting results as extreme as the data assuming H_0 is true. Small values make one more likely to reject H_0.
- α is determined by **design**, and is the maximum probability the experimenter is willing to accept of rejecting a true H_0.
- Reject H_0 if p-value < α for ALL MODELS.

35

Graphic where decision is to Reject Ho

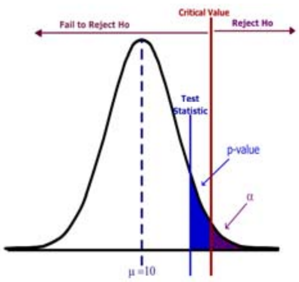
- Ho: μ = 10
  Ha: μ > 10
- Design: Critical Value is determined by significance level α.
- Data Analysis: p-value is determined by Test Statistic
- Test Statistic falls in Rejection Region.
- p-value (blue) < α (purple)
- Reject Ho.
- Strong statement: Data supports Alternative Hypothesis.



36

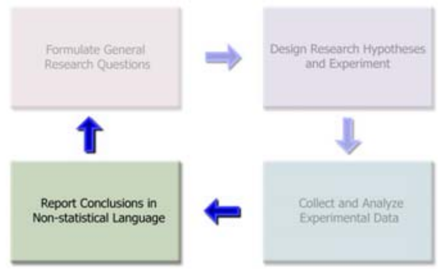
Graphic where decision is Fail to Reject Ho

- Ho: μ = 10
  Ha: μ > 10
- Design: Critical Value is determined by significance level α.
- Data Analysis: p-value is determined by Test Statistic
- Test Statistic falls in Non-rejection Region.
- p-value (blue) > α (purple)
- Fail to Reject Ho.
- Weak statement: Data is inconclusive and does not support Alternative Hypothesis.



37

Hypotheses Testing – Procedure 4



38

Conclusions

- Conclusions need to
  - Be consistent with the results of the Hypothesis Test.
  - Use language that is clearly understood in the context of the problem.
  - Limit the inference to the population that was sampled.
  - Report sampling methods that could question the integrity of the random sample assumption.
  - Conclusions should address the potential or necessity of further research, sending the process back to the first procedure.

39

Conclusions need to be consistent with the results of the Hypothesis Test.

- Rejecting Ho requires a strong statement in support of Ha.
- Failing to Reject Ho does NOT support Ho, but requires a weak statement of insufficient evidence to support Ha.
- Example:
  - The researcher wants to support the claim that, on average, students send more than 1000 text messages per month
  - Ho: μ=1000 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.

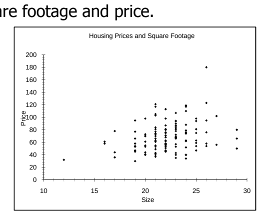
40

Conclusions need to use language that is clearly understood in the context of the problem.

- Avoid technical or statistical language.
- Refer to the language of the original general question.
- Compare these two conclusions from a test of correlation between home prices square footage and price.

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.



41

Conclusions need to limit the inference to the population that was sampled.

- If a survey was taken of a sub-group of population, then the inference applies to the subgroup.
- 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."

source: FDA Consumer Magazine – Jan/Feb 2003

42

Conclusions need to report sampling methods that could question the integrity of the random sample assumption.

- Be aware of how the sample was obtained. Here are some examples of pitfalls:
  - 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.
  - Sampling that didn't occur over the weekend may exclude many full time workers.
  - Self-selected and unverified polls (like ratemyprofessors.com) could contain immeasurable bias.

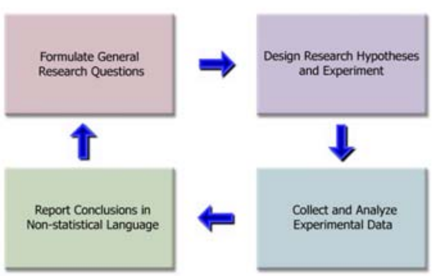
43

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.
  - Example: A math department modifies is 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.

44

Procedures of Hypotheses Testing



45

EXAMPLE – General Question

- 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.

46

EXAMPLE – Design Experiment

- A sample of n=36 bottles will be selected hourly and the contents weighed.
- Ho: $\mu=16$ Ha: $\mu \neq 16$
- The Statistical Model will be the one population test of mean using the Z Test Statistic.
- This model will be appropriate since the sample size insures the sample mean will have a Normal Distribution (Central Limit Theorem)
- We will choose a significance level of α = 5%

47

EXAMPLE – Conduct Experiment

- Last hour a sample of 36 bottles had a mean weight of 15.88 ounces.
- From past data, assume the population standard deviation is 0.5 ounces.
- Compute the Test Statistic

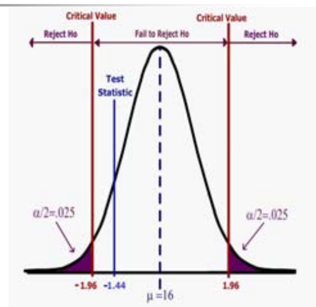
$$Z = [15.88 - 16]/[.5/\sqrt{36}] = -1.44$$

- For a two tailed test, The Critical Values are at Z = ±1.96

48

Decision – Critical Value Method

- This two-tailed test has two Critical Value and Two Rejection Regions
- The significance level (α) must be divided by 2 so that the sum of both purple areas is 0.05
- The Test Statistic does not fall in the Rejection Regions.
- Decision is **Fail to Reject Ho.**



49

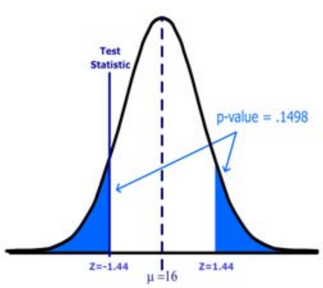
Computation of the p-Value

- One-Tailed Test: p-Value = P{z ≥ absolute value of the computed test statistic value}
- Two-Tailed Test: p-Value = 2P{z ≥ absolute value of the computed test statistic value}
- Example: Z= 1.44, and since it was a two-tailed test, then p-Value = 2P {z ≥ 1.44} = 0.0749) = .1498. Since .1498 > .05, do not reject H_0.

50

Decision – p-value Method

- The p-value for a two-tailed test must include all values (positive and negative) more extreme than the Test Statistic.
- p-value = .1498 which exceeds α = .05
- Decision is **Fail to Reject Ho.**



51

Example - Conclusion

- 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.
- We recommend continued monitoring of the machine during different employee shifts to account for the possibility of potential human error.

52

Hypothesis Testing Design

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

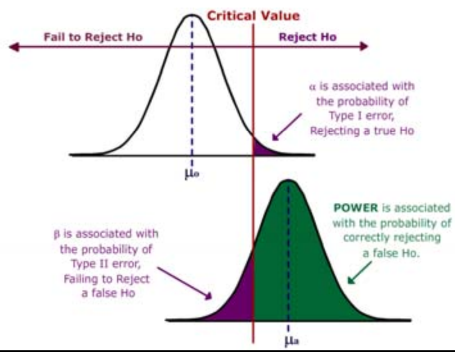
53

Statistical Power and Type II error

	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

54

Graph of "Four Outcomes"



55

Statistical Power (continued)

- Power is the probability of rejecting a false Ho, when $\mu = \mu_a$
- Power depends on:
  - Effect size $|\mu_o - \mu_a|$
  - Choice of α
  - Sample size
  - Standard deviation
  - Choice of statistical test

56

Statistical Power 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 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

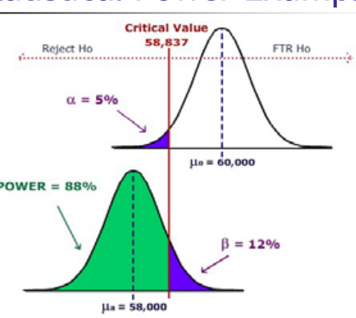
57

Statistical Power Example

- Set up the test
  - Ho: μ >= 60,000 miles
  - Ha: μ < 60,000 miles
  - α = 5%
- Determine the Critical Value
  - Reject Ho if $\overline{X} > 58{,}837$
- Calculate β and Power
  - β = 12%
  - Power = 1 – β = 88%

58

Statistical Power Example



59

New Models, 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.
- 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.

60

Testing for the Population Mean: Population Standard Deviation Unknown

- The test statistic for the one sample case is given by:

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

- The degrees of freedom for the test is n-1.
- 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.

61

Decision Rules

- Like the normal distribution, the logic for one and two tail testing is the same.
- For a two-tail test using the *t*-distribution, you will reject the null hypothesis when the value of the test statistic is greater than $t_{df,\alpha/2}$ or if it is less than - $t_{df,\alpha/2}$
- For a left-tail test using the *t*-distribution, you will reject the null hypothesis when the value of the test statistic is less than $-t_{df,\alpha}$
- For a right-tail test using the *t*-distribution, you will reject the null hypothesis when the value of the test statistic is greater than $t_{df,\alpha}$

62

Example – one population test of mean, σ unknown

- 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.)
- Design a hypotheses where the alternative claim would be the humerus bones were not from Species A.
- Determine the power of this test if the bones actually came from Species B (assume a standard deviation of 0.7)
- Conduct the test using at a 5% significance level and state overall conclusions.

63

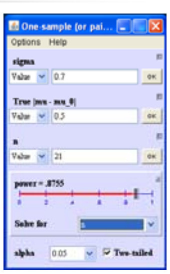
Example – Designing Test

- Research Hypotheses
  - Ho: The humerus bones are from Species A
  - Ha: The humerus bones are not from Species A
- In terms of the population mean
  - Ho: $\mu = 9.6$
  - Ha: $\mu \neq 9.6$
- Significance level
  - $\alpha = .05$
- Test Statistic (Model)
  - t-test of mean vs. hypothesized value.

64

Example - Power Analysis

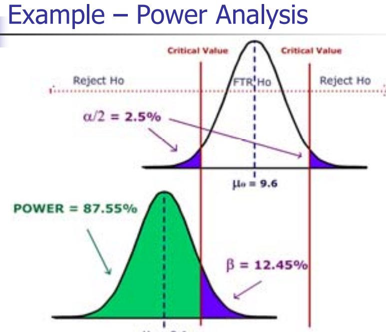
- Information needed for Power Calculation
  - $\mu_o = 9.6$ (Species A)
  - $\mu_a = 9.1$ (Species B)
  - Effect Size $= | \mu_o - \mu_a | = 0.5$
  - $\sigma = 0.7$ (given)
  - $\alpha = .05$
  - n = 21 (sample size)
  - Two tailed test
- Results using online Power Calculator*
  - Power =.8755
  - $\beta = 1 - \text{Power} = .1245$
  - 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%



*source: Russ Lenth, University of Iowa – http://www.stat.uiowa.edu/~rlenth/Power/

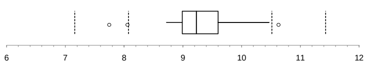
65

Example – Power Analysis



66

Example – Output of Data Analysis



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)

P-value = .0308
α =.05
Since p-value < α
Ho is rejected and we support Ha.

67

Example - Conclusions

- Results:
  - The evidence supports the claim (pvalue<.05) that the humerus bones are not from Species A.
- Sampling Methodology:
  - We are assuming since the bones were unearthed in the same location, they came from the same species.
- Limitations:
  - A small sample size limited the power of the test, which prevented us from making a more definitive conclusion.
- Further Research
  - Test if the bone are from Species B or another unknown species.
  - Test to see if bones are the same age to support the sampling methodology.

68

Tests Concerning Proportion

- Proportion: A fraction or percentage that indicates the part of the population or sample having a particular trait of interest.
- The population proportion is denoted by p.
- The sample proportion is denoted by $\hat{p}$ where

$$\hat{p} = \frac{\text{number of successes in the sample}}{\text{number sampled}}$$

69

Test Statistic for Testing a Single Population Proportion

- If sample size is sufficiently large, $\hat{p}$ has an approximately normal distribution. This approximation is reasonable if np(1-p)>5

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$p = population \quad proportion$$
$$\hat{p} = sample \quad proportion$$

70

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 size so that:
  - The test can be run at the 5% significance level.
  - 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.

71

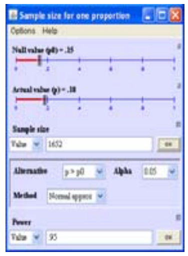
Example – Designing Test

- Research Hypotheses
  - Ho: The new letter is not more effective.
  - Ha: The new letter is more effective.
- In terms of the population proportion
  - Ho: p = 0.15
  - Ha: p > 0.15
- Significance level
  - α =.05
- Test Statistic (Model)
  - Z-test of proportion vs. hypothesized value.

72

Example - Power Analysis

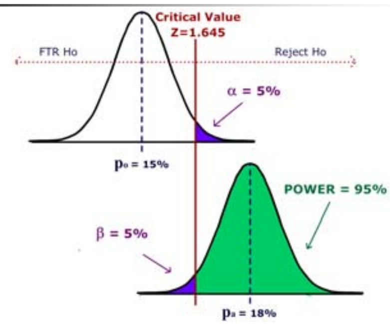
- Information needed for Sample Size Calculation
  - p_o = 0.15 (current letter)
  - p_a = 0.18 (potential new letter)
  - Effect Size =| p_o - p_a | = 0.03
  - Desired Power = 0.95
  - α = .05
  - One tailed test
- Results using online Power Calculator*
  - Sample size = 1652
  - The charity should send out 1652 new solicitation letters to potential donors and run the test.



*source: Russ Lenth, University of Iowa – http://www.stat.uiowa.edu/~rlenth/Power/

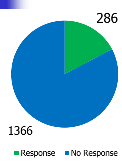
73

Example – Power Analysis



74

Example – Output of Data Analysis



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)

- P-value = .0042
- α =0.05
- Since p-value < α, Ho is rejected and we support Ha.

75

EXAMPLE
Critical Value Alternative Method

- Critical Value =1.645 (95th percentile of the Normal Distribution.)
- H_0 is rejected if Z > 1.645
- Test Statistic:

$$Z = \frac{\left(\frac{286}{1652} - .15\right)}{\sqrt{\frac{(.15)(.85)}{1652}}} = 2.63$$

- Since Z = 2.63 > 1.645, H_0 is rejected. The new letter is more effective.

76

Example - Conclusions

- Results:
  - The evidence supports the claim (pvalue<.01) that the new letter is more effective.
- Sampling Methodology:
  - 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.
- Limitations:
  - The letters needed to be sent in a specific time period, so we were not able to control for seasonal or economic factors.
- Further Research
  - Test both solicitation methods over the entire year to eliminate seasonal effects.
  - Send the old letter to another random sample to create a control group.

77

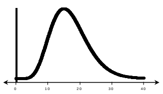
Test for Variance or Standard Deviation vs. Hypothesized Value

- 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 σ are tested with the χ^2 distribution.
- Examples of Hypotheses:
  - Ho: σ = 10 Ha: $\sigma \neq$ 10
  - Ho: σ^2 = 100 Ha: σ^2 > 100
- The sample variance s^2 is used in calculating the Test Statistic.

78

Test Statistic uses χ^2 distribtion

s^2 is the test statistic for the population variance. Its sampling distribution is a χ^2 distribution with n-1 d.f.



$$\chi^2 = \frac{(n-1)s^2}{\sigma_o^2}$$

79

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, shown here:

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	

- The test will be run at 1% significance level.

