

# Cumulative Notes

## AP Statistics



**Prepared by: Earl L. Whitney, FSA, MAAA**

**Version 1.0**

**April 27, 2014**

## AP Statistics Cumulative Notes Table of Contents

Page	Description
6	<b>AP Statistics Formula Sheet</b>
8	<b>Part 1: Exploring Data</b>
8	Variable Types
8	5-Number Summary
8	Other Terms to Know
9	Frequency Distribution
9	Cumulative Frequency Distribution
10	Center, Shape and Spread
11	Types of Plots
13	Marginal Distributions
14	Normal Distribution
15	<b>Part 2: Exploring Bivariate Data</b>
15	Definitions
17	Formulas Relating to the Coefficient of Correlation of a Sample
18	Linear Combinations of Parameters and Statistics
19	Types of Regression Models
20	<b>Chapter 10: Straightening Data In a Scatterplot</b>
20	Selecting a Good Re-Expression Model
22	Re-Expression Models Based on the Scatterplot
23	Re-Expression Models Based on the Residuals
24	Description of Individual Models
30	<b>Chapter 12: Sample Surveys</b>
30	Types of Samples
30	Types of Bias
31	Table of Key Statistics and Parameters
31	The Valid Survey
32	<b>Chapter 13: Experiments and Observational Studies</b>
32	Observation Study
32	Experiment
32	Principles of Experimental Design
33	Diagramming an Experiment
33	Elements of an Experiment
33	Other Terms Relating to Experimentation

## AP Statistics Cumulative Notes Table of Contents

Page	Description
34	<b>Part 4: Probability</b>
34	Key Definitions
36	Geometric Probability Model
36	Binomial Probability Model
36	Normal Probability Model
37	Key Formulas
<b>38</b>	<b>Chapter 18: Sampling Distribution Models</b>
38	Central Limit Theorem
38	Key Assumptions
39	Symbols and Formulas for Proportions
39	Symbols and Formulas for Population Means
40	<b>Chapter 19: Confidence Intervals for Proportions</b>
40	Explaining a Confidence Interval
40	Critical Values
40	Standard Deviation vs. Standard Error
40	Key Assumptions
41	<b>Chapter 20: Hypothesis Testing for Proportions</b>
41	Hypotheses
41	P-Value
42	Four Steps of Hypothesis Testing
43	Z-Values and P-Values for Various Tests - 95% Confidence
44	<b>Chapter 21: More About Tests and Intervals</b>
44	<b>a</b> -Value
44	Rules for Rejecting $H_0$
44	<b>b</b> -Value
44	Power
44	Effect Size
45	95% Confidence Interval (When the Success/Failure Condition Fails)
45	Summary of p's and n's
46	Type I and Type II Errors
46	Interaction of <b>a</b> , <b>b</b> , and Power
47	<b>Chapter 22: Comparing Two Proportions</b>
47	Confidence Interval for the Difference of Two Proportions
47	Key Assumptions and Conditions
48	Hypothesis Testing

## AP Statistics Cumulative Notes Table of Contents

Page	Description
49	<b>Chapter 23: Inferences About Means</b>
49	Sample of Means
49	Confidence Interval for the Difference of Two Proportions
50	Using z vs. t
50	Key Assumptions and Conditions
50	Sample Size for a t-Distribution
50	Special Considerations
50	Language for Confidence Intervals
51	Hypothesis Testing
51	Relationship Between Intervals and Tests
51	Finding the Required Sample Size
51	Finding the Required Sample Size - Example
53	<b>Chapter 24: Comparing Means</b>
53	Boxplots
53	Sample of Means
53	Confidence Interval
54	Key Assumptions and Conditions
54	Sample Size for a t-Distribution
54	Special Considerations
54	Hypothesis Testing
55	Pooled t-Interval and t-Test
56	<b>Chapter 25: Inferences About Paired Data</b>
56	Pairing
56	Sample of Means
56	Confidence Interval
57	Key Assumptions and Conditions
57	Hypothesis Testing
58	<b>Statistical Inference Summary Chart</b>
59	<b>Chapter 26: Comparing Counts – The Distribution</b>
59	Uses and Characteristics
59	Hypothesis tests
60	Testing Goodness of Fit
62	Testing Homogeneity or Independence

## AP Statistics Cumulative Notes Table of Contents

Page	Description
65	<b>Chapter 27: Inferences for Regression</b>
65	Moel of the Idealized Line
65	Visualization
65	Key Assumptions and Conditions
66	Work Order
66	Calculations
67	Thoughts about the Size of the Standard Error
68	Hypothesis Tests and Confidence Intervals for Linear Regression on the TI-84

## AP Statistics Formula Sheet

### Formulas

#### (I) Descriptive Statistics

$$\bar{x} = \frac{\sum x_i}{n}$$

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

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

$$\hat{y} = b_0 + b_1x$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

$$b_1 = r \frac{s_y}{s_x}$$

$$s_{b_1} = \frac{\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

#### (II) Probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

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

$$\text{Var}(X) = \sigma_x^2 = \sum (x_i - \mu_x)^2 p_i$$

If  $X$  has a binomial distribution with parameters  $n$  and  $p$ , then:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mu_x = np$$

$$\sigma_x = \sqrt{np(1-p)}$$

$$\mu_{\hat{p}} = p$$

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

If  $\bar{x}$  is the mean of a random sample of size  $n$  from an infinite population with mean  $\mu$  and standard deviation  $\sigma$ , then:

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

## AP Statistics Formula Sheet

### (III) Inferential Statistics

Standardized test statistic:  $\frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$

Confidence interval:  $\text{statistic} \pm (\text{critical value}) \cdot (\text{standard deviation of statistic})$

#### Single-Sample

Statistic	Standard Deviation of Statistic
Sample Mean	$\frac{\sigma}{\sqrt{n}}$
Sample Proportion	$\sqrt{\frac{p(1-p)}{n}}$

#### Two-Sample

Statistic	Standard Deviation of Statistic
Difference of sample means	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ <p style="text-align: center;">Special case when <math>\sigma_1 = \sigma_2</math></p> $\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Difference of sample proportions	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ <p style="text-align: center;">Special case when <math>p_1 = p_2</math></p> $\sqrt{p(1-p)} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$$\text{Chi-square test statistic} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

## Exploring Data

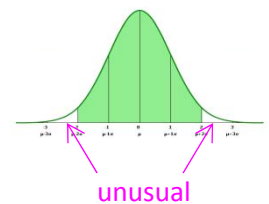
### Variable Types:

- A **quantitative variable** is one that takes on numerical values. Examples: age, weight, salary, GPA, temperature, etc.
- A **categorical variable** is one that is not quantitative. Examples: dog breed, hair color, high school attended, political affiliation, etc.

### 5-Number Summary (and related terms)

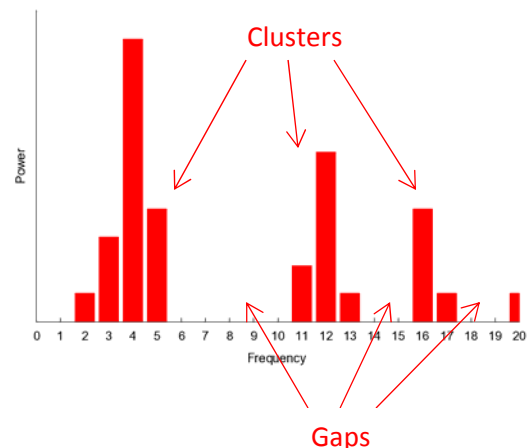
Arrange the quantitative data from low to high. Then:

- **Minimum:** The lowest value.
- **1<sup>st</sup> Quartile (Q1):** The median of the lower half of the values. Note: if there are an odd number of values, exclude the median from the lower half of the values in calculating Q1.
- **Median:** The middle value. Note: if there are two middle values, the median is the mean of those two values.
- **3<sup>rd</sup> Quartile (Q3):** The median of the upper half of the values. Note: if there are an odd number of values, exclude the median from the upper half of the values in calculating Q3.
- **Maximum:** The highest value.
- **Interquartile Range (IQR):**  $IQR = Q3 - Q1$
- **Range:**  $Range = Maximum - Minimum$
- **Outliers:** Values that are more than 1.5 IQR's above Q3 or 1.5 IQR's below Q1. Note: in a Normal Distribution, data are considered outliers if they are more than two Standard Deviations away from the mean. These data are often referred to as "unusual."



### Other Terms to Know

- **Mean:** The arithmetic average of the values.
- **Mode:** The value or values that occur most often.
- **Cluster:** A natural subgroup of data which lie close together.
- **Gap:** A break in the data; i.e., a range where no values exist.

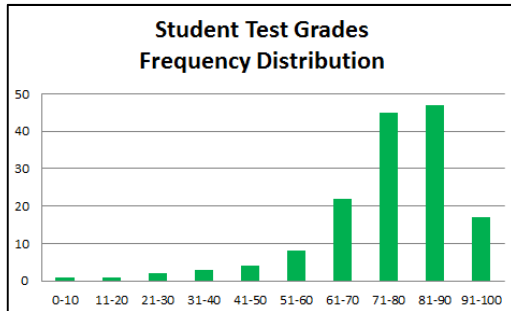




## Frequency Distribution

Data (both quantitative and categorical) are often displayed in a frequency distribution table. A sample frequency distribution table is shown at right.

To get a visual representation of the data, it may be plotted on a graph. For a discrete variable, the graph typically takes the form of a histogram or a bar chart. For a continuous distribution (e.g., the Normal Distribution), it will typically take the form of a continuous curve.

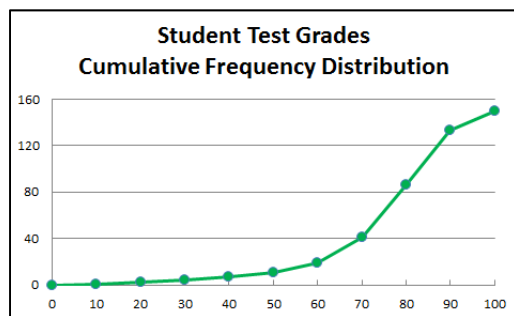


Scores	Number of Students
0 to 10	1
11 to 20	1
21 to 30	2
31 to 40	3
41 to 50	4
51 to 60	8
61 to 70	22
71 to 80	45
81 to 90	47
91 to 100	17
Total	150

## Cumulative Frequency Distribution (Ogive)

Quantitative data may also be displayed as a cumulative frequency distribution. The points plotted on the graph represent the total accumulated frequencies associated with each  $x$ -value.

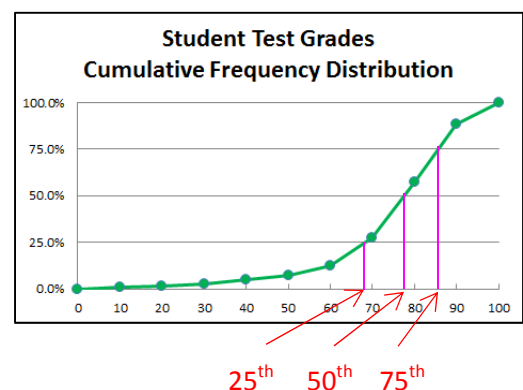
The graph of a cumulative frequency distribution is called an **ogive**.



Scores	Cumulative Number of Students
0 to 10	1
0 to 20	2
0 to 30	4
0 to 40	7
0 to 50	11
0 to 60	19
0 to 70	41
0 to 80	86
0 to 90	133
0 to 100	150
Total	150

It is relatively easy to determine the median and the first and third percentiles (Q1 and Q3) from an ogive:

- Modify the  $y$ -axis to be cumulative percentages instead of cumulative counts.
- Read off the 25<sup>th</sup>, 50<sup>th</sup> (median) and 75<sup>th</sup> percentile values from the  $x$ -axis (see the magenta lines on the graph at right).



## Center, Shape and Spread

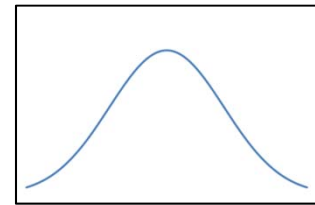
Whenever talking about a distribution of data, the key points to consider are **center, shape and spread**. Note that the Normal Distribution is **unimodal** and **symmetric**.

**Center.** Measures central tendency for a distribution:

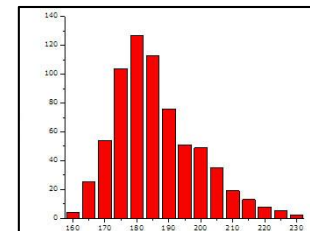
- **Mean:** The arithmetic average of the values in the distribution.
- **Median:** The middle value when the values are arranged from low to high. Note: if there are two middle values, the median is the mean of those two values.
- **Mode:** The value or values that occur most often. In particular, the number of modes in a set of data is a useful thing to know. Distributions may be unimodal (one mode), bi-modal (two modes), etc.
- Note: In a Normal Distribution, the mean, median and mode are all the same value.

**Shape.** Considerations relative to the shape of a distribution:

- **Symmetric:** A symmetric curve's left and right sides are close to mirror images of each other.
- **Bell-Shaped:** Also, called **mound-shaped**) a bell-shaped curve looks like a bell. The Normal Curve is the prototypical curve of this type.
- **Skewed:** A skewed curve has a tail to the right or left. If the tail is on the right the curve is said to be skewed to the right; if the tail is on the left the curve is said to be skewed to the left.



**Symmetric and Bell-Shaped**



**Skewed Right**

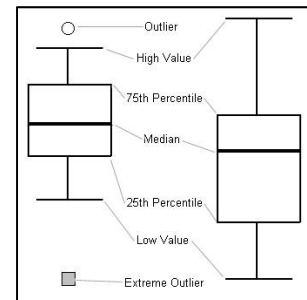
**Spread.** The spread of a distribution describes how varied the data are relative to its mean. This is typically measured by the standard deviation of the distribution.

- For a population,  $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}}$ , where  $\mu$  is the population mean and  $n$  is the size of the population.
- For a sample,  $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ , where  $\bar{x}$  is the sample mean and  $n$  is the number of observations in the sample.

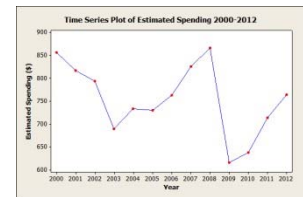
## Types of Plots

**Boxplot:** A boxplot (also called **box-and-whiskers plot**) is used to display **quantitative data**. It relies on a 5-number summary (see the chart below) for the variable being considered. It requires us to know the **minimum**, **1<sup>st</sup> quartile**, **median**, **3<sup>rd</sup> quartile**, and **maximum values** of the variable. Outliers may also be shown separately.

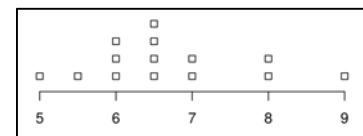
Min	Q1	Median	Q3	Max
13.0	15.0	16.5	18.0	22.0



**Timeplot:** A timeplot is used to display **quantitative data** which can be measured over time. Generally, the  $x$ -axis is represented as time, and the  $y$ -variable is the one being measured at various points in time.



**Dotplot:** A dotplot is used to display **quantitative data**. It places a dot perpendicular to an axis for each case in the data. It's like a stem-and-leaf plot but with dots instead of numbers.



**Stemplot:** A stemplot (also called a stem-and-leaf plot) is used to display quantitative data. It organizes the data in a table so that a portion of each value is on the left (the stems) and the balance of each value is on the right (the leaves). Advantages of a stem-and-leaf plot:

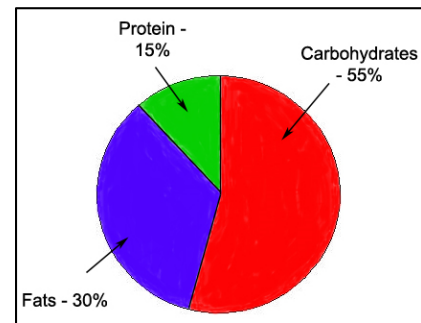
- It approximates the shape of a histogram if the bin widths are the same between the two displays.
- It preserves the actual values of the data, so that if we need them to calculate specific statistics like the median, the 1<sup>st</sup> quartile, or the 3<sup>rd</sup> quartile, we can do it without a lot of additional work.

12, 23, 35, 23, 14, 25, 32, 18	
Stem	leaf
1	2, 4, 8
2	3, 3, 5
3	2, 5

**Histogram:** A histogram is used to display **quantitative data**. It slices up data into equal-width bins and counts the number of occurrences of the variable in each bin.



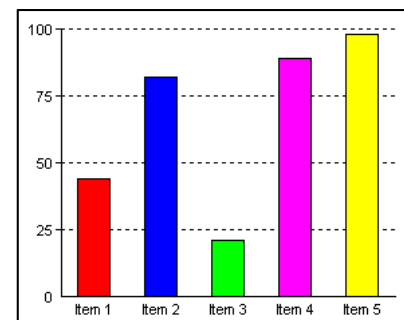
**Pie Chart:** A pie chart (also called a **circle graph**) is used to display **categorical data**. They show the set of possibilities for a variable within a circle. The circle is sliced into pieces whose areas are proportional to the fraction of the total represented by each possible value of the variable.



## Bar Charts (Bar Graphs)

Bar charts are generally used to display **categorical data**, but may also be used to display **quantitative data**. They typically show rectangular bars of equal width. The bars may be displayed either vertically (most common) or horizontally. The height or length of each bar is proportional to the values represented. In fancier charts, the bars may be shown as three-dimensional figures and may be prisms, cylinders or even cones.

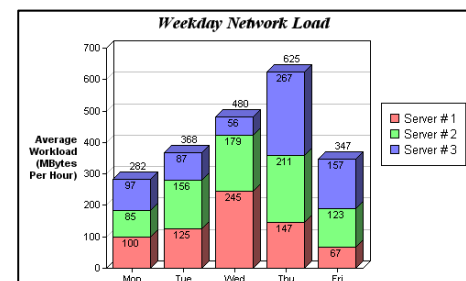
**Simple Bar Chart:** Displays a single set of bars representing data in a number of categories. This type of chart is typically used to compare frequencies of a categorical variable for a single population.



