Author: Brenda Gunderson, Ph.D., 2015

Statistics 250
Lab Workbook

Fall 2015

Weekly Labs and Supplements
Used in all lab sections of Stat 250

Dr. Brenda Gunderson
Department of Statistics
University of Michigan

Table of Contents

Material	Page
Note to Students and Supplements	**1**
Supplement 1: R Commands Summary	**2**
Supplement 2: Notation Sheet	**4**
Supplement 3: Name That Scenario	**6**
Supplement 4: Interpretation Examples	**8**
Supplement 5: Summary of the Main *t*-Tests	**10**
Supplement 6: Regression Output in R	**12**
Lab 1: Describing Data with Graphs and Numbers	**15**
Lab 2: Probability and Random Variables	**19**
Lab 3: Confidence Intervals for a Population Proportion	**27**
Lab 4: Hypothesis Testing for a Population Proportion	**33**
Lab 5: Understanding Normal and Random Data	**39**
Lab 6: Learning about a Population Mean	**47**
Lab 7: Paired Data Analysis	**53**
Lab 8: Comparing Two Means	**59**
Lab 9: One-Way Analysis of Variance (ANOVA)	**67**
Lab 10:Exploring Linear Regression	**75**
Lab 11:Regression Inference	**81**
Lab 12:Chi-Square Tests	**85**

Note to Students

Welcome to Statistics 250 at the University of Michigan!

This is the first summer term in which R and R Commander will be used as the software package for Stats 250. Some of the reasons why we made this switch are:

- The ability to use R is a valuable skill recognized by employers.
- Other Statistics courses use R and this will make for an easier transition into these next courses.
- R is a free, open source software that can be downloaded onto student machines, so students can have access to it any time on their personal devices and won't have to use Virtual Sites.

This lab workbook is designed for you to use in lab and as extra preparation for exams. In the workbook, you will find the following materials:

Supplemental Material – great summaries for reference throughout the term:

1. R Commands Reference
2. Notation Sheet
3. Name That Scenario
4. Interpretation Examples
5. Summary of T-tests
6. Regression Output in R

Weekly Labs (numbered 1 to 12) – each lab contains the follow parts:

- **Lab Background** – objective and brief overview material, which is good to take a couple minutes to read *before* you come to lab each week.
- **Warm-Up Activity** – quick questions for you to do before the In-Lab Project, usually a quick review of concepts you have seen in lecture.
- **ILP (In-Lab Project)** – one or more activities you will work on in lab, in groups.
- **Cool-Down Activity** –questions for you to do after the ILP for further reflection and application of the concepts covered in the ILP.

The Labs are designed to be interactive and to provide you with a complete example for each concept. Completing the corresponding PreLab assignment (a link to video instructions for PreLabs will be on Canvas and the Stat 250 YouTube channel) and reading the upcoming lab background overview before lab each week is a good way to prepare for the various lab activities.

Good luck in Statistics 250!
-- The Stat 250 Instructors and GSIs

Special Thanks to the Statistics Graduate Students
Kit Clement
Sean Pikosz
Daniel Walter
For their substantial contributions to transition and modernize the Lab Materials to the Awesome R computing package

Supplement 1: R Commands Summary

By Lab – For Quick Reference

Lab 1 – Bar Charts, Histograms, Numerical Summaries, Boxplots

Open a data file after loading R Commander: Data > Load data set

To produce a **Histogram:** Graphs > Histogram

To generate **Descriptive Statistics:** Statistics > Summaries > Numerical summaries

To produce a **Bar Chart:** Graphs > Bar Graph

To produce a **Boxplot**: Graphs > Boxplot

Lab 5 – Time Plots, QQ Plots

To produce a **Sequence or Time Plot** for the variable named "VARIABLE" in the data set "DATA"
you must type these two lines of code into the **R Script** box:

```
plot(DATA$VARIABLE, type =”l”, main="Normal QQ Plot of variable
by name")
```

Note that you can find the dataset name in blue text at the top. To find variable names, click **View data set** and look at the top row. To create the plot, highlight the above code and click the **Submit** button.

To produce a **QQ Plot:** you can use the built in option under **Graphs > Quantile-comparison plot**
Or you can make a QQ plot for the variable "VARIABLE" in the data set "DATA" by typing these two lines of code into the **R Script** box:

```
qqnorm(DATA$VARIABLE, main="Normal QQ Plot of variable by
name")
qqline(DATA$VARIABLE)
```

Then highlight this code and click the **Submit** button.

Lab 6 – One-Sample t Procedures for a Population Mean

To perform a **One-Sample T Test for a population mean and obtain a confidence interval:** Statistics > Means > Single-sample t-test

Lab 7 – Paired t Procedures

To perform a **Paired T Test and obtain a confidence interval:** Statistics > Means > Paired t-test

To **compute Differences:** Data > Manage variables in active data set > Compute new variable.

Lab 8 – Independent Samples t Procedures

To perform **Levene's Test**: Statistics > Variances > Levene's Test

To perform a **Two-Samples T Test and obtain a confidence interval**: Statistics > Means > Independent samples t-test

Lab 9 – One-way Analysis of Variance (ANOVA)

To perform an **ANOVA:** Statistics > Means > One-Way ANOVA

Lab 10 and 11 – Linear Regression

To produce the **correlation (R)** for all pairs of variables: Statistics > Summaries > Correlation matrix

To produce a **Scatterplot:** Graphs > Scatterplot

To perform a **Linear Regression:** Statistics > Fit models > Linear regression

To produce a **Residual plot and QQ Plot of residuals**, first make sure you have the correct model selected, then follow**:** Models > Graphs > Basic diagnostic plots

Lab 12 – Chi-Square Tests

To perform a **Goodness of Fit Test:** Statistics > Summaries > Frequency distributions. Make sure to check the box to run a goodness of fit test, and then you can specify the null probabilities.

To perform a **Test of Independence:** Statistics > Contingency tables > Two-way table

To perform a **Test of Homogeneity**: Statistics > Contingency tables > Two-way table

Supplement 2: Notation Sheet

The table below defines important notations, including that used by R, which you will come across in the course. This is not an exhaustive list, but it is a fairly comprehensive overview of the “strange letters” used in the course.
Note: Blank cells mean there is no corresponding notation.