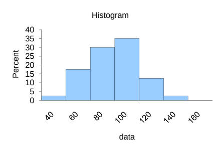
80

Example – Designing Test

- Research Hypotheses
  - Ho: Standard deviation for test scores equals 30.
  - Ha: Standard deviation for test scores is less than 30.
- In terms of the population variance
  - Ho: $\sigma^2 = 900$
  - Ha: $\sigma^2 < 900$
- Significance level
  - $\alpha = .01$
- Test Statistic (Model)
  - χ^2-test of variance vs. hypothesized value.

81

Example – Output of Data Analysis



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)

- p-value = .0054
- $\alpha = 0.01$
- Since p-value < α, Ho is rejected and we support Ha.

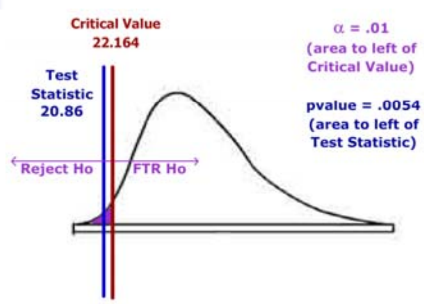
82

EXAMPLE
Critical Value Alternative Method

- Critical Value =22.164 (1st percentile of the Chi-square Distribution.)
- H_0 is rejected if $\chi^2 < 22.164$
- Test Statistic: $\chi^2 = \frac{(40)(469.426)}{900} = 20.86$
- Since Z = 20.86< 22.164, H_0 is rejected. The claim that the standard deviation is under 30 is supported.

83

Example – Decision Graph



84

Example - Conclusions

- Results:
  - The evidence supports the claim (pvalue<.01) that the standard deviation for 8th grade test scores is less than 30.
- Sampling Methodology:
  - The 41 test scores were the results of the recently administered exam to the 8th grade students.
- Limitations:
  - Since the exams were for the current class only, there is no assurance that future classes will achieve similar results.
- Further Research
  - Compare results to other schools that administered the same exam.
  - Continue to analyze future class exams to see if the claim is holding true.

85

Math 10

Part 7

Two Population Inference

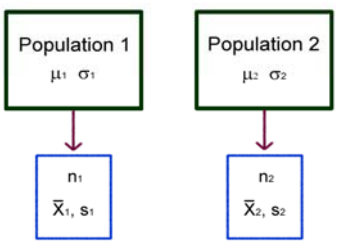
1

Comparing two population means

- Four models
  - Independent Sampling
    - Large Sample or known variances
      - **Z - test**
    - The 2 population variances are equal
      - **Pooled variance t-test**
    - The 2 population variances are unequal
      - **t-test for unequal variances**
  - Dependent Sampling
      - **Matched Pairs t-test**

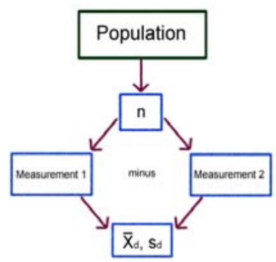
2

Independent Sampling



3

Dependent sampling



4

Difference of Two Population means

- $\overline{X}_1 - \overline{X}_2$ is Random Variable
- $\overline{X}_1 - \overline{X}_2$ is a point estimator for μ1−μ2
- The standard deviation is given by the formula $\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}$
- If n1 and n2 are sufficiently large, $\overline{X}_1 - \overline{X}_2$ follows a normal distribution.

5

Difference between two means – large sample Z test

- If both n_1 and n_2 are over 30 and the two populations are independently selected, this test can be run.
- Test Statistic:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

6

Example 1

- Are larger houses more likely to have pools?
- The housing data square footage (size) was split into two groups by pool (Y/N).
- Test the hypothesis that the homes with pools have more square feet than the homes without pools. Let α = .01

7

EXAMPLE 1 - Design

$$H_o: \mu_1 \le \mu_2 \quad H_a: \mu_1 > \mu_2$$
$$H_o: \mu_1 - \mu_2 \le 0 \quad H_a: \mu_1 - \mu_2 > 0$$

α=.01

$$Z = (\overline{X}_1 - \overline{X}_2)/(\sqrt{\sigma_1 / n_1 + \sigma_2 / n_2})$$

H_0 is rejected if $Z>2.326$

8

EXAMPLE 1 *Data*

- Population 1
Size with pool
- Sample size = 130
- Sample mean = 26.25
- Standard Dev = 6.93

- Population 2
Size without pool
- Sample size = 95
- Sample mean = 23.04
- Standard Dev = 4.55

9

EXAMPLE 1 *DATA*

$$Z = \frac{(26.25 - 23.04) - 0}{\sqrt{\frac{6.93^2}{130} + \frac{4.55^2}{95}}} = 4.19$$

- Decision: Reject Ho
- Conclusion: Homes with pools have more mean square footage.

10

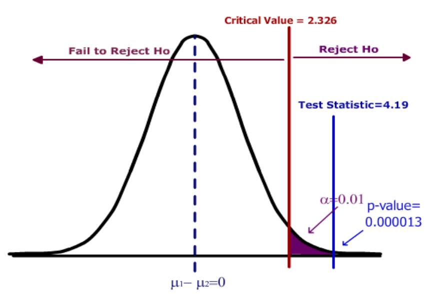
EXAMPLE 1 *p-value method*

- Using Technology Reject Ho if the p-value < α

	Sq ft with pool	*Sq ft no pool*
Mean	26.25	23.04
Std Dev	6.93	4.55
Observations	130	95
Hypothesized Mean Difference	0	
Z	4.19	
p-value	**0.0000137**	

11

EXAMPLE 1 – Results/Decision



12

Pooled variance t-test

- To conduct this test, three assumptions are required:
  - The populations must be normally or approximately normally distributed (or central limit theorem must apply).
  - The sampling of populations must be **independent.**
  - The **population variances** must be **equal**.

13

Pooled Sample Variance and Test Statistic

- Pooled Sample Variance: $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$
- Test Statistic: $t = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

$df = n_1 + n_2 - 2$

14

EXAMPLE 2

- A recent EPA study compared the highway fuel economy of domestic and imported passenger cars.
- A sample of 12 imported cars revealed a mean of 35.76 mpg with a standard deviation of 3.86.
- A sample of 15 domestic cars revealed a mean of 33.59 mpg with a standard deviation of 2.16 mpg.
- At the .05 significance level can the EPA conclude that the mpg is higher on the imported cars? *(Let subscript 2 be associated with domestic cars.)*

15

EXAMPLE 2

- : $H_o: \mu_1 \le \mu_2 \quad H_a: \mu_1 > \mu_2$
- : α=.05
- : $t = (\overline{X}_1 - \overline{X}_2)/(s_p\sqrt{1/n_1 + 1/n_2})$
- : H_0 is rejected if t>1.708, *df=25*
- : t=1.85 H_0 is rejected. Imports have a higher mean mpg than domestic cars.

16

t-test when variances are not equal.

- Test statistic: $t' = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$
- Degrees of freedom: $df = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\left[\dfrac{(s_1^2/n_1)^2}{(n_1 - 1)} + \dfrac{(s_2^2/n_2)^2}{(n_2 - 1)}\right]}$
- This test (also known as the Welch-Aspin Test) has **less power** then the prior test and should only be used when it is clear the population variances are different.

17

EXAMPLE 2

- : $H_o: \mu_1 \le \mu_2 \quad H_a: \mu_1 > \mu_2$
- : α=.05
- : t' test
- : H_0 is rejected if t>1.746, *df*=16
- : t'=1.74 H_0 is not rejected. There is insufficient sample evidence to claim a higher mpg on the imported cars.

18

Using Technology

- Decision Rule: Reject H_o if pvalue<α
- Megastat: Compare Two Independent Groups
- Use Equal Variance or Unequal Variance Test
- Use Original Data or Summarized Data

domestic 29.8 33.3 34.7 37.4 34.4 32.7 30.2 36.2 35.5 34.6 33.2 35.1 33.6 31.3 31.9

import 39.0 35.1 39.1 32.2 35.6 35.5 40.8 34.7 33.2 29.4 42.3 32.2

19

Megastat Result – Equal Variances

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)

20

Megastat Result – Unequal Variances

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)

21

Hypothesis Testing - Paired Observations

- Independent samples are samples that are not related in any way.
- Dependent samples are samples that are paired or related in some fashion.
  - For example, if you wished to buy a car you would look at the *same* car at two (or more) *different* dealerships and compare the prices.
- Use the following test when the samples are dependent:

22

Hypothesis Testing Involving Paired Observations

$$t = \frac{\overline{X}_d - \mu_d}{s_d / \sqrt{n}}$$

- where $\overline{X}_d$ is the average of the differences
- s_d is the standard deviation of the differences
- n is the number of pairs (differences)

23

EXAMPLE 3

- 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?

24

Example 3 – continued

- •Data for Hertz

$\overline{X}_1 = 46.67$

$s_1 = 5.23$

- •Data for Avis

$\overline{X}_2 = 44.87$

$s_2 = 5.62$

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

25

Example 3 - continued

By taking the difference of each pair, variability (measured by standard deviation) is reduced.

$\overline{X}_d = 1.80$

$s_d = 2.513$

$n = 15$

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

26

EXAMPLE 3 *continued*

- $H_0: \mu_d = 0 \qquad H_1: \mu_d \neq 0$
- α=.05
- Matched pairs t test, df=14
- H_0 is rejected if t<-2.145 or t>2.145
- $t = (1.80)/[2.513/\sqrt{15}] = 2.77$
- Reject H_0.
- There is a difference in mean price for compact cars between Hertz and Avis. Avis has lower mean prices.

27

Megastat Output – Example 3

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)

28

Characteristics of F-Distribution

- There is a "family" of *F* Distributions.
- Each member of the family is determined by two parameters: the numerator degrees of freedom and the denominator degrees of freedom.
- *F* cannot be negative, and it is a continuous distribution.
- The *F* distribution is positively skewed.
- Its values range from 0 *to* ∞. As $F \rightarrow \infty$ the curve approaches the X-axis.

29

Test for Equal Variances

- For the two tail test, the test statistic is given by:

$$F = \frac{S_i^2}{S_j^2}$$

- s_i^2 *and* s_j^2 are the sample variances for the two populations.
- There are 2 sets of degrees of freedom: n_i-1 for the numerator, n_j-1 for the denominator

30

EXAMPLE 4

- A stockbroker at brokerage firm, reported that the mean rate of return on a sample of 10 software stocks was 12.6 percent with a standard deviation of 4.9 percent.
- The mean rate of return on a sample of 8 utility stocks 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?

31

Test Statistic depends on Hypotheses

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 α / 2 table*

32

EXAMPLE 4 ***continued***

- : $H_o:\sigma_1 \le \sigma_2 \quad H_a:\sigma_1 > \sigma_2$
- : α =.05
- : F-test
- : H_0 is rejected if F>3.68, df=(9,7)
- : F=4.9²/3.5² =1.96 → Fail to Reject H_0.
- There is insufficient evidence to claim more variation in the software stock.

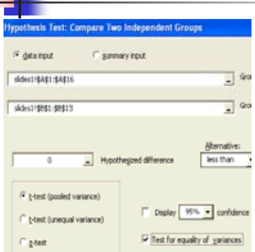
33

Excel Example

- Using Megastat – Test for equal variances under two population independent samples test and click the box to test for equality of variances
- The default p-value is a two-tailed test, so take one-half reported p-value for one-tailed tests
- Example – Domestic vs Import Data
- $H_o:\sigma_1 = \sigma_2 \quad H_a:\sigma_1 \ne \sigma_2$
- α =.10
- Reject Ho means use unequal variance t-test
- FTR Ho means use pooled variance t-test

34

Excel Output



F-test for equality of variance

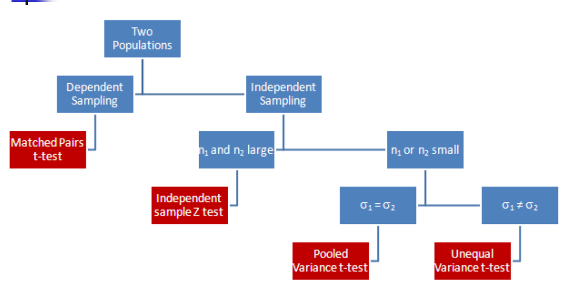
14.894	variance: import
4.654	variance: domestic
3.20	F
.0438	p-value

pvalue <.10, Reject Ho

Use unequal variance t-test to compare means.

35

Compare Two Means Flowchart



36

Math 10

M Geraghty

Part 8

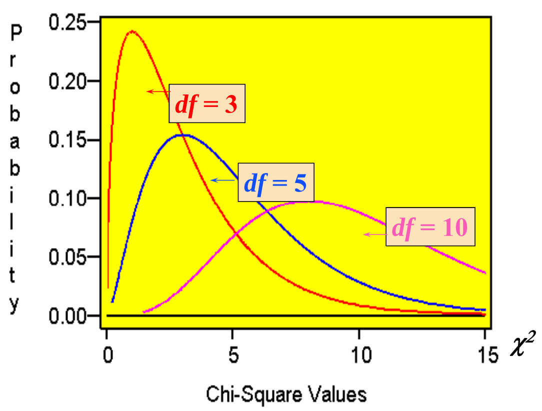
Chi-square and ANOVA tests

1

Characteristics of the Chi-Square Distribution

- The major characteristics of the chi-square distribution are:
  - It is positively skewed
  - It is non-negative
  - It is based on degrees of freedom
  - When the degrees of freedom change a new distribution is created

2



Goodness-of-Fit Test: Equal Expected Frequencies

- Let O_i and E_i be the observed and expected frequencies respectively for each category.
- H_0: there is no difference between Observed and Expected Frequencies
- H_a: there is a difference between Observed and Expected Frequencies
- The test statistic is: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
- The critical value is a chi-square value with (k-1) degrees of freedom, where k is the number of categories

4

EXAMPLE 1

- The following data on absenteeism was collected from a manufacturing plant. At the .01 level of significance, test to determine whether there is a difference in the absence rate by day of the week.

Day	Frequency
Monday	95
Tuesday	65
Wednesday	60
Thursday	80
Friday	100

5

EXAMPLE 1 *continued*

- Assume equal expected frequency: (95+65+60+80+100)/5=80

Day	O	E	(O-E)^2/E
Mon	95	80	2.8125
Tues	65	80	2.8125
Wed	60	80	5.0000
Thur	80	80	0.0000
Fri	100	80	5.0000
Total	**400**	**400**	**15.625**

6

EXAMPLE 1 *continued*

- H_o: there is no difference between the observed and the expected frequencies of absences.
- H_a: there is a difference between the observed and the expected frequencies of absences.
- Test statistic: chi-square=$\Sigma(O-E)^2/E$=15.625
- Decision Rule: reject H_o if test statistic is greater than the critical value of 13.277. (4 df, α=.01)
- Conclusion: reject H_o and conclude that there is a difference between the observed and expected frequencies of absences.

7

Goodness-of-Fit Test: Unequal Expected Frequencies

EXAMPLE 2

The U.S. Bureau of the Census (2000) indicated that 54.4% of the population is married, 6.6% widowed, 9.7% divorced (and not re-married), 2.2% separated, and 27.1% single (never been married).

A sample of 500 adults from the San Jose area showed that 270 were married, 22 widowed, 42 divorced, 10 separated, and 156 single.

At the .05 significance level can we conclude that the San Jose area is different from the U.S. as a whole?