**Multiple Bar Chart:** Displays multiple sets of bars for data in a number of categories. This type of chart is typically used to compare frequencies of a categorical variable for two or more populations.



**Stacked Bar Chart:** Displays bars that are broken into categories. This type of chart is typically used to compare parts to the whole for a categorical variable.



## Marginal Distributions

A **Marginal Distribution** is one where we consider a subset of the data presented that relates specifically to the question being asked. In this question, we must look at only the total line for the “Like Dogs” variable. Whether people like cats is immaterial for this question. The table, then, collapses to the following values (blue), from which we can calculate the required percentages (green):

		Like Dogs		Total
		Yes	No	
Like Cats	Yes	194	21	215
	No	110	10	120
Total		304	31	335

Likes Dogs		
Yes	No	Total
304	31	335
90.7%	9.3%	

Note that the percentages should add up to 100%. If they do not, there should be a clearly identifiable reason why this is. If there is no clear reason, you may have made a math error; check your work.

**Independence** can best be tested by calculating the marginal distribution percentages for each row of data (excluding the total row). Use the  $\chi^2$  test to formally check for Independence.

- Start with the data given (blue); then show the marginal distribution percentages (magenta) for each row.

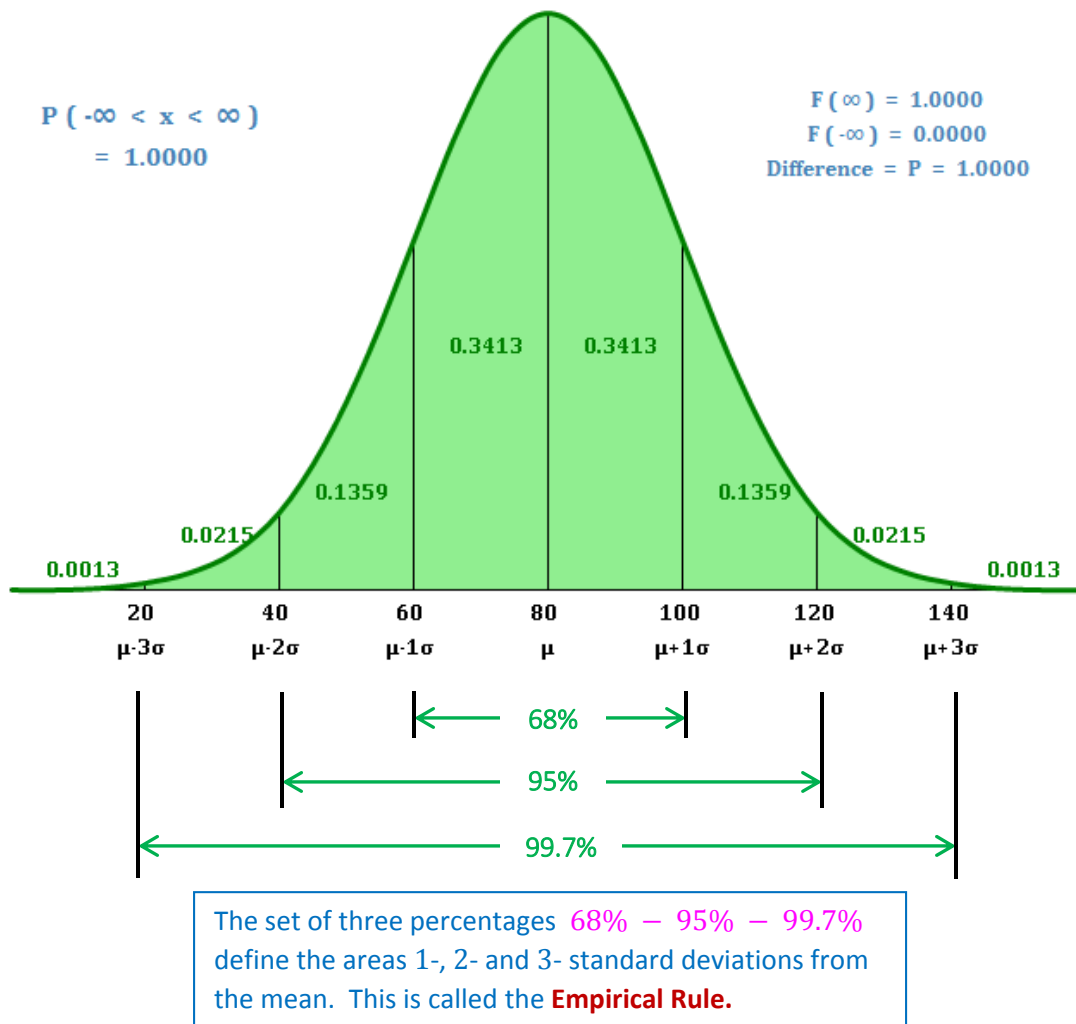
		Likes Dogs		Total
		Yes	No	
Likes Cats	Yes	194	21	215
	No	110	10	120

		Likes Dogs		Total
		Yes	No	
Likes Cats	Yes	90.2%	9.8%	100%
	No	91.7%	8.3%	100%

- Based on the percentages shown, there is little difference between the percentages in the two rows. So we can say, based on our limited sample, that it appears that whether someone likes dogs does not depend on whether they like cats, and vice versa. Therefore, the variables appear to be independent.

## Normal Distribution

Normal Distribution with  $\mu = 80$  and  $\sigma = 20$



**Z-Score:** The z-score is the number of standard deviations above or below the mean represented by the data value. That is,  $z = \frac{x - \mu}{\sigma}$ .

**Points of Inflection** on the Normal Curve occur at  $z = \pm 1$ . Alternatively, we can say that the points of inflection occur **one standard deviation above and below the mean**. This can be proven using elementary Calculus.

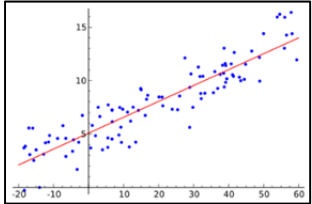
### A note on notation:

Use the Greek symbols  $\mu$  and  $\sigma$  for a known total population distribution (e.g., if we are given that a particular distribution is known to be Normal).

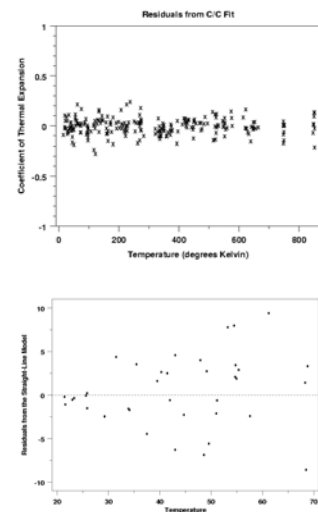
Use the English letters  $\bar{x}$  and  $s$  when we are provided with data that is a sample (i.e., subset) of a population for which the distribution is unknown.

## Exploring Bivariate Data

### Definitions

- **Association** is a relationship between two variables that results in them being statistically dependent. In Answer E) above, there may be an **association** between the categorical variables “gender” and “political party.” However, there cannot be a **correlation** between them because of the specific conditions required for correlation (listed above). In this case, they are not quantitative variables.
- **Correlation** is a measure of the **strength** and **direction** of the **linear relationship** between two **quantitative** variables. Note that **Correlation does not indicate cause and effect**. It merely indicates that two variables tend to move in tandem.
  - The term **Correlation** is often used as a shortcut to reference the **Correlation Coefficient** (i.e., the value of  $r$ ).
  - Both variables must be **quantitative** in order to calculate a correlation coefficient.
  - The correlation coefficient has **values from -1.00 and 1.00 only**.
  - A correlation coefficient of **zero** indicates that there is **no linear relationship** between two variables.
  - The correlation coefficient **does not have units**. It is a number only.
- **Lurking Variable**. A variable that is the cause of two other variables being correlated.
- **Confounding**. When we are unsure which variable is causing an effect, we say that the variables are **confounded**.
- **Scatterplot**. A scatterplot is a plot of x- and y-values on a set of axes (blue points in the diagram). It is used to display data through which a regression line (red line in the diagram) may be plotted.
 
- **Linear Regression** is the process whereby a straight line is fitted to a set of data points so as to **minimize the square errors** between the actual y-values and the y-values on the regression line.
  - The general linear regression equation in z-form is:  $\hat{z}_y = r \cdot z_x$ .
  - The general linear regression equation in slope-intercept form is:  $\hat{y} = a + bx$ .

- Residuals** are the differences between observed and predicted values (i.e.,  $y - \hat{y}$ ). Residuals are the variations in the data that are **not explained by the model**. **Residuals should appear to be randomly distributed, and centered around zero**. If not, their scatter plot would show some kind of pattern, and the researcher may conclude that they have not adequately explained the association between the two variables.  $e = y - \hat{y}$
- Influential Points** are those whose omission results in a very different model. Note that Influential Points do not necessarily have large residuals. They are influential because their  $x$ -values are far from the mean ( $\bar{x}$ ) and, in calculating  $r$ , we use the term  $(x - \bar{x})^2$ . Any time you multiply something by the square of a large number, it can have a big impact, even if the thing being multiplied is not, itself, large.
- Leverage** refers to the outsized effect on a regression analysis caused by a point with an  $x$ -value far from the mean. This occurs because the term  $(x - \bar{x})^2$  in the formula for the correlation coefficient is very large for this kind of point.
- Extrapolation** is extending the results of a regression beyond the values for which the scatterplot was developed. Extrapolation is generally a bad idea because it can easily generate erroneous results. You cannot assume that the relationship between a pair of variables will extend beyond the values for which the scatterplot was developed.
- Predictive Power.** The ability of the model to “predict” a  $y$ -value based on a given an  $x$ -value. **A model with high predictive power will have residuals close to zero**. Consider the two residual plots shown at right. The top one has randomly distributed residuals that are close to zero. The bottom one has randomly distributed residuals but they are not close to zero. Both appear to be good models. However, the top plot is indicative of a model that has high predictive power, while the bottom one is not.
- Regression to the Mean** is the phenomenon where “each predicted  $y$  tends to be closer to its mean (in standard deviations) than its corresponding  $x$  was.” (Bock, p. 174.) **Values that are further from their means are more likely to be influential.**





## Formulas Relating to the Coefficient of Correlation of a Sample

$$n = \text{sample size} \quad \bar{x} = \frac{\sum x}{n} \quad \bar{y} = \frac{\sum y}{n}$$

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

$$s_y^2 = \frac{\sum(y - \bar{y})^2}{n - 1} = \frac{\sum y^2 - \frac{(\sum y)^2}{n}}{n - 1} \quad s_y = \sqrt{s_y^2}$$

$$z_x = \frac{x - \bar{x}}{s_x} \quad z_y = \frac{y - \bar{y}}{s_y}$$

Covariance:  

$$s_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1}$$

## Formulas for $r$ (Pearson's Coefficient of Correlation)

$$r = \frac{\sum z_x z_y}{n - 1}$$

$$r = \frac{s_{xy}}{s_x s_y}$$

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1) s_x s_y}$$

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Note: All of these formulas for  $r$  are equivalent. That is, they all produce the same value of  $r$ .

Additionally, if  $(n - 1)$  is replaced by  $n$  throughout all of the calculations, the value of  $r$  is unchanged.

The Coefficient Of Determination,  $R^2$  is calculated as  $r^2$ . So,  $R^2 = r^2$ . (i.e.,  $R$  and  $r$  are the same guy.)

## Least Squares Line

Equation of the line:  $\hat{y} = a + bx$  where:  $b = r \frac{s_y}{s_x}$  and  $a = \bar{y} - b\bar{x}$

z-form:  $z_y = r \cdot z_x$

## Residuals

For each value of  $x$ , there is a residual error from the least squares line:  $e = y - \hat{y}$ .

The standard deviation of the residuals is:  $s_e = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}}$

We define  $(1 - r^2)$  to be the variation left in the residuals.

Note that:  
 $\sum y = \sum \hat{y}$ , so  
 $\sum e = 0$ , or  $\bar{e} = 0$

## Linear Combinations of Parameters and Statistics

Add a constant $k$ to all data values in a distribution		Multiply all values in a distribution by a constant $k$	
<b>Center:</b>		<b>Center:</b>	
Mean ( $\mu$ or $\bar{x}$ )	Add $k$	Mean ( $\mu$ or $\bar{x}$ )	Multiply by $k$
Median ( $M$ )	Add $k$	Median ( $M$ )	Multiply by $k$
Mode	Add $k$	Mode	Multiply by $k$
<b>Spread:</b>		<b>Spread:</b>	
Range	Unchanged	Range	Multiply by $k$
IQR	Unchanged	IQR	Multiply by $k$
Std Dev ( $\sigma$ or $s$ )	Unchanged	Std Dev ( $\sigma$ or $s$ )	Multiply by $k$

**Theorem:** When  $y = a + bx$ , where  $a$  and  $b$  are constants, it is true that  $z_y = z_x$ .

**Proof:** Let:  $y = a + bx$ , where  $a$  and  $b$  are constants

$z_x$  be the z-score for variable  $x$

$z_y$  be the z-score for variable  $y$

Notice from the table above (using statistical notation) that:

$\bar{y} = a + b\bar{x}$  (means are affected when adding a constant, as well as when multiplying by a constant)

$s_y = b \cdot s_x$  (standard deviations are affected when multiplying by a constant, but not when adding a constant)

Then,

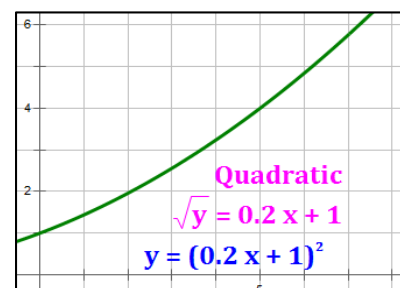
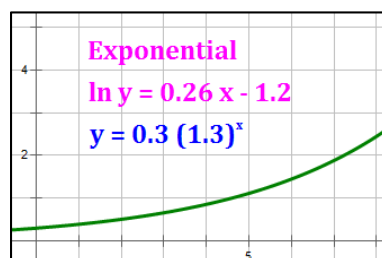
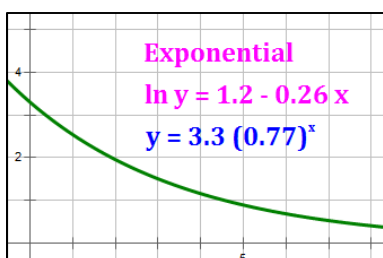
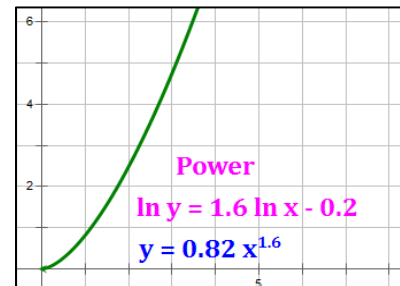
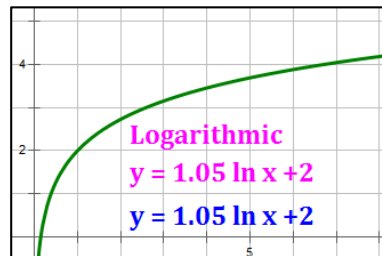
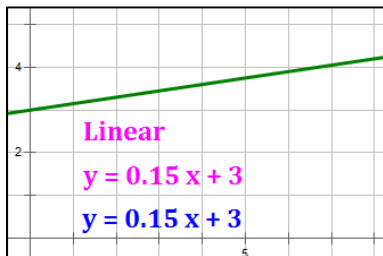
$$z_y = \frac{y - \bar{y}}{s_y} = \frac{(a + bx) - (a + b\bar{x})}{b \cdot s_x} = \frac{bx - b\bar{x}}{b \cdot s_x} = \frac{b(x - \bar{x})}{b \cdot s_x} = \frac{(x - \bar{x})}{s_x} = z_x$$

Therefore, when  $y = a + bx$ , where  $a$  and  $b$  are constants, it is true that  $z_y = z_x$ .

## Types of Regression Models

- **Linear** means increases or decreases of the **same amount** each year. This is the correct answer.
- **Exponential** refers to a model that follows the pattern of a logarithmic curve. The model values increase or decrease by the **same factor, or percentage**, each year.
- **Logarithmic** refers to a model that follows the pattern of a logarithmic curve.
- **Power** refers to a model where **the logarithm of  $y$  is plotted as a linear function of the logarithm of  $x$** , e.g.,  $\ln y = 1.6 \cdot \ln x - 0.2$ . This is roughly equivalent to plotting  $y$  against a power of  $x$ , e.g.,  $y = 0.82 \cdot x^{1.6}$  (these two equations are roughly equivalent because  $e^{-0.2} \sim 0.82$ ). Hence the name "Power."
- **Quadratic** refers to a model that follows the pattern of a **second degree curve**.

Examples of the models mentioned above are provided below (they are defined in linear regression form in magenta). Equivalent "y =" forms are given in dark blue. Note that the two forms are the same for the linear and logarithmic models.



## **Straightening Data in a Scatterplot Selecting a Good Re-Expression Model**

### **What Is All This Stuff?**

Here's what is included:

- Graphs of the three main patterns of data points that the student is likely to encounter in scatter plots, along with suggestions on which re-expression models to try.
- Graphs of residual plots that the student is likely to encounter, along with suggestions on which re-expression models to try.
- Scatter plots developed using various model types, their associated regression lines, re-expressions used to make the model linear, and some ideas on when to use the re-expressions.

### **Explanation**

In an effort to assist the student in selecting a model to use in “straightening” data in a scatterplot, I developed a series of models described in the following pages. Using the patterns in the scatterplots and residual plots generated by these models, I created the summaries shown.

While it is not possible to identify and chart every pattern that may arise, I am hopeful that reviewing the information in this package may help the student identify patterns and quickly identify which re-expressions make the most sense to try, depending on the pattern observed in a scatterplot that is not nearly linear. If the student runs into a different pattern than those shown, I am hopeful that they will be able to identify a similar pattern in the summaries and, thereby, select an appropriate model.

