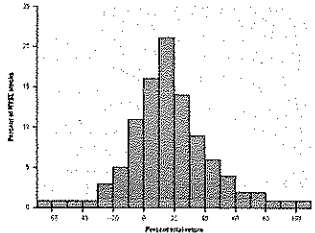
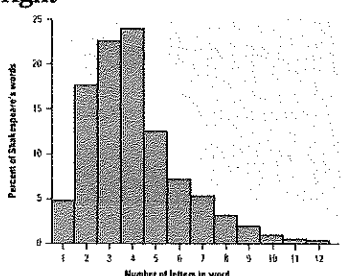
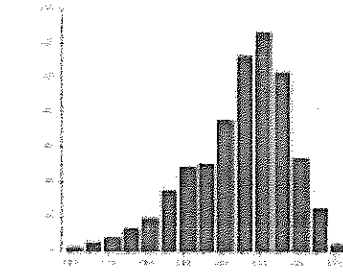
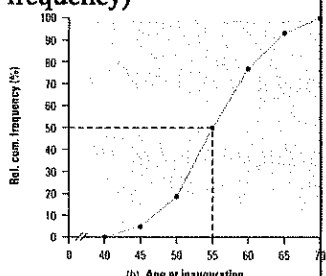
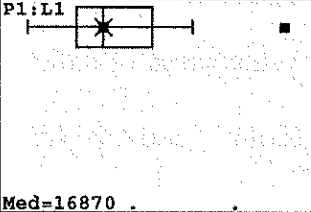
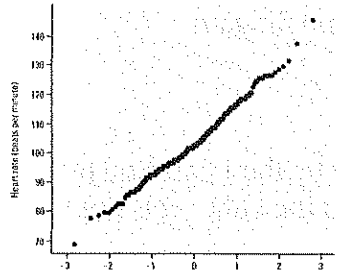

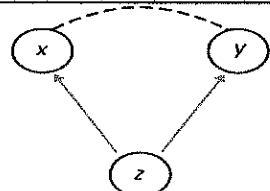
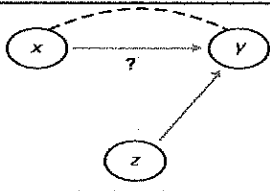


## Important Concepts not on the AP Statistics Formula Sheet

### Part I:

<p>IQR = <math>Q_3 - Q_1</math>                      Test for an outlier:                      1.5(IQR) above <math>Q_3</math> or below <math>Q_1</math>                      The calculator will run the test for you as long as you choose the boxplot with the outlier on it in STATPLOT</p>	<p>Linear transformation:  <b>Addition:</b> affects center NOT spread                      adds to <math>\bar{x}</math>, M, <math>Q_1</math>, <math>Q_3</math>, IQR                       not <math>\sigma</math>  <b>Multiplication:</b> affects both center and spread                      multiplies <math>\bar{x}</math>, M, <math>Q_1</math>, <math>Q_3</math>, IQR, <math>\sigma</math></p>	<p>When describing data:                      describe center, spread, and shape.                       Give a 5 number summary or mean and standard deviation when necessary.</p>	<p>Histogram:                      fairly symmetrical unimodal</p> 
<p>skewed right</p> 	<p>Skewed left</p> 	<p>Ogive (cumulative frequency)</p> 	<p>Boxplot (with an outlier)</p> 
<p>Stem and leaf</p> <p>Treasury bills</p> <pre> 0   9 1   0 2 5 5 6 6 6 8 2   1 5 7 7 9 3   0 1 1 3 5 5 8 9 9 4   2 4 7 7 8 5   1 1 2 2 2 5 6 6 7 8 7 9 6   2 4 5 6 9 7   2 7 8 8   0 4 8 9   8 10   4 5 11   3 12   13   14   7                     </pre> <p>(b)</p>	<p>Normal Probability Plot</p>  <p>The 80<sup>th</sup> percentile means that 80% of the data is below that observation.</p>	$z = \frac{x - \text{mean}}{\text{standard dev}}$ <p style="text-align: center;">or</p> $z = \frac{x - \mu}{\sigma}$ <p>HOW MANY STANDARD DEVIATIONS AN OBSERVATION IS FROM THE MEAN</p> <p>68-95-99.7 Rule for Normality  <math>N(\mu, \sigma)</math>  <math>N(0, 1)</math> Standard Normal</p>	<p><math>r</math>: correlation coefficient, The strength of the linear relationship of data. Close to 1 or -1 is very close to linear</p> <p><math>r^2</math>: coefficient of determination. How well the model fits the data. Close to 1 is a good fit. "Percent of variation in y described by the LSRL on x"</p>
<p>residual = <math>y - \hat{y}</math></p> <p>residual = observed - predicted</p> <p><math>y = a + bx</math>                      Slope of LSRL(b): rate of change in y for every unit x</p> <p>y-intercept of LSRL(a): y when x = 0</p>	<p>Exponential Model:  <math>y = ab^x</math> take log of y</p> <p>Power Model:  <math>y = ax^b</math> take log of x and y</p>	<p>Explanatory variables explain changes in response variables.                      EV: x, independent                      RV: y, dependent</p>	<p>Lurking Variable: A variable that may influence the relationship between two variables. LV is not among the EV's</p>
<p>Confounding: two variables are confounded when the effects of an RV cannot be distinguished.</p>	 <p style="text-align: center;">Causation (a)</p>	 <p style="text-align: center;">Common response (b)</p>	 <p style="text-align: center;">Confounding (c)</p>

## Regression in a Nutshell

Given a Set of Data:

Data:

NEA change (cal):	-94	-57	-29	135	143	151	245	355
Fat gain (kg):	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
NEA change (cal):	392	473	486	535	571	580	620	690
Fat gain (kg):	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

Enter Data into L<sub>1</sub> and L<sub>2</sub> and run **8:Linreg(a+bx)**

The regression equation is:

$$\text{predicted fat gain} = 3.5051 - 0.00344(\text{NEA})$$

**y-intercept:** Predicted fat gain is 3.5051 kilograms when NEA is zero.

**slope:** Predicted fat gain decreases by .00344 for every unit increase in NEA.

**r: correlation coefficient**

$$r = -0.778$$

Moderate, negative correlation between NEA and fat gain.

**r<sup>2</sup>: coefficient of determination**

$$r^2 = 0.606$$

60.6% of the variation in fat gained is explained by the Least Squares Regression line on NEA.

The linear model is a moderate/reasonable fit to the data. It is not strong.

**The residual plot** shows that the model is a reasonable fit; there is not a bend or curve, There is approximately the same amount of points above and below the line. There is No fan shape to the plot.

**Predict the fat gain that corresponds to a NEA of 600.**

$$\text{predicted fat gain} = 3.5051 - 0.00344(600)$$

$$\text{predicted fat gain} = 1.4411$$

**Would you be willing to predict the fat gain of a person with NEA of 1000?**

No, this is extrapolation, it is outside the range of our data set.