8

EXAMPLE 2 *continued*

Status	O	E	$\sum \frac{(O-E)^2}{E}$
Married	270	272	0.015
Widowed	22	33	3.667
Divorced	42	48.5	0.871
Separated	10	11	0.091
Single	156	135.5	3.101
Total	500	500	7.745

9

EXAMPLE 2 *continued*

- **Design:** H_o: p_1=.544 p_2=.066 p_3=.097 p_4=.022 p_5=.271
  H_a: at least one p_i is different
- α=.05
- Model: Chi-Square Goodness of Fit, df=4
- H_o is rejected if χ^2 > 9.488
- **Data:** χ^2 = 7.745, Fail to Reject Ho
- **Conclusion:** Insufficient evidence to conclude San Jose is different than the US Average

10

Contingency Table Analysis

- Contingency table analysis is used to test whether two traits or variables are related.
- Each observation is classified according to two variables.
- The usual hypothesis testing procedure is used.
- The *degrees of freedom* is equal to: (number of rows-1)(number of columns-1).
- The expected frequency is computed as: Expected Frequency = (row total)(column total)/grand total

11

EXAMPLE 3

- In May 2014, Colorado became the first state to legalize the recreational use of marijuana.
- A poll of 1000 adults were classified by gender and their opinion about same-sex marriage.
- At the .05 level of significance, can we conclude that gender and the opinion about legalizing marijuana for recreational use are dependent events?

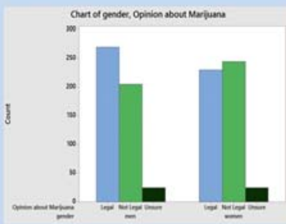
Marijuana should be	Men	Women	Total
Legal	270	230	500
Not Legal	205	245	450
Unsure	25	25	50
Total	**500**	**500**	**1000**

12

EXAMPLE 3 *continued*

```
Rows: Opinion about Marijuana
Columns: gender

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

	men	women	All
Legal	270	230	500
	250	250	
	1.600	1.600	
Not Legal	205	245	450
	225	225	
	1.778	1.778	
Unsure	25	25	50
	25	25	
	0.000	0.000	
All	500	500	1000



13

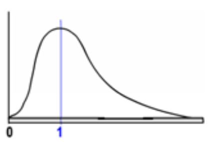
EXAMPLE 3 *continued*

- **Design:** H_o: Gender and Opinion are independent.
  H_a: Gender and Opinion are dependent.
- α=.05
- Model: Chi-Square Test for Independence, df=2
- H_o is rejected if $\chi^2 > 5.99$
- **Data:** $\chi^2 = 6.756$, Reject Ho
- **Conclusion:** Gender and opinion are dependent variables. Men are more likely to support legalizing marijuana for recreational use.

14

Characteristics of F-Distribution

- There is a "family" of *F* Distributions.
- Each member of the family is determined by two parameters: the numerator degrees of freedom and the denominator degrees of freedom.
- *F* cannot be negative, and it is a continuous distribution.
- The *F* distribution is positively skewed.
- Its values range from 0 *to* ∞ . As $F \to \infty$ the curve approaches the X-axis.



15

Underlying Assumptions for ANOVA

- The *F* distribution is also used for testing the equality of more than two means using a technique called analysis of variance (ANOVA). ANOVA requires the following conditions:
  - The populations being sampled are normally distributed.
  - The populations have equal standard deviations.
  - The samples are randomly selected and are independent.

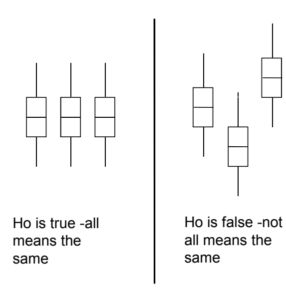
16

Analysis of Variance Procedure

- The Null Hypothesis: the population means are the same.
- The Alternative Hypothesis: at least one of the means is different.
- The Test Statistic: F=(between sample variance)/(within sample variance).
- Decision rule: For a given significance level α , reject the null hypothesis if *F* (computed) is greater than *F* (table) with numerator and denominator degrees of freedom.

17

ANOVA – Null Hypothesis



18

ANOVA NOTES

- If there are k populations being sampled, then the df (numerator)=k-1
- If there are a total of n sample points, then df (denominator) = n-k
- The test statistic is computed by:F=[(SS_F)/(k-1)]/[(SS_E)/(N-k)].
- SS_F represents the factor (between) sum of squares.
- SS_E represents the error (within) sum of squares.
- Let T_C represent the column totals, n_c represent the number of observations in each column, and ΣX represent the sum of all the observations.
- These calculations are tedious, so technology is used to generate the **ANOVA table.**

19

Formulas for ANOVA

$$SS_{Total} = \Sigma(X^2) - \frac{(\Sigma X)^2}{n}$$

$$SS_{Factor} = \Sigma\left(\frac{T_c^2}{n_c}\right) - \frac{(\Sigma X)^2}{n}$$

$$\mathrm{SS_{Error}} = \mathrm{SS_{Total}} - \mathrm{SS_{Factor}}$$

20

ANOVA Table

Source	SS	df	MS	F
Factor	SS_{Factor}	k-1	SS_F/df_F	MS_F/MS_E
Error	SS_{Error}	n-k	SS_E/df_E	
Total	SS_{Total}	n-1		

21

EXAMPLE 4

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 for sample of five days.
- 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?

22

Example 4

	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

23

Example 4 *continued*

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

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

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

24

Example 4 *continued*

ANOVA TABLE

Source	SS	df	MS	F
Factor	76.25	2	38.125	39.10
Error	9.75	10	0.975	
Total	86.00	12		

25

EXAMPLE 4 ***continued***

- **Design:** H_o: $\mu1=\mu2=\mu3$
  H_a: Not all the means are the same
- $\alpha=.05$
- Model: One Factor ANOVA
- H_0 is rejected if F>4.10
- **Data:** Test statistic: F=[76.25/2]/[9.75/10]=39.1026
- H_0 is rejected.
- **Conclusion:** There is a difference in the mean number of pizzas sold at each pizzeria.

26

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

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

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

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

27

Post Hoc Comparison Test

- Used for pairwise comparison
- Designed so the **overall** signficance level is 5%.
- Use technology.
- Refer to **Tukey Test** Material in Supplemental Material.

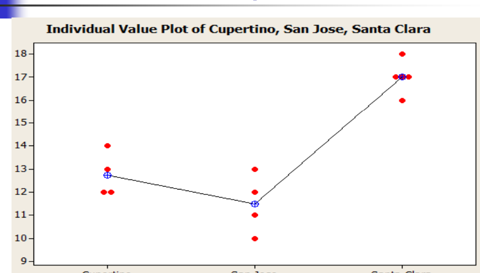
28

Post Hoc Comparison Test

```
Grouping Information Using Tukey Method

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

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

29

Post Hoc Comparison Test



30

ANOVA - Tukey's HSD Test

Application: One-way ANOVA – pair-wise comparison of means.

Requirements: Model is ussually balanced, which means that the sample size in each population should be the same. The samples taken in each population are called **replicates**. Each population is called a **treatment**. (Note: There are methods of approximating this model if the design is not balanced, but we will not cover them.)

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

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

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

q = value from studentized range table.

MSE = Mean Square Error from ANOVA table

n_c = number of replicates per treatment

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

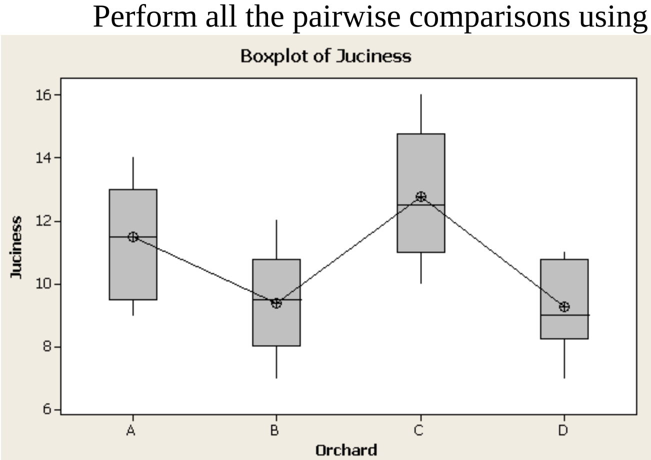
Note: Minitab will group differences into families by assigning letters. Pairs that do not share a common letter are significantly different pairs.

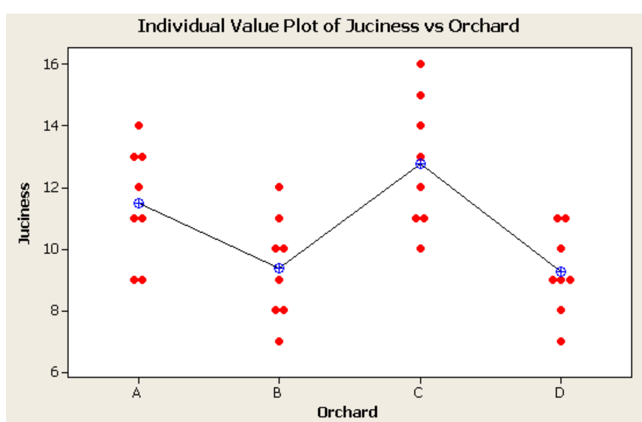
Example:
Valencia oranges were tested for juiciness at 4 different orchards. Eight oranges were sampled from each orchard, and the total ml of juice per 20 gms of orange was calculated:

Orchard A:	**Orchard B:**	**Orchard C:**	**Orchard D:**
11,13,12,14, 9,13,11,9	10,9,8,10, 11,12,7,8	13,15,14,11, 12,10,16,11	9,7,11,9, 9,11,10,8

SS Total =158.469	SS Between=69.594

a. Test for a difference in Orchards using alpha = .05
Perform all the pairwise comparisons using Tukey's Test and an overall risk level of 5%.





One-way ANOVA: Orchard A, Orchard B, Orchard C, Orchard D

```
Source  DF      SS     MS     F      P
Factor   3   69.59  23.20  7.31  0.001
Error   28   88.88   3.17
Total   31  158.47

S = 1.782   R-Sq = 43.92%   R-Sq(adj) = 37.91%


                             Individual 95% CIs For Mean Based on Pooled StDev
Level      N    Mean  StDev    +---------+---------+---------+---------
Orchard A  8  11.500  1.852                 (-------*-------)
Orchard B  8   9.375  1.685     (-------*-------)
Orchard C  8  12.750  2.121                         (-------*-------)
Orchard D  8   9.250  1.389    (-------*-------)
                               +---------+---------+---------+---------
                             8.0       9.6      11.2      12.8

Pooled StDev = 1.782


Grouping Information Using Tukey Method

           N    Mean  Grouping
Orchard C  8  12.750  A
Orchard A  8  11.500  A B
Orchard B  8   9.375    B
Orchard D  8   9.250    B

Means that do not share a letter are significantly different.


Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons

Individual confidence level = 98.92%


Orchard A subtracted from:

            Lower  Center  Upper     +---------+---------+---------+---------
Orchard B  -4.556  -2.125  0.306          (-------*-------)
Orchard C  -1.181   1.250  3.681                     (-------*-------)
Orchard D  -4.681  -2.250  0.181         (--------*-------)
                                     +---------+---------+---------+---------
                                  -6.0      -3.0       0.0       3.0


Orchard B subtracted from:

            Lower  Center  Upper     +---------+---------+---------+---------
Orchard C   0.944   3.375  5.806                            (-------*-------)
Orchard D  -2.556  -0.125  2.306                (--------*-------)
                                     +---------+---------+---------+---------
                                  -6.0      -3.0       0.0       3.0


Orchard C subtracted from:

            Lower  Center   Upper     +---------+---------+---------+---------
Orchard D  -5.931  -3.500  -1.069     (-------*-------)
                                      +---------+---------+---------+---------
                                   -6.0      -3.0       0.0       3.0
```

Math 10

Correlation and Regression

Part 9 Slides

1

Mathematical Model

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

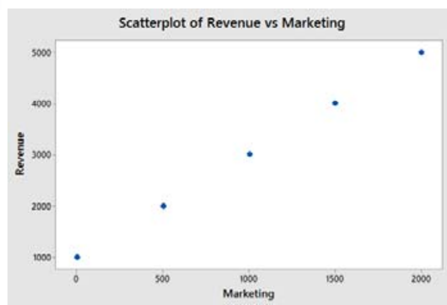
2

Mathematical Model - Table

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

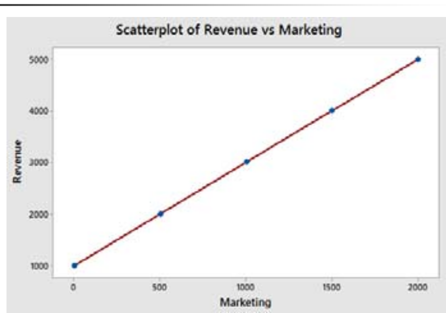
3

Mathematical Model - Scatterplot



4

Mathematical Model - Linear



5

Mathematical Linear Model

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

6

Statistical Model

- You have a small business producing custom t-shirts.
- Without marketing, your business has revenue (sales) of $1000 per week.
- Every dollar you spend marketing will increase revenue by an expected value of 2 dollars.
- Let variable X represent amount spent on marketing and let variable Y represent revenue per week.
- Let ε represent the difference between Expected Revenue and Actual Revenue (Residual Error)
- Write a statistical model that relates X to Y

7

Statistical Model - Table

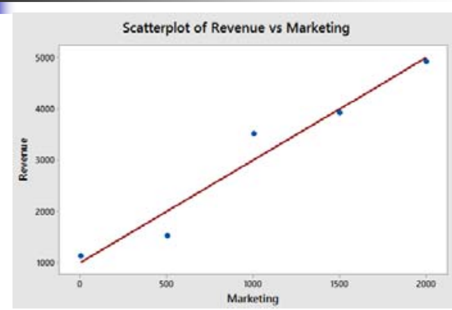
X=Marketing	Expected Revenue	Y=Actual Revenue	ε=Residual Error
$0	$1000	$1100	+$100
$500	$2000	$1500	-$500
$1000	$3000	$3500	+$500
$1500	$4000	$3900	-$100
$2000	$5000	$4900	-$100

8

Statistical Model - Scatterplot

9

Statistical Model - Linear



10

Statistical Linear Model

Regression Model

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

Y: Dependent Variable

X: Independent Variable

β_0: Y-intercept

β_1: Slope

ε: $Normal(0, \sigma)$

Example

$Y = 1000 + 2X + \varepsilon$

Y: Revenue

X: Marketing

β_0: \$1000

β_1: \$2 per \$1 marketing

11

Regression Analysis

- Purpose: to determine the regression equation; it is used to predict the value of the dependent variable (Y) based on the independent variable (X).
- Procedure: select a sample from the population and list the paired data for each observation; draw a scatter diagram to give a visual portrayal of the relationship; determine the regression equation.

12

Simple Linear Regression Model

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

Y : *Dependent Variable*

X : *Independent Variable*

β_0 : *Y – intercept*

β_1 : *Slope*

ε : *Normal* $(0, \sigma)$

13

Estimation of Population Parameters

- From sample data, find statistics that will estimate the 3 population parameters
- Slope parameter
  - b_1 will be an estimator for β_1
- Y-intercept parameter
  - b_o 1 will be an estimator for β_o
- Standard deviation
  - s_e will be an estimator for σ

14

Regression Analysis

- the regression equation: $\hat{Y} = b_0 + b_1 X$, where:
- $\hat{Y}$ is the average predicted value of *Y* for any *X*.
- b_0 is the Y-intercept, or the estimated *Y* value when *X=0*
- b_1 is the slope of the line, or the average change in $\hat{Y}$ for each change of one unit in *X*
- the least squares principle is used to obtain b_1 and b_0

$$SSX = \Sigma X^2 - \tfrac{1}{n}(\Sigma X)^2$$

$$SSY = \Sigma Y^2 - \tfrac{1}{n}(\Sigma Y)^2$$

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

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

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

15

Assumptions Underlying Linear Regression

- For each value of X, there is a group of Y values, and these Y values are *normally distributed*.
- The *means* of these normal distributions of Y values all lie on the straight line of regression.
- The *standard deviations* of these normal distributions are equal.
- The Y values are statistically independent. This means that in the selection of a sample, the Y values chosen for a particular X value do not depend on the Y values for any other X values.