### **Why Is it Difficult to Pick a Model?**

Rarely will it be obvious that a single model is much better than every other model. There will often be more than one model which makes a good choice in “straightening” data. Why is this? The graphs on the next page illustrate why it is so difficult.

Figure 1, below, summarizes a number of models that work with scatterplots where the data are simultaneously rising and flattening. Notice that all of these models have the same general pattern. And ... there are more models that have similar patterns.

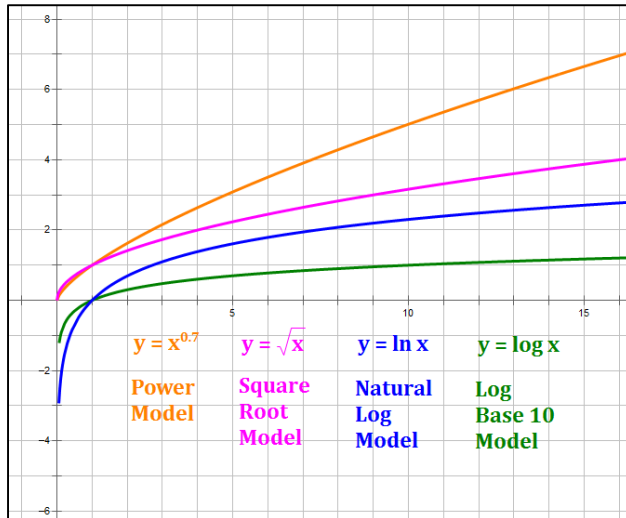


Figure 1

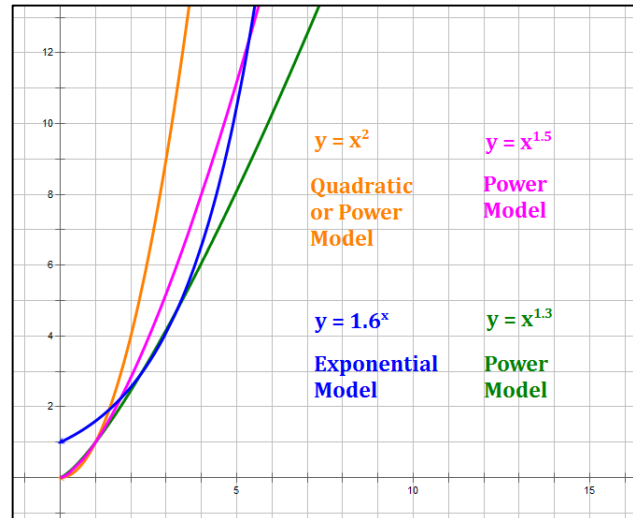


Figure 2

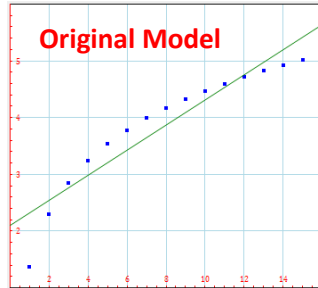
Figure 2, above, summarizes a number of models that work with scatterplots where the data are simultaneously rising and concave up. Notice that all of these models have the same general pattern. Again ... there are more models that have similar patterns.

Finding the right model is difficult because there are so many possibilities. And, another danger lurks. It is not always best to find the model with the best fit of the data (i.e., the highest  $R^2$ ). Selecting the right model often involves understanding the nature of the data what gives it its general shape. In hard science (e.g., chemistry, physics), fitting with the proper model depends on the underlying scientific nature of the data. In the social sciences (e.g., human behavior), however, it is often more of a challenge to select the right model for a given set of data.

### The Log Models – A Safe Haven

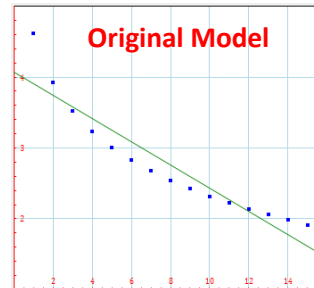
It is noteworthy that the three models that are most common are the three involving logarithms – the logarithmic model, the power model and the exponential model. The student should become proficient at using these models and knowing when each should be used. When in doubt, try the Power model; it is very Power-ful.

## Re-Expression Models to Consider Based on the Pattern of Points in a Scatterplot



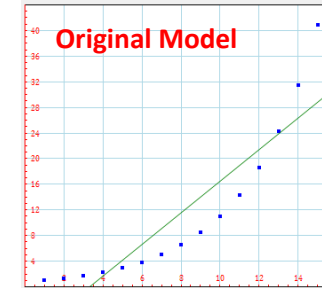
Models to consider:

- Logarithmic  
 $\hat{y} = a + b \log x$
- Square Root  
 $\hat{y}^2 = a + bx$
- Power  
 $\log \hat{y} = a + b \log x$



Models to consider:

- Logarithmic  
 $\hat{y} = a + b \log x$
- Exponential  
 $\log \hat{y} = a + bx$
- Reciprocal  
 $\frac{1}{\hat{y}} = a + bx$



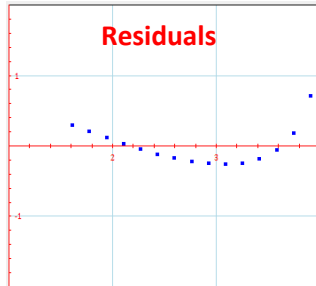
Models to consider:

- Quadratic (for counts)  
 $\sqrt{\hat{y}} = a + bx$
- Exponential  
 $\log \hat{y} = a + bx$
- Power  
 $\log \hat{y} = a + b \log x$

### Summary of When to Use Specific Models

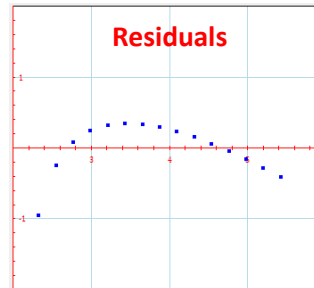
- **Logarithmic Model:**  $\hat{y} = a + b \log x$  Use when values flatten out on the right.
- **Exponential Model:**  $\log \hat{y} = a + bx$  Use when values increase or decrease at a constant percentage rate.
- **Power Model:**  $\log \hat{y} = a + b \log x$  Similar to the Square Root and Quadratic Models, but allows powers other than  $\frac{1}{2}$  or 2. A very nice feature is that the regression determines the appropriate power (i.e., the value of  $b$ ) to be used in the model.
- **Square Root Model:**  $\hat{y}^2 = a + bx$  For values that look like a square root function.
- **Quadratic Model:**  $\sqrt{\hat{y}} = a + bx$  Start here when counts are involved.
- **Reciprocal Model:**  $\frac{1}{\hat{y}} = a + bx$  Use when values are ratios, like “miles per hour.” Alternatively, invert the ratio and try a linear model.

## Re-Expression Models to Consider Based on a Graph of the Residuals



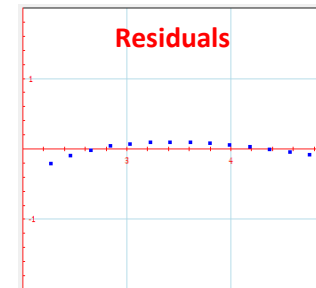
Models to consider:

- Logarithmic  
 $\hat{y} = a + b \log x$



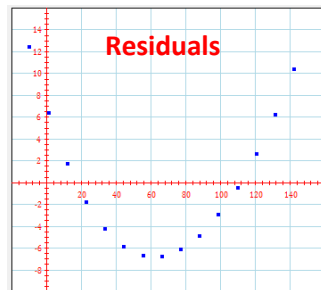
Models to consider:

- Logarithmic  
 $\hat{y} = a + b \log x$



Models to consider:

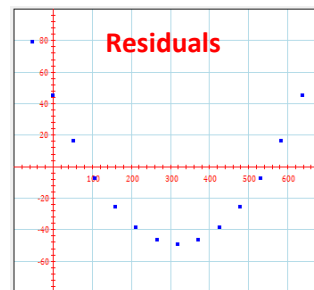
- Square Root  
 $\hat{y}^2 = a + bx$
- Power  
 $\log \hat{y} = a + b \log x$



Bottom  
Tilts Left

Models to consider:

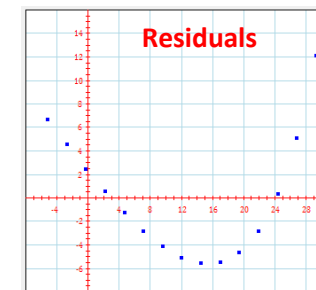
- Power  
 $\log \hat{y} = a + b \log x$



Bottom  
Centered

Models to consider:

- Quadratic  
 $\sqrt{\hat{y}} = a + bx$
- Power  
 $\log \hat{y} = a + b \log x$



Bottom  
Tilts Right

Models to consider:

- Exponential  
 $\log \hat{y} = a + bx$
- Reciprocal  
 $\frac{1}{\hat{y}} = .024 + .06x$

## Straightening Data in a Scatterplot Description of Individual Models

The following pages provide the background material used to create the summaries on pages 3 and 4. It is not necessary to study the detail for each of these models, but some familiarity with situations in which they should be used (green boxes) is recommended.

I attempted to choose models that the student may encounter on the AP Exam. Other models are certainly possible, so this list should not be considered exhaustive. Two models are shown on each page. Here is an explanation of what is shown for each model.

- **Top Line:** Name of the model and the function I used to generate the data points. The specific function I used is not of major importance to the student; rather, the pattern of the points in the scatterplot and in the residual plot should be noted.
- **Top Two Graphs:** These are the scatterplot and residual plot generated by the function chosen. These are very important and represent the kind of patterns the student should look for when choosing a re-expression model. Note that, for consistency, I used  $x$ -values from 1 to 15 in every model. The regression equation associated with the scatterplot is also shown.

Note that the residual plots shown in this packet are based on the definition of residual plots in your textbook. A point is plotted for each point in the original scatterplot. The abscissa (i.e., horizontal coordinate) of each point is the predicted value,  $\hat{y}$ , associated with the corresponding point in the original scatterplot, and the ordinate (i.e., vertical coordinate) of each point is the residual,  $y - \hat{y}$ , associated with that point. So, you can think of each ordered pair in the residual plot as having the coordinates:  $(\hat{y}, y - \hat{y})$ .

- **Action Line:** A description of the action to be taken by the student to produce a data re-expression that will attempt to “straighten” the data.
- **Bottom Graph:** A scatterplot of the re-expressed data. Note that the re-expressed data for each model in this packet all lie on straight lines; this results directly from the manner in which I created the original scatterplots. *In your work, you will want the re-expression to produce 1) a scatterplot with points that are close to a straight line, and 2) a residual plot with randomly distributed points that have no apparent pattern.*

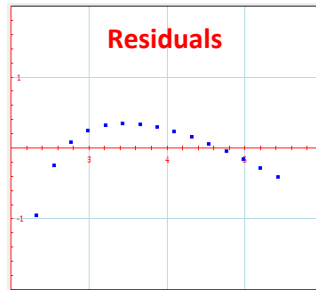
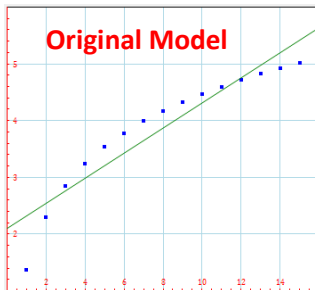
I show the general description of the re-expression model being used and the regression equation for the re-expressed data to the right of the graph. Note that this regression equation is equivalent to the sample function shown in the top line; this also results directly from the manner in which I created the original scatterplots.

- **Green Box:** Comments on when the model should be used and anything else I found particularly interesting in developing this packet.



## Logarithmic Model

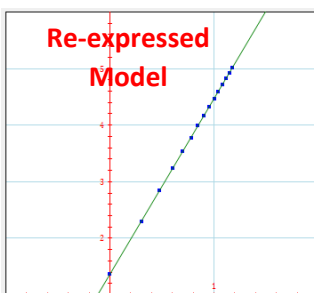
sample function:  $y = 3.1 \log(x) + 1.37$



Regression Equation:

$$\hat{y} = 2.10 + 0.22x$$

Action: Change  $x$ -axis values from  $x$  to  $\log(x)$



Model:  $\hat{y} = a + b \log x$

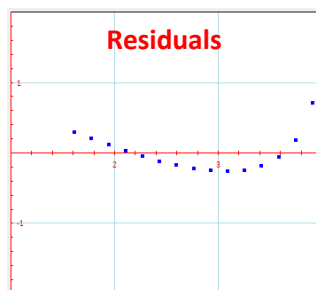
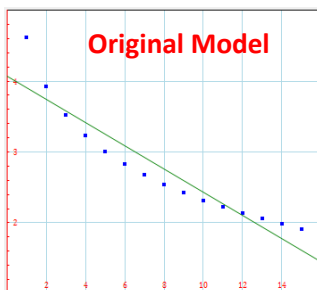
Regression Equation:

$$\hat{y} = 1.37 + 3.10 \cdot \log x$$

Use when values flatten out on the right.

## Logarithmic Model

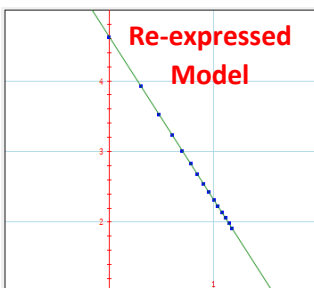
sample function:  $y = -2.3 \log(x) + 4.62$



Regression Equation:

$$\hat{y} = 4.07 - 0.16x$$

Action: Change  $x$ -axis values from  $x$  to  $\log(x)$



Model:  $\hat{y} = a + b \log x$

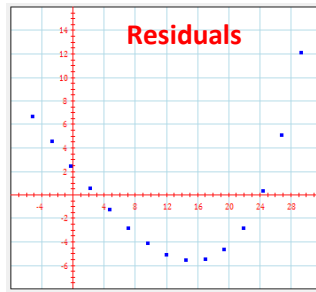
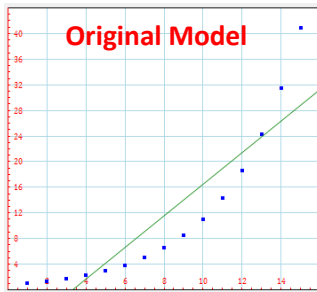
Regression Equation:

$$\hat{y} = 4.62 - 2.30 \cdot \log x$$

Use when values flatten out on the right.

## Exponential Model

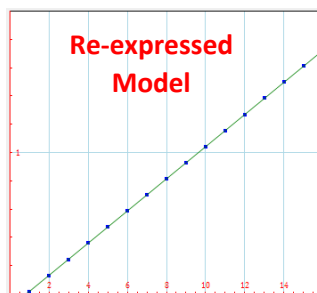
sample function:  $y = 0.8 \cdot (1.3)^x$



Regression Equation:

$$\hat{y} = -8.11 + 2.46x$$

Action: Change y-axis values from  $y$  to  $\log(y)$



Model:  $\log \hat{y} = a + bx$

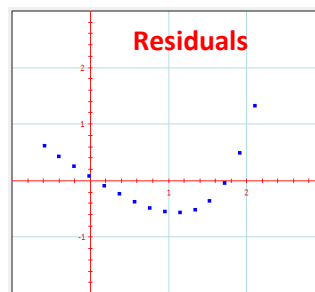
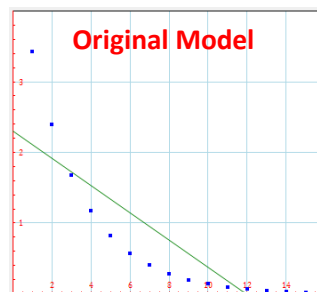
Regression Equation:

$$\log \hat{y} = -0.1 + 0.11x$$

Use when values increase or decrease at a constant percentage rate.

## Exponential Model

sample function:  $y = 4.9 \cdot (0.7)^x$



Regression Equation:

$$\hat{y} = 2.30 - 0.19x$$

Action: Change y-axis values from  $y$  to  $\log(y)$



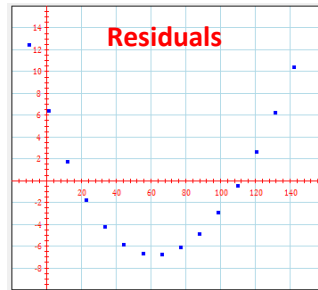
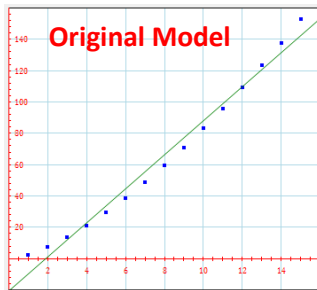
Model:  $\log \hat{y} = a + bx$

Regression Equation:

$$\log \hat{y} = 0.69 - 0.15x$$

Use when values increase or decrease at a constant percentage rate.

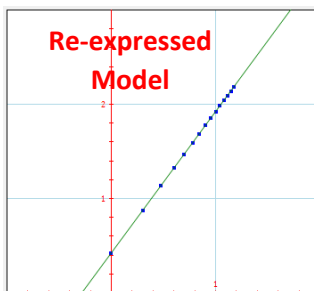
**Power Model** sample function:  $y = 2.63 \cdot x^{1.5}$



Regression Equation:

$$\hat{y} = -20.72 + 10.88x$$

**Action:** Change  $x$ -axis values from  $x$  to  $\log(x)$  AND  $y$ -axis values from  $y$  to  $\log(y)$



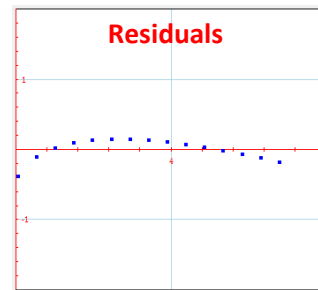
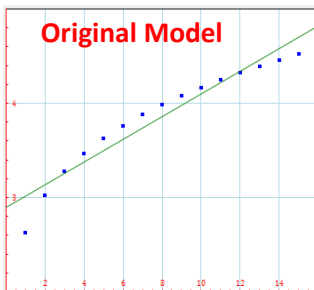
$$\text{Model: } \log \hat{y} = a + b \log x$$

Regression Equation:

$$\log \hat{y} = 0.42 + 1.50 \cdot \log x$$

Similar to the Square Root and Quadratic Models, but allows powers other than  $\frac{1}{2}$  or 2. A very nice feature is that the regression determines the appropriate power (i.e., value of  $b$ ) to be used in the model.