Name	Population Notation	Sample Notation	Notation used in R Commander
Summary Measures			
Mean	μ (read as “mu”)	$\bar{x}$ (*x*-bar)	Mean
Proportion	p	$\hat{p}$ (*p*-hat)	
Standard deviation	σ (sigma)	s	Varies, often “sd”
Variance	σ^2	s^2	Variance
Sample size		n	*n (sometimes N)*
Confidence Intervals			
Multipliers		z^* (*z*-star)	
		t^* (*t*-star)	
Margin of error		m, m.e.	
Hypothesis Testing			
Test statistics Note: *t*, *F*, and χ^2 statistics have degrees of freedom (abbreviated df) associated with them. Look for these on your Formula Card.		z	
		t	t
		F	F
		χ^2 (chi-square)	Chi-square
Significance level		α (alpha)	
***p*-value**		*p*-value	Pr(*) (the star will depend on what test is being used)

Name	Population Notation	Sample Notation	Notation used in R
Analysis of Variance (abbreviated ANOVA)			
Sum of squares for groups		SSG	Row labeled with the grouping variable, column labeled Sum Sq
Sum of squares for error		SSE	Row labeled Residuals, column labeled Sum Sq
Mean square for groups		MSG	Row labeled with the grouping variable, column labeled Mean Sq
Mean square error		MSE	Row labeled Residuals, column labeled Mean Sq
Regression			
Response (dependent) variable	*y*	y	(given by name of *y*-variable)
Predicted (estimated) response	E*(y)* (expected value of *y*)	$\hat{y}$ (*y*-hat)	
Explanatory (independent) variable	*x*	*x*	(given by name of *x*-variable)
y-intercept	β_o (beta-not)	b_o	B (look in the row labeled (Intercept))
Slope	β_1 (beta-one)	b_1	B (look in the row labeled with the name of the x-variable)
Coefficient of correlation		*r*	Values in Correlation Matrix
Coefficient of determination		r^2	Multiple-R Squared
Error terms vs Residuals	ε (error terms)	*e* (residuals)	Unstandardized residuals

Supplement 3: Name That Scenario

The first thing to do in any research inference problem is determine what type of inference problem it is. This will help in deciding what procedure/formulas are appropriate to use. The following questions can help you determine the data scenario you are working with.

Please note, when answering, "How many variables are there?" do not count the variable which defines the populations (if there is more than one population).

❑ How many populations are there?

One **Two** **More than two**

❑ How many variables are there?

One **Two**

❑ What type of variable(s)?

Categorical **Quantitative**

❑ Then use the following table to determine which type of inference would be appropriate for this scenario.

Note the corresponding parameter is in parentheses, where appropriate.

		Number of Populations		
Number of Variables and Type		**One**	**Two**	**More Than Two**
One	**Categorical**	□ 1-sample inference for population proportion (p) (Labs 3-4) □ Chi-square: Goodness of Fit (Lab 12)	□ 2 indep. samples inference for the difference between 2 population proportions $(p_1 – p_2)$ □ Chi-square: Homogeneity (Lab 12)	□ Chi-square: Homogeneity (Lab 12)
	Quantitative	□ 1-sample inference for population mean (μ) (Lab 6) □ Paired samples inference for a population mean difference (μ_D) (Lab 7)	□ 2 indep. samples inference for the difference between 2 population means $(\mu_1 - \mu_2)$ (Lab 8)	□ ANOVA (μ_i – where there is one μ_i for each population) (Lab 9)
Two	**Categorical** (relationship)	□ Chi-square: Independence (Lab 12)		
	Quantitative (relationship)	□ Regression (β_1) (Labs 10-11)		

Supplement 4: Interpretation Examples

In 1980, Bausch and Lomb Corporation developed a new type of extended-life contact lens made of silicone, which it claimed had a useful life of more than 4 years. During the research and development period, a random sample of 6 contact wearers was asked to wear the new contact lenses and record how long they lasted. The average useful life of the six pairs of lenses was 4.6 years, with a standard deviation of 0.49 years.

a. **Interpretation of the Standard Deviation *s*:**
An **estimate** of the **average distance** of the observed useful lives of these lenses **from their mean** useful life of 4.6 years is ***about*** 0.49 years.

Note: if given the true population standard deviation (σ) this becomes:

The **average distance** of the observed useful lives of these lenses **from their mean** useful life of 4.6 years is ***about*** 0.49 years.

b. Calculate the value of the **standard error of the mean**. $SE(\overline{X}) = \frac{s}{\sqrt{n}} = \frac{0.49}{\sqrt{6}} = 0.200$

Interpretation:
The standard error is an **estimate** of the **average distance** of all the possible sample means from the true population mean (roughly). In context: An **estimate** for the **average distance** of xbar (sample averages of contact life from samples of size 6) from the **population mean useful life**, μ, *is* ***roughly*** 0.20 years.

c. Construct a **90% confidence interval** for the population mean life of all such silicone-based lenses:

$$4.6 \pm (2.015)(0.200) \Rightarrow (4.197, 5.003)$$

Interpretation of the <u>Interval</u>:
This interval provides a range of reasonable values for the population mean useful life, μ. We would estimate the **population mean** useful life, μ, to be between 4.197 years and 5.003 years, with 90% confidence.

Interpretation of the 90% Confidence <u>Level</u>:
If we repeatedly took new samples of the same size (computing new 90% confidence intervals each time), we would expect 90% of these resulting intervals to contain the population mean life, μ.

d. State the hypotheses to test the claim made by Bausch and Lomb about their new contact lens; that is, test if the population mean useful life is more than 4 years.

$H_0 : \mu = 4$, $H_a : \mu > 4$, with an observed t-test statistic of

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{4.6 - 4}{0.200} \cong 3.00.$$

The ***p*-value** for this test is the probability of getting a *t*-test statistic at least as extreme as the observed test statistic, **assuming the null hypothesis is true**. So we have the *p*-value = Prob(T >= 3 | H_0=True) found under the *t*(5) distribution. This *p*-value turns out to be equal to 0.015.

Interpretation of the value of the test statistic *t = 3.00* in terms of a distance: The observed sample mean was 3 average distances (i.e. 3 standard errors) above the hypothesized mean of 4. In other words, since the standard error for xbar was .2 it took 3 of them to get from 4 (value under null) to 4.6 (test statistic value)

Interpretation of the resulting *p-value of 0.015*: If the null hypothesis was true (the population mean useful life is just 4 years) and this procedure (study) was repeated many times, we would expect to see a *t*-test statistic value of 3.00 or larger in only 1.5% of the repetitions. Thus are data are somewhat unusual under the null hypothesis theory, providing evidence for the alternative theory that the population mean useful life is greater than 4 years.

e. At a 10% significance level, what is the **decision**?
Reject *H_0* since the *p*-value is less than 0.10.

f. What is the **conclusion**? There is sufficient evidence to conclude that the **population mean** useful life of the new lenses is greater than 4 years.

NOTE: These interpretations can be extended to the any test and confidence interval, adjusting for the different parameters, different directions of extreme, different test statistics, etc.

Supplement 5: Summary of the Main *t*-Tests

The three inference scenarios presented in Labs 6, 7, 8 are: **one-sample *t* procedures, paired *t* procedures, and two independent samples *t* procedures**. Data exploration is always essential to determining whether the model you want to use is appropriate. That is, we need to check the assumptions. (Recall that checking assumptions is the second step in performing a hypothesis test.)