16

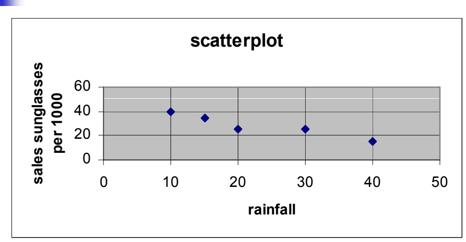
Example

- X = Average Annual Rainfall (Inches)
- Y = Average Sale of Sunglasses/1000
  - Make a Scatterplot
  - Find the least square line

X	10	15	20	30	40
Y	40	35	25	25	15

17

Example *continued*



18

Example *continued*

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

19

Example *continued*

- Find the least square line
  - SSX = 580
  - SSY = 380
  - SSXY = -445
  - b_1 = -.767
  - b_0 = 45.647
  - $\hat{Y}$ = 45.647 - .767X

20

The Standard Error of Estimate

- The standard error of estimate measures the scatter, or dispersion, of the observed values around the line of regression
- The formulas that are used to compute the standard error:

$$SSR = b_1 \cdot SSXY$$

$$SSE = \sum\left(Y - \hat{Y}\right)^2 = SSY - SSR$$

$$MSE = SSE/(n-2)$$

$$s_e = \sqrt{MSE}$$

21

Example *continued*

- Find SSE and the standard error:
  - SSR = 341.422
  - SSE = 38.578
  - MSE = 12.859
  - s_e = 3.586

X	Y	Y'	$(Y\text{-}Yhat)^2$
10	40	37.97	4.104
15	35	34.14	0.743
20	25	30.30	28.108
30	25	22.63	5.620
40	15	14.96	0.002
		Total	**38.578**

22

Correlation Analysis

- Correlation Analysis: A group of statistical techniques used to measure the strength of the relationship (correlation) between two variables.
- Scatter Diagram: A chart that portrays the relationship between the two variables of interest.
- Dependent Variable: The variable that is being predicted or estimated. "Effect"
- Independent Variable: The variable that provides the basis for estimation. It is the predictor variable. "Cause?" (Maybe!)

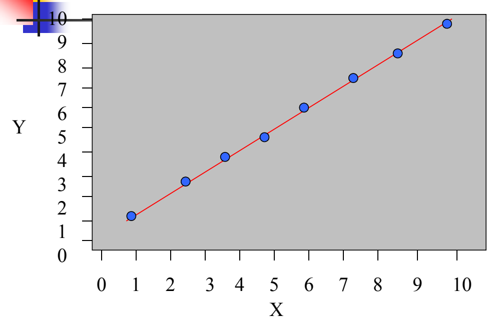
23

The Coefficient of Correlation, r

- The Coefficient of Correlation (r) is a measure of the **strength** of the relationship between two variables.
  - It requires interval or ratio-scaled data (variables).
  - It can range from -1.00 to 1.00.
  - Values of -1.00 or 1.00 indicate perfect and strong correlation.
  - Values close to 0.0 indicate weak correlation.
  - Negative values indicate an inverse relationship and positive values indicate a direct relationship.

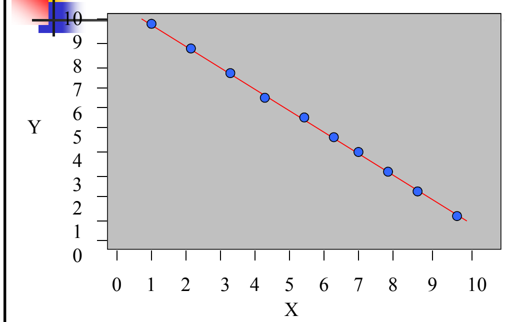
24

Perfect Positive Correlation



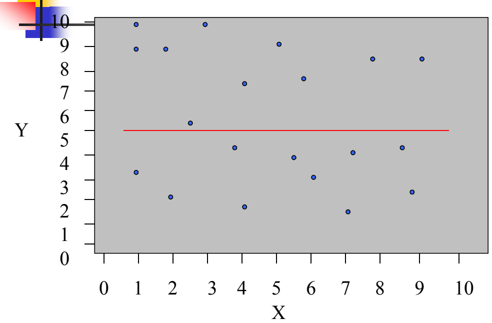
25

Perfect Negative Correlation



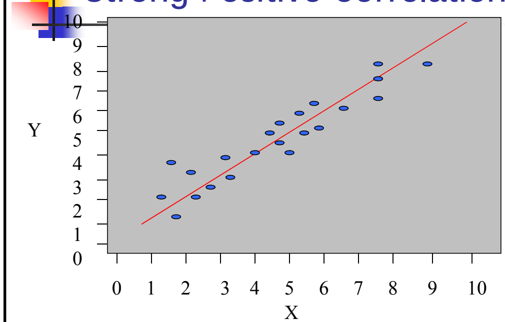
26

Zero Correlation



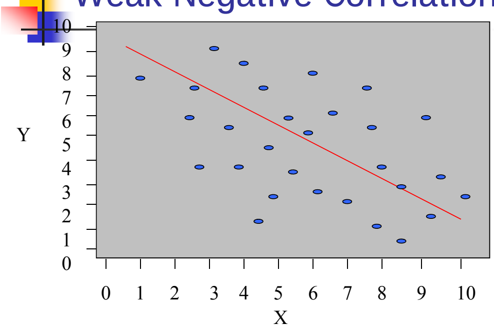
27

Strong Positive Correlation



28

Weak Negative Correlation



29

Causation

- Correlation does not necessarily imply causation.
- There are 4 possibilities if X and Y are correlated:
  1. X causes Y
  2. Y causes X
  3. X and Y are caused by something else.
  4. Confounding - The effect of X and Y are hopelessly mixed up with other variables.

30

Causation - Examples

- City with more police per capita have more crime per capita.
- As Ice cream sales go up, shark attacks go up.
- People with a cold who take a cough medicine feel better after some rest.

31

r^2: Coefficient of Determination

- r^2 is the proportion of the total variation in the dependent variable Y that is explained or accounted for by the variation in the independent variable X.
- The coefficient of determination is the square of the coefficient of correlation, and ranges from 0 to 1.

32

Formulas for r and r^2

$$r = \frac{SSXY}{\sqrt{SSX \cdot SSY}} \qquad r^2 = \frac{SSR}{SSY}$$

$$SSX = \Sigma X^2 - \tfrac{1}{n}(\Sigma X)^2$$
$$SSY = \Sigma Y^2 - \tfrac{1}{n}(\Sigma Y)^2$$
$$SSXY = \Sigma XY - \tfrac{1}{n}(\Sigma X \cdot \Sigma Y)$$
$$SSR = SSY - \left(SSXY^2 / SSX\right)$$

33

Example

- X = Average Annual Rainfall (Inches)
- Y = Average Sale of Sunglasses/1000

X	10	15	20	30	40
Y	40	35	25	25	15

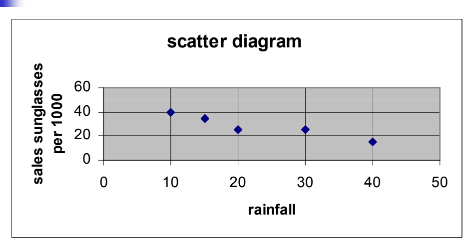
34

Example *continued*

- Make a Scatter Diagram
- Find r and r^2

35

Example *continued*



36

Example *continued*

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

- SSX = 3225 - 115^2/5 = 580
- SSY = 4300 - 140^2/5 = 380
- SSXY= 2775 - (115)(140)/5 = -445

37

Example *continued*

- r = -445/sqrt(580 x 330) = -.9479
  - Strong negative correlation
- r^2 = .8985
  - About 89.85% of the variability of sales is explained by rainfall.

38

Characteristics of F-Distribution

- There is a "family" of *F* Distributions.
- Each member of the family is determined by two parameters: the numerator degrees of freedom and the denominator degrees of freedom.
- *F* cannot be negative, and it is a continuous distribution.
- The *F* distribution is positively skewed.
- Its values range from 0 *to* ∞ . As $F \rightarrow \infty$ the curve approaches the X-axis.

39

Hypothesis Testing in Simple Linear Regression

- The following Tests are equivalent:
  - H_0: X and Y are uncorrelated
  - H_a: X and Y are correlated
  - H_0: $\beta_1 = 0$
  - H_a: $\beta_1 \neq 0$
- Both can be tested using ANOVA

40

ANOVA Table for Simple Linear Regression

Source	SS	df	MS	F
Regression	SSR	1	SSR/dfR	MSR/MSE
Error/Residual	SSE	n-2	SSE/dfE	
TOTAL	SSY	n-1		

41

Example *continued*

- Test the Hypothesis H_o: $\beta_1 = 0$, α=5%

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

- Reject Ho p-value < α

42

Confidence Interval

- The confidence interval for the mean value of Y for a given value of X is given by:

$$\hat{Y} \pm t \cdot s_e \cdot \sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{SSX}}$$

- Degrees of freedom for t =n-2

43

Prediction Interval

- The prediction interval for an individual value of Y for a given value of X is given by:

$$\hat{Y} \pm t \cdot s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{SSX}}$$

- Degrees of freedom for t =n-2

44

Example *continued*

- Find a 95% Confidence Interval for Sales of Sunglasses when rainfall = 30 inches.
- Find a 95% Prediction Interval for Sales of Sunglasses when rainfall = 30 inches.

45

Example *continued*

- 95% Confidence Interval

$$22.63 \pm 6.60$$

- 95% Confidence Interval

$$22.63 \pm 13.18$$

46

Using Minitab to Run Regression

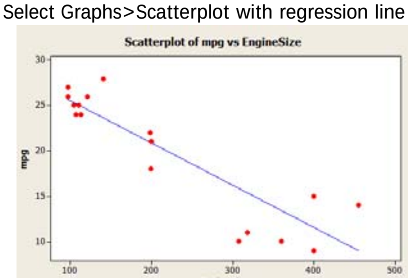
- Data shown is engine size in cubic inches (X) and MPG (Y) for 20 cars.

x	y		x	y
400	15		104	25
455	14		121	26
113	24		199	21
198	22		360	10
199	18		307	10
200	21		318	11
97	27		400	9
97	26		97	27
110	25		140	28
107	24		400	15

47

Using Minitab to Run Regression

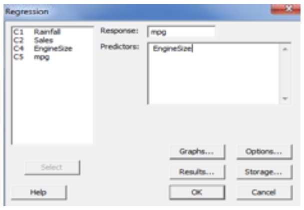
Select Graphs>Scatterplot with regression line



48

Using Minitab to Run Regression

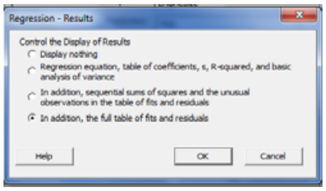
Select Statistics>Regression>Regression, then choose the Response (Y-variable) and model (X-variable)



49

Using Minitab to Run Regression

Click the results box, and choose the fits and residuals to get all predictions.



50

Using Minitab to Run Regression

The results at the beginning are the regression equation, the intercept and slope, the standard error of the residuals, and the r^2

```
The regression equation is
mpg = 30.2 - 0.0466 EngineSize


Predictor        Coef   SE Coef      T      P
Constant       30.203     1.361  22.20  0.000
EngineSize  -0.046598  0.005378  -8.66  0.000


S = 2.95688   R-Sq = 80.7%   R-Sq(adj) = 79.6%
```

51

Using Minitab to Run Regression

Next is the ANOVA table, which tests the significance of the regression model.

```
Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  656.42  656.42  75.08  0.000
Residual Error  18  157.38    8.74
Total           19  813.80
```

52

Using Minitab to Run Regression

Finally, the residuals show the potential outliers.

Obs	EngineSize	mpg	Fit	SE Fit	Residual	St Resid
1	400	15.000	11.564	1.167	3.436	1.26
2	455	14.000	9.001	1.421	4.999	1.93
3	113	24.000	24.937	0.880	-0.937	-0.33
4	198	22.000	20.976	0.673	1.024	0.36
5	199	18.000	20.930	0.672	-2.930	-1.02
6	200	21.000	20.883	0.671	0.117	0.04
7	97	27.000	25.683	0.939	1.317	0.47
8	97	26.000	25.683	0.939	0.317	0.11
9	110	25.000	25.077	0.891	-0.077	-0.03
10	107	24.000	25.217	0.902	-1.217	-0.43
11	104	25.000	25.357	0.913	-0.357	-0.13
12	121	26.000	24.565	0.853	1.435	0.51
13	199	21.000	20.930	0.672	0.070	0.02
14	360	10.000	13.427	0.998	-3.427	-1.23
15	307	10.000	15.897	0.807	-5.897	-2.07R
16	318	11.000	15.385	0.842	-4.385	-1.55
17	400	9.000	11.564	1.167	-2.564	-0.94
18	97	27.000	25.683	0.939	1.317	0.47
19	140	28.000	23.679	0.792	4.321	1.52
20	400	15.000	11.564	1.167	3.436	1.26

53

How to Choose a Model

a. Categorical or Numeric Data?
b. One, Two or many populations?
c. Test of mean, proportion, standard deviation, or something else?
d. Independent or dependent sampling?
e. Large or small sample size?
f. One or two tailed test?