**Residual:**

observed y - predicted y

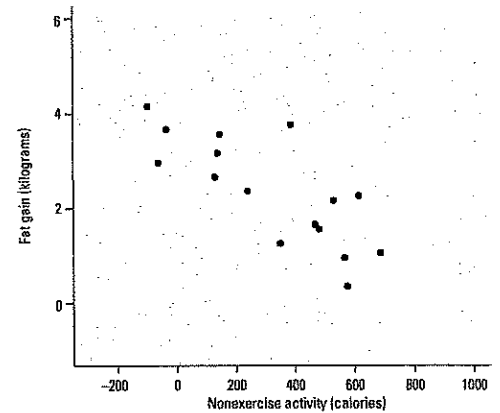
**Find the residual for an NEA of 473**

First find the predicted value of 473:

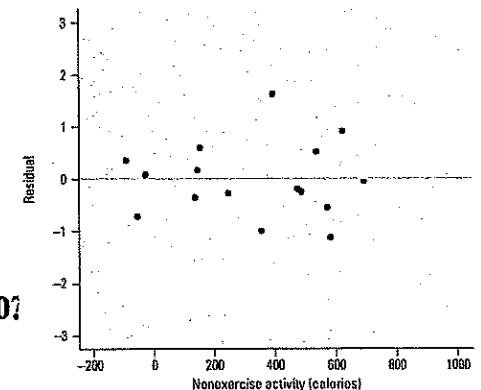
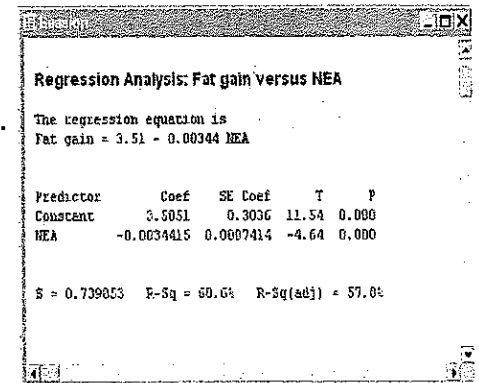
$$\text{predicted fat gain} = 3.5051 - 0.00344(473)$$

$$\text{predicted fat gain} = 1.87798$$

$$\text{observed} - \text{predicted} = 1.7 - 1.87798 = -0.17798$$



Minitab



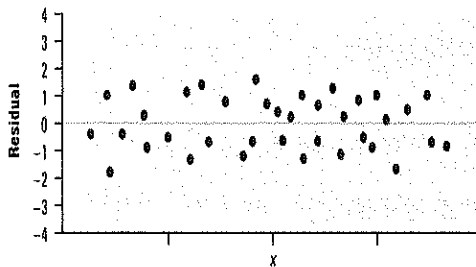
**Transforming Exponential Data  $y = ab^x$**

Take the log or ln of y.  
The new regression equation is:  
 $\log(y) = a + bx$

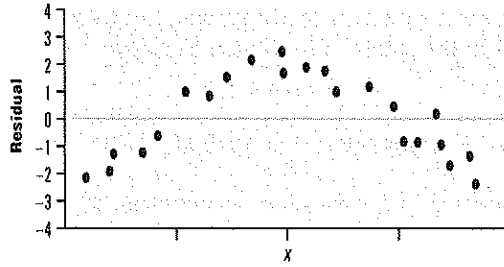
**Transforming Power Data  $y = ax^b$**

Take the log or ln of x and y.  
The new regression equation is:  
 $\log(y) = a + b \log(x)$

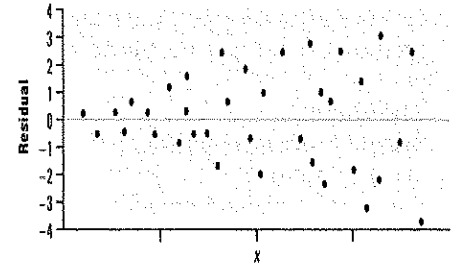
**Residual Plot examples:**



Linear model is a Good Fit



Curved Model would be a good fit



Fan shape loses accuracy as x increases

**Inference with Regression Output:**

Regression Analysis				
The regression equation is				
IQ = 91.3 + 1.49 Crycount				
Predictor	Coef	StDev	T	P
Constant	91.268	8.934	10.22	0.000
Crycount	1.4929	0.4870	3.07	0.004
S = 17.50		R-Sq = 20.7%		
estimates $\sigma$		SE <sub>b</sub>		
We usually ignore this part.				

**Construct a 95% Confidence interval for the slope of the LSRL of IQ on cry count for the 20 babies in the study.**

Formula:  $df = n - 2 = 20 - 2 = 18$

$$b \pm t^* SE_b$$

$$1.4929 \pm (2.101)(0.4870)$$

$$1.4929 \pm 1.0232$$

$$(0.4697, 2.5161)$$

**Find the t-test statistic and p-value for the effect cry count has on IQ.**

From the regression analysis  $t = 3.07$  and  $p = 0.004$

Or

$$t = \frac{b}{SE_b} = \frac{1.4929}{0.4870} = 3.07$$

$s = 17.50$

This is the standard deviation of the residuals and is a measure of the average spread of the deviations from the LSRL.

## Part II: Designing Experiments and Collecting Data:

### Sampling Methods:

#### The Bad:

**Voluntary sample.** A voluntary sample is made up of people who decide for themselves to be in the survey.

Example: Online poll

**Convenience sample.** A convenience sample is made up of people who are easy to reach.

Example: interview people at the mall, or in the cafeteria because it is an easy place to reach people.

#### The Good:

**Simple random sampling.** Simple random sampling refers to a method in which all possible samples of  $n$  objects are equally likely to occur.

Example: assign a number 1-100 to all members of a population of size 100. One number is selected at a time from a list of random digits or using a random number generator. The first 10 selected without repeats are the sample.

**Stratified sampling.** With stratified sampling, the population is divided into groups, based on some characteristic. Then, within each group, a SRS is taken. In stratified sampling, the groups are called **strata**.

Example: For a national survey we divide the population into groups or strata, based on geography - north, east, south, and west. Then, within each stratum, we might randomly select survey respondents.

**Cluster sampling.** With cluster sampling, every member of the population is assigned to one, and only one, group. Each group is called a cluster. A sample of clusters is chosen using a SRS. Only individuals within sampled clusters are surveyed.

Example: Randomly choose high schools in the country and only survey people in those schools.

Difference between cluster sampling and stratified sampling. With stratified sampling, the sample includes subjects from each stratum. With cluster sampling the sample includes subjects only from sampled clusters.

**Multistage sampling.** With multistage sampling, we select a sample by using combinations of different sampling methods.

Example: Stage 1, use cluster sampling to choose clusters from a population. Then, in Stage 2, we use simple random sampling to select a subset of subjects from each chosen cluster for the final sample.

**Systematic random sampling.** With systematic random sampling, we create a list of every member of the population. From the list, we randomly select the first sample element from the first  $k$  subjects on the population list. Thereafter, we select every  $k$ th subject on the list.

Example: Select every 5<sup>th</sup> person on a list of the population.

### Experimental Design:

A well-designed experiment includes design features that allow researchers to eliminate extraneous variables as an explanation for the observed relationship between the independent variable(s) and the dependent variable.

**Experimental Unit or Subject:** The individuals on which the experiment is done. If they are people then we call them subjects

**Factor:** The explanatory variables in the study

**Level:** The degree or value of each factor.

**Treatment:** The condition applied to the subjects. When there is one factor, the treatments and the levels are the same.