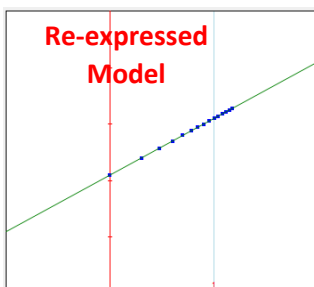
**Power Model** sample function:  $y = 2.63 \cdot x^{0.2}$



Regression Equation:

$$\hat{y} = 2.89 + 0.12x$$

**Action:** Change  $x$ -axis values from  $x$  to  $\log(x)$  AND  $y$ -axis values from  $y$  to  $\log(y)$



$$\text{Model: } \log \hat{y} = a + b \log x$$

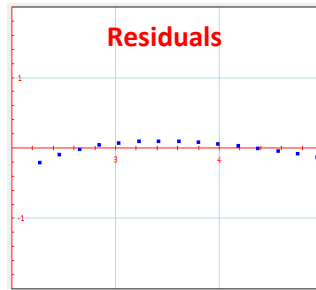
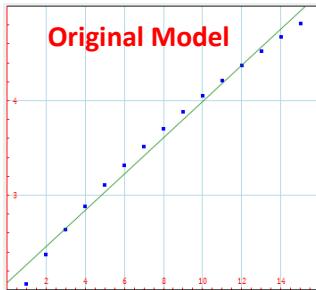
Regression Equation:

$$\log \hat{y} = 0.42 + 0.2 \cdot \log x$$

Similar to the Square Root and Quadratic Models, but allows powers other than  $\frac{1}{2}$  or 2. A very nice feature is that the regression determines the appropriate power (i.e., value of  $b$ ) to be used in the model.

## Square Root Model

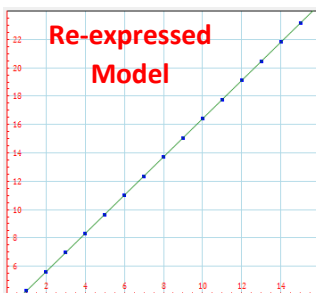
sample function:  $y = \sqrt{1.35x + 2.91}$



Regression Equation:

$$\hat{y} = 2.08 + 0.19x$$

Action: Change y-axis values from  $y$  to  $y^2$



Model:  $\hat{y}^2 = a + bx$

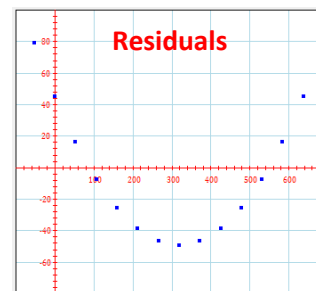
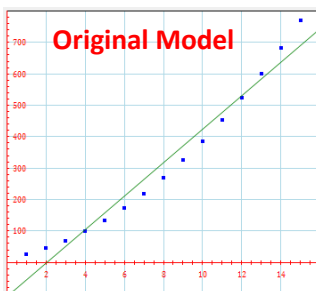
Regression Equation:

$$\hat{y}^2 = 2.91 + 1.35x$$

For values that look like a square root function.

## Quadratic Model

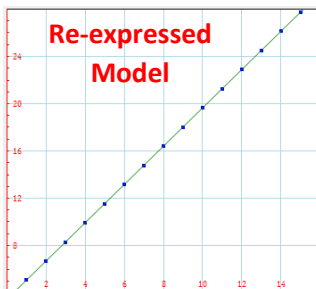
sample function:  $y = (1.62x + 3.44)^2$



Regression Equation:

$$\hat{y} = -107.14 + 53.14x$$

Action: Change y-axis values from  $y$  to  $\sqrt{y}$



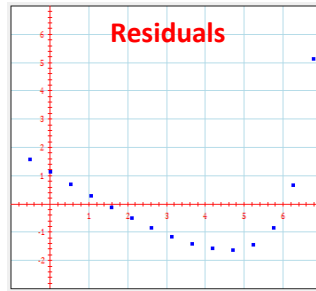
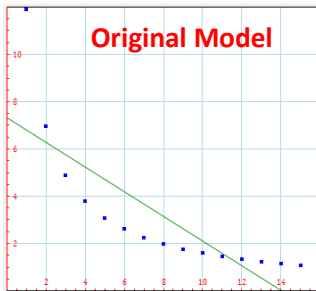
Model:  $\sqrt{\hat{y}} = a + bx$

Regression Equation:

$$\sqrt{\hat{y}} = 3.44 + 1.62x$$

Start here when counts are involved.

**Reciprocal Model** sample function:  $y = \frac{50}{3x+1.2}$



Regression Equation:

$$\hat{y} = 7.31 - 0.52x$$

Action: Change  $x$ -axis values from  $y$  to  $\frac{1}{y}$



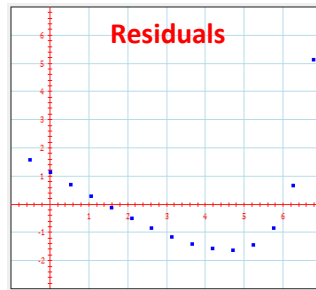
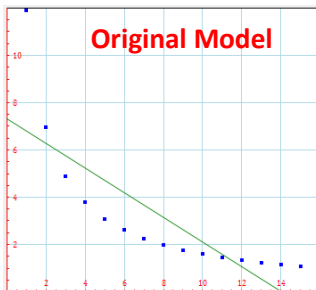
Model:  $\frac{1}{\hat{y}} = a + bx$

Regression Equation:

$$\frac{1}{\hat{y}} = .024 + .06x$$

Use when values are ratios, like "miles per hour."  
Alternatively, invert the ratio and try a linear model.

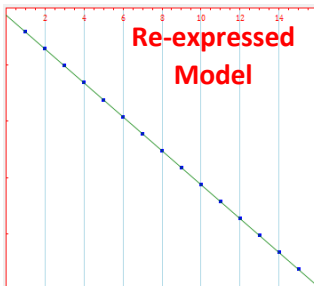
**Negative Reciprocal Model** sample function:  $y = \frac{50}{3x+1.2}$



Regression Equation:

$$\hat{y} = 7.31 - 0.52x$$

Action: Change  $x$ -axis values from  $y$  to  $\frac{-1}{y}$



Model:  $\frac{-1}{\hat{y}} = a + bx$

Regression Equation:

$$\frac{-1}{\hat{y}} = -.024 - .06x$$

Like the Reciprocal Model, but preserves the direction of the original curve.

## Chapter 12: Sample Surveys

### Terms and Notes

**Sample:** a subset of a population that is examined in order to determine information about the entire population.

**Types of Samples:** Note that all statistical sampling approaches have the common goal that chance, rather than human choice, is used to select the sample.

- **Cluster Sample:** a sampling approach in which entire groups (i.e., clusters) are chosen at random; a census is taken of each cluster. *Each cluster should be representative of the entire population.* All clusters should be heterogeneous and similar to each other. The problem with cluster samples is that the clusters are often not homogeneous and representative.
- **Convenience Sample:** a sample of individuals who are conveniently available. Convenience samples often fail to be representative.
- **Multistage Sample:** a sampling approach that combines several sampling methods. Example: stratify the country by geographic region; randomly select cities from each region; interview a cluster of residents from each city. Care should be taken at each step not to introduce bias.
- **Simple Random Sample (SRS):** a sample of size  $n$  in which each set of  $n$  elements has an equal chance of being selected. This is the standard against which other sampling methods are measured.
- **Stratified Random Sample:** the population is divided into subgroups (i.e., strata), and random samples are taken from each subgroup. This is better than a simple random sample if the strata are relatively homogeneous and different from each other. It results in reduced sampling variability, and can point out differences in responses among groups.
- **Systematic Sample:** individuals are selected systematically from a sampling frame (e.g., every 10<sup>th</sup> person). Can be representative if there is no relationship between the order of the sampling frame and the variables of interest.

**Randomization:** each member of a population is given a fair, random chance of selection in the sample. This reduces bias in a sample.

**Biased Sample:** one that over- or under-emphasizes some characteristics of the population. It is caused by poor design and is not reduced as sample size increases.

#### Types of Bias

- **Voluntary Response Bias:** occurs when sample participants are self-selected volunteers (i.e., those willing to participate).
- **Undercoverage Bias:** occurs when some members of the population are inadequately covered in a sample.
- **Nonresponse Bias:** occurs when respondents to a survey differ in meaningful ways from non-respondents.
- **Response Bias:** occurs when the question is asked in such a way that it influences the response.

**Sample Size:** the number of individuals in a sample.

**Required sample size** does **NOT** depend on the size of the population (as long as the population is large enough and our sample is less than 10% of the population).

**Representative Sample:** A sample whose statistics accurately reflect the corresponding population parameters.

**Sampling Frame:** a list of individuals from which the sample is drawn.

**Sampling Variability:** the natural tendency of randomly drawn samples to differ from one another.  
Note: sampling variability is not a problem.

**Pilot:** A small trial run of a survey used to determine if the questions are clear.

**Population:** the entire group of individuals that we hope to learn about.

**Census:** examination of information about every member of a population. This is the best approach when the population is small and accessible.

Why not do a census all the time?

- Difficult or expensive to complete.
- Populations rarely stand still. A census takes time and the population changes during it.
- A census is more complex than a sample.

**Parameter:** a descriptive measure (using a numerical value) of the population, e.g.,  $\mu$ ,  $\sigma$ . Also called a **population parameter**.

**Statistic:** a descriptive measure (using a numerical value) of a sample, e.g.,  $\bar{x}$ ,  $s$ . Also called a **sample statistic**.

Key Statistics and Parameters		
Name	Sample Statistic	Population Parameter
Mean	$\bar{x}$	$\mu$ (mu)
Standard Deviation	$s$	$\sigma$ (sigma)
Correlation	$r$	$\rho$ (rho)
Regression Coefficient	$b$	$\beta$ (beta)
Proportion	$\hat{p}$	$p$

### The Valid Survey

- What do I want to know?
- Am I asking the right respondents (i.e., do I have the right sampling frame)?
- Am I asking the right questions? Ask only questions that help you learn what you want to know. Be specific. In each question, either give a set of alternative answers (i.e., multiple choice) or ask for a numerical response, if possible. Ask questions in a neutral way (i.e., avoid bias).
- What will I do with the answers: will they address what I want to know?

## Chapter 13: Experiments and Observational Studies

### Terms and Notes

**Observational Study:** a study on data in which there is no manipulation of factors (i.e., variables). The subjects need not be random. Typically used to study differences between groups of subjects.

- **Retrospective Study:** The previous conditions or behaviors of subjects are studied.
- **Prospective Study:** Subjects are followed to observe future outcomes. No treatments are applied, marking the difference between a prospective study and an experiment.

**Matching:** Subjects who are similar in ways not under study are matched and placed in separate groups, to reduce random variation among the groups.

**Experiment:** A process in which the experimenter ...

- Actively and deliberately manipulates factor levels to create treatments,
- Randomly assigns subjects to the treatments, and
- Compares responses across treatment levels.

An experiment is the only way to (statistically) claim a cause and effect relationship.

#### Related Terms:

- **Random Assignment:** assigning experimental units to treatment groups at random.
- **Factor:** a variable whose levels are manipulated by the experimenter. If more than one variable is manipulated, the experiment has multiple factors.
- **Response Variable:** The outcome of the experiment; it is compared over the various treatment levels to draw conclusions about what was learned.
- **Experimental Unit:** The **subjects** or **participants** on whom the experiment is performed.
- **Levels:** The specific values that the experimenter chooses for a factor.
- **Treatment:** The process, intervention, or other controlled circumstance applied in the experiment. Treatments are created by varying the levels of one or more factors.

#### Principles of Experimental Design (“Control what you can and randomize the rest.”)

Required elements of an experiment

- **Control** sources of variation that are not being studied, but will have an impact on the results. This makes conditions as similar as possible for all treatment groups. *Note: the results of the study cannot be generalized (i.e., extrapolated) to other levels of the controlled variables.*
- **Replicate** over as many experimental units as possible. Just as for simulations, replication reduces random variability in an experiment.
- **Randomize** experimental units to treatments. This equalizes the effects of unknown or uncontrollable sources of variation.
- **Block** to reduce the effects of things that cannot be controlled. Blocking is like stratifying a sample; we separate dissimilar experimental units into separate blocks, then randomize by treatment within each block. In effect, the experiment is run separately on each block (see Figure 2 – next page); the results are combined across blocks for purposes of study.



## Diagramming an Experiment

Tomato plant example from the BVD textbook, pp. 299, 305.

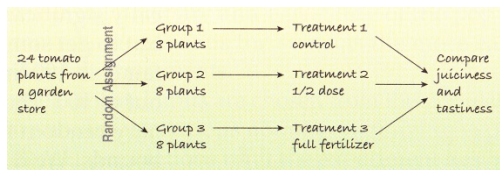


Figure 1: Experiment Diagram

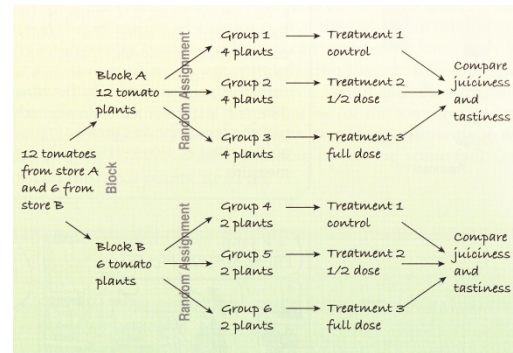


Figure 2: Experiment with Randomized Block Design

## Elements of an Experiment

Setting up the experiment

- **Plan:** State what you want to know.
- **Response:** Specify the response variable.
- **Treatments:** Specify the factor levels and treatments.
- **Experimental Units:** Specify the experimental units.
- **Experimental Design:** Control, replicate and randomly assign (also, block if it helps).
- **Draw a Diagram:** See Figures 1 and 2, above.
- **Perform the Experiment**
- **Describe the Results:** And, determine whether the results are statistically significant

The best experiments are usually:

- Randomized
- Double-blind
- Comparative
- Placebo-controlled

Consider using box plots to help compare the results of the various treatment groups.

## Other Terms Relating to Experimentation

- **Statistically Significant:** Observed differences that are too large to believe that they are the result of random fluctuation.
- **Control Group:** The group against which experimental results will be compared. This group receives one of: **null treatment** (i.e., no treatment), **placebo treatment** (i.e., “fake” treatment), or the **control treatment** (i.e., a treatment which is well understood).
- **Blinding:** Keeping individuals from knowing how the subjects are assigned to treatment groups. Anyone who is unaware of the assignments is said to be **blinded**. Blinding is used to reduce bias.
  - **Class A:** those who can *influence* the results (e.g., subjects, administrators, technicians).
  - **Class B:** those who *evaluate* the results (e.g., judges, treating physicians, statisticians).
    - **Single Blind:** When one of the above groups, but not both, is blinded.
    - **Double Blind:** When both of the above groups are blinded.
- **Placebo:** A “fake” treatment which is known to have no effect.
- **Placebo Effect:** The tendency in human subjects to exhibit a response even though they have received a placebo. This is often 20% or more of the subjects.
- **Blocking:** Separating dissimilar groups into blocks, and performing the experiment on each.
- **Completely Randomized Design:** Each subject has an equal chance of receiving each treatment.
- **Confounding:** When the effects of two or more factors cannot be separated.

## Part 4 – Probability

### Key Definitions

**Random Phenomenon:** a phenomenon about which we know what outcomes could happen, but not which particular values will happen.

**Trial:** a single occurrence of a random phenomenon.

**Outcome:** the value measured, observed or reported from a trial. In essence, the result of the trial.

**Event:** A collection of outcomes that is a subset of the sample space. We identify events so that we can assign probabilities to them, typically denoting them with bold capital letters (e.g., **A**, **B**, **C**).

**Sample Space:** The collection of all possible outcomes, typically denoted **S**.

**Bernoulli Trial:** a random event with the following properties:

- There are exactly two possible outcomes.
- The probability of success is the same for each trial.
- The trials are independent. (Note: if the population is finite, this condition is deemed to be met if the sample is random and the sample size is less than 10% of the total population.)

**Disjoint (or Mutually Exclusive):** Two events are disjoint when they cannot happen at the same time.

**Independent:** Two events,  $A$  and  $B$ , are independent if learning that one event occurs does not change the probability that the other event occurs. More formally,  $A$  and  $B$  are independent whenever  $P(B|A) = P(B)$ .

**Drawing With Replacement:** After an individual is drawn, it goes back into the pool, and may be drawn again.

**Drawing Without Replacement:** After an individual is drawn, it does NOT go back into the pool, so it cannot be drawn again.

**10% Condition:** When sampling without replacement, trials are not independent. However, if the sample is less than 10% of the total population, it is “independent enough.” We say that such a sample is “deemed independent.” (*This is the same condition mentioned under Bernoulli trial.*)

**Law of Large Numbers:** the relative frequencies of repeated independent events approaches the true relative frequencies of those events as the number of trials increases.

**Law of Averages:** The often-quoted Law of Averages does not exist and can lead a student to an incorrect conclusion.

**Probability:** the likelihood that an event will occur. For any event  $\mathbf{A}$ ,  $0 \leq P(\mathbf{A}) \leq 1$ .

**Empirical Probability:** Probability estimated based on repeated trial or frequency of occurrence (i.e., a simulation). Also called **Experimental Probability**. *When estimating probability using empirical methods, the student should always include the wording "According to this model ..."*

**Theoretical Probability:** Probability determined from mathematical processes, without the use of repeated trials.

**Personal Probability:** Subjective probability based on a person's experience and belief. *Personal probabilities are notoriously erroneous. Stay away from them and rely on the mathematics.*

**Legitimate Probability Assignment:** An assignment of probabilities to outcomes in a sample space is legitimate (but not necessarily accurate) if:

- each probability is between 0 and 1, and
- the sum of the probabilities for all outcomes is 1.