- **One Population Tests**

  - Numeric Data

    1. **Z test for population mean vs. hypothesized value** (Part 6 slides)
       - Test of mean, population standard deviation known
       - $Ho: \mu = 10\ Ha: \mu \neq 10,\ Ho: \mu \leq 10\ Ha: \mu > 10,\ Ho: \mu \geq 10\ Ha: \mu < 10$
       - $Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$ Degrees of freedom – Not Applicable

    2. **t test for population mean vs. hypothesized value** (Part 6 slides)
       - Test of mean, population standard deviation unknown
       - $Ho: \mu = 10\ Ha: \mu \neq 10,\ Ho: \mu \leq 10\ Ha: \mu > 10,\ Ho: \mu \geq 10\ Ha: \mu < 10$
       - $t = \frac{\overline{X} - \mu_0}{s / \sqrt{n}}$ Degrees of freedom = n-1

    3. **χ^2 test for variance vs. hypothesized value (Part 6 slides)**
       - Test of standard deviation or variance
       - $Ho: \sigma = 10\ Ha: \sigma \neq 10,\ Ho: \sigma \leq 10\ Ha: \sigma > 10,\ Ho: \sigma \geq 10\ Ha: \sigma < 10$
       - $\chi^2 = \frac{s^2(n-1)}{\sigma^2}$ Degrees of freedom = n-1

  - Categorical Data

    4. **Z test for proportion vs. hypothesized value (Part 6 slides)**
       - Two choices (Yes/No) - Test of population proportion
       - $Ho: p = 0.5\ Ha: p \neq 0.5,\ Ho: p \leq 0.5\ Ha: p > 0.5,\ Ho: p \geq 0.5\ Ha: p < 0.5$
       - $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{(p_o)(1-p_o)}{n}}}$ Degrees of freedom = not applicable

    5. **χ^2 Goodness of fit test (Part 8 Slides)**
       - Multiple choices (k)- Test of multiple proportions
       - $Ho: p_1 = 0.4\ \ p_2 = 0.1\ \ p_3 = 0.5\ \ Ha: \textit{At least one } p_i \textit{ is different}$
       - $\chi^2 = \sum \frac{(O-E)^2}{E}$ Degrees of freedom = k-1

- **Two or more Population Tests**
  - Numeric Data - One scale variable with two or more populations (factor variable)

6. **Independent Samples: Z-test (Part 7 Slides)**
   - Comparing 2 Means – Large Sample Size(n_1,n_2>30) or population standard deviation known
   - $Ho: \mu_1 = \mu_2 \; Ha: \mu_1 \neq \mu_2, \; Ho: \mu_1 \leq \mu_2 Ha: \mu_1 > \mu_2, \; Ho: \mu_1 \geq \mu_2 \; Ha: \mu_1 < \mu_2$
   - $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ Degrees of freedom – Not Applicable

7. **Independent Sample t-test with equal variances (pooled variance t test) (Part 7 Slides)**
   - Comparing 2 Means – Not Large Sample Sizes, assume $\sigma_1 = \sigma_2$
   - $Ho: \mu_1 = \mu_2 \; Ha: \mu_1 \neq \mu_2, \; Ho: \mu_1 \leq \mu_2 Ha: \mu_1 > \mu_2, \; Ho: \mu_1 \geq \mu_2 \; Ha: \mu_1 < \mu_2$
   - $t = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ Degrees of freedom = n_1+n_2-2 (More power)

8. **Independent Sample t-test with** unequal variances (Part 7 Slides)
   - Comparing 2 Means – Not Large Sample Sizes, assume $\sigma_1 \neq \sigma_2$
   - $Ho: \mu_1 = \mu_2 \; Ha: \mu_1 \neq \mu_2, \; Ho: \mu_1 \leq \mu_2 Ha: \mu_1 > \mu_2, \; Ho: \mu_1 \geq \mu_2 \; Ha: \mu_1 < \mu_2$
   - $t = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ degrees of freedom < n_1+n_2-2 (Less power)

9. **Dependent Sampling – Matched Pairs (Part 7 Slides)**
   - Comparing 2 Means – Look at differences of measurements
   - $Ho: \mu_d = 0 \, Ha: \mu_d \neq 0, \; Ho: \mu_d \leq 0 Ha: \mu_d > 0, \; Ho: \mu_d \geq 0 \, Ha: \mu_d < 0$
   - $t = \dfrac{\overline{X}_d - \mu_d}{s_d / \sqrt{n}}$ Degrees of freedom = n-1

10. **F test of Variances (Part 7 Slides)**
    - Comparing 2 Variances
    - $Ho:\sigma_1 = \sigma_2\ Ha:\sigma_1 \neq \sigma_2,\ Ho:\sigma_1 \leq \sigma_2 Ha:\sigma_1 > \sigma_2,\ Ho:\sigma_1 \geq \sigma_2\ Ha:\sigma_1 < \sigma_2$
    - $F = \frac{s_1^2}{s_2^2}\ or\ F = \frac{s_2^2}{s_1^2}$ Degrees of freedom = n_1-1, n_2-1 or n_2-1, n_1-1
    - Use this test to help choose between models 7 and 8 above.

11. **One Factor Analysis of Variance (Part 8 Slides)**
    - Comparing 3 or more Means – (ANOVA) – F test
    - $Ho:\mu_1 = \mu_2 = \mu_3 = ... = \mu_k Ha: at\ least\ one = \mu_i\ is\ different$
    - (ANOVA table) $F = \frac{MSfactor}{MSerror}$ Degrees of freedom = k-1, n-k
    - Post Hoc - Pairwise comparisons – **Tukey's HSD test (Part 8 Slides)**
    - Categorical variable is Factor, Numeric Variable is Response

- **Categorical Data –Comparing 2 or more variables**

12. **χ^2 Test for Independence (Part 8 Slides)**
    - Test for a relationship between two variables (A and B) in a contingency table -
    - Ho: A and B are Independent Ha: A and B are dependent
    - $\chi^2 = \sum \frac{(O-E)^2}{E}$ Degrees of freedom = (rows-1)(columns-1)
    - E = (Row Total)(Column Total)/Grand Total

- **Two numeric variables (X,Y) – bivariate data**

13. **Correlation Coefficient (Part 9 Slides)**
    Strength of Relationship between two variables
    - correlation between two numeric variables
    - X and Y are not correlated Ha: X and Y are correlated
    - (ANOVA table) $F = \frac{MSregression}{MSerror}$ Degrees of freedom = 1, n-2

14. **Simple Linear Regression – F test (Part 9 Slides)**
    - Significance and prediction of Linear fit between two variables
    - $Ho: slope = 0 Ha: slope \neq 0$
    - (ANOVA table) $F = \frac{MSregression}{MSerror}$ Degrees of freedom = 1, n-2

Math 10 - Homework 0 **Name:_______________________________**
Course Syllabus and Materials

Go to the class website at http://nebula2.deanza.edu/~mo (or simply Google "mo de anza").
Find the link for Math 10 Sec 28, and then find the Syllabus:

Questions about the Syllabus

1. What are the required materials?

2. How many homework assignments are there?

3. What assignments can be turned in as a group?

4. What is the course policy on late homework and labs?

Questions about the Calendar (on the same page as the Syllabus)

5. What assignments are due in the first two weeks?

6. What are the dates of the midterms?

7. What is the date of the final exam for your section?

8. How many labs are there?

9. What are the deadlines for dropping and withdrawing from the course?

Exploring the Website – Find the link for Math 10 Handouts and open the PDF file for Part 1. I highly recommend you print this out and bring it to class to take notes.

10. What is the Topic for Part 1?

11. Do you any comments about improving these handouts?

Frequently Asked Questions – Find the "FAQ" at the top of the page. If Flash is not working, you can use the menu sidebar on the home page.

12. Find the question that starts "I need help in this class…" What are three things you can do if you need help?

13. Read the two questions that have two do with cheating and sign and date the following statement:

I have read the course policy on cheating in both the syllabus and the Frequently Asked Questions . I understand and agree to the terms as outlined in these policies.

____ ____________________________

Signature Date

Please write any Comments or Questions about the Course Policies here:

Math 10 - Homework 1

1. Identify the following data by type (categorical, discrete, continuous) and level (nominal, ordinal, interval, ratio)

    a. Number of tickets sold at a rock concert.

    b. Make of automobile.

    c. Age of a fossil.

    d. Temperature of a nuclear power plant core reactor.

    e. Number of students who transfer to private colleges.

    f. Cost per unit at a state University.

    g. Letter grade on an English essay.

2. A poll was taken of 150 students at De Anza College. Each student was asked how many hours they work outside of college. The students were interviewed in the morning between 8Am and 11 AM on a Thursday. The sample mean for these 150 students was 9.2 hours.

    a. What is the Population?

    b. What is the Sample?

    c. Does the 9.2 hours represent a statistic or parameter? Explain.

    d. Is the sample mean of 9.2 a reasonable estimate of the mean number of hours worked for all students at De Anza? Explain any possible bias.

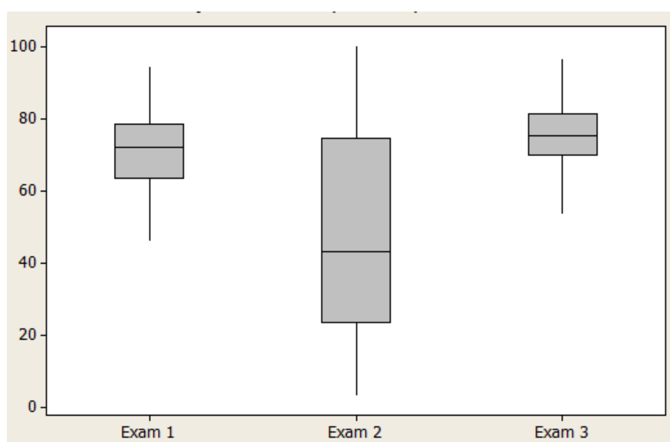
3. The box plots represent the results of three exams for 40 students in a Math course.

    a. Which exam has the highest median?

    b. Which exam has the highest standard deviation?

    c. For Exam 2, how does the median compare to the mean?

    d. In your own words, compare the exams.



4. The following average daily commute time (minutes) for residents of two cities.

City A	2	4	4	4	4	5	7	9	13	14	16	16	16	18	19	19
	21	21	21	27	30	35	37	38	47	48	50	59	70	72	87	97
City B	29	38	38	40	40	48	48	50	52	52	54	55	56	57	57	58
	58	58	59	59	59	62	62	63	66	66	67	69	69	71	75	89

a. Construct a back-to back stem and leaf diagram and interpret the results.

b. Find the quartiles and interquartile range for each group.

c. Calculate the 80th percentile for each group.

d. Construct side-by-side box plots and compare the two groups.

e. For each group, determine the z-score for a commute of 75 minutes. For which group would a 75 minute commute be more unusual.

5. The February 10, 2009 Nielsen ratings of 20 TV programs shown on commercial television, all starting between 8 PM and 10 PM, are given below:

2.1	2.3	2.5	2.8	2.8	3.6	4.4
4.5	5.7	7.6	7.6	8.1	8.7	10.0
10.2	10.7	11.8	13.0	13.6	17.3	

a. Graph a stem and leaf plot with the tens and ones units making up the stem and the tenths unit being the leaf.

b. Group the data into intervals of *width* 2, starting the 1st interval at 2 and obtain the *frequency* of each of the intervals.

c. Graphically depict the grouped frequency distribution in (b) by a histogram.

d. Obtain the *relative frequency*, *%* and *cumulative frequency* and *cumulative relative frequency* for the intervals in (b)

e. Construct an *ogive* of the data. Estimate the median and quartiles.

f. Obtain the sample *mean* and the *median*. Compare the median to the ogive.

g. Do you believe that the data is symmetric, right-skewed or left skewed?

h. Determine the sample *variance* and *standard deviation*.

i. Assuming the data are bell shaped, between what two numbers would you expect to find 68% of the data

6. The following data represents recovery time for 16 patients (arranged in a table to help you out)

count	Days (X)	$X - \overline{X}$	$(X - \overline{X})^2$	Z Score
#1	2			
#2	3			
#3	4			
#4	4			
#5	5			
#6	5			
#7	5			
#8	5			
#9	5			
#10	6			
#11	6			
#12	7			
#13	7			
#14	8			
#15	8			
#16	16			
Totals				

a. Calculate the sample mean and median

b. Use the table to calculate the variance and standard deviation.

c. Use the range of the data to see if the standard deviation makes sense. (Range should be between 3 and 6 standard deviations)

d. Using the empirical rule between what two numbers should you expect to see 68% of the data? 95% of the data? 99.7% of the data?

e. Calculate the Z-score for observation. Do you think any of these data are outliers?

7. The following data represents the heights (in feet) of 20 almond trees in an orchard.

14	14	14	14	15	18	18	20	21	21
22	24	25	25	25	27	27	29	31	45

a. Construct a box plot of the data.

b. Do you think the tree with height of 45 feet is an outlier? Use both methods we covered in class to justify your answer.

8. Rank the following correlation coefficients from weakest to strongest.

   .343, -.318, .214, -.765, 0, .998, -.932, .445

9. If you were trying to think of factors that affect health care costs:

   a. Choose a variable you believe would be positively correlated with health care costs.

   b. Choose a variable you believe would be negatively correlated with health care costs.

   c. Choose a variable you believe would be uncorrelated with health care costs.

Math 10 - Homework 2

1. A student has a 90% chance of getting to class on time on Monday and a 70% chance of getting to class on time on Tuesday. Assuming these are **independent** events, determine the following probabilities:

    a. The student is on time both Monday and Tuesday.

    b. The student is on time at least once (Monday or Tuesday).

    c. The student is late both days.

2. A class has 10 students, 6 females and 4 males. 3 students will be sampled without replacement for a group presentation.

    a. Construct a tree diagram of all possibilities (there will be 8 total branches at the end)

    b. Find the following probabilities:

        i. All male students in the group presentation.

        ii. Exactly 2 female students in the group presentation.

        iii. At least 2 female students in the group presentation.

3. 20% of professional cyclists are using a performance enhancing drug. A test for the drug has been developed that has a 60% chance of correctly detecting the drug(true positive). However, the test will come out positive in 2% of cyclists who do not use the drug (false positive).

    a. Construct a tree diagram where the first set of branches are cylcists with and without the drug, and the 2nd set is whether or not they test positive.

    b. From the tree diagram create a contingency table.

    c. What percentage of cyclists will test positive for the drug?

    d. If a cyclist tests positive, what is the probability that the cyclist really used the drug?

4. We wish to determine the morale for a certain company. We give each of the workers a questionnaire and from their answers we can determine the level of their morale, whether it is 'Low', 'Medium ' or 'High'; also noted is the 'worker type' for each of the workers. For each worker type, the frequencies corresponding to the different levels of morale are given below.

WORKER MORALE

Worker Type	Low	Medium	High
Executive	1	14	35
Upper Management	5	30	65
Lower Management	5	40	55
Non-Management	354	196	450

a. We randomly select 1 worker from this population. What is the probability that the worker selected

- is an executive?
- is an executive with medium morale?
- is an executive or has medium morale?
- is an executive, given the information that the worker has medium morale.

b. Given the information that the selected worker is an executive, what is the probability that the worker

- has medium morale?
- has high morale?

c. Are the following events independent or dependent? Explain your answer:

- is an executive', 'has medium morale', are these independent?
- is an executive', 'has high morale', are these independent?

Math 10 - Homework 3

Additional Problems:

1. Explain the difference between population parameters and sample statistics. What symbols do we use for the mean and standard deviation for each of these?

2. Consider the following probability distribution function of the random variable X which represents the number of people in a group(party) at a restaurant:

X	P(X)				
1	.10				
2	.25				
3	.20				
4	.20				
5	.10				
6	.05				
7	.05				
8	.05				

   a. Find the population mean of X.

   b. Find the population variance and standard deviation of X.

   c. Find the probability that the next party will be over 4 people.

   d. Find the probability that the next three parties (assuming independence) will each be over 4 people.

3. 10% of all children at large urban elementary school district have been diagnosed with learning disabilities. 10 children are randomly and independently selected from this school district.

   a. Let X = the number of children with learning disabilities in the sample. What type of random variable is this?

   b. Find the mean and standard deviation of X.

   c. Find the probability that exactly 2 of these selected children have a learning disability.

   d. Find the probability that at least 1 of these children has a learning disability.

   e. Find the probability that less than 3 of these children have a learning disability.

4. A general statement is made that an error occurs in 10% of all retail transactions. We wish to evaluate the truthfulness of this figure for a particular retail store, say store A. Twenty transactions of this store are randomly obtained. Assuming that the 10% figure also applies to store A and let X be the number of retail transactions with errors in the sample

    a. The probability distribution function (pdf) of X is binomial. Identify the parameters n and p.

    b. Calculate the expected value of X

    c. Calculate the variance of X

    d. Find the probability exactly 2 transactions sampled are in error.

    e. Find the probability at least 2 transactions sampled are in error.

    f. Find the probability that no more than one transaction is in error.

    g. Would it be unusual if 5 or more transactions were in error?