**Control.** Control refers to steps taken to reduce the effects of other variables (i.e., variables other than the independent variable and the dependent variable). These variables are called **lurking variables**.

Control involves making the experiment as similar as possible for subjects in each treatment condition. Three control strategies are control groups, placebos, and blinding.

**Control group.** A control group is a group that receives no treatment

**Placebo.** A fake or dummy treatment.

**Blinding:** Not telling subjects whether they receive the placebo or the treatment

**Double blinding:** neither the researchers or the subjects know who gets the treatment or placebo

**Randomization.** Randomization refers to the practice of using chance methods (random number tables, flipping a coin, etc.) to assign subjects to treatments.

**Replication.** Replication refers to the practice of assigning each treatment to many experimental subjects.

**Bias:** when a method systematically favors one outcome over another.

### Types of design:

**Completely randomized design** With this design, subjects are randomly assigned to treatments.

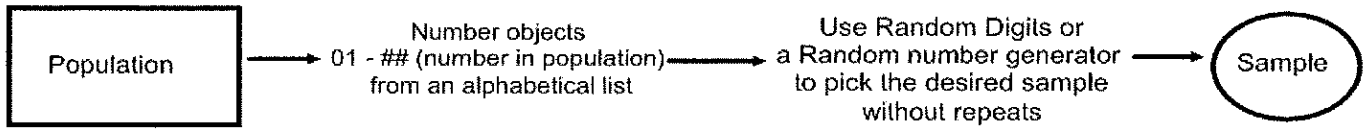
**Randomized block design,** the experimenter divides subjects into subgroups called **blocks**. Then, subjects within each block are randomly assigned to treatment conditions. Because this design reduces variability and potential confounding, it produces a better estimate of treatment effects.

**Matched pairs design** is a special case of the randomized block design. It is used when the experiment has only two treatment conditions; and subjects can be grouped into pairs, based on some blocking variable. Then, within each pair, subjects are randomly assigned to different treatments. **In some cases** you give two treatments to the same experimental unit. That unit is their own matched pair!

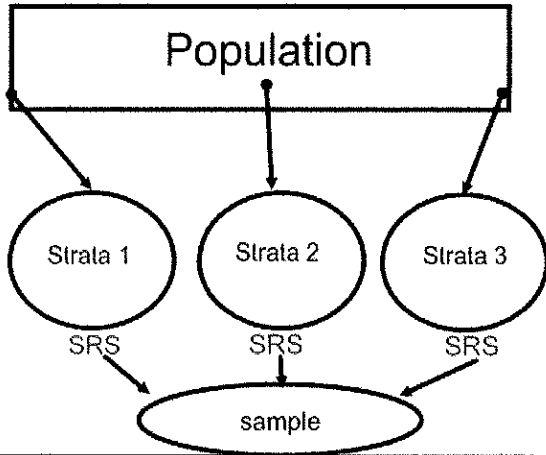
Part II in Pictures:

Sampling Methods

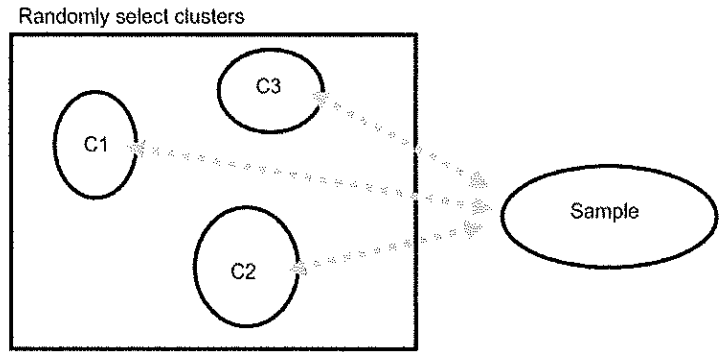
**Simple Random Sample:** Every group of n objects has an equal chance of being selected. (Hat Method!)



**Stratified Random Sampling:**  
Break population into strata (groups) then take an SRS of each group.



**Cluster Sampling:**  
Randomly select clusters then take all Members in the cluster as the sample.

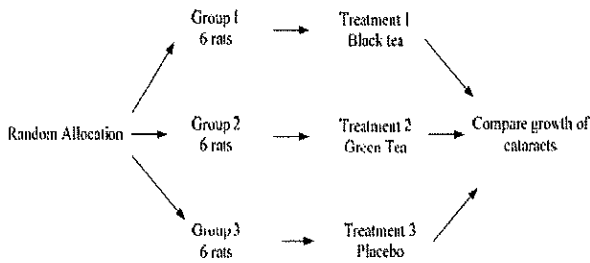


**Systematic Random Sampling:**  
Select a sample using a system, like selecting every third subject.

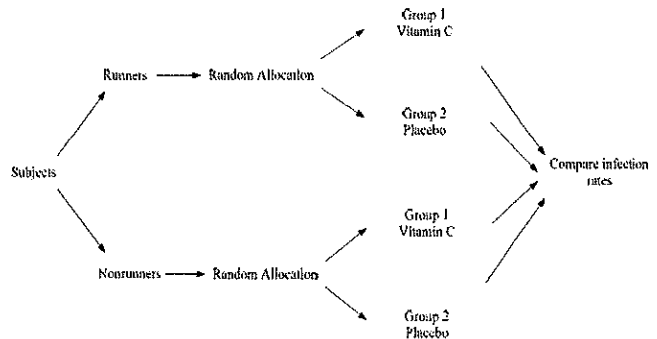


Experimental Design:

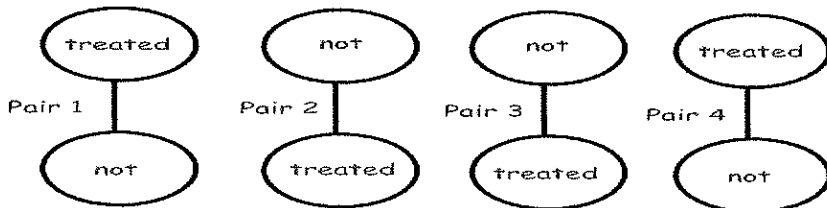
**Completely Randomized Design:**



**Randomized Block Design:**



**Matched Pairs Design:**

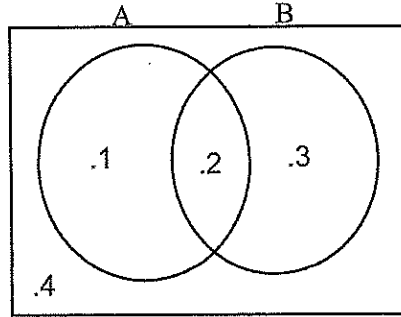


**Part III: Probability and Random Variables:**

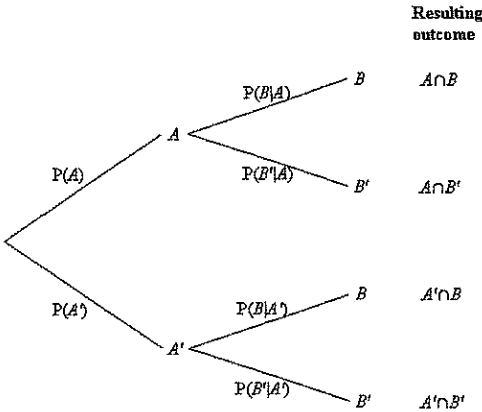
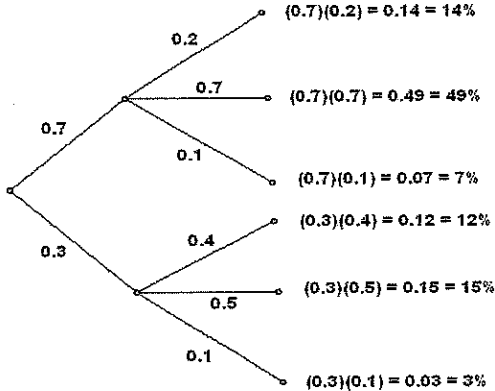
Counting Principle:  
 Trial 1: a ways  
 Trial 2: b ways  
 Trial 3: c ways ...  
 There are  $a \times b \times c$  ways to do all three.  
 $0 \leq P(A) \leq 1$   
 $1 - P(A) = P(A^c)$