**Random Variable:** A variable that takes any of several numerical values as a result of a random event. Random variables are denoted by a capital letter (not bolded), such as  $X$  or  $Y$ .

**Continuous Random Variable:** A random variable that can take any numeric value within a range of values.

**Discrete Random Variable:** A random variable that can take a number of distinct outcomes which are not continuous.

**Probability Model:** A function that associates a probability with each value of a random variable.

**Expected Value:** The mean value of a random variable. It is denoted  $\mu$  or  $E(X)$ .

$$\mu = E(X) = \sum x_i \cdot P(x_i)$$

**Variance:** The expected value of the squared deviation from the mean. It is denoted  $\sigma^2$  or  $Var(X)$ .

$$\sigma^2 = Var(X) = \sum (x_i - \mu)^2 \cdot P(x_i)$$

**Standard Deviation:** The square root of the variance. It describes the spread of values in the model, and is denoted  $\sigma$  or  $SD(X)$ .

$$\sigma = SD(X) = \sqrt{Var(X)}$$

**Geometric Probability Model:** A model that tells us how many Bernoulli trials are required to achieve the first success. This model uses only one parameter,  $p$ , the probability of success.

Model definition:  $\text{Geom}(p)$

$$p = P(\text{success}) \quad q = 1 - p = P(\text{failure})$$

$X = \text{number of trials until first success}$  (note: this is a whole number)

$$P(X = x) = q^{x-1} \cdot p$$

$$\mu = E(X) = \frac{1}{p}$$

$$\sigma^2 = \text{Var}(X) = \frac{q}{p^2}$$

**Binomial Probability Model:** A model that tells us the number of successes in  $n$  Bernoulli trials. This model uses two parameters:  $p$ , the probability of success, and  $n$ , the number of trials.

Model definition:  $\text{Binom}(n, p)$

$n = \text{number of trials}$

$$p = P(\text{success}) \quad q = 1 - p = P(\text{failure})$$

$X = \text{number of successes in } n \text{ trials}$  (note: this is a whole number)

$$P(X = x) = {}_n C_x \cdot p^x \cdot q^{n-x} \quad {}_n C_x = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

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

$$\sigma^2 = \text{Var}(X) = npq$$

### Using the Normal Model as an Approximation of the Binomial Model

**Success/Failure Condition:** A Binomial Model is approximately normal if we expect at least 10 successes and 10 failures. That is, if  $np \geq 10$  and  $nq \geq 10$ .

**Combining Normal Models (by addition or subtraction):** When two independent continuous random variables ( $X$  and  $Y$ ) are Normal models, so is their sum or difference. Further:

$$\mu_{X+Y} = \mu_X + \mu_Y$$

$$\mu_{X-Y} = \mu_X - \mu_Y$$

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$$

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$$

Note that the sign in both variance formulas is "+".

## Part 4 – Probability

### Key Formulas

For any sample space,  $\mathbf{S}$ ,  $P(\mathbf{S}) = 1$

Let  $\mathbf{A}^c$  be the complement of  $\mathbf{A}$ , then  $P(\mathbf{A}^c) = 1 - P(\mathbf{A})$

Addition Rule: If  $\mathbf{A}$  and  $\mathbf{B}$  are disjoint events, then  $P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B})$

Multiplication Rule: If  $\mathbf{A}$  and  $\mathbf{B}$  are independent events, then  $P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A}) \cdot P(\mathbf{B})$

#### General Rules:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

#### Bayes' Rule:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B^c) \cdot P(B^c)}$$

#### Expected Value, Variance and Standard Deviation

$$\mu = E(X) = \sum x_i \cdot P(x_i)$$

$$\sigma^2 = \text{Var}(X) = \sum (x_i - \mu)^2 \cdot P(x_i)$$

$$\sigma = \sqrt{\text{Var}(X)}$$

$$E(X + Y) = E(X) + E(Y)$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

$$E(X - Y) = E(X) - E(Y)$$

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$$

Sometimes called the  
"Pythagorean Theorem  
of Statistics."

*Variance addition and subtraction formulas assume  $X$  and  $Y$  are independent. Note that you add the variances in both formulas.*

$$E(X + c) = E(X) + c$$

$$\text{Var}(X + c) = \text{Var}(X)$$

$$E(aX) = a \cdot E(X)$$

$$\text{Var}(aX) = a^2 \cdot \text{Var}(X)$$

$a$  and  $c$  are constants.

#### Calculator Functions (Binomial Distribution)

**2ND – DISTR – binompdf( – ENTER:** Enter  $n$ ,  $p$  and  $x$  to get the probability of  $x$  successes in the Binomial Distribution  $\text{Binom}(n, p)$ .

**2ND – DISTR – binomcdf( – ENTER:** Enter  $n$ ,  $p$  and  $x$  to get the probability of  $x$  or fewer successes in the Binomial Distribution  $\text{Binom}(n, p)$ .

**MATH – PROB – nCr – ENTER:** Enter  $n$  and  $r$  to get the value of the  ${}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$

## Chapter 18: Sampling Distribution Models

**Central Limit Theorem:** Also called the **Fundamental Theorem of Statistics**. The mean of a random sample is a random variable whose sampling distribution can be approximated using a Normal Model. As the size of the sample ( $n$ ) increases, the model becomes more Normal.

**Note:** The distribution of a random variable does NOT necessarily approach a Normal Model as  $n$  increases; it will only do this if the underlying population distribution is Normal. A [distribution of the mean of samples of the same size always approaches a Normal Model as  \$n\$  increases](#), regardless of the distribution of the original variable.

**Sampling Distribution:** distribution of results from simulating proportions from all samples. (Note: samples must be random and are typically of the same size. If they are not, the math gets more complicated.) The sampling distribution can be shown graphically in a histogram.

**Sampling Error:** Also called **Sampling Variability**, this term describes the differences in the values of  $\hat{p}$  from sample to sample.

### Key Assumptions and Conditions (RITSS):

- **Randomization:** Data should be sampled at random or generated from a properly randomized experiment.
  - If performing an experiment, subjects must be randomly assigned to treatments.
  - If performing a survey, use a simple random sample.
  - Alternatively, use another design that is free of bias.
- **Independence:** Sample values must be independent of each other.
- **Ten-Percent Condition:**  $n \leq 10\%$  of the population.
- **Sample Size**
  - $n$  must be sufficiently large in order to use the Normal Model. This requirement is called the **Large Enough Sample Condition**.
  - The required value of  $n$  depends on the nature of the distribution; in particular, the number of modes and the Skewness of the distribution affect the required value of  $n$ .
  - Larger values of  $n$  are required for proportions near 0% or 100%.
- **Success/Failure Condition:**  $n$  must be of the right size to allow the analysis to proceed. It must be true that  $n\hat{p} \geq 10$  and  $n\hat{q} \geq 10$ .

### Symbols and Formulas for Proportions:

- $n$  = number of observations in one sample (used to calculate one instance of a proportion)
- $p$  = probability of success (theoretical proportion)
- $\hat{p}$  = observed proportion
- $\mu(\hat{p}) = p$  is the theoretical mean of the proportions
- $SD(\hat{p}) = \sigma(\hat{p}) = \sqrt{\frac{pq}{n}}$  is the theoretical standard deviation of the proportion
- $N(\mu, \sigma)$  = General Normal Model with mean  $\mu$  and standard deviation  $\sigma$
- $N(\hat{p}, \sigma(\hat{p}))$  = Specific Normal Model of the observed proportions
- $z = \frac{\hat{p} - p}{SD(\hat{p})} = \frac{(\text{observed proportion}) - (\text{expected proportion})}{(\text{proportion standard deviation})}$

### Symbols and Formulas for Population Means:

- $n$  = number of observations in one sample
- $\mu$  = the population mean (theoretical mean)
- $\bar{x}$  = sample mean (observed mean)
- $\mu_{\bar{x}} = \mu$  is the theoretical mean of the means, which is the mean
- $SD(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$  is the theoretical standard deviation of the mean
- $N(\mu, \sigma)$  = General Normal Model with mean  $\mu$  and standard deviation  $\sigma$
- $N(\mu, \sigma_{\bar{x}})$  = Specific Normal Model of the mean
- $z = \frac{\bar{x} - \mu}{SD(\bar{x})} = \frac{(\text{sample mean}) - (\text{theoretical mean})}{(\text{standard deviation of the mean})}$

## Chapter 19: Confidence Intervals for Proportions

**Confidence Interval:**  $\hat{p} \pm ME = \hat{p} \pm z^* \cdot SE(\hat{p})$

- $\hat{p}$  = sample proportion
- $\hat{q} = 1 - \hat{p}$
- $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$  is the Standard Error of  $\hat{p}$
- $z^*$  = Critical Value
- Margin of Error:  $ME = z^* \cdot SE(\hat{p})$  (note: if not able to determine  $\hat{p}$ , use  $\hat{p} = 0.5$ )

**Recall:**

$$P(|z - \mu| \leq 1\sigma) \sim 68\%$$

$$P(|z - \mu| \leq 2\sigma) \sim 95\%$$

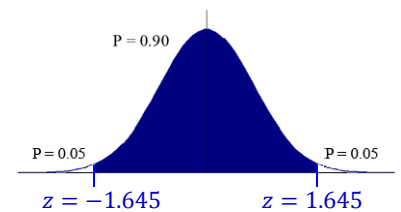
$$P(|z - \mu| \leq 3\sigma) \sim 99.7\%$$

**Explaining a Confidence Interval:** What does a confidence interval tell us?

Example: A 95% confidence interval tells us that we will capture the true value of the proportion in 95% of the confidence intervals developed from samples of the same size. Know this explanation!

### Critical Values

2-Tail Confidence Interval	Critical Value ( $z^*$ )	2-Tail Confidence Interval	Critical Value ( $z^*$ )
90%	1.645	98%	2.326
95%	1.960	99%	2.576



### Standard Deviation vs. Standard Error

- Use **Standard Deviation** when we know or hypothesize the value of the **population proportion,  $p$** .
- Use **Standard Error** when we estimate the value of  $p$  using the **observed (i.e., sample) proportion,  $\hat{p}$** .

**Key Assumptions and Conditions (RITSS):** See notes on Chapter 18.

**Width of Confidence Interval:** The higher the confidence level, the wider the confidence interval.



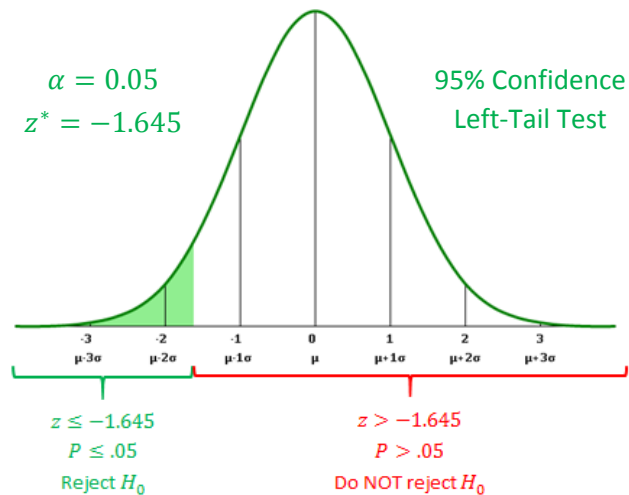
## Chapter 20: Hypothesis Testing for Proportions

### Hypotheses

- $H_0 = \text{Null Hypothesis}$ . This is the claim being tested. Generally, it will be the condition including an equal sign.
- $H_A = \text{Alternative Hypothesis}$ . This is what we conclude if  $H_0$  fails the test.
- Our goal is to determine that  $H_0$  is not correct given the hypothesis level of the test.
- Conclusions:
  - If  $H_0$  fails the test, we reject  $H_0$ , conclude that  $H_0$  is unlikely to be true, and, therefore,  $H_A$  is likely to be true.
  - If  $H_0$  passes the test, fail to reject  $H_0$ ; we cannot conclude anything.
  - Always identify the **P-value** of the test in your conclusion. For example, "We conclude that ... with 95% confidence."
  - Always check the **RITSS** conditions to determine if the Normal Model can be used.
- Note that we use  $SD(p)$  instead of  $SE(\hat{p})$  because we are hypothesizing the value of  $p$  in the population (note:  $p$  is a parameter because it relates to the population;  $\hat{p}$  is a statistic because it relates to the sample).

**P-Value:** the probability that the observed statistic value (or a more extreme value) could occur in a sample of the same size by natural sampling variation if  $H_0$  were correct.

In a Normal Model,  $N(p, SD(\hat{p}))$ , it is the area under the curve based associated with the calculated **z-statistic**.



### What do Hypotheses Look Like?

- 1-Sided Test
  - $H_0: p = p_0$
  - $H_A: p < p_0$  or  $H_A: p > p_0$
- 2-Sided Test
  - $H_0: p = p_0$
  - $H_A: p \neq p_0$

$p_0$  is the hypothesized value of the proportion.

$$SD(\hat{p}) = \sigma(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}}$$

is the theoretical standard deviation of the proportion.

## Standard Deviation vs. Standard Error

- Use **Standard Deviation** when we know or hypothesize the value of the **population proportion,  $p$** . Use **SD for hypothesis testing**.
- Use **Standard Error** when we estimate the value of  $p$  using the **observed (i.e., sample) proportion,  $\hat{p}$** . **Do NOT use SE for hypothesis testing**.

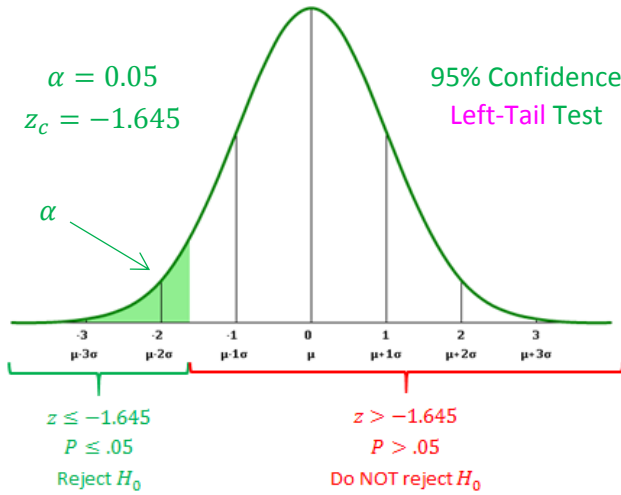
## Four Steps of Hypothesis Testing

1. **Hypotheses:** Set up  $H_0$  and  $H_A$  as shown above.
2. **Model:**
  - a. You are **modeling the sampling distribution of the proportion**. You need to memorize the words in green; you will be using them a lot.
  - b. Check the RITSS assumptions and conditions.
    - i. If they are met you say **“Because the conditions are satisfied, I can model the sampling distribution of the proportion.”**
    - ii. If they are NOT met you say **“Because the conditions are NOT satisfied, I cannot proceed with the test.”**
  - c. The test we are using is called a **one-proportion z-test**. We use the *Standard Normal Model* ( $\mu = 0, \sigma = 1$ ) to obtain a **P-value**. Use the statistic:

$$z = \frac{\hat{p} - p_0}{SD(\hat{p})} \quad \text{where} \quad SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}}$$

3. **Mechanics:** Perform the test to obtain a **P-value**.
  - a. The **P-value** is the probability that the observed statistic value (or a more extreme value) could occur if  $H_0$  were correct.
  - b. Put another way, the **P-value** is the probability that the observed results could occur if  $H_0$  is true. *This is a conditional probability.*
  - c. If the **P-value** is small enough, we reject  $H_0$ , and accept  $H_A$ .
4. **Conclusion:** The conclusion must state whether we **reject** or **fail to reject** the Null Hypothesis,  $H_0$ .
  - a. If we reject  $H_0$ , we are tacitly accepting the Alternative Hypothesis. Nevertheless, we use the language **“reject the Null Hypothesis.”**
  - b. If we **fail to reject  $H_0$** , we still do not accept it. We may revise our hypothesis or revise the  **$\alpha$ -value** and do another test.
  - c. Report the **P-value** associated with the test in order to identify the strength or your conclusion.

**Z-Values and P-Values for Various Tests – 95% Confidence**

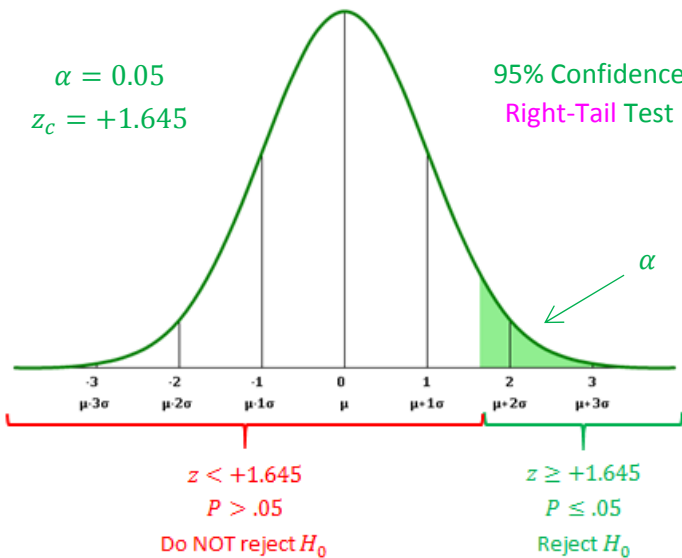
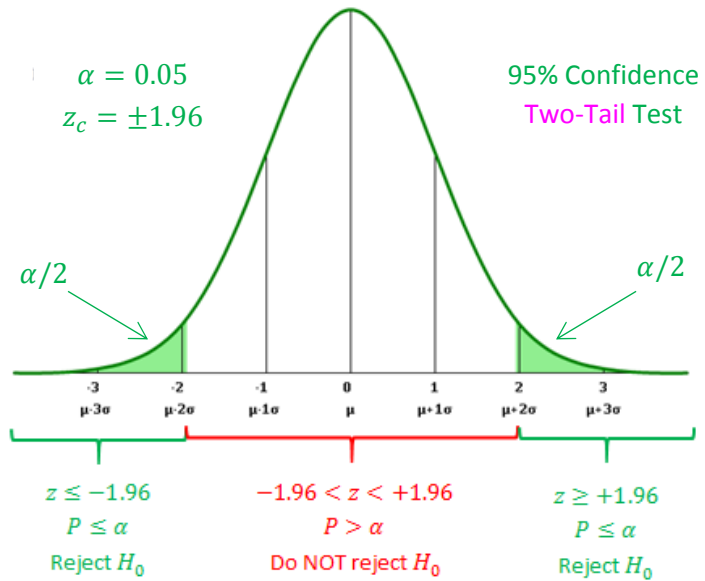


$\alpha$  is the significance level of the test (e.g., for 95% confidence,  $\alpha = .05$ ).