5. A newspaper finds a mean of 4 typographical errors **per** page. Assume the errors follow a Poisson distribution.

    a. Let X equal the number of errors on one page. Find the mean and standard deviation of this random variable.

    b. Find the probability that exactly three errors are found on one page.

    c. Find the probability that no more than 2 errors are found on one page.

    d. Find the probability that no more than 2 errors are found on **two** pages.

6. Major accidents at a regional refinery occur on the average once every five years. Assume the accidents follow a Poisson distribution.

    a. How many accidents would you expect over 10 years?

    b. Find the probability of no accidents in the next 10 years.

    c. Find the probability of no accidents in the next 20 years.

7. 20% of the people in a California town consider themselves vegetarians. If 20 people are randomly sampled, find the probability that:

    a. Exactly 3 are vegetarians.
    b. At least 3 are vegetarians.
    c. At most 3 are vegetarians

8. Cargo ships arrive at a loading dock at a rate of 2 per day. The dock has the capability of handling 3 arrivals per day. How many days per month (assume 30 days in a month) would you expect the dock being unable to handle all arriving ships? (Hint: first find the probability that more than 3 ships arrive and then use that probability to find the expected number of days in a month too many ships arrive.)

Math 10 - Homework 4

1. A ferry boat leaves the dock once per hour. Your waiting time for the next ferryboat will follow a uniform distribution from 0 to 60 minutes.

   a. Find the mean and variance of this random variable.
   b. Find the probability of waiting more than 20 minutes for the next ferry.
   c. Find the probability of waiting exactly 20 minutes for the next ferry.
   d. Find the probability of waiting between 15 and 35 minutes for the next ferry.
   e. Find the conditional probability of waiting at least 10 more minutes after you have already waited 15 minutes.
   f. Find the probability of waiting more than 45 minutes for the ferry on 3 consecutive independent days.

2. The cycle times for a truck hauling concrete to a highway construction site are uniformly distributed over the interval 50 to 70 minutes.
   a. Find the mean and variance for cycle times.
   b. Find the 5^{th} and 95^{th} percentile of cycle times.
   c. Find the interquartile range.
   d. Find the probability the cycle time for a randomly selected truck exceeds 62 minutes.
   e. If you are given the cycle time exceeds 55 minutes, find the probability the cycle time is between 60 and 65 minutes.

3. The amount of gas in a car's tank (X) follows a Uniform Distribution where the minimum is zero and the maximum is 12 gallons.
   a. Find the mean and median amount of gas in the tank.
   b. Find the variance and standard deviation of gas in the tank.
   c. Find the probability that there is more than 3 gallons in the tank.
   d. Find the probability that there is between 4 and 6 gallons in the tank.
   e. Find the probability that there is exactly 3 gallons in the tank
   f. Find the 80^{th} percentile of gas in the tank.

4. A normally distributed population of package weights has a *mean* of 63.5 g and a *standard deviation* of 12.2 g.

   a. What percentage of this population weighs 66 g or more?
   b. What percentage of this population weighs 41 g or less?
   c. What percentage of this population weighs between 41 g and 66 g?
   d. Find the 60^{th} percentile for distribution of weights.
   e. Find the three quartiles and the interquartile range.
   f. If you sample 49 packages, find the probability the sample mean is over 66 g. Compare this answer to part a.

5. Assume the expected waiting time until the next RM (Richter Magnitude) 7.0 or greater earthquake somewhere in California follows an exponential distribution with $\mu = 10$ years.

   a. Find the probability of waiting 10 or more years for the next RM 7.0 or greater earthquake.

   b. Determine the median waiting time until the next RM 7.0 or greater earthquake.

6. High Fructose Corn Syrup (HFCS) is a sweetener in food products that is linked to obesity and type II diabetes. The mean annual consumption in the United States in 2008 of HFCS was 60 lbs with a standard deviation of 20 lbs. Assume the population follows a Normal Distribution.

   a. Find the probability a randomly selected American consumes more than 50 lbs of HFCS per year.

   b. Find the probability a randomly selected American consumes between 30 and 90 lbs of HFCS per year.

   c. Find the 80th percentile of annual consumption of HFCS.

   d. In a sample of 40 Americans how many would you expect consume more than 50 pounds of HFCS per year.

   e. Between what two numbers would you expect to contain 95% of Americans HFCS annual consumption?

   f. Find the quartiles and Interquartile range for this population.

   g. A teenager who loves soda consumes 105 lbs of HFCS per year. Is this result unusual? Use probability to justify your answer.

7. State in your own words the 3 important parts of the Central Limit Theorem.

8. For women aged 18-24, systolic blood pressures (in mmHg) are normally distributed with $\mu=114.8$ and $\sigma=13.1$.

   a. Find the probability a woman aged 18-24 has systolic blood pressure exceeding 120.

   b. If 4 women are randomly selected, find the probability that their mean blood pressure exceeds 120.

   c. If 40 women are randomly selected, find the probability that their mean blood pressure exceeds 120.

   d. If the pdf for systolic blood pressure did NOT follow a normal distribution, would your answer to part c change? Explain.

9. The following data represents 20 random samples from a discrete uniform distribution S={1,2,3,4,5,6,7,8,9}. The sample mean $(\overline{X})$ was calculated for each group:

																					$(\overline{X}$
Sample 01	7	8	4	1	9	8	4	6	4	5	9	5	9	3	7	6	5	8	7	5	**6.00**
Sample 02	4	7	9	2	5	4	8	3	6	6	6	6	1	7	1	4	2	6	8	9	**5.20**
Sample 03	4	3	2	7	9	8	3	7	6	3	4	1	9	8	6	1	5	8	8	7	**5.45**
Sample 04	7	7	5	7	8	1	4	7	4	9	3	2	5	9	8	1	1	8	9	1	**5.30**
Sample 05	3	6	4	4	8	3	1	3	9	1	4	9	2	7	6	2	4	2	5	7	**4.50**
Sample 06	9	6	1	2	2	1	2	5	6	5	3	2	4	6	2	7	6	7	7	2	**4.25**
Sample 07	3	4	3	7	7	5	8	2	4	6	1	7	5	7	8	7	3	1	7	8	**5.15**
Sample 08	8	2	1	7	8	7	9	3	4	7	4	6	1	7	8	5	2	2	6	4	**5.05**
Sample 09	2	7	3	6	4	1	9	7	2	4	9	4	1	6	5	6	8	6	5	3	**4.90**
Sample 10	8	7	7	7	8	1	3	2	4	6	8	8	9	1	9	5	2	6	1	8	**5.50**
Sample 11	2	1	3	3	5	7	1	2	6	7	7	6	1	2	9	6	9	3	9	5	**4.70**
Sample 12	9	5	7	2	3	7	5	1	7	6	1	8	6	3	6	5	9	2	8	7	**5.35**
Sample 13	2	7	7	9	5	8	8	2	3	6	4	2	8	2	6	7	7	4	6	7	**5.50**
Sample 14	1	5	2	1	3	4	6	5	5	1	5	9	7	8	5	9	5	4	1	9	**4.75**
Sample 15	3	9	7	7	4	8	8	8	6	7	8	1	9	3	4	3	3	7	4	3	**5.60**
Sample 16	1	7	1	1	3	6	9	9	3	5	8	3	5	8	8	1	4	9	1	9	**5.05**
Sample 17	3	5	5	1	1	7	6	5	6	2	5	6	5	5	3	7	2	2	5	8	**4.45**
Sample 18	7	9	7	9	8	5	9	5	3	8	7	6	4	2	6	3	5	3	3	6	**5.75**
Sample 19	5	2	6	2	1	4	3	7	8	7	7	1	5	1	1	5	7	9	2	3	**4.30**
Sample 20	4	7	9	1	1	5	9	1	9	7	3	4	1	6	1	8	9	6	1	2	**4.70**

a. Consider the sample mean (last column) as a random variable and group the data into the following categories and make a histogram:

Interval for $(\overline{X})$	Frequency	Rel Freq
(4.05 to 4.50)		
(4.55 to 5.00)		
(5.05 to 5.50)		
(5.55 to 6.00)		
Total		

b. Describe the shape of the data

c. Calculate the sample mean and standard deviation of these 20 values.

d. For this discrete uniform distribution, $\mu = 5$ and $\sigma = 2.58$. Based on the Central Limit Theorem, what would the mean and standard deviation of the sample mean random variable be? How does this compare with sample mean and standard deviation results from part c?

Math 10 - Homework 5

1. The average number of years of post secondary education of employees in an industry is 1.5. A company claims that this *average* is higher for its employees. A random sample of 16 of its employees has an *mean* of 2.1 years of post secondary education with a *standard deviation* of 0.6 years.

    a. Find a 95% confidence interval for the **mean** number years of post secondary education for the company's employees. How does this compare with the industry value?

    b. Find a 95% confidence interval for the **standard deviation** of number years of post secondary education for the company's employees.

2. When polling companies report a margin of error, they are referring to a 95% confidence interval. Go to the website www.pollingreport.com and verify the stated margins of error for 2 polls.

Constructing Confidence Intervals In Exercises 3 and4 you are given the sample mean and the sample standard deviation. Assume the random variable is normally distributed and use a t-distribution to construct a 95% confidence interval for the population mean μ. What is the margin of error of the confidence interval?

3. ***Repair Costs: Microwaves*** In a random sample of five microwave ovens, the mean repair cost was $75.00 and the standard deviation was $12.50.

4. ***Repair Costs: Computers*** In a random sample of seven computers, the mean repair cost was $100.00 and the standard deviation was $42.50.

5. You did some research on repair costs of microwave ovens and found that the standard deviation is σ = $15. Repeat Exercise 3, using a **normal distribution** with the appropriate calculations for a standard deviation that is known. Compare the results.

6. ***Mini-Soccer Balls*** A soccer ball manufacturer wants to estimate the mean circumference of mini-soccer balls within 0.15 inch. Assume that the population of circumferences is normally distributed.

    (a) Determine the minimum sample size required to construct a 99% confidence interval for the population mean. Assume the population standard deviation is 0.20 inch.

    (b) Repeat part (a) using a standard deviation of 0.10 inch. Which standard deviation requires a larger sample size? Explain.

    (c) Repeat part (a) using a confidence level of 95%. Which level of confidence requires a larger sample size? Explain.

7. If all other quantities remain the same, how does the indicated change affect the minimum sample size requirement (Increase, Decrease or No Change)?

    (a) Increase in the level of confidence

    (b) Increase in the error tolerance

(c) Increase in the standard deviation

8. ***Stressful Travel:*** In a survey of 3224 U.S. adults, 1515 said flying is the most stressful form of travel. Construct a 95% confidence interval for the proportion of all adults who say flying is the most stressful form of travel.

9. ***Accidents and Alcohol:*** A study of 2008 traffic fatalities found that 800 of the fatalities were alcohol related. Find a 99% confidence interval for the population proportion and explain what it means.

10. ***Happy at Work?*** In a survey of 1003 U.S. adults, 662 would be happy spending the rest of their career with their current employer. Construct a 90% confidence interval for the proportion who would be happy staying with their current employer. Does this result surprise you?

11. ***Computer Repairs*** You wish to estimate, with 95% confidence and within 3.5% of the true population, the proportion of computers that need repairs or have problems by the time the product is three years old
    a. No preliminary estimate is available. Find the minimum sample size needed.

    b. Find the minimum sample size needed, using a prior study that found that 19% of computers needed repairs or had problems by the time the product was three years old.

    c. Compare the results from parts (a) and (b).

12. ***Lawn Mower*** A lawn mower manufacturer is trying to determine the **standard deviation** of the life of one of its lawn mower models. To do this, it randomly selects 12 lawn mowers that were sold several years ago and finds that the sample standard deviation is 3.25 years. Use a 99% level of confidence to find a confidence interval for **standard deviation.**

13. ***Monthly Income*** The monthly incomes of 20 randomly selected individuals who have recently graduated with a bachelor's degree in social science have a sample standard deviation of $107. Use a 95% level of confidence to find a confidence interval for **standard deviation.**

14. Read the attached article on the CBS News poll regarding the birth control pill.

   a. What would the point estimator be for the proportion of adults who believe the pill has made women's lives better.

   b. What is the sample size for this study?

   c. What is the margin of error for this poll as reported in the article. Assuming a 95% level of confidence, verify this poll by calculation.

May 7, 2010

Poll: Most Say The Pill Improved Women's Lives

Most Americans Consider “The Pill” One of the Most Significant Medical Advances of the Last 50 Years, a CBS News Poll Finds

(CBS)

(CBS) *Poll analysis by Jennifer De Pinto.*

More than half the public -- including most women -- believes the birth control pill has been one of the most significant medical developments of the last half century, a new CBS News poll finds.

Most Americans say "the pill" has had an impact on American society and on women’s lives in particular, and credit it with helping women enter the work force.

The birth control pill was approved by the Food and Drug Administration in 1960. Today, 52 percent of Americans say it has been one of the most significant medical developments of the last 50 years, according to the poll, conducted on May 4th and 5th.

Four in five Americans think the birth control pill has had at least some effect on American society overall, including 41 percent who say it’s had a great deal of impact.

Even more, 54 percent, think the birth control pill has had a great deal of impact on women’s lives in particular.

The Pill: Women’s Lives Made Better

Most Americans say women’s lives were changed for the better because of the birth control pill. Only a quarter think it made no difference, and even fewer say the pill made women’s lives worse.

Men (59 percent), women (54 percent), and women who have ever taken the pill (54 percent) say that women’s lives were improved as a result of the birth control pill.

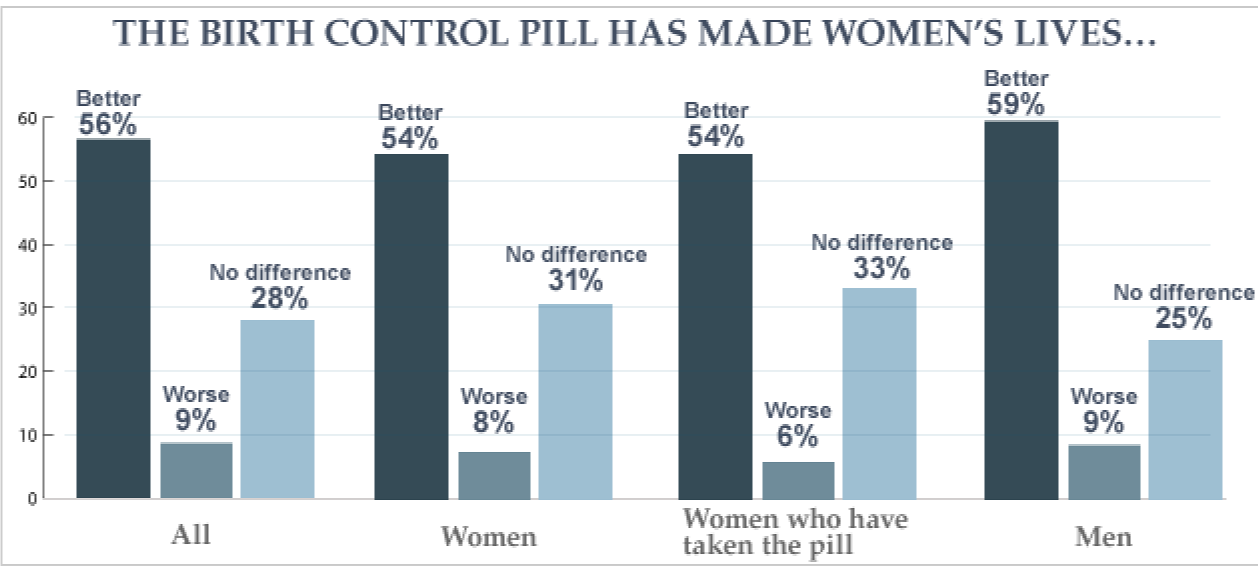
More specifically, Americans think the birth control pill helped women enter the work force: 57

percent say the pill made it easier for women to have jobs and careers outside the home.