A and B are **disjoint** or **mutually exclusive** if they have no events in common.  
 Roll two die: DISJOINT rolling a 9 rolling doubles  
 Roll two die: not disjoint rolling a 4 rolling doubles

A and B are **independent** if the outcome of one does not affect the other.  
**Mutually Exclusive events CANNOT BE Independent**



For Conditional Probability use a TREE DIAGRAM:



$P(A) = 0.3$   
 $P(B) = 0.5$   
 $P(A \cap B) = 0.2$   
 $P(A \cup B) = 0.3 + 0.5 - 0.2 = 0.6$   
 $P(A|B) = 0.2/0.5 = 2/5$   
 $P(B|A) = 0.2/0.3 = 2/3$

For Binomial Probability:  
**Look for x out of n trials**  
 1. Success or failure  
 2. Fixed n  
 3. Independent observations  
 4. p is the same for all observations

$P(X=3)$  Exactly 3  
 use  $\text{binompdf}(n,p,3)$   
 $P(X \leq 3)$  at most 3  
 use  $\text{binomcdf}(n,p,3)$  (Does 3,2,1,0)  
 $P(X \geq 3)$  at least 3 is  $1 - P(X \leq 2)$   
 use  $1 - \text{binomcdf}(n,p,2)$

Normal Approximation of Binomial:  
 for  $np \geq 10$  and  $n(1-p) \geq 10$   
 the X is approx  $N(np, \sqrt{np(1-p)})$

Discrete Random Variable: has a countable number of possible events (Heads or tails, each .5)  
 Continuous Random Variable: Takes all values in an interval: (EX: normal curve is continuous)  
 Law of large numbers. As n becomes very large  $\bar{x} \rightarrow \mu$

Linear Combinations:

$\mu_{a+bx} = a + b\mu_x$

$\mu_{X+Y} = \mu_x + \mu_y$

$\sigma_{a+bx}^2 = b^2 \sigma_x^2$

$\sigma_{X+Y}^2 = \sigma_x^2 + \sigma_y^2$

$\sigma_{X-Y}^2 = \sigma_x^2 + \sigma_y^2$

Geometric Probability:  
**Look for # trial until first success**

1. Success or Failure
2. X is trials until first success
3. Independent observations
4. p is same for all observations

$P(X=n) = p(1-p)^{n-1}$   
 $\mu$  is the expected number of trials until the first success or  $\frac{1}{p}$

$\sigma^2 = \frac{1-p}{p^2}$

$P(X > n) = (1-p)^n = 1 - P(X \leq n)$

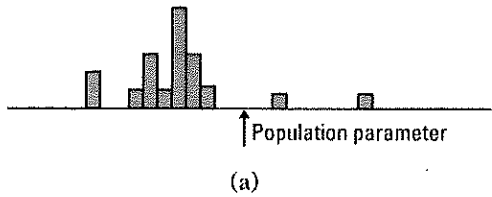
**Sampling distribution:** The distribution of all values of the statistic in all possible samples of the same size from the population.

**Central Limit Theorem:** As  $n$  becomes very large the sampling distribution for  $\bar{x}$  is approximately NORMAL

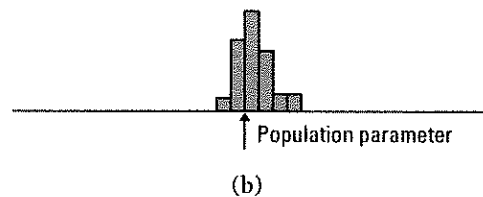
Use ( $n \geq 30$ ) for CLT

Low Bias: Predicts the center well  
Low Variability: Not spread out

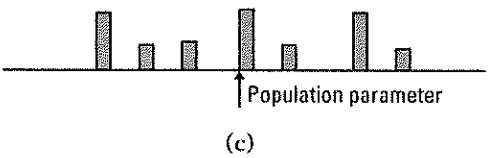
High Bias: Does not predict center well  
High Variability: Is very spread out



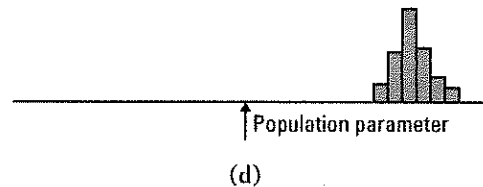
High bias, High Variability



Low Bias, Low Variability



Low Bias, High Variability



High Bias, Low Variability

See other sheets for Part IV

ART is my BFF

Type I Error: Reject the null hypothesis when it is actually True

Type II Error: Fail to reject the null hypothesis when it is False.

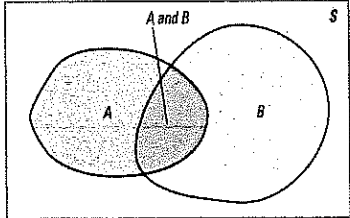
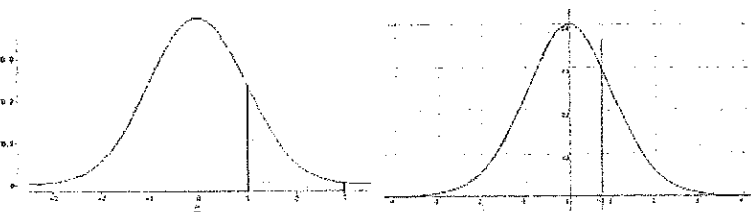
**ESTIMATE – DO A CONFIDENCE INTERVAL**

**EVIDENCE - DO A TEST**

Paired Procedures	Two Sample Procedures
<ul style="list-style-type: none"> <li>• Must be from a matched pairs design:</li> <li>• Sample from one population where each subject receives two treatments, and the observations are subtracted. <b>OR</b></li> <li>• Subjects are matched in pairs because they are similar in some way, each subject receives one of two treatments and the observations are subtracted</li> </ul>	<ul style="list-style-type: none"> <li>• Two independent samples from two different populations <b>OR</b></li> <li>• Two groups from a randomized experiment (each group would receive a different treatment) Both groups may be from the same population in this case but will randomly receive a different treatment.</li> </ul>

## Major Concepts in Probability

For the expected value (mean,  $\mu_x$ ) and the  $\sigma_x$  or  $\sigma_x^2$  of a probability distribution use the formula sheet