$P$  is the probability associated with the value of  $z$  calculated from the sample.

$z_c$  is the critical value of  $z$  used in the test.

In a 2-tail test, the  $P$ -value is **two times** the probability you would calculate in a one-tail test.



## Chapter 21: More About Tests and Intervals

**P-Value:** the probability that the observed statistic value (or a more extreme value) could occur if  $H_0$  were correct. The P-value measures the strength of evidence against the null hypothesis; the smaller the P-value, the greater the evidence that  $H_0$  should be rejected.

**Explaining a P-value:** The P-value (e.g., 1.7% if  $P = .017$ ) is the probability of seeing the observed statistic value, or one more extreme, in a sample of the same size by natural sampling variation (i.e., just by chance) if the null hypothesis were true. Know this explanation!

**$\alpha$ -Value:** the **significance level** of the test.

- $\alpha$  is the threshold used for the test.
- $\alpha$  = the probability of rejecting a correct  $H_0$ .
- In a 2-tail test, we have probabilities of  $\alpha/2$  in each tail.
- In a 1-tail test,  $\alpha$  is the probability in the lone tail.

When we reject  $H_0$  at a given level of  $\alpha$  (e.g.,  $\alpha = .05$ ), we say the results are statistically significant at the  $\alpha$  (e.g., 5%) level.

**Rules for Rejecting  $H_0$  based on  $P$  and  $\alpha$ .**

- If  $P < \alpha$ , reject  $H_0$ . Wording: the data provide sufficient evidence to reject the null hypothesis. Also, report  $P$  and  $\alpha$ .
- If  $P \geq \alpha$ , do not reject  $H_0$ . Wording: the data have failed to provide sufficient evidence to reject the null hypothesis. Also, report  $P$  and  $\alpha$ .

Note: When we reject  $H_0$ , it is good practice to provide a confidence level around the sample value of  $\hat{p}$  or  $\tilde{p}$ .

**$\beta$ -Value:** the probability that a test fails to reject a false null hypothesis.

**Power:** the probability that the test correctly rejects a false null hypothesis ( $= 1 - \beta$ ).

**Effect Size:** according to the BVD textbook, **Effect Size** is the distance between the null hypothesis value  $p_0$  and the true value,  $p$ , i.e.,  $|p - p_0|$ .

Note: More generally, the **Effect Size** is the difference between two means (e.g., treatment minus control) divided by the standard deviation of the two conditions. [This is not likely to be on a test.] See the following website for additional information:

[www.bwgriffin.com/gsu/courses/edur9131/content/Effect\\_Sizes\\_pdf5.pdf](http://www.bwgriffin.com/gsu/courses/edur9131/content/Effect_Sizes_pdf5.pdf).

**95% Confidence Interval (When the Success/Failure Condition Fails):**  $\tilde{p} \pm z^* \cdot \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}}$

- $\hat{p} = \frac{x \text{ successes}}{n \text{ observations}}$  sample proportion
- $\tilde{p} = \frac{x+2}{n+4}$  (note: if not able to determine  $\tilde{p}$ , use:  $\tilde{p} = 0.5$ )
- $\tilde{q} = 1 - \tilde{p}$
- $\tilde{n} = n + 4$
- $z^*$  = Critical Value for the Confidence Interval (e.g.,  $z^* \sim 1.96$  for a 95% CI)
- Generalized Formula for any confidence interval:

$$\tilde{p} = \frac{x+(z^*)^2/2}{n+(z^*)^2} \quad \text{Note: this will not be tested in this course}$$

$$\text{Using } z^* \sim 2 \text{ for a 95\% CI gives: } \tilde{p} = \frac{x+(z^*)^2/2}{n+(z^*)^2} = \frac{x+(2)^2/2}{n+(2)^2} = \frac{x+2}{n+4}$$

### Summary of $p$ 's and $n$ 's

	$p$	$n$
<b>Population</b>	$p$ = population proportion.	$n$ = population size.
<b>Hypothesis Testing</b>	$p_0$ = test value for population proportion. Ex: $H_0: p = p_0$ .	
<b>Sample</b>	$\hat{p} = \frac{x}{n}$ = sample proportion (the proportion observed in the sample). $x$ is the number of successes and $n$ is the sample size.	$n$ = sample size.
<b>Small Sample</b>	$\tilde{p} = \frac{x+2}{n+4}$ = sample proportion used when the success/failure test fails. $x$ is the number of successes and $n$ is the sample size.	$\tilde{n} = n + 4$ = the pseudo-sample size
<b>Critical Value</b>	$p^*$ = the critical value of the proportion corresponding to a specified value of $\alpha$ .	
<b>Probability</b>	$P$ -value = probability that $\hat{p}$ and/or $\tilde{p}$ could occur if $H_0$ were correct	

### Type I and Type II Errors

- **Type I Error:** Reject the null hypothesis when it is true.
- **Type II Error:** Do not (i.e., fail to) reject the null hypothesis when it is false.

Type I and Type II Errors		
Action Taken	$H_0$ is TRUE	$H_0$ is FALSE
Do Not Reject $H_0$	Correct action	Type II Error = $\beta$
Reject $H_0$	Type I Error = $\alpha$	Correct action ← Power

### Interaction of $\alpha$ , $\beta$ and Power

Interaction of $\alpha$ , $\beta$ and Power			
$n$ unchanged			
$\alpha \uparrow \Rightarrow$		$\beta \downarrow$	Power $\uparrow$
$\alpha \downarrow \Rightarrow$		$\beta \uparrow$	Power $\downarrow$
$n \uparrow \Rightarrow$	$\alpha \downarrow$	$\beta \downarrow$	Power $\uparrow$
$n \downarrow \Rightarrow$	$\alpha \uparrow$	$\beta \uparrow$	Power $\downarrow$

**Note:** A lower value of  $\alpha$  will provide a higher “standard of proof.” That is, rejecting a null hypothesis using a lower value of  $\alpha$  produces stronger results.

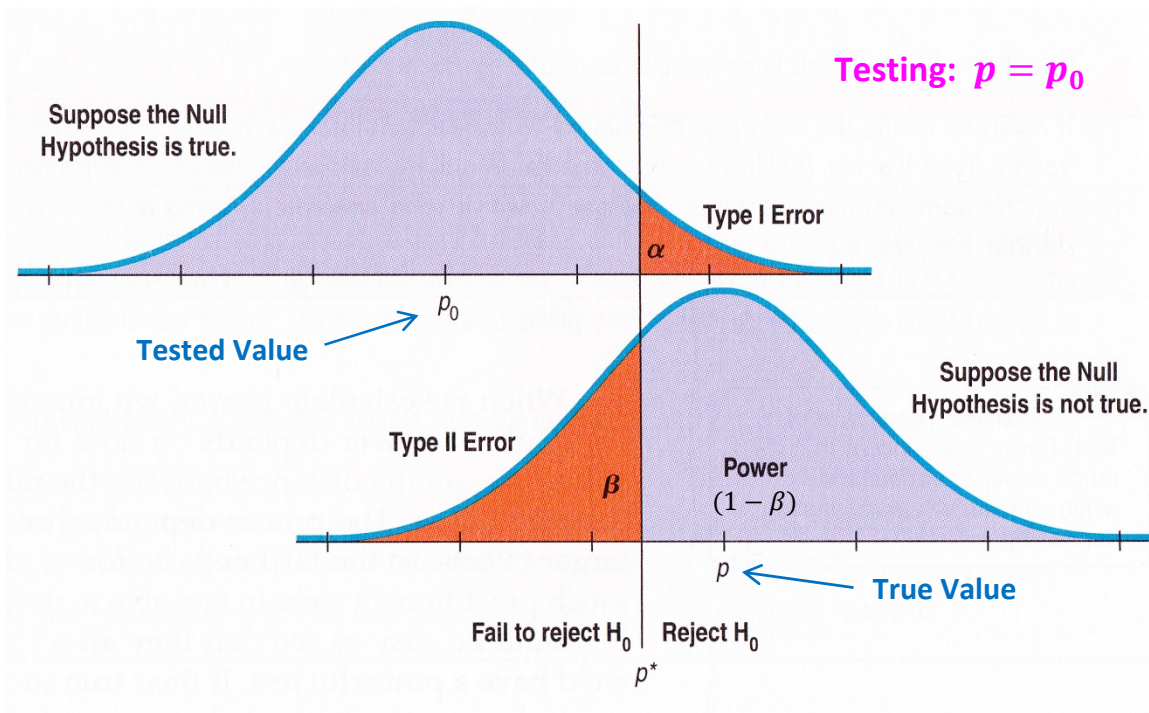


Illustration from page 494 of the BVD textbook.

## Chapter 22: Comparing Two Proportions

### Confidence Interval for the Difference of Two Proportions $(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot SE(\hat{p}_1 - \hat{p}_2)$

- **Mean:**  $\mu(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$
- **Variance:**  $Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}$
- **Standard Deviation:**  $SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$
- **Standard Error:**  $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$
- **Two-Proportion Confidence Interval:**  $(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot SE(\hat{p}_1 - \hat{p}_2)$

### Key Assumptions and Conditions for a Difference of Two Proportions (RITSS):

- **Randomization:** Data within each group should be sampled at random or generated from a properly randomized experiment.
- **Independence**
  - Sample values within each group must be independent of each other.
  - **The groups must be independent of each other. (NEW CONDITION)**
- **Ten-Percent Condition:**  $n \leq 10\%$  of the population for each group.
- **Sample Size**
  - $n$  must be sufficiently large in order to use the Normal Model. This requirement is called the **Large Enough Sample Condition**.
  - The required value of  $n$  depends on the nature of the distribution; in particular, the number of modes and the Skewness of the distribution affect the required value of  $n$ .
  - Larger values of  $n$  are required for proportions near 0% or 100%.
- **Success/Failure Condition:**  $n$  must be of the right size to allow the analysis to proceed. It must be true that  $n\hat{p} \geq 10$  and  $n\hat{q} \geq 10$  for each group. **Note, if we are given the counts of successes and failures in each sample, we can satisfy this condition simply by checking whether each of the values is 10 or greater.**

## Hypothesis Testing for the Difference of Two Proportions

- $H_0: p_1 - p_2 = 0$  (note: the two samples are drawn from the same population)
- $H_A: p_1 - p_2 \neq 0$  or  $H_A: p_1 - p_2 < 0$  or  $H_A: p_1 - p_2 > 0$
- $\hat{p}_1 = \frac{x_1}{n_1} = \frac{\text{successes in Sample 1}}{\text{observations in Sample 1}}$        $\hat{p}_2 = \frac{x_2}{n_2} = \frac{\text{successes in Sample 2}}{\text{observations in Sample 2}}$
- $\hat{p}_{pooled} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{\text{total successes}}{\text{total observations}}$

- **Success/Failure Condition:** Use  $\hat{p}_{pooled}$  and  $\hat{q}_{pooled}$  in the calculations. Make sure:  
 $n_1 \hat{p}_{pooled} \geq 10$      $n_1 \hat{q}_{pooled} \geq 10$      $n_2 \hat{p}_{pooled} \geq 10$      $n_2 \hat{q}_{pooled} \geq 10$

Note: as a proxy for these four tests, observe whether the number of successes and failures in each sample are 10 or greater.

- Use Standard Error in the calculation since there is no Standard Deviation available.

**Standard Error:**  $SE_{pooled}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_{pooled} \hat{q}_{pooled}}{n_1} + \frac{\hat{p}_{pooled} \hat{q}_{pooled}}{n_2}}$

- **Two-Proportion Test Statistic:**  $z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{pooled}(\hat{p}_1 - \hat{p}_2)}$



## Chapter 23: Inferences About Means

### Sample of Means:

- $n = \text{number of observations}$  in one sample
- $\mu = \text{the population mean}$  (theoretical mean)
- $\bar{x} = \text{sample mean}$  (observed mean)
- $SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}$  is the theoretical **standard deviation of the population mean** (you will use this only when the standard deviation of the population is known, i.e., it will not be used in this chapter)
- $s = \text{sample standard deviation}$  (observed standard deviation – with  $(n - 1)$  in the denominator)
- $SE(\bar{x}) = \frac{s}{\sqrt{n}}$  is the **standard error** of the sample mean (use this when the standard deviation of the population is unknown, in which case you calculate the standard deviation of the sample,  $s$ . This is what will be used in this chapter)
- **t-distribution:**  $t_{n-1}(\mu, s) = \text{General Student's t Model with } df = (n - 1) \text{ degrees of freedom, mean } \mu \text{ and standard deviation } s$
- **t-value:**  $t_{n-1} = \frac{\bar{x} - \mu}{SE(\bar{x})} = \frac{(\text{sample mean}) - (\text{theoretical mean})}{(\text{standard error of the mean})}$

If the sample size is large enough, the sample of means approximates a Normal Distribution.

### Confidence Interval:

- **One-sample t-interval for the mean:**  $\bar{x} \pm t_{n-1}^* \cdot SE(\bar{x})$
- **Standard Error:**  $SE(\bar{x}) = \frac{s}{\sqrt{n}}$
- **Degrees of Freedom:**  $df = n - 1$
- **High  $n$ :** As  $n \rightarrow \infty$ , t-distribution  $\rightarrow$  Normal Distribution
- **On the TI-84 Calculator:**
  - **STAT – CALC – 1-Var Stats – L<sub>1</sub>** to get the required statistics for **L<sub>1</sub>**
  - **$tcdf(\text{lower}, \text{upper}, df)$**  you have the  $t$ -value; you want the desired area
  - **$invT(\text{area}, df)$**  you have the area; you want the desired  $t$ -value

### Using $z$ vs. $t$

- If  $\sigma$  is known, use  $z$  and  $\sigma$
- If  $\sigma$  is **NOT** known, use  $t$  and  $s$

### Key Assumptions and Conditions for a Difference of Two Proportions (RITN):

- **Randomization:** Data should be sampled at random or generated from a properly randomized experiment.
- **Independence:** Sample values are independent of each other.
- **Ten-Percent Condition:**  $n \leq 10\%$  of the population for each group.
- **Nearly Normal Condition:** the source population should be nearly Normal, i.e., unimodal and symmetric. Check this for the sample using:
  - A histogram
  - A plot of Normal scores

### Size of Sample Required to Use $t$ -Distribution

- $n < 15$ : Population should be normally distributed
- $15 \leq n \leq 40$ : Population should be nearly normal (unimodal and symmetric)
- $n > 40$ : Generally safe unless the population is very skewed (also see special considerations below)

### Special Considerations Are Required For (MOBS):

- **M**ultimodal data (separate the data so that it becomes unimodal)
- **O**utliers (do analysis with and without outliers)
- **B**ias (remove the bias or use special techniques)
- **S**kewed data (increase the value of  $n$ , or use special techniques)

### Two Sets of Language for a Confidence Interval (p. 542 BVD Textbook)

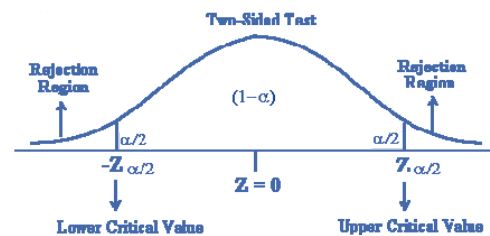
- A 95% confidence interval tells us that we will capture the true value of the statistic (identify which statistic) in 95% of the confidence intervals developed from samples of the same size.
- I am 95% confident that the true value of the statistic (state which statistic) is between the lower and upper values of my confidence interval.

## Hypothesis Testing (One-sample $t$ -test)

- $H_0: \mu = \mu_0$
- $H_A: \mu \neq \mu_0$  or  $H_A: \mu < \mu_0$  or  $H_A: \mu > \mu_0$
- **$t$ -value:**  $t_{n-1} = \frac{\bar{x} - \mu}{SE(\bar{x})} = \frac{(\text{sample mean}) - (\text{theoretical mean})}{(\text{standard error of the mean})}$
- **Standard Error:**  $SE(\bar{x}) = \frac{s}{\sqrt{n}}$
- **Degrees of Freedom:**  $df = n - 1$

## Relationship between Intervals and Tests

A “**Level  $C$** ” **Confidence Interval** contains all of the possible null hypothesis values that would NOT be rejected by a 2-sided hypothesis test with  $\alpha = 1 - c$ .



## Finding the Required Sample Size $n$ for a Given Margin of Error (ME)

- Use the formula:  $ME = t_{n-1}^* \cdot \frac{s}{\sqrt{n}}$ .
- **Note:** the book (BVD) substitutes  $z^*$  for  $t_{n-1}^*$ , using the formula  $ME = z^* \cdot \frac{s}{\sqrt{n}}$  which does not provide a precise value of  $n$ . The process described below is better, but you may not be required to know it for this course.
- This process should be **iterative** because the  $n$  appears twice in the equation we need to solve – once in the calculation of  $t_{n-1}^*$  and once in the denominator of  $\frac{s}{\sqrt{n}}$ .
- **Steps (example on next page):** **Note:** the book (BVD) stops after Step 1.
  1. Assume a preliminary value of  $n$  to calculate the value of  $t_{n-1}^*$ . Alternatively, replace  $t_{n-1}^*$  with  $z^*$  in the first calculation (i.e., in the first iteration, use the  $z$ -interval formula  $ME = z^* \cdot \frac{s}{\sqrt{n}}$  to calculate a preliminary of  $n$ ).
  2. Leave the  $n$  in the denominator because that is what we are trying to calculate.
  3. Calculate the value of  $n$  in the denominator from the formula:  $= t_{n-1}^* \cdot \frac{s}{\sqrt{n}}$ .
  4. Use the calculated value of  $n$  to calculate a new value of  $t_{n-1}^*$ .
  5. Repeat Steps 3 and 4 until you hone in on a single value of  $n$ .