The *t* procedures have the following general assumptions:

1. Each sample is a **random sample** – (the observations can be viewed as realizations of independent and identically distributed random variables). In the paired *t* procedures, the **differences** are assumed a random sample.

2. Each sample is drawn from a **normal population**, that is, the response variable has a normal distribution for each population. In the paired *t* procedures, the population of **differences** is assumed to have a normal distribution. In the two-sample case, both populations of responses are assumed to have normal distributions.

   You need normality of the underlying population for the response in order to have normality for the sample mean. In the case where you do not have a normal population, you can still have normality of the sample mean if you have a large enough sample size (most texts state that a sample size of *at least 25-30* is required). Thus we will accept *at least 25* as large enough to assume CLT holds for non-normal populations.

3. For the two independent samples *t* procedures, we also assume that the **two samples are independent**. We also need to assess whether the **two population variances** can be assumed **equal** in order to decide between the pooled and the unpooled *t* tests.

Graphical tools can be used to check these assumptions (see Labs 1 and 5 for more details about these various graphs).

Time Plots (or Sequence Plots): If your quantitative data have been gathered over time, then a time plot can be used to determine if the underlying process that generated that time dependent data ***appears*** to be stable. For example, in paired design problems we assume our set of differences calculated from the paired observations (d_1, d_2, ..., d_n) are a random sample. To check this, the values should be plotted by time to see if it is plausible that all values randomly came from one parent population. If that was the case the graph would be stable, with no patterns and constant mean/variance.

Remember:

#1 Time or Sequence plots are useful for checking stability only when the data are ordered in some sense. If there is no inherent order to the data, a sequence plot should not be made.

#2 If a Time plot makes sense to be examined and does show evidence of instability, it would not make sense to treat those observations as being a random sample; thus it would not be appropriate to make a histogram, QQ plot, or boxplot of the observations. No statistical procedure taught in this course is appropriate for non-stable data.

Histograms: Histograms are especially useful for displaying the distribution of a quantitative response variable. You could make a histogram of the observations in a one-sample problem, of the differences in a matched pairs design, and of each of the two samples separately in the independent samples design. Examine the histogram for evidence of strong departures from normality, such as bimodality or extreme outliers. Since you are just plotting data (just a sample and not the entire population of responses), your histogram may not look *perfectly* bell-shaped or normal.

QQ plots: QQ plots (or quantile plots or normal probability plots) are generally better than histograms for assessing if a normal model is appropriate. If the points in a QQ plot fall *approximately* in a straight line (with a positive slope) then the normal model assumption is reasonable.

Boxplots: Boxplots are most useful for assessing the validity of the assumption of **equality of population variances** in **the two independent samples design**. We would see if the IQRs (shown graphically by the length of the boxes) are comparable, and also compare the overall ranges. If they do have comparable lengths or sizes (they do not need to be lined up), then we have support that the equality of population variances assumption is reasonable. We would also want to compare the two sample standard deviations themselves, and Levene's test of equality of the two population variances may also be available.

Name that Scenario Practice for the Three T Tests:

Having just reviewed the three main *t*-test inference scenarios, you should understand the testing procedures and be able to interpret the results of a test. However, it is important to know when each scenario applies. Read each of the following inference scenarios and determine which of the three *t*-test procedures would be most appropriate: the one-sample *t*-test, the paired *t*-test, or the two-independent samples *t*-test.

1. A researcher is studying the effect of a new teaching technique for middle school students. One class of 30 students is taught using the new technique and their mean score on a standardized test is compared to the mean score of another class of 27 students who were taught using the old technique.

2. A company claims that the economy size version of their product contains 32 ounces. A consumer group decides to test the claim by examining a random sample of 100 economy size boxes of the product, since they have received reports that the boxes contain less than the 32 ounces claimed.

3. At some universities, athletic departments have come under fire for low academic achievement among their athletes. An athletic director decides to test whether or not athletes do in fact have lower GPAs. A random sample of 200 student athletes and a random sample of 500 non-athlete students are taken and their GPAs are recorded.

4. As part of a biology project, some high school students compare heart rates of 40 of their classmates before and after running a mile. They want to see if the heart rate of students their age is faster after running a mile than before, on average.

5. A hospital is studying patient costs; they decide to follow 500 surgery patients' hospital and medical bills for a year after surgery, and compare them to the estimated costs provided to the patients before surgery. They want to see if the estimated and actual costs are comparable on average.

6. A chemical process requires that no more than 23 grams of an ingredient be added to a batch before the first hour of the process is complete. An analyst feels that due to current settings more than 23 grams may actually be added. If the analyst is correct, the settings need to be altered and recent batches recalled. A random sample of 25 batches is obtained from the machine that is supposed to add the ingredient. The measurements are used to test the analyst's claim.

Supplement 6: Regression Output in R

There are several different pieces of output for regression. In this example, we will be using the **dentistry.Rdata** data set. In these models, the explanatory x variable is **DNA,** and the response y variable is **PLAQUE**.

In some situations, we may have many potential predictors of our response variable – here there is just one potential explanatory variable, DNA. To analyze the correlation potential predictors to our response variable, we can create a **Scatterplot matrix**. We see the following matrix for our variables here:

```
             DNA    PLAQUE
DNA    1.0000000 0.8557985
PLAQUE 0.8557985 1.0000000
```

This matrix shows us the **correlation coefficient**, r, for all pairs of variables. The correlation coefficient measures the strength of the linear association between the two variables. The closer it is to +1 or -1, the stronger the linear association.

We choose a pair by picking a column and row for each variable, and checking the value for that column and row pair. We can see that each pair is listed twice in the matrix (DNA-PLAQUE and PLAQUE-DNA), and that each variable is perfectly correlated with itself (r = 1). The main information here that we gather here is that the correlation of our model for predicting plaque using DNA is 0.856.

Next, we can generate our model, and R will give us a summary of the model, which looks like this:

```
Call:
lm(formula = PLAQUE ~ DNA, data = dentistry)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7639 -3.5107 -0.9454  4.0531  6.2532 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.54830    8.19299  -0.067  0.94829   
DNA          0.16685    0.03566   4.679  0.00158 **
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 4.851 on 8 degrees of freedom
Multiple R-squared:  0.7324,    Adjusted       R-squared:
0.6989 
F-statistic: 21.89 on 1 and 8 DF,  p-value: 0.001584
```

The summary starts with a **Call** line – this just tells you what model you are looking at. Here, **PLAQUE ~ DNA** is saying that we are predicting PLAQUE using the explanatory variable DNA. Next, we see that R gives us some quartiles for our residuals. We might find the median residual especially useful, as having a median of -0.9454 here tells us that the majority of residuals are negative.