That number rises to 69 percent among Americans age 45 and over -- an age group more likely to have felt the impact of the pill when it was first developed and put on the market. Among women age 45 and older that figure is 64 percent.

By contrast, 53 percent of younger Americans say the birth control pill had no effect on the ability of women to work outside the home.

Among working women, 55 percent say the birth control poll has made it easier for women to enter the workforce.



(CBS)

Family Life and Attitudes Toward Sex

Roughly half of Americans say the birth control pill has improved American family life, while a third doesn't think it has had much effect.

Religion has some impact on these views. Among Catholics, whose church opposes non-natural forms of birth control, just 38 percent believe the birth control pill has improved American family life. That figure is 52 percent among Protestants.

Eight in ten Americans think the birth control pill has affected Americans' attitudes toward sex, including 51 percent who say it impacted those attitudes a great deal.

The Pill: Safety and Effectiveness

The poll finds public concerns about the safety of the birth control pill have diminished over time.

In 1966, six years after the pill was approved by the FDA, fewer than half of Americans - 43 percent - told a Gallup Poll that birth control pills could be used safely without danger to a person's health.

That number has risen to 64 percent today.

Among women, 58 percent now think the birth control pill can be used safely, as do a similar percentage of women who have ever taken it.

Nearly half of women think the birth control pill is just as safe as other forms of birth control, and another 20 percent believe the pill is safer. Still, one in five thinks it is less safe. Views are similar among women who have ever taken birth control pills.

More than eight in 10 Americans (including 82 percent of women) say birth control pills are effective. In a 1966 Gallup Poll, a smaller number of Americans (though still a 61 percent majority) thought the birth control pill was effective.

Some medical research has been done on a contraceptive for men similar to that of the birth control pill. A majority of women do *not* think most men would take birth control pills if they were available.

In contrast, two-thirds of men think most men would take the pill if it were available.

This poll was conducted among a random sample of 591 adults nationwide, interviewed by telephone May 4-5, 2010. Phone numbers were dialed from random digit dial samples of both standard land-line and cell phones. The error due to sampling for results based on the entire sample could be plus or minus four percentage points. The error for subgroups is higher.

This poll release conforms to the Standards of Disclosure of the National Council on Public Polls.

Math 10 - Homework 6

Part A

1. What are the two types of hypotheses used in a hypothesis test? How are they related?

2. Describe the two types of error possible in a hypothesis test decision.

True or False?
In Exercises 3-8, determine whether the statement is true of false. If it is false, rewrite it as a true statement.

3. In a hypothesis test, you assume the alternative hypothesis is true.

4. A statistical hypothesis is a statement about a sample.

5. If you decide to reject the null hypothesis, you can support the alternative hypothesis.

6. The level of significance is the maximum probability you allow for rejecting a null hypothesis when it is actually true.

7. A large P-value in a test will favor a rejection of the null hypothesis.

8. If you want to support a claim, write it as your null hypothesis.

Stating Hypotheses
In Exercises 9-14, use the given statement to represent a
claim. Write its complement and state which is Ho and which is *Ha.*

9. $p > .65$

10. $\mu \leq 128$

11. $\sigma^2 \neq 5$

12. $\mu = 1.2$

13. $p \geq 0.45$

14. $\sigma < 0.21$

Think about the context of the claim. *Determine* whether you want to support or reject the claim.

a. State the null and alternative hypotheses in words.
b. Write the null and alternative hypotheses in appropriate symbols
c. Describe in words Type I error (the consequence of rejecting a true null hypothesis.)
d. Describe in words Type II error (the consequence of failing to reject a false null hypothesis.)

15. You represent a chemical company that is being sued for paint damage to automobiles. You want to support the claim that the mean repair cost per automobile is about \$650. How would you write the null and alternative hypotheses?

16. You are on a research team that is investigating the mean temperature of adult humans. The commonly accepted claim is that the mean temperature is about 98.6°F. You want to show that this claim is false. How would you write the null and alternative hypotheses?

17. A light bulb manufacturer claims that the mean life of a certain type of light bulb is at least 750 hours. You are skeptical of this claim and want to refute it.

18. As stated by a company's shipping department, the number of shipping errors per million shipments has a standard deviation that is less than 3. Can you support this claim?

19. A research organization reports that 33% of the residents in Ann Arbor, Michigan are college students. You want to reject this claim.

20. The results of a recent study show that the proportion of people in the western United States who use seat belts when riding in a car or truck is under 84%. You want to support this claim.

PART B – Hypothesis Testing Procedure

21. In your work for a national health organization, you are asked to monitor the amount of sodium in a certain brand of cereal. You find that a random sample of 82 cereal servings has a mean sodium content of 232 milligrams with a standard deviation of 10 milligrams. At $\alpha = 0.01$, can you conclude that the mean sodium content per serving of cereal is over 230 milligrams?

(a) (DESIGN) State your Hypothesis	**(d) (DESIGN)** Determine decision rule (pvalue method) **(e) (DATA)** Conduct the test and **circle** your decision Reject Ho Fail to Reject Ho **(f) (CONCLUSION)** State your overall conclusion in language that is clear, relates to the original problem and is consistent with your decision.
(b) (DESIGN) State Significance Level of the test and explain what it means,	
(c) (DESIGN) Determine the statistical model (test statistic)	

22. A tourist agency in Florida claims the mean daily cost of meals and lodging for a family of four traveling in Florida is $284. You work for a consumer protection advocate and want to test this claim. In a random sample of 50 families of four traveling in Florida, the mean daily cost of meals and lodging is $292 and the standard deviation is $25. At $\alpha = 0.05$, do you have enough evidence to reject the agency's claim?

(a) (DESIGN) State your Hypothesis	**(d) (DESIGN)** Determine decision rule (critical value method) **(e) (DATA)** Conduct the test and **circle** your decision Reject Ho Fail to Reject Ho **(f) (CONCLUSION)** State your overall conclusion in language that is clear, relates to the original problem and is consistent with your decision.
(b) (DESIGN) State Significance Level of the test and explain what it means.	
(c) (DESIGN) Determine the statistical model (test statistic)	

23. An environmentalist estimates that the mean waste recycled by adults in the United States is more than 1 pound per person per day. You want to test this claim. You find that the mean waste recycled per person per day for a random sample of 12 adults in the United States is 1.2 pounds and the standard deviation is 0.3 pound. At a = 0.05, can you support the claim?

(d) (DESIGN) State your Hypothesis	**(d) (DESIGN)** Determine decision rule (critical value method)
(e) (DESIGN) State Significance Level of the test and explain what it means.	**(e) (DATA)** Conduct the test and **circle** your decision Reject Ho Fail to Reject Ho
(f) (DESIGN) Determine the statistical model (test statistic)	**(f) (CONCLUSION)** State your overall conclusion in language that is clear, relates to the original problem and is consistent with your decision.

24. A government association claims that 44% of adults in the United States do volunteer work. You work for a volunteer organization and are asked to test this claim. You find that in a random sample of 1165 adults, 556 do volunteer work. At $a = 0.05$, do you have enough evidence to reject the association's claim?

<table>
<tr>
<td>(a) (DESIGN) State your Hypothesis</td>
<td rowspan="3">(d) (DESIGN) Determine decision rule (pvalue method)

(e) (DATA) Conduct the test and circle your decision

Reject Ho   Fail to Reject Ho

(f) (CONCLUSION) State your overall conclusion in language that is clear, relates to the original problem and is consistent with your decision.</td>
</tr>
<tr>
<td>(b) (DESIGN) State Significance Level of the test and explain what it means,</td>
</tr>
<tr>
<td>(c) (DESIGN) Determine the statistical model (test statistic)</td>
</tr>
</table>

PART C – total hypothesis testing

25. The geyser Old Faithful in Yellowstone National Park is claimed to erupt for on average for about three minutes. Thirty-six observations of eruptions of the Old Faithful were recorded (time in minutes)

1.8	1.98	2.37	3.78	4.3	4.53
1.82	2.03	2.82	3.83	4.3	4.55
1.88	2.05	3.13	3.87	4.43	4.6
1.9	2.13	3.27	3.88	4.43	4.6
1.92	2.3	3.65	4.1	4.47	4.63
1.93	2.35	3.7	4.27	4.47	6.13

Sample mean = 3.394 minutes. Sample standard deviation = 1.168 minutes

Test the hypothesis that the mean length of time for an eruption is 3 minutes answering ALL the following questions:

A. General Question
    a. Why do you think this test is being conducted?
B. Design
    a. State the null and alternative hypotheses
    b. What is the appropriate test statistic/model?
    c. What is significance level of the test?
    d. What is the decision rule?
C. Conduct the test
    a. Are there any unusual observations that question the integrity of the data or the assumptions of the model? (additional problem only)
    b. Is the decision to reject or fail to reject Ho?
D. Conclusions - State a one paragraph conclusion that is consistent with the decision using language that is clearly understood in the context of the problem. Address any potential problems with the sampling methods and address any further research you would conduct.

PART D– Definitions and Power

26. Define the following terms:

    a. Parameter
    b. Statistic
    c. Statistical Inference
    d. Hypothesis
    e. Hypothesis Testing
    f. Null Hypothesis (Ho)
    g. Alternative Hypothesis (Ha)
    h. Type I Error
    i. Type II Error
    j. Level of Significance (α)
    k. Beta (β)
    l. Statistical Model
    m. Test Statistic
    n. Model Assumptions
    o. Critical value(s)
    p. Rejection Region
    q. p-value
    r. Decision Rule
    s. Power
    t. Effect Size

27. A study claims more than 60% of students text-message frequently. In a poll of 1000 students, 660 students said they text message frequently. Can you support the study's claim? Conduct the test with α = 1%

28. 15 I-pod users were asked how many songs were on their I-pod. Here are the summary statistics of that study:

    $\overline{X} = 650$ $\qquad$ $s = 200$

    a. Can you support the claim that the number of songs on a user's I-pod is different from 500? Conduct the test with α= 5% .

    b. Can you support the claim that the population standard deviation is under 300? Conduct the test with α= 5% .

29. Consider the design procedure in the test you conducted in Question 28a. Suppose you wanted to conduct a Power analysis if the population mean under Ha was actually 550. Use the online Power calculator to answer the following questions.

    a. Determine the Power of the test.

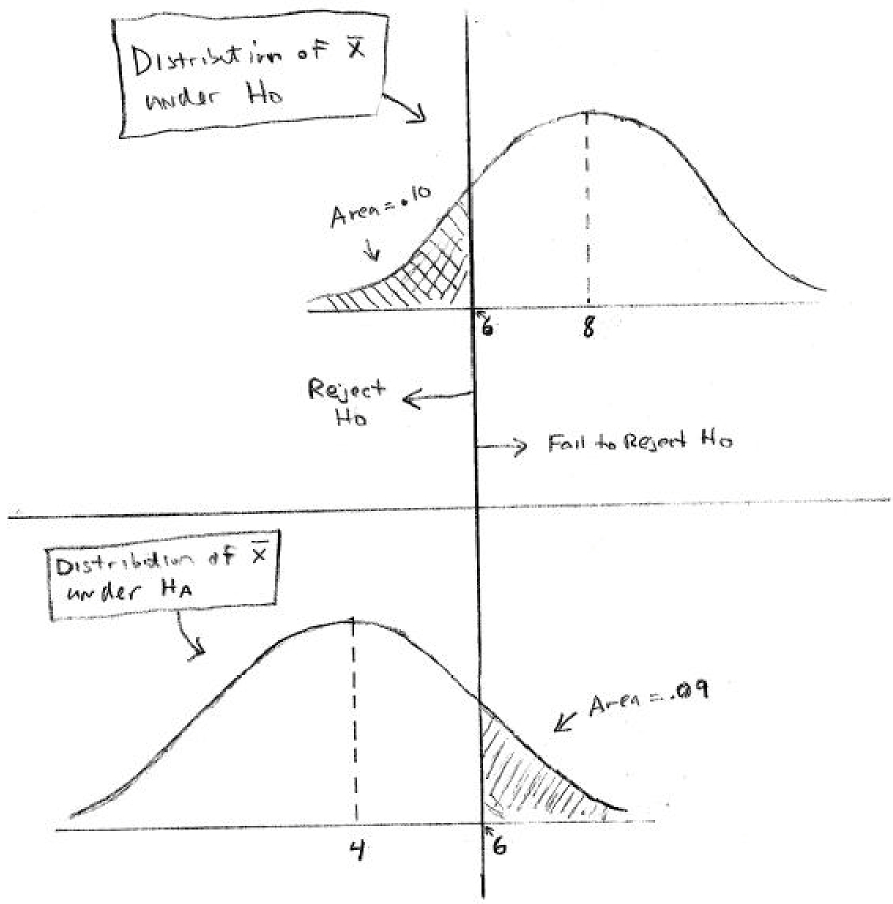
    b. Determine Beta.

    c. Determine the sample size needed if you wanted to conduct the test in Question 28a with 95% power.

30. The drawing shown diagrams a hypothesis test for population mean design under the Null Hypothesis (top drawing) and a specific Alternative Hypothesis (bottom drawing). The sample size for the test is 200.

a. State the Null and Alternative Hypotheses

b. What are the values of μ_0 and μ_a in this problem?



c. What is the significance level of the test?

d. What is the Power of the test when the population mean = 4?

e. Determine the probability associated with Type I error.

f. Determine the probability associated with Type II error.

g. Under the Null Hypothesis, what is the probability the sample mean will be over 6?

h. If the significance level was set at 5%, would the power increase, decrease or stay the same?

i. If the test was conducted, and the p-value was .085, would the decision be Reject or Fail to Reject the Null Hypothesis?

j. If the sample size was changed to 100, would the shaded on area on the bottom (Ha) graph increase, decrease or stay the same?

Math 10 - Homework 7

1. What is the difference between two samples that are dependent and two samples that are independent? Give an example of two dependent samples and two independent samples.

2. What conditions are necessary in order to use the dependent samples t-test for the mean of the difference of two populations?

In Problems 3-10, classify the two given samples as independent or dependent. Explain your reasoning.

3. Sample 1: The SAT scores for 35 high school students who did not take an SAT preparation course
Sample 2: The SAT scores for 40 high school students who did take an SAT preparation course

4. Sample 1: The SAT scores for 44 high school students
Sample 2: The SAT scores for the same 44 high school students after taking an SAT preparation course

5. Sample 1: The weights of 51 adults
Sample 2: The weights of the same 51 adults after participating in a diet and exercise program for one month

6. Sample 1: The weights of 40 females
Sample 2: The weights of 40 males

7. Sample 1: The average speed of 23 powerboats using an old hull design
Sample 2: The average speed of 14 powerboats using a new hull design

8. Sample 1: The fuel mileage of 10 cars
Sample 2: The fuel mileage of the same 10 cars using a fuel additive

9. The table shows the braking distances (in feet) for each of four different sets of tires with the car's anti-lock braking system (ABS) on and with ABS off. The tests were done on ice with cars traveling at 15 miles per hour.