Binomial Probability	Simple Probability (and, or, not)
<p>Fixed Number of Trials Probability of success is the same for all trials Trials are independent</p> <p>If X is B(n,p) then (ON FORMULA SHEET) Mean <math>\mu_x = np</math> Standard Deviation <math>\sigma_x = \sqrt{np(1-p)}</math> For Binomial probability use <math>P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}</math> or use: Exactly: <math>P(X = x) = \text{binompdf}(n, p, x)</math> At Most: <math>P(X \leq x) = \text{binomcdf}(n, p, x)</math> At least: <math>P(X \geq x) = 1 - \text{binomcdf}(n, p, x-1)</math> More than: <math>P(X &gt; x) = 1 - \text{binomcdf}(n, p, x)</math> Less Than: <math>P(X &lt; x) = \text{binomcdf}(n, p, x-1)</math></p> <p>You may use the normal approximation of the binomial distribution when <math>np \geq 10</math> and <math>n(1-p) \geq 10</math>. Use then mean and standard deviation of the binomial situation to find the Z score.</p>	<p>Finding the probability of multiple simple events. Addition Rule: <math>P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)</math> Multiplication Rule: <math>P(A \text{ and } B) = P(A)P(B A)</math></p> <p>Mutually Exclusive events CANNOT be independent A and B are <b>independent</b> if the outcome of one does not affect the other. A and B are <b>disjoint</b> or <b>mutually exclusive</b> if they have no events in common. Roll two die: DISJOINT rolling a 9 rolling doubles</p> <div style="text-align: center;">  </div> <p>Roll two die: NOT disjoint rolling a 4 rolling doubles</p> <p><b>Independent: <math>P(B) = P(B A)</math></b> <b>Mutually Exclusive: <math>P(A \text{ and } B) = 0</math></b></p>
Geometric Probability	Conditional Probability
<p>You are interested in the amount of trials it takes UNTIL you achieve a success. Probability of success is the same for each trial Trials are independent</p> <p>Use simple probability rules for Geometric Probabilities.</p> <p><math>P(X=n) = p(1-p)^{n-1}</math>      <math>P(X &gt; n) = (1-p)^n = 1 - P(X \leq n)</math> <math>\mu_x</math> is the expected number of trials until the first success or <math>\frac{1}{p}</math></p>	<p>Finding the probability of an event given that another event has already occurred.</p> <p>Conditional Probability: <math>P(B A) = \frac{P(A \cap B)}{P(A)}</math></p> <p>Use a two way table or a Tree Diagram for Conditional Problems. Events are Independent if <math>P(B A) = P(B)</math></p>
Normal Probability	
<p><b>For a single observation from a normal population</b></p> <p><math>P(X &gt; x) = P(z &gt; \frac{x - \mu}{\sigma})</math>      <math>P(X &lt; x) = P(z &lt; \frac{x - \mu}{\sigma})</math></p> <div style="text-align: center;">  </div> <p>To find <math>P(x &lt; X &lt; y)</math> Find two Z scores and subtract the probabilities (upper - lower)</p> <p>Use the table to find the probability or use <math>\text{normalcdf}(\text{min}, \text{max}, 0, 1)</math> after finding the z-score</p>	<p><b>For the mean of a random sample of size n from a population.</b></p> <p>When <math>n &gt; 30</math> the sampling distribution of the sample mean <math>\bar{x}</math> is approximately Normal with:</p> <p><math>\mu_{\bar{x}} = \mu</math> <math>\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}</math></p> <p>If <math>n &lt; 30</math> then the population should be Normally distributed to begin with to use the z-distribution.</p> <p><math>P(\bar{X} &gt; x) = P(z &gt; \frac{\bar{x} - \mu}{\sigma/\sqrt{n}})</math>      <math>P(\bar{X} &lt; x) = P(z &lt; \frac{\bar{x} - \mu}{\sigma/\sqrt{n}})</math></p> <p>To find <math>P(x &lt; X &lt; y)</math> Find two Z scores and subtract the probabilities (upper - lower) Use the table to find the probability or use <math>\text{normalcdf}(\text{min}, \text{max}, 0, 1)</math> after finding the z-score</p>



### Binomial Probability

Mr. K is shooting three point jump shots. Mr. K has a career shooting percentage of 80%. Mr. K is going to shoot 30 three pointers during a practice session.

X: number of threes made, X is B(30, 0.6)

$$\mu_X = np = 30(.6) = 18$$

$$\sigma_X = \sqrt{np(1-p)} = \sqrt{30(.60)(.40)} = 2.683$$

The probability that Mr. K makes exactly 20 is:

$$P(X = 20) = \text{binompdf}(30, 0.6, 20) = 0.1152$$

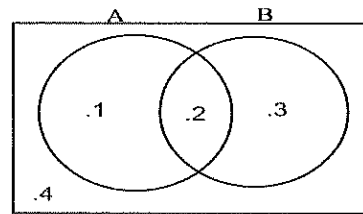
The probability that Mr. K makes at most 20 is:

$$P(X \leq 20) = \text{binomcdf}(30, 0.6, 20) = 0.8237$$

The probability the Mr. K makes at least 20 is:

$$P(X \geq 20) = 1 - \text{binomcdf}(30, 0.6, 19) = 1 - 0.7085 = 0.2915$$

### Simple Probability (and, or, not):



$$P(A) = 0.3$$

$$P(B) = 0.5$$

$$P(A \cap B) = 0.2$$

$$P(A \cup B) = 0.3 + 0.5 - 0.2 = 0.6$$

$$P(A|B) = 0.2/0.5 = 2/5$$

$$P(B|A) = 0.2/0.3 = 2/3$$

$$P(A^c) = 1 - 0.3 = 0.7$$

### Geometric Probability

The population of overweight manatees is known to be 40% You select a random Manatee and weigh it, and then you repeat the selection until one is overweight.

Find the probability that the fifth manatee you choose is overweight.

$$P(X = 5) = (\text{notover})^4 (\text{over}) = (0.60)^4 (0.40) = .05184$$

Find the probability that it takes more than five attempts to find an overweight manatee.

$$P(X > 5) = (\text{notoverweight})^5 = (0.60)^5 = 0.07776$$

How many manatees would you expect to choose before you found one to be overweight?

$$\mu_X = \frac{1}{p} = \frac{1}{0.4} = 2.5$$

### Conditional Probability with a Tree Diagram

Of adult users of the Internet:

29% are 18-29

47% are 30-49

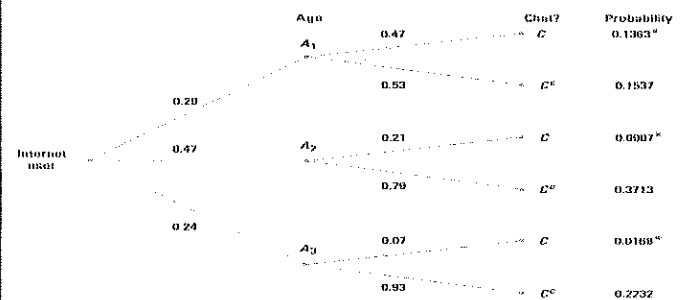
24% are over 50

47% of the 18-29 group chat

21% of the 30-49 group chat

7% of the 50 and over group chat.