Next, we see the **Coefficients** table, which gives us a wealth of information. In this section, the least square estimates for the regression line are given. These estimated regression coefficients are found under the column labeled **Estimate**. The estimated slope is next to the independent variable name (in this example it is DNA), and the estimated intercept is next to **(Intercept)**. So, b_0 is the coefficient for the variable (Constant), and b_1 is the coefficient for the independent variable x in the model. The next column heading is **Std. Error**, which provides the corresponding **standard error** of each of the least squares estimates. Also produced in this table, are the ***t*-test statistics** in the column labeled **t value** and **Pr(>|t|)**, which reports the two-sided ***p*-values** for these *t*-test statistics.

In the last few lines of output, we get our standard deviation, R-squared, and F-statistic. The **Residual standard error** gives the value of *s*, the estimate of the population standard deviation σ. The next line gives two values, Multiple R-squared and Adjusted R-squared. We ignore the Adjusted and just look at the Multiple R-squared. This value, which is the square of the correlation has a useful interpretation in regression. It is often called the **coefficient of determination**, or r^2, and measures the proportion of the variation in the response that can be explained by the linear regression of *y* on *x*. Thus, it is a measure of how well the linear regression model fits the data. The final line is an F-statistic, which also gives us a way to test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, but it only allows for a two-sided test. This F-statistic comes from an ANOVA table, which we can generate separately using **Model > Hypothesis tests > ANOVA table**. Make sure to select the **Type I** option to get the ANOVA table into a familiar format.

```
Analysis of Variance Table

Response: PLAQUE
          Df Sum Sq Mean Sq F value   Pr(>F)   
DNA        1 515.14  515.14  21.894 0.001584 **
Residuals  8 188.23   23.53                    
```

We see that this ANOVA gives us the same F-statistic as before (F = 21.89). It also gives us some measures of variance within our model, the **Regression Sum of Squares** (SSModel = 515.14), and leftover residual variance, or the **Residual Sum of Squares** (SSRes = 188.23). We can use this to calculate r^2, or the proportion of variability in plaque that can be explained by its linear relationship with DNA, by taking the model variability and dividing by the total variability – r^2=515.14/(515.14+188.23) = .7324. Another value we can get again is an estimate of our total variability σ, or the **residual standard error**, by taking the square root of the MSRes = 23.53, much like we did for ANOVA to find the estimate of the pooled standard deviation.

Finally, the ratio of the Mean Squares provides the ***F*** statistic which tests if the slope is significantly different from zero (i.e. if there is a significant non-zero linear relationship between the two variables – H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$.) The **Pr(>F)** is the corresponding *p*-value for the *F* test of these hypotheses. In simple linear regression, the 2-sided *t*-test in the **Coefficients** output for the slope is equivalent to the ANOVA *F*-test. Notice that the square of the *t*-statistic for testing about the slope is equal to the *F*-statistic in the ANOVA table, and the corresponding *p*-values are the same.

Interpretation of estimated slope b_1:
According to our regression model, we **estimate** that increasing DNA by one unit has the effect of increasing the predicted plaque by .167 units.

Interpretation of r^2:
According to our model, **73% of variation** in plaque levels can be accounted for by its **linear relationship** with DNA.

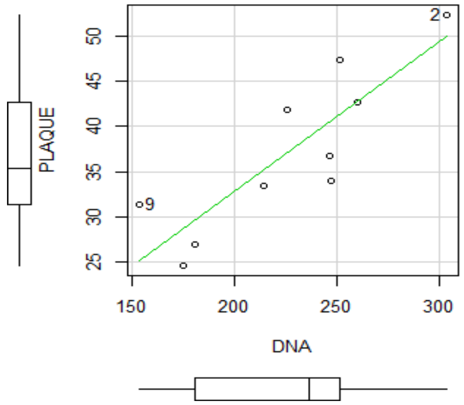
Decision for test of a significant linear relationship:
Since the *p*-value = .002 is less than the significance level α = .05, we can reject the null hypothesis that the population slope, β_1, equals 0.

Conclusion: There is sufficient evidence to conclude that in the linear model for plaque based on DNA the population slope, β_1, does not equal zero. Hence, it appears that DNA is a significant linear predictor of plaque.

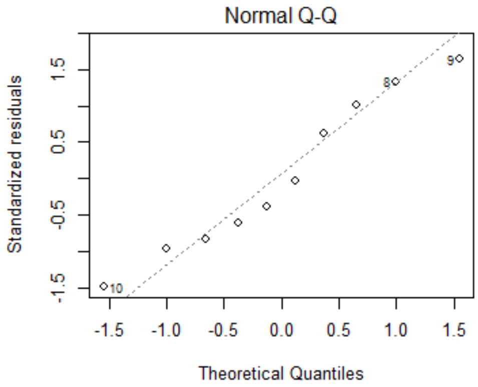
Checking the Simple Linear Regression Assumptions

Here is a summary of some graphical procedures that are useful in detecting departures from the assumptions underlying the simple linear regression model.

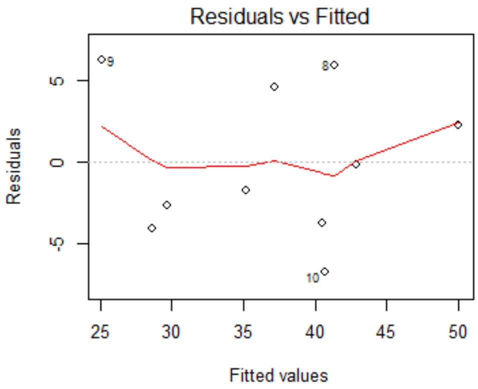
1. LINEARITY: Do a scatter plot of *y* versus *x*.
The plot should appear to be roughly linear.



2. **NORMALITY:** Examine a QQ plot of the residuals to check on the assumption of normality for the population (true) error terms. An example QQ plot is shown below.



3. **CONSTANT VARIANCE (or STANDARD DEVIATION)** of the population (true) error terms: Make a plot of the residuals versus *the fitted y values (ŷ)*. This plot is called a **residuals vs fitted plot**. The residuals represent what is left over after the linear model has been fit. The residuals vs fitted plot should be a random scatter of points in roughly a horizontal band, with no apparent pattern. An example residuals vs fitted plot is shown at the right. Sometimes this plot can also reveal departures from linearity (i.e. that the regression analysis is not appropriate due to lack of a linear relationship).



Lab 1: Describing Data with Graphs and Numbers

Objective: In this module, you will use some graphical and numerical tools to summarize the distribution for a quantitative variable or response – a histogram, a boxplot, mean, median, standard deviation, and interquartile range (IQR). You will also be introduced to side-by-side boxplots for comparing two or more distributions and bar charts for summarizing categorical data. These techniques can be very useful at the start of data analysis to get a feel for the data.