Tire Set	**1**	**2**	**3**	**4**
Braking distance with ABS	**42**	**55**	**43**	**61**
Braking distance without ABS	**58**	**67**	**59**	**75**

10. The table shows the heart rates (in beats per minute) of five people before exercising and after.

Person	**1**	**2**	**3**	**4**	**5**
Heart Rate before Exercising	**42**	**55**	**43**	**61**	**65**
Heart Rate after Exercising	**58**	**67**	**59**	**75**	**90**

11. In a study testing the effects of an herbal supplement on blood pressure DATA in men, 11 randomly selected men were given an herbal supplement for 15 weeks. The following measurements are for each subject's diastolic blood pressure taken before and after the 15-week treatment period. At $\alpha = .10$, can you support the claim that systolic blood pressure was lowered?

(a) (DESIGN) State your Hypothesis

(b) (DESIGN) State Significance Level of the test and explain what it means,

(c) (DESIGN) Determine the statistical model (test statistic)

(d) DESIGN) Determine decision rule (pvalue method)

(e) (DATA) Conduct the test and **circle** your decision

Before	After
123	124
109	97
112	113
102	105
98	95
114	119
119	114
112	114
110	121
117	118
130	133

Hypothesis Test: Paired Observations

0.000	hypothesized value
113.273	mean Before
113.909	mean After
-0.636	mean difference (Before - After)
5.870	std. dev.
1.770	std. error
11	n
10	df
-0.36	t
.6367	p-value (one-tailed, upper)
.3633	p-value (one-tailed, lower)
.7266	p-value (two-tailed)

Reject Ho **Fail to Reject Ho**

(f) (CONCLUSION) State your overall conclusion in language that is clear, relates to the original problem and is consistent with your decision.

12. A random sample of 25 waiting times (in minutes) before patients saw a medical professional in a hospital's minor emergency department had a standard deviation of 0.7 minute. After a new admissions procedure was implemented, a random sample of 21 waiting times had a standard deviation of 0.5 minute. At $\alpha = .10$, can you support the hospital's claim that the standard deviation of the waiting times has decreased?

(a) (DESIGN) State your Hypothesis

(b) (DESIGN) State Significance Level of the test and explain what it means.

(c) (DESIGN) Determine the statistical model (test statistic)

(d) (DESIGN) Determine decision rule (critical value method)

(e) (DATA) Conduct the test and **circle** your decision

Reject Ho   Fail to Reject Ho

(f) (CONCLUSION) State your overall conclusion in language that is clear, relates to the original problem and is consistent with your decision.

13. An engineer wants to compare the tensile strengths of steel bars that are produced using a conventional method and an experimental method. (The tensile strength of a metal is a measure of its ability to resist tearing when pulled lengthwise.) To do so, the engineer randomly selects steel bars that are manufactured using each method and records the following tensile strengths (in Newtons per square millimeter). At $\alpha = .10$, can the engineer claim that the experimental method produces steel with greater mean tensile strength? Should the engineer recommend using the experimental method? First use the F test to determine whether or not to use equal variances in choosing the model.

Experimental	395	389	421	394	407	411	389	402	422	416	402	408	400	386	411	405	389
Conventional	362	352	380	382	413	384	400	378	419	379	384	388	372	383			

Hypothesis Test: Independent Groups (t-test, pooled variance)

Experimental	Conventional	
402.76	384.00	mean
11.34	17.70	std. dev.
17	14	n

29	df
18.765	difference (Experimental - Conventional)
211.416	pooled variance
14.540	pooled std. dev.
5.248	standard error of difference
0	hypothesized difference
3.58	t
.0012	p-value (two-tailed)
.0006	p-value (one-tailed, upper)
.9994	p-value (one-tailed, lower)

Hypothesis Test: Independent Groups (t-test, unequal variance)

Experimental	onventional	
402.76	384.00	mean
11.34	17.70	std. dev.
17	14	n

21	df
18.765	difference (Experimental - Conventional)
5.472	standard error of difference
0	hypothesized difference
3.43	t
.0025	p-value (two-tailed)
.0013	p-value (one-tailed, upper)
.9987	p-value (one-tailed, lower)

F-test for equality of variance

313.23	variance: Conventional
128.69	variance: Experimental
2.43	F
.0944	p-value

(a) (DESIGN) State your Hypothesis	**(d) (DESIGN)** Determine decision rule (pvalue method) **(e) (DATA)** Conduct the test and **circle** your decision Reject Ho Fail to Reject Ho **(f) (CONCLUSION)** State your overall conclusion in language that is clear, relates to the original problem and is consistent with your decision.
(b) (DESIGN) State Significance Level of the test and explain what it means,	
(c) (DESIGN) Determine the statistical model (test statistic)	

Math 10 - Homework 8

1. A bicycle safety organization claims that fatal bicycle accidents are uniformly distributed throughout the week. The table shows the day of the week for which 911 randomly selected fatal bicycle accidents occurred. At α= 0.10, can you reject the claim that the distribution is uniform?

(a) (DESIGN) State your Hypothesis

(b) (DESIGN) State Significance Level of the test and explain what it means.

(c) (DESIGN) Determine the statistical model . Determine decision rule (critical value method)

(d) (DATA) Conduct the test and **circle** your decision

Survey	**Observe**	p_i	**Expected**	**ChiSq**
Sunday	118			
Monday	119			
Tuesday	127			
Wednesday	137			
Thursday	129			
Friday	146			
Saturday	135			
Total	**911**			

Reject Ho Fail to Reject Ho

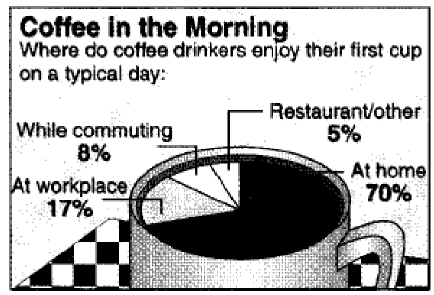
(e) (CONCLUSION) State your overall conclusion in language that is clear, relates to the original problem and is consistent with your decision.

2. Results from a survey five years ago asking where coffee drinkers typically drink their first cup of coffee are shown in the graph. To determine whether this distribution has changed, you randomly select 581 coffee drinkers and ask each where they typically drink their first cup of coffee. The results are shown in the table. Can you conclude that there has been a change in the claimed or expected distribution? Use α= 0.05.

(a) (DESIGN) State your Hypothesis

(b) (DESIGN) State Significance Level of the test and explain what it means.

(c) (DESIGN) Determine the statistical model . Determine decision rule (critical value method)



(d) (DATA) Conduct the test and **circle** your decision

Survey	Observe	p_i	Expected	ChiSq
Home	389			
Work	110			
Commute	55			
Rest/Other	27			
Total	**581**			

Reject Ho   Fail to Reject Ho

(e) (CONCLUSION) State your overall conclusion in language that is clear, relates to the original problem and is consistent with your decision

3. In a recent SurveyUSA poll, 500 Americans adults were asked if marijuana should be legalized. The results of the poll were cross tabulated as shown in the contingency tables below. Conduct **two** tests for independence to determine if opinion about legalization of marijuana is dependent on gender or age

	Male	Female
Should be Legal	123	90
Should Not be Legal	127	160

	18-34	35-54	55+
Should be Legal	95	83	48
Should Not be Legal	65	126	83

4. 1000 American adults were recently polled on their opinion about effect of recent stimulus bill and the economy. The results are shown in the following contingency table, broken down by gender:

	Stimulus will hurt economy	**Stimulus will help the economy**	**Stimulus will have no effect**	**TOTAL**
Male	150	150	200	**500**
Female	100	200	200	**500**
TOTAL	**250**	**350**	**400**	**1000**

a) Are gender and opinion on the stimulus dependent variables? Test using α =1%.
b) Give a possible explanation for the conclusion you came up with in part a.

5. A clinical psychologist completed a study on hyperactivity in children using one-way ANOVA. The model was balanced with **5 replicates per treatment**. The factor was 3 types of school district (urban, rural and suburban). Unfortunately, hackers broke into the psychologist's computer and wiped out all the data. All that remained was a fragment of the ANOVA table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F statistic	Critical Value of F for α= .05	Decision
Factor	7000					
Error						
Total	9000					

Fill in the table and conduct the hypotheses test that compares mean level of hyperactivity in the 3 types of districts. **Explain your results.**

6. A sociologist was interested in commute time for workers in the Bay Area. She categorized commuters by 4 regions (North Bay, South Bay, East Bay and Peninsula) and designed a balanced model with 8 replicates per region. Data is round trip commute time in minutes . The results and ANOVA output are shown on the next page:

   a. Test the Null Hypothesis that all regions have the same mean commute time at a significance level of 5%. State your decision in non-statistical language.

   b. Conduct **all** pairwise comparisons at an overall significance level of 5%.

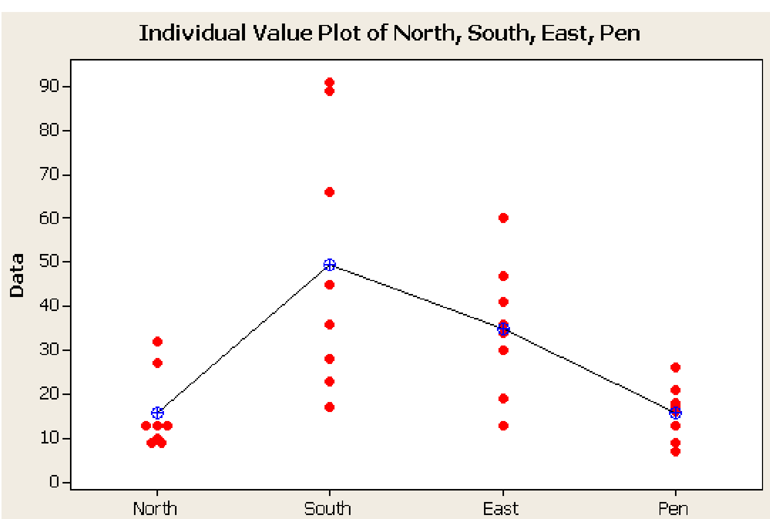
   c. Explain the results of this experiment as if you were addressing a transportation committee. What would you recommend?

MINITAB OUTPUT

North	South	East	Pen
13	91	41	17
9	45	30	16
10	28	60	13
13	17	34	26
27	89	47	7
13	36	13	9
9	23	19	21
32	66	36	18

One-way ANOVA: North, South, East, Pen

```
Source  DF     SS    MS     F      P
Factor   3   6392  2131  7.14  0.001
Error   28   8356   298
Total   31  14748

S = 17.28   R-Sq = 43.34%   R-Sq(adj) = 37.27%


                        Individual 95% CIs For Mean Based on
                        Pooled StDev
Level  N   Mean  StDev  --------+---------+---------+---------+-
North  8  15.75   8.76  (--------*-------)
South  8  49.38  29.22                       (-------*-------)
East   8  35.00  14.99             (-------*--------)
Pen    8  15.88   6.20  (--------*-------)
                        --------+---------+---------+---------+-
                               15        30        45        60

Pooled StDev = 17.28

Grouping Information Using Tukey Method

       N   Mean  Grouping
South  8  49.38  A
East   8  35.00  A B
Pen    8  15.88    B
North  8  15.75    B

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



North South East Pen

7. People who are concerned about their health may prefer hot dogs that are low in salt and calories. The data contains data on the calories and sodium contained in each of 54 major hot dog brands. The hot dogs are classified by type: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat). Minitab output is attached for two different hypothesis tests.

A test for a difference in **calories** due to hot dog type will be performed.

i. Design the test.

ii. Fill in the missing information in the ANOVA table on the next page.

iii. Conduct the test with an overall confidence level of 5%, including pairwise comparisons.

One-way ANOVA: Calories versus Type

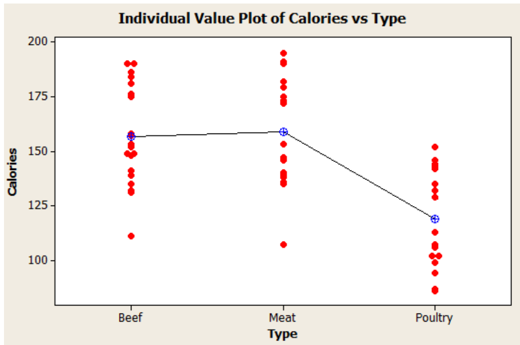
Source	DF	SS	MS	F	p-value
Type	_______	17692	__________	__________	0.000
Error	_______	28067	__________		
Total	_______	45759			

112 128 144 160

Grouping Information Using Tukey Method

Type	N	Mean	Grouping
Meat	17	158.71	A
Beef	20	156.85	A
Poultry	17	118.76	B

Means that do not share a letter are significantly different.



Math 10 – Homework 9 Name_________________________________

Q1 A manager is concerned that overtime (measured in hours) is contributing to more sickness (measured in sick days) among the employees. Data records for 10 employees were sampled with the following results:

Partial Computer Output are attached on next page

a) Find the least square line where Sick Days is dependent on Overtime. Interpret the slope.

b) Test the hypothesis that the regression model is significant (α = .10)

c) Find and interpret the r^2, coefficient of determination. (Blank Box)

d) Find he estimate of standard deviation of the residual error. (Blank Box)

e) What would your prediction of sick days be for an employee who works 100 hours overtime.

f) Analyze the residuals and determine which pair of data is the most unusual.

g) Explain why this model would not be appropriate for an employee who works 500 hours overtime.

Overtime	20	25	35	39	43	55	67	113	135	155
Sick Days	0	0	2	7	3	5	4	11	7	9

Regression Analysis

r^2		n	10
r		k	1
Std. Error		Dep. Var.	**SickDays**

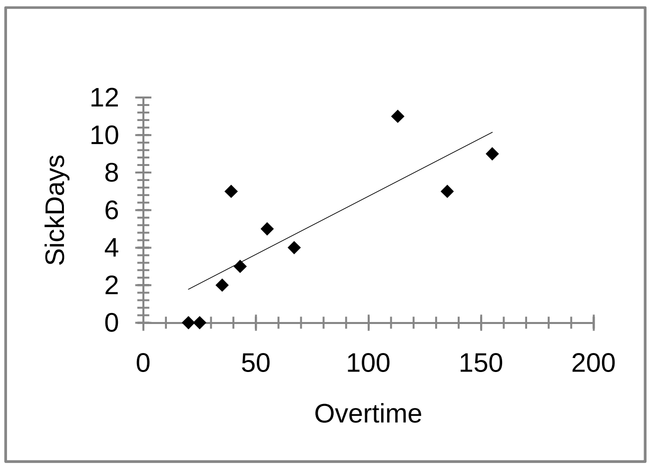
ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Regression	80.6944	1	80.6944	15.05	.0047
Residual	42.9056	8	5.3632		
Total	123.6000	9			

Regression output

					confidence interval	
variables	*coefficients*	*std. error*	*t (df=8)*	*p-value*	*95% lower*	*95% upper*
Intercept	0.5369	1.3207	0.407	.6950	-2.5086	3.5824
Overtime	0.0621	0.0160	3.879	.0047	0.0252	0.0989

Observation	*SickDays*	*Predicted*	*Residual*
1	0.0	1.8	-1.8
2	0.0	2.1	-2.1
3	2.0	2.7	-0.7
4	7.0	3.0	4.0
5	3.0	3.2	-0.2
6	5.0	3.9	1.1
7	4.0	4.7	-0.7
8	11.0	7.5	3.5
9	7.0	8.9	-1.9
10	9.0	10.2	-1.2



Predicted values for: SickDays

		95% Confidence Intervals		*95% Prediction Intervals*		
Overtime	*Predicted*	*lower*	*upper*	*lower*	*upper*	*Leverage*
100	6.742	4.696	8.788	1.023	12.461	0.147
500	31.564	15.563	47.564	14.696	48.432	8.977