**Example: Finding the Required Sample Size  $n$  for a Given Margin of Error (ME)**

**Note:** The book (BVD) describes a non-iterative process, which does not generate a precise value of  $n$ . The process laid out in this example goes beyond that to calculate a more precise value of  $n$ .

Assume we are given the following:  $ME = 3\%$   $CI = 95\%$   $s = 0.066$ . Find the required value of  $n$ .

**Step 1:** Begin with the  $z$ -interval formula:  $ME = z^* \cdot \frac{s}{\sqrt{n}}$ . The  $z^*$ -value for a 95% confidence interval is  $z^* = 1.960$ .

**Step 2:** The formula becomes:  $0.03 = 1.96 \cdot \frac{0.066}{\sqrt{n}}$  or  $n = \left(1.96 \cdot \frac{0.066}{0.03}\right)^2$

**Step 3:** Solve to get  $n = 18.59$  Remember to round up!  $n = 19$

**Step 4:** Using the new value of  $n = 19$ , calculate a 2-tail  $t$ -value:  $t_{df=18}^* = 2.101$

**Step 5:** Calculate  $n$  in the formula  $ME = t_{n-1}^* \cdot \frac{s}{\sqrt{n}}$

$$0.03 = 2.101 \cdot \frac{0.066}{\sqrt{n}} \quad \text{or} \quad n = \left(2.101 \cdot \frac{0.066}{0.03}\right)^2$$

Solve to get  $n = 21.36$  Remember to round up!  $n = 22$

**Repeat:**

Using the new value of  $n = 22$ , calculate a 2-tail  $t$ -value:  $t_{df=21}^* = 2.080$

$$0.03 = 2.080 \cdot \frac{0.066}{\sqrt{n}} \quad \text{or} \quad n = \left(2.080 \cdot \frac{0.066}{0.03}\right)^2$$

Solve to get  $n = 20.93$  Remember to round up!  $n = 21$

**Repeat:**

Using the new value of  $n = 21$ , calculate a 2-tail  $t$ -value:  $t_{df=20}^* = 2.086$

$$0.03 = 2.086 \cdot \frac{0.066}{\sqrt{n}} \quad \text{or} \quad n = \left(2.086 \cdot \frac{0.066}{0.03}\right)^2$$

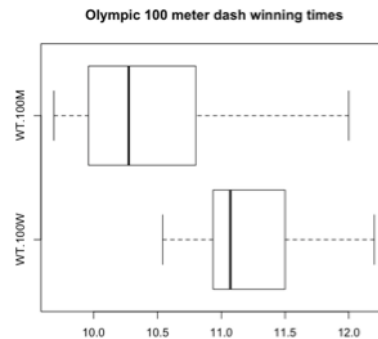
Solve to get  $n = 21.06$  Remember to round up!  $n = 22$

We now find ourselves in a repeating loop between  $n = 21$  and  $n = 22$ . In such a case, we would use the higher number to assure we get the proper confidence level. In this example, we would use  $n = 22$ .

## Chapter 24: Comparing Means

### Boxplots

**Draw boxplots** to compare the distributions of the two samples. The analysis may need to be run twice if there are outliers – once with and once without outliers.



### Sample of Means:

- Means:**

$$\bar{x}_1 = \text{mean of Sample 1}$$

$$\bar{x}_2 = \text{mean of Sample 2}$$

$$\mu(\bar{x}_1 - \bar{x}_2) = \bar{x}_1 - \bar{x}_2$$

- Variance:**  $Var(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

- Standard Deviation:**  $SD(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

- Standard Error:**  $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Not used in this chapter.

### Confidence Interval:

- Two-sample t-interval** for the difference of means:  $(\bar{x}_1 - \bar{x}_2) \pm t_{df}^* \cdot SE(\bar{x}_1 - \bar{x}_2)$

- Standard Error:**  $SE(\bar{x}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Margin of Error (ME)

- Degrees of Freedom:**  $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \cdot \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \cdot \left(\frac{s_2^2}{n_2}\right)^2}$

- On the TI-84 Calculator:**

- **STAT – TESTS – 2-SampTTest:** Enter the requested information and hit **ENTER**
- Read the **Degrees of Freedom** and **P-Value** from the output

### Key Assumptions and Conditions for a Difference of Two Proportions (RITN):

- **Randomization:** Data within each group should be sampled at random or generated from a properly randomized experiment.
- **Independence:** Sample values within each group are independent of each other. In addition the two groups must be independent of each other. Data must not be paired (paired data – e.g., data on husbands and wives – are handled in Chapter 25).
- **Ten-Percent Condition:**  $n \leq 10\%$  of the population for each group.
- **Nearly Normal Condition:** the source population for each sample should be nearly Normal, i.e., unimodal and symmetric. Check this for the sample using:
  - A histogram
  - A plot of Normal scores

### Size of Sample Required to Use $t$ -Distribution

- $n < 15$ : Population should be normally distributed
- $15 \leq n \leq 40$ : Population should be nearly normal (unimodal and symmetric)
- $n > 40$ : Generally safe unless the population is very skewed (also see special considerations below)

### Special Considerations Are Required For (MOBS):

- **Multimodal data** (separate the data so that it becomes unimodal)
- **Outliers** (do analysis with and without outliers)
- **Bias** (remove the bias or use special techniques)
- **Skewed data** (increase the value of  $n$ , or use special techniques)

### Hypothesis Testing (Two-sample $t$ -test)

- $H_0: \mu_1 - \mu_2 = \Delta_0$  Usually,  $\Delta_0$  will be equal to zero.
- $H_A: \mu_1 - \mu_2 \neq \Delta_0$  or  $H_A: \mu_1 - \mu_2 < \Delta_0$  or  $H_A: \mu_1 - \mu_2 > \Delta_0$
- **$t$ -value:**  $t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{SE(\bar{x}_1 - \bar{x}_2)} = \frac{\text{(difference in means)} - \text{(tested difference)}}{\text{(standard error of the mean)}}$
- **Standard Error:**  $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- **Degrees of Freedom:** see formula on previous page

### Pooled $t$ -Interval and $t$ -Test

- **Additional assumption required:** the two samples must have equal variances. This can happen, for example, in a randomized comparative experiment where the two samples are from the same population.

- **The Pooled  $t$ -Interval and  $t$ -Test are rarely used.**

- $$s_{pooled}^2 = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

- $$SE_{pooled} = \sqrt{\frac{s_{pooled}^2}{n_1} + \frac{s_{pooled}^2}{n_2}} = s_{pooled}^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- **Confidence Interval:**  $(\bar{x}_1 - \bar{x}_2) \pm t_{df}^* \cdot SE_{pooled}(\bar{x}_1 - \bar{x}_2)$

- **Hypothesis Test:**  $t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{SE_{pooled}(\bar{x}_1 - \bar{x}_2)} \quad df = n_1 + n_2 - 2$

Note: usually,  $\Delta_0$  will be equal to zero.

## Chapter 25: Inferences About Paired Data

When two samples are **not independent of each other**, they are called **Paired Data**. When paired data are analyzed, we look at **the differences of the paired values** and use the techniques described in Chapter 23.

### Pairing

- From an experiment: a type of blocking
- From an observational study: a type of matching

### Sample of Means:

- **$n$  = number of paired observations** of differences (data which are not paired are excluded from the Paired Data analysis)
- **$\mu_d$  = theoretical mean** of the differences in the underlying populations
- **$\bar{d}$  = sample mean of the differences of the pairs** (observed mean of differences)
- **$s_d$  = sample standard deviation** (observed standard deviation of the differences – with  $(n - 1)$  in the denominator)
- **$SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$**  is the **standard error** of the mean of the differences
- **Tested Difference in Means:**  $\Delta_0$ . Usually this value is zero; i.e,  $\Delta_0 = 0$ .
- **$t$ -value:**  $t_{n-1} = \frac{\bar{d} - \Delta_0}{SE(\bar{d})} = \frac{(\text{mean difference}) - (\text{tested mean difference})}{(\text{standard error of the mean difference})}$

### Confidence Interval:

- **Paired  $t$ -interval for the mean:**  $\bar{d} \pm t_{n-1}^* \cdot SE(\bar{d})$
- **Standard Error:**  $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$
- **Degrees of Freedom:**  $df = n - 1$
- **On the TI-84 Calculator:**
  - Enter the two samples into  $L_1$  and  $L_2$ ; then  $(L_3 = L_1 - L_2)$  or  $(L_3 = L_2 - L_1)$
  - **STAT – CALC – 1-Var Stats –  $L_3$**  to get the required statistics for  $L_3$
  - **$tcdf(\text{lower}, \text{upper}, df)$**  you have the  $t$ -value; **you want the desired area**
  - **$invT(\text{area}, df)$**  you have the area; **you want the desired  $t$ -value**



### Key Assumptions and Conditions for a Difference of Two Proportions (RITN):

- **Randomization:** Data within each group should be sampled at random or generated from a properly randomized experiment.
- **Independence:** Sample values are independent of each other. However, the two samples must NOT be independent of each other.
- **Ten-Percent Condition:**  $n \leq 10\%$  of the population for each group.
- **Nearly Normal Condition:** the source population should be nearly Normal, i.e., unimodal and symmetric. Check this for the sample using:
  - A histogram
  - A plot of Normal scores
- **Paired Data:** the data must pair naturally. You cannot decide to pair independent samples. Whether or not data are paired is generally easy to determine if you know how the data were collected.

### Hypothesis Testing (Paired $t$ -test)

- **$H_0: \mu_d = \Delta_0$**  Usually,  $\Delta_0$  will be equal to zero.
- **$H_A: \mu_d \neq \Delta_0$**  or  **$H_A: \mu_d < \Delta_0$**  or  **$H_A: \mu_d > \Delta_0$**
- **$t$ -value:**  $t_{n-1} = \frac{\bar{d} - \Delta_0}{SE(\bar{d})} = \frac{(\text{sample mean}) - (\text{theoretical mean})}{(\text{standard error of the mean})}$
- **Standard Error:**  $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$
- **Degrees of Freedom:**  $df = n - 1$

**Sample Size Requirements** and **Special Considerations** are the same as they are in Chapters 23 and 24.

## Statistical Inference – Summary

Proportions	Difference of Proportions	$\sigma$ known	$\sigma$ not known		
		Means	Means	Difference of Means	Paired Mean
Randomization Independence within sample Ten-Percent Sample Size Success/Failure	Randomization Independence between and within samples Ten-Percent Sample Size Success/Failure	Randomization Independence within sample Ten-Percent Sample Size Success/Failure	Randomization Independence within sample Ten-Percent Nearly Normal	Randomization Independence between and within samples Ten-Percent Nearly Normal	Randomization Independence within sample Ten-Percent Nearly Normal Paired data
<b>Hypothesis Testing</b>					
$H_0: p = p_0$	$H_0: p_1 - p_2 = 0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu_1 - \mu_2 = \Delta_0$	$H_0: \mu_d = \Delta_0$
$SD = \sqrt{\frac{p_0 q_0}{n}}$	$SE_p = \sqrt{\frac{\hat{p}_p \hat{q}_p}{n_1} + \frac{\hat{p}_p \hat{q}_p}{n_2}} \quad (1)$	$SE = \frac{\sigma}{\sqrt{n}}$	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (2)$	$SE = \frac{sd}{\sqrt{n}}$
$z = \frac{\hat{p} - p_0}{SD}$	$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_p}$	$z = \frac{\bar{x} - \mu_0}{SE}$	$t_{df} = \frac{\bar{x} - \mu_0}{SE}$	$t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{SE} \quad (3)$	$t_{df} = \frac{\bar{d} - \Delta_0}{SE(\bar{d})} \quad (3)$
<b>Confidence Interval</b>					
$\hat{p} \pm z^* \cdot SE$	$(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot SE$	$\bar{x} \pm z^* \cdot SE$	$\bar{x} \pm t_{df}^* \cdot SE$	$(\bar{x}_1 - \bar{x}_2) \pm t_{df}^* \cdot SE$	$\bar{d} \pm t_{df}^* \cdot SE$
$SE = \sqrt{\frac{\hat{p}\hat{q}}{n}}$	$SE = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$	$SE = \frac{\sigma}{\sqrt{n}}$	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$SE(\bar{d}) = \frac{sd}{\sqrt{n}}$
			$df = n - 1$	$df \quad (4)$	$df = n - 1$

## Notes:

(1)  $\hat{p}_p$  = the pooled value of  $\hat{p}$ ;  $\hat{p}_p = \frac{\text{total successes}}{\text{total observations}}$

(2) If  $\sigma$  is the same in the two populations, pooled values of  $s$  and  $SE$  may be used, but pooling is not required.

(3) Usually,  $\Delta_0 = 0$ .

$$(4) df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \cdot \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \cdot \left(\frac{s_2^2}{n_2}\right)^2}$$

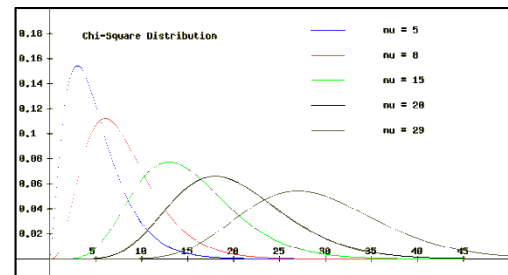
## Chapter 26: Comparing Counts – The $\chi^2$ Distribution

### Uses of the $\chi^2$ Distribution

- **Goodness of Fit:** One-dimensional data. **One variable** compared to a set of **expected values**.
- **Homogeneity:** Two-dimensional data. **One variable** compared on **two or more populations**.
- **Independence:** Two-dimensional data. **Two or more variables** compared on **one population**. The table used for independence tests is called a **contingency table**.

### $\chi^2$ Characteristics

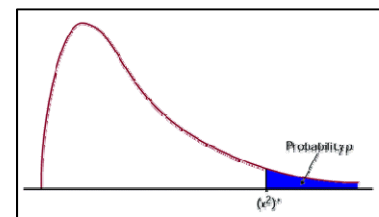
- $\chi^2$  is a **family of curves** that vary depending on the degrees of freedom.
- $\chi^2$  is used to compare data on **categorical variables**, NOT quantitative variables.
- The  $\chi^2$  distribution is **skewed right**.
- The number of **degrees of freedom** is denoted as either  **$df$**  or  **$\nu$  (nu)** – the Greek letter for “n”.
- **$df$**  depends on the **number of categories** in the analysis, NOT on sample size.
- As  **$df$**  approaches  $\infty$ , the  **$\chi^2$  distribution approaches a Normal Distribution**.
- There are **no Confidence Intervals**. Hypothesis tests only.
- Large samples can be problematic.



$df$  is often denoted with the Greek letter  $\nu$  (nu)

### Hypothesis Tests

- Hypothesis tests are one-sided only.
- Large value of  $\chi^2$  results in a low P-value and implies  $H_0$  should be rejected.
- Small P-values do not imply causation between variables.
- If  $H_0$  is rejected, we will need to look further to see the reasons behind the rejection.



## Testing Goodness-of-Fit

### Calculating the $\chi^2$ Statistic

- $$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$
- **Obs** = Observed count. **Exp** = Expected count. The summation is calculated over all cells (i.e., categories) in the sample.
- **Degrees of Freedom:**  $df = n - 1$ , where  $n$  is the number of categories and does not depend on sample size.

M&M's $\chi^2$ Example		Sample contains 103 Milk Chocolate M&M's					
Color	Percent Distribution (from company)	Expected Number of M&M's	Observed Number of M&M's	Residual (= Observed - Expected)	Squared Residual	Chi-Squared Component = $\frac{(Obs - Exp)^2}{Exp}$	Standardized Residual = $\frac{Obs - Exp}{\sqrt{Exp}}$
Blue	<b>24.0%</b>	24.7	<b>22</b>	-2.7	7.40	<b>0.30</b>	-0.55
Brown	<b>13.0%</b>	13.4	<b>11</b>	-2.4	5.71	<b>0.43</b>	-0.65
Green	<b>16.0%</b>	16.5	<b>20</b>	3.5	12.39	<b>0.75</b>	0.87
Orange	<b>20.0%</b>	20.6	<b>17</b>	-3.6	12.96	<b>0.63</b>	-0.79
Red	<b>13.0%</b>	13.4	<b>15</b>	1.6	2.59	<b>0.19</b>	0.44
Yellow	<b>14.0%</b>	14.4	<b>18</b>	3.6	12.82	<b>0.89</b>	0.94
Total	100.0%	103.0	103	0		<b>3.19</b>	0.26

Note: Data input into the model is shown in **bold blue**. All other values are calculated.

$H_0$ : The company's stated distribution of M&M's by color is correct.  
 $H_A$ : The company's stated distribution of M&M's by color is **not** correct.

Calculated  $\chi^2$  value: **3.19**  
P-Value (df = 5): **0.671**  
**Result: Fail to reject  $H_0$ .**

### Key Assumptions and Conditions for a Goodness-of-Fit Test (RITSC):

- **Randomization:** Data should be sampled at random or generated from a properly randomized experiment.
- **Independence:** Sample values are independent of each other.
- **Ten-Percent Condition:**  $n \leq 10\%$  of the population for each group.
- **Sample Counts:** The expected count in each cell should be 5 or more. Note that this deals with the expected count, NOT the observed count.
- **Counted Data Condition (NEW):** data must be counts for the categories of a categorical variable, NOT a quantitative variable.

## Hypotheses

- $H_0$ : There distribution of items by category is consistent with what is claimed (as represented by the expected distribution of items by category).
- $H_A$ : There distribution of items by category is NOT consistent with what is claimed.