Overview: Two graphs that can be used to summarize the distribution for a single quantitative variable or response are a **histogram** and a **boxplot**. Each graph provides different information about the distribution. When used properly, graphs can be a very effective way to summarize data. Data on a single quantitative variable should first be examined graphically. The overall shape of the distribution and existence of outliers can generally be used to assess if the data appear to be coming from a relatively homogenous population. If so, then various numerical summaries may be used to characterize the center of the distribution (such as mean and median) and the spread of the distribution (such as the standard deviation and the IQR). For categorical variables, a **bar chart** can be used to display the number falling in each category (frequency distribution).

Histograms: A histogram displays the distribution of a quantitative variable by showing the frequency (count) or percent of the values that are in various classes. The classes are typically intervals of numbers that cover the full range of the variable. Histograms can be used to assess the **symmetry** and **modality** of a single distribution or for comparing the relative locations and shapes of several distributions.

Boxplots: One plot that can detect extreme observations or outliers is the **boxplot.** A boxplot is a graphical representation of the five-number summary, namely, the minimum, first quartile, median, third quartile, and maximum of the data. The centerline of the box marks the median or the 50th percentile. The sides of the box show the first (lower) quartile, Q1, and the third (upper) quartile, Q3. Thus a boxplot shows the overall range (maximum – minimum) and the **interquartile range** (IQR = Q3 – Q1). A modified boxplot uses a rule for identifying values that are extraordinary compared to the others (**outliers** or **outside values**). Circles (o) are used to denote outliers and asterisks (*) to denote extreme outliers if any are present. Any point below Q1 – (1.5 x IQR) or above Q3 + (1.5 x IQR) is considered an outlier. Extreme outliers are points below Q1 – (2 x IQR) or above Q3 + (2 x IQR). *Box plots cannot tell you the shape of the distribution.*

Side-by-side Boxplots: These plots are helpful for comparing two or more distributions with respect to the five-number summary. For example, suppose you are interested in comparing the distribution of a variable, such as the salary of the employees of a certain company. If you have information on sex for the group, you might be interested in comparing the distribution of salary of females with respect to males. In this case, the side-by-side boxplot will be an important part of the descriptive analysis of the data set involved.

Bar Charts: One way to display the number or frequency distribution for a categorical variable is with a bar chart. A bar chart shows the percentage of items that fall into each **category** or value of a **categorical variable**. It displays a bar for each category with the height of each bar equal to the number, the proportion, or the percentage of items in that category**.** If the categories have no inherent order, we could rearrange the bars in the graph in any way we like. In such cases, the shape of the bar graph would have no bearing on its interpretation.

Measures of Center: Measures of center are numerical values that tend to report the middle of a set of data. The two that we will focus on are the mean and the median.

1. **Mean:** The mean of a set of *n* observations is simply the sum of the observations divided by the number of observations, *n*.

2. **Median:** The median of a set of observations, ordered from smallest to largest, is a value such that at least half of the observations are less than or equal to that value and at least half the observations are greater than or equal to that value.

Measures of Variation or Spread: Measures of variation include the IQR and standard deviation. These numerical summaries describe the amount of spread that is found among the data, with larger values indicating more variability.

a. **Standard Deviation:** Standard deviation is a measure of the spread of the observations from the mean. It is actually the square root of an average of the squared deviations of the observations from the mean. *We can think of the standard deviation as approximately an average distance of the observations from the mean.*

b. **IQR:** The IQR measures the spread of the middle 50% of the data. It is defined as the difference between the 3rd quartile (Q3) and the 1st quartile (Q1). These quartiles are also called the 75th and 25th percentiles, respectively. IQR = Q3 – Q1.

Warm-Up: Mean and Median

We are interested in analyzing Lebron James' scoring output by game. We have his scoring output for 5 games which we have arranged from lowest to highest: 6, 24, 28, 34, 36. There are two measures of center we could report.

Which measure would be better to report? Median Mean

Which is the most likely value for the mean? 8 12 26

What is the median? 24 28 34

ILP: Visualizing and Exploring a Data Set

In this activity, you will learn how to create graphs and obtain descriptive statistics for a data set using R.

Task: The data set **employee.Rdata** contains information on employees at a company. Explore possible questions this data could be used to address. Create appropriate graphs and obtain descriptive statistics for current salary, and discuss the results.

1. First, we must start up R Commander within R. To do this from within R, select **Packages > Load package** and from the list, scroll down and select **Rcmdr** and click OK. This will boot up the **R Commander window.**

2. To obtain the data set, go to Canvas, and find the "Data Sets" folder under the "Files" tool. Select **employee.Rdata** and save it to a directory of your choice. To open the **employee.Rdata** data set from within R Commander, click on the **Data** menu at the top, and then select **Load data set**. Open **employee.Rdata** from the directory you saved the file in.

3. We now see that our data is loaded, as we can see **employee** in the active data set box (next to the R logo). Before we begin analyzing the data, let's first view what the data set looks like – to do this, click on **View data set**. Here, you can see the variables in the data set and their values. The first variable you should see is ID.
   What is the second variable present in the data set?
   What type of variable is it?
   What is the eighth variable present in the data set?
   What type of variable is it?

4. Create a histogram for current salary. Use the graphs menu – **Graphs > Histogram** and select SALARY, then click OK.

> Note: All homework and prelab assignments will require that students provide an appropriate title and their name on each R chart or output. For histograms, click on the **Options** tab and enter your title there.

   Draw a quick sketch of the histogram and describe what the histogram shows about the distribution of current salaries. A good description will have information about the shape, general center, variability, and if relevant, a comment about potential outliers.

5. You would like to compare the distribution of salary for minorities versus non-minorities. Generate histograms again, but this time, click the **Plot by groups** button and select MINORITY as the grouping variable. Note: Since the dataset contains far more non-minorities, it may be useful to go to **Options** and select **Percentages** for the axis scaling. You can also change the number of bins (or bars) if you are interested in seeing a finer detail. Compare and contrast the distribution of salary for the two groups – can we use the same descriptions for both histograms?

6. Obtain a boxplot for current salary. Use: **Graphs > Boxplot** and select the SALARY variable again. Make a quick sketch of this boxplot, and describe what the boxplot shows about the distribution of current salary. What do the various lines on the boxplot represent?

7. As we did with histograms, we can also use side-by-side boxplots to compare the distributions. Return to the dialog for boxplots, and again click the **Plot by groups** button and select MINORITY as the grouping variable.

   How does the distribution for current salary compare for minorities versus non-minorities?

8. Numerical summaries may also be obtained for any quantitative variable. Basic descriptive summaries can be obtained via **Statistics > Summaries > Numerical Summaries**. The summaries can then be found in the output box. Fill in the basic summary measures for current salary (some require hand calculation).

   Mean: Median: Standard Deviation:

   Q1: Q3: IQR: Q3-Q1 =

   Min: Max: Range: Max-Min =

Cool-down: Which Measure of Center to Report?