Find the probability that a randomly selected adult chats



### Normal Probability

The weight of manatees follows a normal distribution with a mean weight of 800 pounds and a standard deviation of 120 pounds.

Find the probability that a randomly selected Manatee weighs more than 1000 pounds:

X is N(800,120)

$$P(X > 1000) = P\left(z > \frac{1000 - 800}{120}\right) = P(z > 1.67) = 0.0475$$

Find the probability that a random sample of 50 manatees has a mean weight more than 1000 pounds:

$$P(\bar{X} > 1000) = P\left(z > \frac{1000 - 800}{120/\sqrt{50}}\right) = P(z > 11.79) \approx 0$$

Even if you did not know the population was normal you could use CLT and assume the sampling distribution is approximately normal.

### Conditional Probability with a two way table:

Table 6.1 Grades awarded at a university, by school

	Grade Level			Total
	A	B	Below B	
Liberal Arts	2,142	1,890	2,268	6,300
Engineering and Physical Sciences	368	432	800	1,600
Health and Human Services	882	630	588	2,100
Total	3,392	2,952	3,656	10,000

$$P(\text{A grade} | \text{liberal arts course}) = 2142 / 6300$$

$$P(\text{Liberal arts course} | \text{A Grade}) = 2142 / 3392$$

$$P(\text{B Grade} | \text{Engineering and PS}) = 432 / 1600$$

$$P(\text{Engineering and PS} | \text{B Grade}) = 432 / 2952$$

Mutually Exclusive vs. Independence

You just heard that Dan and Annie who have been a couple for three years broke up. This presents a problem, because you're having a big party at your house this Friday night and you have invited them both. Now you're afraid there might be an ugly scene if they both show up. When you see Annie, you talk to her about the issue, asking her if she remembers about your party. She assures you she's coming. You say that Dan is invited, too, and you wait for her reaction. If she says, "That jerk! If he shows up I'm not coming. I want nothing to do with him!", they're **mutually exclusive**. If she says, "Whatever. Let him come, or not. He's nothing to me now.", they're **independent**.

Mutually Exclusive and Independence are two very different ideas

<p align="center"><b><u>Mutually Exclusive (disjoint):</u></b>  <math>P(A \text{ and } B) = 0</math></p> <p>Events A and B are mutually exclusive if they have no outcomes in common. That is A and B cannot happen at the same time.</p> <p>Example of <b>mutually exclusive (disjoint)</b>:                  A: roll an odd on a die                  B: roll an even on a die</p> <p>Odd and even <b>share no outcomes</b>  <math>P(\text{odd and even}) = 0</math>                  Therefore, they are mutually exclusive.</p> <p>Example of <b>not mutually exclusive (joint)</b>:                  A: draw a king                  B: draw a face card</p> <p>King and face card <b>do share outcomes</b>. All of the kings are face cards.  <math>P(\text{king and face card}) = 4/52</math>                  Therefore, they are not mutually exclusive.</p>	<p align="center"><b><u>Independence:</u></b>  <math>P(B) = P(B A)</math></p> <p>Events A and B are independent if knowing one outcome does not change the probability of the other. That is knowing A does not change the probability of B.</p> <p>Examples of <b>independent</b> events:                  A: draw an ace                  B: draw a spade</p> <p><math>P(\text{Spade}) = 13/52 = 1/4</math>  <math>P(\text{Spade}   \text{Ace}) = 1/4</math>                  Knowing that the drawn card is an ace <b>does not change</b> the probability of drawing a spade</p> <p>Examples that are <b>dependent</b> (not independent):                  A: roll a number greater than 3                  B: roll an even</p> <p><math>P(\text{even}) = 3/6 = 1/2</math>  <math>P(\text{even}   \text{greater than } 3) = 2/3</math>                  Knowing the number is greater than three <b>changes</b> the probability of rolling an even number.</p>
---	---

<p>Mutually Exclusive events cannot be independent</p> <p><b>Mutually exclusive and dependent</b></p> <p>A: Roll an even                  B: Roll an odd</p> <p>They share no outcomes and knowing that it is odd changes the probability of it being even.</p>	<p>Independent events cannot be Mutually Exclusive</p> <p><b>Independent and not mutually exclusive</b></p> <p>A: draw a black card                  B: draw a king</p> <p>Knowing it is a black card does not change the probability of it being a king and they do share outcomes.</p>	<p>Dependent Events may or may not be mutually exclusive</p> <p><b>Dependent and mutually exclusive</b></p> <p>A: draw a queen                  B: draw a king                  Knowing it is a queen changes the probability of it being a king and they do not share outcomes.</p> <p><b>Dependent and not mutually exclusive</b></p> <p>A: Face Card                  B: King                  Knowing it is a face card changes the probability of it being a king and they do share outcomes.</p>
---	--	--

<p>If events are mutually exclusive then:  <math>P(A \text{ or } B) = P(A) + P(B)</math>                  If events are not mutually exclusive use the general rule:  <math>P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)</math></p>	<p>If events are independent then:  <math>P(A \text{ and } B) = P(A)P(B)</math>                  If events are not independent then use the general rule:  <math>P(A \text{ and } B) = P(A)P(B A)</math></p>
---	--

## Interpretations from Inference

### Interpretation for a Confidence Interval:

I am C% confident that the true parameter (mean  $\mu$  or proportion  $p$ ) lies between # and #.

**INTERPRET IN CONTEXT!!**

### Interpretation of C% Confident:

Using my method, If I sampled over and over again, C% of my intervals would contain the true parameter (mean  $\mu$  or proportion  $p$ ).

**NOT:** The parameter lies in my interval C% of the time. It either does or does not!!

If  $p < \alpha$  I **reject** the null hypothesis  $H_0$  and I have **sufficient/strong** evidence to support the alternative hypothesis  $H_a$

**INTERPRET IN CONTEXT in terms of the alternative.**

If  $p > \alpha$  I **fail to reject** the null hypothesis  $H_0$  and I have **insufficient/poor** evidence to support the alternative hypothesis  $H_a$

**INTERPRET IN CONTEXT in terms of the alternative.**

### Evidence Against $H_0$

#### **P-Value**

"Some"  $0.05 < \text{P-Value} < 0.10$

"Moderate or Good"  $0.01 < \text{P-Value} < 0.05$

"Strong"  $\text{P-Value} < 0.01$

### Interpretation of a p-value:

The probability, assuming the null hypothesis is true, that an observed outcome would be as extreme or more extreme than what was actually observed.

### Duality: Confidence intervals and significance tests.

If the hypothesized parameter lies **outside** the C% confidence interval for the parameter I can **REJECT**  $H_0$

If the hypothesized parameter lies **inside** the C% confidence interval for the parameter I **FAIL TO REJECT**  $H_0$

### Power of test:

The probability that at a fixed level  $\alpha$  test will reject the null hypothesis when and alternative value is true.