## Goodness-of-Fit on the TI-84 Calculator

### Doing It All on the Calculator

- Enter **Observed Values** in a list, e.g.,  $L_1$ .
  - **STAT – EDIT** – Enter values in list form.
- Enter **Expected Values** in a list, e.g.,  $L_2$ .
  - Calculate expected values as  $Exp = (Expected\ \%) \cdot (Total\ Observations)$
  - **STAT – EDIT** – Enter values in list form.
- **STAT – TESTS – D:  $\chi^2$ GOF-Test**: Enter the requested information and hit **ENTER**
  - Enter the locations of the Observed and Expected Values (e.g.,  $L_1$  and  $L_2$ ).
  - Enter the number of degrees of freedom ( $df = n - 1$ ), where  $n$  is the number of categories, NOT the number of observations.
- Read the **P-Value** from the output. Determine from the P-value whether you will reject or fail to reject  $H_0$ .
- Read the **contributions to the  $\chi^2$  statistic** from the output (note: these are the values in magenta in the example on the previous page).

### Calculating the P-Value Directly from the $\chi^2$ Statistic

- **2ND – DISTR – 8:  $\chi^2$ cdf**: Enter the requested information and hit **ENTER**
- Note: an upper limit of 1000 is generally sufficient to obtain an accurate value of  $\chi^2$ .

### Next Steps: What if You Reject $H_0$ ?

- Look at the **components of the  $\chi^2$  statistic** in order to determine the sources of the rejection.
- Large components of the  $\chi^2$  Statistic should be investigated further. What can we glean from these values? They are likely to be the most significant reasons why the null hypothesis was rejected.

## Testing Homogeneity or Independence

The mechanics for testing homogeneity or independence are the same. Both approaches calculate  $\chi^2$  Statistic by comparing a two-dimensional table of observed data to a table of expected values. The key is calculating the table of expected values that represent homogeneous or independent data. An example is provided on the next page.

### What's the Difference Between Homogeneity and Independence?

- **Both are based on two-dimensional data.**
- **Homogeneity:** A test to see if two populations share a common characteristic. In a test for homogeneity, **one variable is compared on two or more populations.**
- **Independence:** A test to see if two variables are associated with each other. In a test for independence, **two or more variables are compared on one population.**

### Calculating the $\chi^2$ Statistic

- $$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$
- **Obs** = Observed count. **Exp** = Expected count. The summation is calculated over all cells (i.e., every cell within the two-dimensional table) in the sample.
- **Degrees of Freedom:**  $df = (R - 1)(C - 1)$ , where **R** is the number of rows of data and **C** is the number of columns of data.

### Key Assumptions and Conditions for Homogeneity or Independence Tests (RITSC):

- **Randomization:** Data should be sampled at random or generated from a properly randomized experiment.
- **Independence:** Sample values are independent of each other.
- **Ten-Percent Condition:**  $n \leq 10\%$  of the population for each group.
- **Sample Counts:** The expected count in each cell should be 5 or more. Note that this deals with the expected count, NOT the observed count.
- **Counted Data Condition (NEW):** data must be counts for the categories of a categorical variable, NOT a quantitative variable.

Note: If you do not want to generalize your conclusions, then the **Randomization** and **Ten-Percent** conditions may be ignored.

Two-Dimensional $\chi^2$ Example		Patients with Infection XYZ	
<p><math>H_0</math>: Prescriptions A, B and C are equally effective at curing patients.  <math>H_A</math>: Prescriptions A, B and C are NOT equally effective at curing patients.</p>			
<b>1. Observed Data (in blue)</b>			
Treatment	Patient Recovers	Patient Does Not Recover	Total
Prescription A	33	17	50 25%
Prescription B	43	17	60 30%
Prescription C	44	46	90 45%
Total	120 60%	80 40%	200
<b>2. Calculation of Expected Values</b>			
Treatment	Patient Recovers	Patient Does Not Recover	Total
Prescription A	$0.25 \cdot 0.60 \cdot 200$	$0.25 \cdot 0.40 \cdot 200$	
Prescription B	$0.30 \cdot 0.60 \cdot 200$	$0.30 \cdot 0.40 \cdot 200$	
Prescription C	$0.45 \cdot 0.60 \cdot 200$	$0.45 \cdot 0.40 \cdot 200$	
Total			
<b>3. Expected Values</b>			
Treatment	Patient Recovers	Patient Does Not Recover	Total
Prescription A	30.0	20.0	50
Prescription B	36.0	24.0	60
Prescription C	54.0	36.0	90
Total	120.0	80.0	200
<b>4. Observed Less Expected</b>			
Treatment	Patient Recovers	Patient Does Not Recover	Total
Prescription A	3.0	-3.0	0
Prescription B	7.0	-7.0	0
Prescription C	-10.0	10.0	0
Total	0	0	0
<b>5. Chi-Square Components</b>			
$(\text{Observed Less Expected})^2 \div \text{Expected}$ $df = (R - 1)(C - 1) = (2 - 1)(3 - 1) = 2$			
Treatment	Patient Recovers	Patient Does Not Recover	Total
Prescription A	0.300	0.450	
Prescription B	1.361	2.042	
Prescription C	1.852	2.778	
Total			$\chi^2$ -Value: 8.782 P-Value: 0.012
<b>6. Standardized Residuals</b>			
$(\text{Observed Less Expected}) \div \sqrt{\text{Expected}}$			
Treatment	Patient Recovers	Patient Does Not Recover	Total
Prescription A	0.548	-0.671	
Prescription B	1.167	-1.429	
Prescription C	-1.361	1.667	
Total			

**Step 1.** Enter the data into a table. Entered data are shown in blue. The dark red totals and the percentages for each row and column are calculated.

**Step 2:** Shows the formulas used to calculate the expected values. Each value in the expected value table is:  $(\text{row percentage}) \cdot (\text{column percentage}) \cdot (\text{total})$ . These are the values we would “expect” if the data were 100% homogeneous or 100% independent.

**Step 3:** Shows the results of the calculations defined in Step 2.

**Step 4:** Calculates the Observed Less Expected Values. Notice that the total for each column and each row is zero.

**Step 5:** Calculates each  $\chi^2$  component as  $\frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$ . Add these values to get the  $\chi^2$  Statistic, used to calculate a P-value:  $\chi^2 \text{cdf}(8.782, 1000, 2) = 0.012$ . Based on this P-value, we would most likely **reject the null hypothesis that all of the prescriptions cure the same percentage of patients.** Prescriptions B and C each contribute large components to the  $\chi^2$  value.

**Step 6:** Get more information by looking at the standardized residuals. Notice that Prescription B’s results are positive and Prescription C’s results are negative. That is, Prescription B is more effective and Prescription C is less effective. **More testing is needed to draw definitive conclusions about A and B.**

## Hypotheses

### Homogeneity

$H_0$ : There are NO differences in choices among different groups (populations).

$H_A$ : There are differences in choices among different groups (populations).

### Independence

$H_0$ : The variables are independent.

$H_A$ : The variables are NOT independent.

## Standardized Residuals

- When  $H_0$  is rejected, examine the standardized residuals to determine the sources of variation from the expected values.
- **Residual formula:**  $c = \frac{\text{Obs} - \text{Exp}}{\sqrt{\text{Exp}}}$
- **$\chi^2$  components:** note that the squares of the residuals are the components of the  $\chi^2$  statistic.
- **Small cells:** If large residuals (positive or negative) exist in cells with small frequencies, the results may be distorted. A possible solution is to combine these cells with others in some logical fashion. This should reduce or eliminate the distortion in the model.

## 2-Dimensional $\chi^2$ Tests on the TI-84 Calculator

### Doing It All on the Calculator

- Enter **Observed Values** in a matrix, e.g., **[A]**.
  - **MATRIX – EDIT** – Specify the dimensions and enter the values in the matrix.
- **STAT – TESTS – C:  $\chi^2$ -Test – CALCULATE**
  - The  **$\chi^2$  value**, the **P-value** and the **df** will be provided on the calculator screen.
  - The expected frequencies will be calculated and stored in matrix **[B]**.
  - Determine from the P-value whether you will reject or fail to reject  $H_0$ .
- **MATRIX – EDIT – [B]** to see the matrix of expected counts.
- There is no easy way to calculate the standardized residuals using the calculator, but these are easy to calculate in Excel (see the example above).

### Calculating the P-Value Directly from the $\chi^2$ Statistic

- **2ND – DISTR – 8:  $\chi^2$ cdf:** Enter the requested information and hit **ENTER**
- Note: an upper limit of 1000 is generally sufficient to obtain an accurate value of  $\chi^2$ .



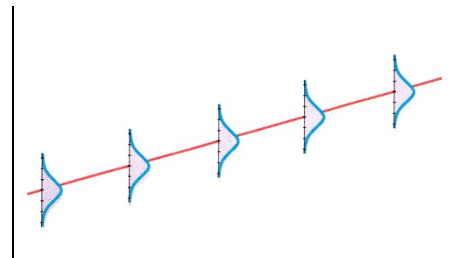
## Chapter 27: Inferences for Regression

### Model of the Idealized Line

- **True Regression Line:**  $y = \beta_0 + \beta_1 x + \varepsilon$
- **Estimate the Regression Line:**
  - **Observed  $y$ -values:**  $\hat{y} = b_0 + b_1 x$ , where  $b_1 = r \frac{s_y}{s_x}$  and  $b_0 = \bar{y} - b_1 \bar{x}$
  - **Residuals:**  $e = y - \hat{y}$
  - **Standard deviation of the residuals:**  $s_e = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$

### Visualization

- Visualize a distribution about the mean for each  $x$ -value. Mean:  $\mu_x = \beta_0 + \beta_1 x$
- **Challenge:** Account for the uncertainty around the calculated sample values of  $b_0, b_1$



### Key Assumptions and Conditions (LINER): *(Note: these are quite different – memorize them)*

- **Linearity:** Straight-enough condition. Requires quantitative data.
- **Independence:** Residuals are independent of each other.
  - Check scatterplot of  $x$  vs.  $e$ . Look for a pattern (which is bad).
- **Normal Population:** Nearly normal and outlier conditions. Note: this is less important as the size of the sample (i.e.,  $n$ ) increases and the Central Limit Theorem takes over.
  - Look at the histogram of residuals and the  $z$ -score plot of the residuals.
  - May need to remove outliers and re-do the analysis.
- **Equal Variance Assumption:** The variability of  $y$  should be about the same for every  $x$ .
  - Check scatterplot of  $x$  vs.  $y$ . Look for a fan pattern (which is bad).
  - Check scatterplot of  $\hat{y}$  vs.  $e$ .
- **Randomization:** The sample should be representative of the larger population.

### Work Order (SFRNI – Acronym: San Francisco Residents Need Income)

- **Scatterplot:** Look at a scatterplot of the observed data. Check the Straight-Enough Condition.
- **Fit Regression Line:** Use calculator, calculate  $e$ 's and  $y$ 's. Note: we can do this with **LinRegTTest**, even if we don't want all of the information just yet.
- **Residual Plot:** Review the residual scatterplot ( **$x$  vs.  $e$** ).
  - If outliers need to be removed, do so and start over.
  - If the curve is not linear, perform a transformation and start over.
  - If time is involved, plot the residuals against time.
- **Nearly Normal Condition:**
  - Review a histogram of the Residuals.
  - Review a Normal Probability Plot (z-score plot) of the Residuals.
- **Inference:** If all conditions are met, go ahead with the Inference Analysis.

### Calculations:

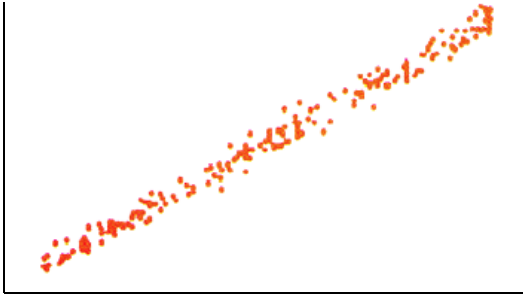
- $t_{n-2} = \frac{b_1 - \beta_1}{SE(b_1)}$        $SE(b_1) = \frac{s_e}{s_x \sqrt{n-1}}$        $s_e = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$
- **95% CI:**  $b_1 \pm t_{n-2}^* \cdot SE(b_1)$
- **Hypothesis Test:**  $H_0: \beta_1 = 0$  with  $t_{n-2} = \frac{b_1}{SE(b_1)}$  and  $df = n - 2$ 
  - Typically, this is a test to see if the slope is zero. If we want to test a non-zero slope, these formulas generalize to:
    - $H_0: \beta_1 = m_0$  with  $t_{n-2} = \frac{b_1 - m_0}{SE(b_1)}$
- **Why  $(n - 2)$  degrees of freedom?** There are two things we do not know about the regression line – slope and intercept. Therefore, we use two fewer degrees of freedom than we have observed values.
- **Can We Test for  $\beta_0$ :** Yes. We could do the same types of calculations for  $\beta_0$ , but finding CI's or hypothesis testing for  $\beta_0$  is usually less interesting.

## Thoughts about the Size of the Standard Error

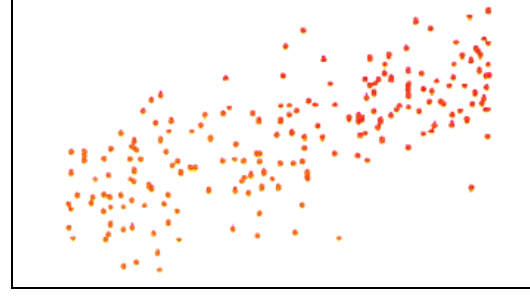
Based on the shape of the scatterplot, we can expect a smaller standard error from:

- **Tighter spread around the regression line**
- **Broader range of x-values**
- **Larger sample size**

$$SE(b_1) = \frac{s_e}{s_x \sqrt{n-1}}$$



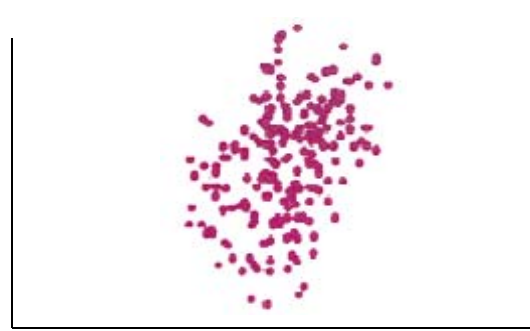
Tighter spread around the regression line ( $s_e \downarrow$ )  
Smaller Standard Error,  $SE(b_1) \downarrow$



Looser spread around the regression line ( $s_e \uparrow$ )  
Greater Standard Error,  $SE(b_1) \uparrow$



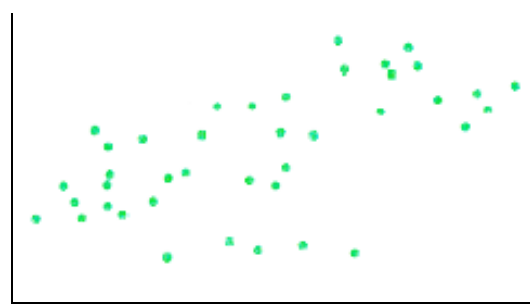
Broader spread of  $x$ -values ( $s_x \uparrow$ )  
Smaller Standard Error,  $SE(b_1) \downarrow$



Narrower spread of  $x$ -values ( $s_x \downarrow$ )  
Greater Standard Error,  $SE(b_1) \uparrow$



Larger sample size ( $n \uparrow$ )  
Smaller Standard Error,  $SE(b_1) \downarrow$



Smaller sample size ( $n \downarrow$ )  
Greater Standard Error,  $SE(b_1) \uparrow$

### Using the TI-84 Calculator (Hypothesis Tests and Confidence Intervals for Linear Regression)

- **Make sure Diagnostic is On:** 2<sup>ND</sup> – CATALOG – DIAGNOSTICON
- **Enter the  $x$ - and  $y$ -values** in lists, e.g., L<sub>1</sub> and L<sub>2</sub>. STAT – EDIT – Enter values in list form.
- **Perform the Linear Regression Test** and **Check the Scatterplot of  $x$ - and  $y$ -values.**

Perform Linear Regression Hypothesis Test	View " $x$ vs $y$ " Scatterplot	Confidence Interval
STAT – TESTS – F: LinRegTTest	2 <sup>ND</sup> STATPLOT – Plot1	STAT – TESTS – G: LinRegTInt
Xlist: L <sub>1</sub>	Type: scatterplot (1 <sup>st</sup> plot)	Xlist: L <sub>1</sub>
Ylist: L <sub>2</sub>	Xlist: L <sub>1</sub>	Ylist: L <sub>2</sub>
Freq: 1	Ylist: L <sub>2</sub>	Freq: 1
$\beta$ and $\rho$ : $\neq 0$	ZOOM – 9	C-Level: .95 (for 95% CI)
RegEQ: Y <sub>1</sub> <sup>(1)</sup>	Want linear scatterplot.	RegEQ: Y <sub>1</sub> <sup>(1)</sup>
Calculate – ENTER <sup>(2)</sup>		Calculate – ENTER

- **Check the Residuals Plots** to make sure the required conditions are met.

Residuals Scatterplot	Residuals Histogram	$z$ -scores vs. Residuals
2 <sup>ND</sup> STATPLOT – Plot1	2 <sup>ND</sup> STATPLOT – Plot1	2 <sup>ND</sup> STATPLOT – Plot1
Type: scatterplot (1 <sup>st</sup> plot)	Type: histogram (3 <sup>rd</sup> plot)	Type: $z$ -scores (6 <sup>th</sup> plot)
Xlist: L <sub>1</sub>	Xlist: RESID <sup>(3)</sup>	Data List: RESID <sup>(3)</sup>
Ylist: RESID <sup>(3)</sup>		Data Axis: Y
ZOOM – 9	ZOOM – 9	ZOOM – 9
Want no obvious pattern or fanning in the scatterplot.	Want histogram to be unimodal and symmetric.	Want plot to be close to a straight line.

#### Notes:

- (1) Enter Y<sub>1</sub> by: VARS – Y-VARS – FUNCTION – Y<sub>1</sub> – ENTER. This tells the calculator to store the regression equation in Y<sub>1</sub>. You can access it by: Y=.
- (2) Lots of info about the test is provided. You can calculate:  $SE = b \div t$ .
- (3) To change the name of the variable in a list: 2<sup>ND</sup> LIST – Names – [List Name] – ENTER