Mark is a Stats 250 GSI who would like to report a measure of center for scores on the first exam. The mean score for his lab section was 77.46 points and the median was 84 points. One of Mark's students did not take the exam and received a zero. Since Mark knows this score will not count against the student, he removes the score of zero from his data.

How will the mean test score change if the grade of 0 is not included?

If there is an outlier test score of 20 points, which measure of center would you recommend that Mark report?

Lab 2: Probability and Random Variables

Objective: The objective of this lab is to become familiar with using the models for random variables and to find the probabilities associated with the models you have learned. The probabilities we compute from these models (for example, p-values in testing theories) will help us make reasonable decisions. You will work with three random variables and the methods used to calculate probability for each variable. You will also become familiar with several concepts that allow for easier calculation of probabilities.

Application: Researchers at University of Michigan Hospital and Baylor University were interested in determining if lack of sleep for teenagers led to an increased incidence of heart disease. These researchers conducted a study in which the sleep patterns for 37 teenagers were recorded in addition to the presence of heart disease for each teenager. The researchers can use the probability rules to determine if lack of sleep and incidence of heart disease are independent events.

Overview: In this lab you will be introduced to several random variables and their models. These variables can be classified as one of two types: a ***discrete random variable***, which has a finite number of outcomes, and a ***continuous random variable*** which has an infinite number of outcomes. You will practice computing probabilities for each of these variables, using concepts such as a ***standardized score*** and a ***Normal Approximation.*** There are several situations that can make computing probabilities easier, such as ***independent events*** and ***mutually exclusive events*** that will also be explored in this lab.

Independent Events: Two events A, B are said to be independent if knowing that one will occur (or has occurred) does not change the probability that the other occurs. In probability notation this can be expressed as P(A|B) = P(A). Going back to the application, the researchers from the study can check the independence of the two events by computing the probability of heart disease for teenagers who sleep less than the recommended amount as well as the probability of heart disease for all teenagers. If the incidence of heart disease for a teen is independent of the amount of sleep they get, the two probabilities will be equal.

Mutually Exclusive: Two events A, B are mutually exclusive (or disjoint) if they do not contain any of the same outcomes. So their intersection is empty.

Random Variables: A random variable assigns a number to each outcome of a random circumstance, or, equivalently, a random variable assigns a number to each unit in a population. The distribution of a random variable is a model that shows us what values are possible for that particular random variable and how often those values are expected to occur (i.e. their probabilities). The model can be expressed as a function or table or picture, depending on the type of variable it is. We will consider two broad classes of random variables: discrete random variables and continuous random variables.

Discrete Random Variable: A discrete random variable, X, is a random variable with a finite or countable number of possible outcomes. The probability distribution function (pdf) for a discrete random variable X is a table or rule that assigns probabilities to the possible values of the X.

Two conditions that must always apply to the probabilities for a discrete random variable are:

Condition 1: The sum of all of the individual probabilities must equal 1.
Condition 2: The individual probabilities must be between 0 and 1.

Binomial Random Variable B(n,p): One discrete random variable is the binomial random variable, which counts the number of times a certain event occurs out of a particular number of observations or trials of a random experiment. A binomial experiment is defined by the following conditions:

1. There are n "trials" where n is determined in advance and is not a random value.
2. There are two possible outcomes on each trial, called "success" (S) and "failure" (F).
3. The outcomes are independent from one trial to the next.
4. The probability of a "success" remains the same from one trial to the next, and this probability is denoted by p. The probability of a "failure" is 1 – p for every trial.

Continuous Random Variable: A continuous random variable, X , takes on all possible values in an interval (or a collection of intervals). The way that we determine probabilities for continuous random variables differs in one important respect from how we determine probabilities for discrete random variables. For a discrete random variable, we can find the probability that the variable X exactly equals a specified value. We can't do this for a continuous random variable. Instead, we are only able to find the probability that X could take on values in an interval. We do this by determining the corresponding area under a curve called the probability density function of the random variable.

So the probability distribution of a continuous random variable is described by a density curve. The probability of an event is the area under the curve for the values of X that make up the event. The probability model for a continuous random variable assigns probabilities to intervals.

Definition: A curve (or function) is called a **Probability Density Curve** if:

1. It lies on or above the horizontal axis.
2. Total area under the curve is equal to 1.

Normal Random Variable N(μ,σ): The family of normal distributions is very important because many variables have this shape and form approximately and many statistics that we use in our inference methods are based on sums or averages which generally have (approximately) a normal distribution.

A normal curve is symmetric, bell-shaped, centered at the mean and its spread is determined by the standard deviation. In fact, the points of inflection on each side of the mean mark the values which are one standard deviation away from the mean.

Standardized Scores: A normal distribution is indexed by its population mean, and its population standard deviation. Recall that the standard deviation is a useful "yardstick" for measuring how far an individual value falls from the mean. The standardized score or z-score is the distance between the observed value and the mean, measured in terms of number of standard deviations. Values that are above the mean have positive z-scores, and values that are below the mean have negative z-scores.

Normal Approximation to the Binomial Distribution: The easier way involves using a normal distribution. The normal distribution can be used to approximate probabilities for other types of random variables, one being binomial random variables when the sample size n is large.

Expected Value: The expected value of a random variable is the mean value of the variable X in the sample space, or population, of possible outcomes. Expected value, denoted by E(X), can also be interpreted as the mean value that would be obtained from an infinite number of observations on the random variable.

Standard Deviation: The standard deviation can be viewed as approximately the average distance of the possible values of X from its mean.

Formula Card

Probability Rules

- **Complement rule**

  $P(A^C) = 1 - P(A)$

- **Addition rule**

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

  For independent events:

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

  For mutually exclusive events: $P(A \text{ or } B) = P(A) + P(B)$

- **Multiplication rule**

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

  For independent events: $P(A \text{ and } B) = P(A)P(B)$

  For mutually exclusive events: $P(A \text{ and } B) = 0$

- **Conditional Probability**

  General: $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$

  For independent events: $P(A \mid B) = P(A)$

  For mutually exclusive events: $P(A \mid B) = 0$

Discrete Random Variables

Mean

$$E(X) = \mu = \sum x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

Standard Deviation

$$s.d.(X) = \sigma = \sqrt{\sum (x_i - \mu)^2 p_i} = \sqrt{\sum \left(x_i^2 p_i\right) - \mu^2}$$

Binomial Random Variables

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

where $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$

Mean

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

Standard Deviation

$$s.d.(X) = \sigma_X = \sqrt{np(1-p)}$$

Normal Random Variables

- $z\text{-score} = \dfrac{\text{observation} - \text{mean}}{\text{standard deviation}} = \dfrac{x-\mu}{\sigma}$
- Percentile: $x = z\sigma + \mu$
- If X has the $N(\mu, \sigma)$ distribution, then the variable

  $Z = \dfrac{X-\mu}{\sigma}$ has the $N(0,1)$ distribution.