## Confidence Intervals

<p><b><u>One Sample Z Interval</u></b>            Use when estimating a single population mean and <math>\sigma</math> is known            Conditions:            -SRS            -Normality: CLT, stated, or plots            -Independence and <math>N \geq 10n</math>            Interval:  <math display="block">\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}</math></p> <p>CALCULATOR:  <b>7: Z-Interval</b></p>	<p><b><u>One Sample t Interval</u></b>            Use when estimating a single mean and <math>\sigma</math> is NOT known            [Also used in a <u>matched paired design</u> for the mean of the difference:  <i>PAIRED t PROCEDURE</i>]            Conditions:            -SRS            -Normality: CLT, stated, or plots            -Independence and <math>N \geq 10n</math>            Interval:  <math display="block">\bar{x} \pm t^* \frac{s}{\sqrt{n}}</math> <math display="block">df = n - 1</math></p> <p>CALCULATOR:  <b>8: T-Interval</b></p>	<p><b><u>One Proportion Z Interval</u></b>            Use when estimating a single proportion            Conditions:            -SRS            -Normality: <math>n\hat{p} \geq 10, n(1 - \hat{p}) \geq 10</math>            -Independence and <math>N \geq 10n</math>            Interval:  <math display="block">\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}</math></p> <p>CALCULATOR:  <b>A: 1-PropZInt</b></p>
<p><b><u>Two Sample Z Interval</u></b>  <b><math>\sigma</math> known (RARELY USED)</b>            Use when estimating the difference between two population means and <math>\sigma</math> is known            Conditions:            -SRS for both populations            -Normality: CLT, stated, or plots for both populations            -Independence and <math>N \geq 10n</math> for both populations            Interval:  <math display="block">(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}</math></p> <p>CALCULATOR:  <b>9: 2-SampZInt</b></p>	<p><b><u>Two Sample t Interval</u></b>  <b><math>\sigma</math> unknown</b>            Use when estimating the difference between two population means and <math>\sigma</math> is NOT known            Conditions:            -SRS for both populations            -Normality: CLT, stated, or plots for both populations            -Independence and <math>N \geq 10n</math> for both populations            Interval:  <math display="block">(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}</math></p> <p style="text-align: center;">Use lower n for df (<math>df = n - 1</math>) or use calculator</p> <p>CALCULATOR:  <b>0: 2-SampTInt</b></p>	<p><b><u>Two Proportion Z Interval</u></b>            Use when estimating the difference between two population proportions.            Conditions:            -SRS for both populations            -Normality:  <math>n_1\hat{p}_1 \geq 10, n_1(1 - \hat{p}_1) \geq 10</math>  <math>n_2\hat{p}_2 \geq 10, n_2(1 - \hat{p}_2) \geq 10</math>            -Independence and <math>N \geq 10n</math> for both populations.            Interval:  <math display="block">(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}</math></p> <p>CALCULATOR:  <b>B: 2-PropZInt</b></p>
<p><b><u>Confidence interval for Regression Slope:</u></b></p> <p>Use when estimating the slope of the true regression line            Conditions:            1. Observations are independent            2. Linear Relationship (look at residual plot)            3. Standard deviation of y is the same (look at residual plot)            4. y varies normally (look at histogram of residuals)            Interval:  <math display="block">b \pm t^* SE_b</math> <math display="block">df = n - 2</math></p> <p>CALCULATOR: <b>LinRegTInt</b> Use technology readout or calculator for this confidence interval.</p>		

## Hypothesis Testing

<p style="text-align: center;"><b>One Sample Z Test</b></p> <p>Use when testing a mean from a single sample and <math>\sigma</math> is known</p> <p><math>H_0: \mu = \#</math>  <math>H_a: \mu \neq \#</math> -or- <math>\mu &lt; \#</math> -or- <math>\mu &gt; \#</math></p> <p>Conditions:                      -SRS                      -Normality: Stated, CLT, or plots                      -Independence and <math>N \geq 10n</math></p> <p>Test statistic:</p> $z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ <p style="text-align: center;"><b>CALCULATOR: 1: Z-Test</b></p>	<p style="text-align: center;"><b>One Sample t Test</b></p> <p>Use when testing a mean from a single sample and <math>\sigma</math> is NOT known                      [Also used in a <u>matched paired design</u> for the mean of the difference:  <u>PAIRED t PROCEDURE</u>]</p> <p><math>H_0: \mu = \#</math>  <math>H_a: \mu \neq \#</math> -or- <math>\mu &lt; \#</math> -or- <math>\mu &gt; \#</math></p> <p>Conditions:                      -SRS                      -Normality: Stated, CLT, or plots                      -Independence and <math>N \geq 10n</math></p> <p>Test statistic: <math>df = n-1</math></p> $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ <p style="text-align: center;"><b>CALCULATOR: 2: T-Test</b></p>	<p style="text-align: center;"><b>One Proportion Z test</b></p> <p>Use when testing a proportion from a single sample</p> <p><math>H_0: p = \#</math>  <math>H_a: p \neq \#</math> -or- <math>p &lt; \#</math> -or- <math>p &gt; \#</math></p> <p>Conditions:                      -SRS                      -Normality: <math>np_0 \geq 10</math>, <math>n(1-p_0) \geq 10</math>                      -Independence and <math>N \geq 10n</math></p> <p>Test statistic:</p> $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ <p style="text-align: center;"><b>CALCULATOR: 5: 1-PropZTest</b></p>
<p style="text-align: center;"><b>Two Sample Z test</b>  <math>\sigma</math> known (RARELY USED)</p> <p>Use when testing two means from two samples and <math>\sigma</math> is known</p> <p><math>H_0: \mu_1 = \mu_2</math>  <math>H_a: \mu_1 \neq \mu_2</math> -or- <math>\mu_1 &lt; \mu_2</math> -or- <math>\mu_1 &gt; \mu_2</math></p> <p>Conditions:                      -SRS for both populations                      -Normality: Stated, CLT, or plots for both populations                      -Independence and <math>N \geq 10n</math> for both populations.</p> <p>Test statistic:</p> $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ <p style="text-align: center;"><b>CALCULATOR: 3: 2-SampZTest</b></p>	<p style="text-align: center;"><b>Two Sample t Test</b>  <math>\sigma</math> unknown</p> <p>Use when testing two means from two samples and <math>\sigma</math> is NOT known</p> <p><math>H_0: \mu_1 = \mu_2</math>  <math>H_a: \mu_1 \neq \mu_2</math> -or- <math>\mu_1 &lt; \mu_2</math> -or- <math>\mu_1 &gt; \mu_2</math></p> <p>Conditions:                      -SRS for both populations                      -Normality: Stated, CLT, or plots for both populations                      -Independence and <math>N \geq 10n</math> for both populations.</p> <p>Test statistic: use lower n for df (<math>df = n-1</math>) or use calculator.</p> $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ <p style="text-align: center;"><b>CALCULATOR: 4: 2-SampTTest</b></p>	<p style="text-align: center;"><b>Two Proportion Z test</b></p> <p>Use when testing two proportions from two samples</p> <p><math>H_0: p_1 = p_2</math>  <math>H_a: p_1 \neq p_2</math> -or- <math>p_1 &lt; p_2</math> -or- <math>p_1 &gt; p_2</math></p> <p>Conditions:                      -SRS for both populations                      -Normality:  <math>n_1 \hat{p}_1 \geq 10</math> <math>n_1(1 - \hat{p}_1) \geq 10</math>  <math>n_2 \hat{p}_2 \geq 10</math> <math>n_2(1 - \hat{p}_2) \geq 10</math>                      Independence and <math>N \geq 10n</math> for both populations.</p> <p>Test statistic:</p> $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1 - \hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ <p style="text-align: center;"><b>CALCULATOR: 6: 2-PropZTest</b></p>
<p style="text-align: center;"><b><math>\chi^2</math> - Goodness of Fit</b>                      L1 and L2</p> <p>Use for categorical variables to see if one distribution is similar to another</p> <p><math>H_0</math>: The distribution of two populations are not different  <math>H_a</math>: The distribution of two populations are different</p> <p>Conditions:                      -Independent Random Sample                      -All expected counts <math>\geq 1</math>                      -No more than 20% of Expected Counts <math>&lt; 5</math></p> <p>Test statistic: <math>df = n-1</math></p> $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ <p style="text-align: center;"><b>CALCULATOR: D: <math>\chi^2</math>GOF-Test(on 84)</b></p>	<p style="text-align: center;"><b><math>\chi^2</math> - Homogeneity of Populations</b>                      r x c table</p> <p>Use for categorical variables to see if multiple distributions are similar to one another</p> <p><math>H_0</math>: The distribution of multiple populations are not different  <math>H_a</math>: The distribution of multiple populations are different</p> <p>Conditions:                      -Independent Random Sample                      -All expected counts <math>\geq 1</math>                      -No more than 20% of Expected Counts <math>&lt; 5</math></p> <p>Test statistic: <math>df = (r-1)(c-1)</math></p> $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ <p style="text-align: center;"><b>CALCULATOR: C: <math>\chi^2</math>-Test</b></p>	<p style="text-align: center;"><b><math>\chi^2</math> - Independence/Association</b>                      r x c table</p> <p>Use for categorical variables to see if two variables are related</p> <p><math>H_0</math>: Two variables have no association (independent)  <math>H_a</math>: Two variables have an association (dependent)</p> <p>Conditions:                      -Independent Random Sample                      -All expected counts <math>\geq 1</math>                      -No more than 20% of Expected Counts <math>&lt; 5</math></p> <p>Test statistic: <math>df = (r-1)(c-1)</math></p> $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ <p style="text-align: center;"><b>CALCULATOR: C: <math>\chi^2</math>-Test</b></p>
<p><b>Significance Test for Regression Slope:</b> <math>H_0: \beta = 0</math> <math>H_a: \beta \neq \#</math> -or- <math>\beta &lt; \#</math> -or- <math>\beta &gt; \#</math></p> <p>Conditions: 1. Observations are independent 2. Linear Relationship (look at residual plot)                      3. Standard deviation of y is the same (look at residual plot) 4. y varies normally (look at histogram of residuals)</p> <p style="text-align: center;"><b>CALCULATOR: LinRegTTest</b> Use technology readout or the calculator for this significance test <math>t = \frac{b}{SE_b}</math> <math>df = n-2</math></p>		