Normal Approximation to the Binomial Distribution

If X has the $B(n, p)$ distribution and the sample size n is large enough (namely $np \geq 10$ and $n(1-p) \geq 10$), then X is approximately $N\left(np, \sqrt{np(1-p)}\right)$.

Warm-Up: Types of Variables and Types of Distribution

1. **Type of Variables:**
Today's typical undergraduate student is often characterized as preferring teamwork, experiential activities, and the use of technology. An ECAR (Educause Center for Applied Research) study was published on technology use among undergraduate students. The study used survey and interviewer data to create a portrait of today's students' experiences with and skill using information technology.

Listed below are some of the response variables that were measured in this study. For each of these determine whether it is categorical, quantitative discrete, or quantitative continuous.

a. Technology ownership: Do you own a computer?

categorical **quantitative discrete** **quantitative continuous**

b. Time (per week in minutes) spent using a computer for writing documents (word processing).

categorical **quantitative discrete** **quantitative continuous**

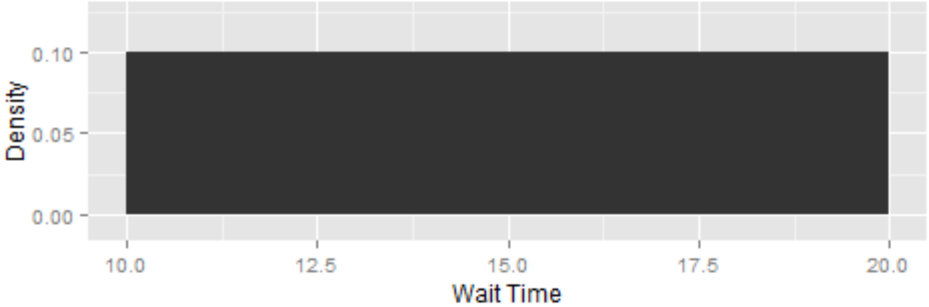
c. Which social networking site(s) are you a member? (Facebook, Myspace, Friendster, etc.)

categorical **quantitative discrete** **quantitative continuous**

2. **Types of Models:**
Identify appropriate model for each of the three variables. (Be complete).

a. Below is the model for the random variable X which represents waiting time to be served at a deli.

X has a ____ distribution

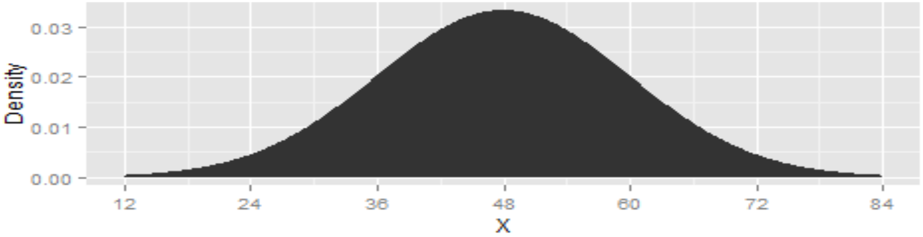


b. Suppose that 45% of Michigan residents own dogs. Let *X* represent the number of Michigan residents with a dog in a random sample of 10 Michigan residents. A success is defined as the resident owning a dog.

X has a ________________________________ distribution

c. Below is the model for the random variable X which represents the score on a Stats 250 Exam (out of 100 points).

X has a ____________________________________ distribution



ILP: Probability and Random Variable Review

Problem 1: Study on Smiling

In a recent study people were observed for about 10 seconds in public places (e.g. malls and restaurants) to determine whether they smiled during the randomly chosen 10-second interval. The table shows the results for comparing Adults (group 1) and Minors (group 2).

	Smile	**No Smile**	**Total**
1 = Adult	3269	3806	**7075**
2 = Minor	4471	4278	**8749**
Total	**7740**	**8084**	**15824**

a. What is the probability that a randomly selected person smiled?

b. The researcher would like to ***assess if smiling status is independent of age group***.

i. To check for independence, the probability found in part (a) should be compared to which of the following probabilities?

P(smiled and adult) ***P(smiled given adult)***

P(adult given smiled) ***P(adult)***

ii. Find the probability selected above and circle the appropriate conclusion.
The probability = ______________________________

Thus it appears that smiling status ***is*** ***is not*** independent of age group.

Problem 2: Summer Trip Length

Did high gas prices keep Americans from hitting the road this past summer? In a nationwide survey of adults, one variable measured was how many days vacationers spent driving on the road on their longest trip. Consider the following (partial) probability distribution for the random variable *X* = the number of days for the longest car trip.

X	4	5	6	7	8
Probability	0.10	0.20	0.25		

a. Suppose the probability of 7 days is twice as likely as the probability of 8 days. Complete the probability distribution for *X*. Show your work.

b. What is the expected number of days for the longest trip?
Include ***symbol, value, and units***.

Problem 3: How Much Time do You Spend Studying Statistics?

A Washington Post article “Is college too easy? As study time falls, debate rises” (May 21, 2012) stated that the amount of time college students actually study has dwindled from an average of 24 hours per week to about 15 hours (based on a survey). A professor of statistics decided to ask all of his current semester students to report the number of hours per week they spend studying his course material (on a regular, non-exam week). The mean for the female students was 10 hours and the standard deviation was 3.5 hours.

a. Consider the following interpretations of the standard deviation and circle those that are correct.

- On average, the number of hours spent studying statistics varied from the mean by about 3.5 hours.
- The average distance between the number of hours spent studying statistics is roughly 3.5 hours.
- The average number of hours spent studying statistics is about 3.5 hours away from the mean.

b. Julie is one of his current students and she studies statistics for about 6 hours per week. What is her corresponding z-score? Then find the probability a randomly selected statistics student studies more than 6 hours per week. Use the Z-table and then use the prob() R script to get a more exact probability and to produce a nice picture and solution.

c. Juan learns that he is in the top 30% of the studying distribution. Based on the distribution, Juan must study at least how many hours per week? Make a hand sketch of what you are trying to find to help show your work.

Problem 4: The Walking Dead

The television series *The Walking Dead* is an American post-apocalyptic horror drama that is based on the comic book series by the same name. The first episode aired in October 2010 and earned a 25% audience share. That is, 25% of all TVs in use during the show time period were tuned to a station airing *The Walking Dead*.

A random sample of 5 people who watched TV during that time period were selected. Find the probability that exactly 2 of the 5 people selected watched *The Walking Dead*.

Cool-Down: True or False

Decide whether the following questions are true or false.

1. If the time to wait for pharmacy help has a uniform distribution from 0 minutes to 30 minutes, then 33% of the customers are expected to wait more than 20 minutes.

   True **False**

2. If X has a Binomial (50, 0.7) distribution, then the criteria to use the normal approximation are met.

   True **False**

3. 68% of all test scores will fall within one standard deviation of the mean test score.