## Notation and Interpretations

IQR	Inner Quartile Range
$\bar{x}$	Mean of a sample
$\mu$	Mean of a population
s	Standard deviation of a sample
$\sigma$	Standard deviation of a population
$\hat{p}$	Sample proportion
p	Population proportion
$s^2$	Variance of a sample
$\sigma^2$	Variance of a population
M	Median
$\Sigma$	Summation
$Q_1$	First Quartile
$Q_3$	Third Quartile
Z	Standardized value – z test statistic
$z^*$	Critical value for the standard normal distribution
t	Test statistic for a t test
$t^*$	Critical value for the t-distribution
$N(\mu, \sigma)$	Notation for the normal distribution with mean and standard deviation
r	Correlation coefficient – strength of linear relationship
$r^2$	Coefficient of determination – measure of fit of the model to the data
$\hat{y} = a + bx$	Equation for the Least Squares Regression Line
a	y-intercept of the LSRL
b	Slope of the LSRL
$(\bar{x}, \bar{y})$	Point the LSRL passes through
$y = ax^b$	Power model
$y = ab^x$	Exponential model
SRS	Simple Random Sample
S	Sample Space
P(A)	The probability of event A
$A^c$	A complement
P(B A)	Probability of B given A
$\cap$	Intersection (And)
U	Union (Or)
X	Random Variable
$\mu_x$	Mean of a random variable
$\sigma_x$	Standard deviation of a random variable
$\sigma_x^2$	Variance of a random variable
$B(n,p)$	Binomial Distribution with observations and probability of success
$\binom{n}{k}$	Combination n taking k
pdf	Probability distribution function
cdf	Cumulative distribution function
n	Sample size
N	Population size
CLT	Central Limit Theorem
$\mu_{\bar{x}}$	Mean of a sampling distribution
$\sigma_{\bar{x}}$	Standard deviation of a sampling distribution
df	Degrees of freedom
SE	Standard error
$H_0$	Null hypothesis-statement of no change
$H_a$	Alternative hypothesis- statement of change
p-value	Probability (assuming $H_0$ is true) of observing a result as large or larger than that observed
$\alpha$	Significance level of a test. P(Type I) or the y-intercept of the true LSRL
$\beta$	P(Type II) or the true slope of the LSRL
$\chi^2$	Chi-square test statistic

z-score (z)	The number of standard deviations an observation is above/below the mean
slope (b)	The change in predicted y for every unit increase on x
y-intercept (a)	Predicted y when x is zero
r (correlation coefficient)	strength of linear relationship. (Strong/moderate/weak) (Positive/Negative) linear relationship between y and x.
$r^2$ (coefficient of determination)	percent of variation in y explained by the LSRL of y on x.
variance ( $\sigma^2$ or $s^2$ )	average squared deviation from the mean
standard deviation ( $\sigma$ or s)	measure of variation of the data points from the mean
Confidence Interval (#,#)	I am C% confident that the true parameter (mean $\mu$ or proportion p) lies between # and #.
C % Confidence (Confidence level)	Using my method, If I sampled repeatedly, C% of my intervals would contain the true parameter (mean $\mu$ or proportion p).
$p < \alpha$	Since $p < \alpha$ I reject the null hypothesis $H_0$ and I have sufficient/strong evidence to conclude the alternative hypothesis $H_a$
$p > \alpha$	Since $p > \alpha$ I fail to reject the null hypothesis $H_0$ and I have do not have sufficient evidence to support the alternative hypothesis $H_a$
p-value	The probability, assuming the null hypothesis is true, that an observed outcome would be as or more extreme than what was actually observed.
Duality-Outside Interval Two sided test	If the hypothesized parameter lies outside the $(1 - \alpha)\%$ confidence interval for the parameter I can REJECT $H_0$ for a two sided test.
Duality-Inside Interval Two sided test	If the hypothesized parameter lies inside the $(1 - \alpha)\%$ confidence interval for the parameter I FAIL TO REJECT $H_0$ for a two sided test.
Power of the test	The probability that a fixed level test will reject the null hypothesis when an alternative value is true
standard error (SE) in general	Estimates the variability in the sampling distribution of the sample statistic.
standard deviation of the residuals (s from regression)	A typical amount of variability of the vertical distances from the observed points to the LSRL
standard error of the slope of the LSRL ( $SE_b$ )	This is the standard deviation of the estimated slope. This value estimates the variability in the sampling distribution of the estimated slope.

# Which Inference Procedure Should I Use?