   True **False**

4.. Police report that 78% of drivers stopped on suspicion of drunk driving are given a breath test, 36% are given a blood test and 22% are given both tests. Do the police administer these two tests independently? Justify your answer with a calculation.

   True **False**

Lab 3: Confidence Intervals for a Population Proportion

Objective: This lab will help you better understand the ideas involved in confidence interval estimation as well as how to interpret both the confidence level and confidence interval for a population proportion. You will construct one-sample confidence intervals for a population proportion, and to check that the conditions necessary for the interval are valid.

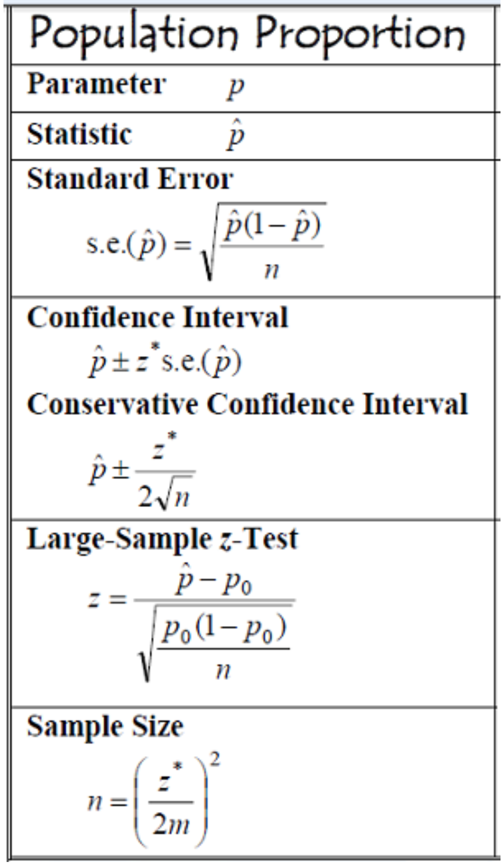
Application: The University of Michigan wants to learn more about the number of computers that are needed by students on campus as they have the resources to add more computers if needed. We have access to the total number of students that are enrolled at the University and we are asked to conduct a study to assess if more computers are needed. Using the data collected from the study, we can estimate the proportion of all Michigan students who don't have a laptop and will require access to an on-campus computer. The University can use this estimate of the population proportion along with the number of total students on campus to estimate the number of computers on campus that are needed.

Overview: Since this population proportion is an unknown number, we are interested in knowing how close a sample proportion (based on the random sample) is expected to be to the true proportion. When we report the estimate, we should provide both the estimate of UM students who don't own a laptop (sample proportion) and a statement that describes the precision of the estimation process.

Confidence intervals (CIs) provide a method of stating both how close the sample proportion is likely to be to the value of a population proportion and the accuracy of it being that close.

The basic structure for any confidence interval is: **estimate ± (multiplier x standard error)**. The "multiplier x standard error" portion is also called the **margin of error** (or error margin). The multiplier used will depend on the confidence level we will use and will be a z multiplier since we are constructing an interval for a population proportion.

In contrast, the **confidence <u>level</u>** is the **proportion of times this <u>method</u> *will* produce an interval that contains the true proportion of UM students without a laptop in repeated random sampling** (if this study were to be repeated over and over).



Population Proportion	
Parameter	p
Statistic	$\hat{p}$
Standard Error $\text{s.e.}(\hat{p}) = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$	
Confidence Interval $\hat{p} \pm z^{*}\text{s.e.}(\hat{p})$ **Conservative Confidence Interval** $\hat{p} \pm \dfrac{z^{*}}{2\sqrt{n}}$	
Large-Sample z-Test $z = \dfrac{\hat{p}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	
Sample Size $n = \left(\dfrac{z^{*}}{2m}\right)^2$	

Confidence Interval (CI) Summary

For Exploring the Population Proportion of UM Students without a Laptop

1. **Interpreting a 95% Confidence Interval itself:** We are 95% confident that the true population proportion of UM students who do not own a laptop lies inside the CI we will compute. The interval provides a range of reasonable values for this population proportion.
2. **Interpreting a 95% Confidence Level used to make the interval:** If the procedure were repeated many times (that is, if we repeatedly took a random sample of the same size, and computed a 95% CI based on each sample), we would expect 95% of the resulting CIs to contain the true population proportion of UM students who do not own a laptop.
3. **Be Careful about how you use *probability* or *chance*:** The **probability** that the true population proportion of UM students without a laptop is located in a particular, already computed confidence interval is either 0 or 1. Both the interval and the population proportion are fixed entities, so either the true proportion is in that particular interval, or it is not.
4. **How Confidence Intervals can Guide Decision Making:**

   Principle 1: A value in the CI we calculate is an "acceptable" or "reasonable" possibility for the population proportion of UM students without a laptop. A value not in a CI can be rejected as a likely value of this population proportion.

   Principle 2: When the CIs for the proportion for two different populations do not overlap, it is reasonable to conclude that the proportions for the two populations are different.

The above summary and interpretations can be used for any problem by adjusting the confidence level and parameter under study, given the context of the problem. Phrases such as "true population proportion of UM students who do not own a laptop" should be adjusted accordingly.

Warm-Up: Confidence Level Interpretation

Background: A study was conducted to learn about eating habits for American families. A random sample of 200 families was selected, and an adult head of household was asked to complete a survey. One question asked was, "Did your family eat dinner together last Sunday night – yes, no?" Based on the results, a 95% (conservative) confidence interval for the proportion of all such families that ate dinner together last Sunday night is given by (0.56, 0.70).

Task: As part of an exam question, students were asked to write a concise statement to explain the meaning of the confidence level of 95% (*'interpret the confidence level)*. Below are some of the answers received. For each, determine if it is a correct or an incorrect interpretation of the 95% confidence **level**.

1. If this survey was repeated many times, we are confident the true population proportion of families that ate dinner together last Sunday night is represented in the data 95% of the time.

   Correct **Not Correct**

2. The 95% confidence level means that with this method, and for similar samples, we would construct many confidence intervals, and of these, 95% would contain the true population proportion of families that ate dinner together last Sunday night.

   Correct **Not Correct**

3. In the long run, the population proportion of families that ate dinner together last Sunday night will be found in the interval 0.56 to 0.70, 95% of the time.

   Correct **Not Correct**

ILP: Calculating a CI for a Population Proportion

In this ILP, we are going to investigate using the sample proportion as a way of estimating the population proportion. For the purpose of this activity, we are going to assume that the lab section is a representative random sample of the UM student body. Follow the steps below to estimate the population proportion.

1. Determine a question to investigate by filling in the following question.

   "What proportion of UM student …

   ____?"

2. Once the question has been determined, the instructor will pose the question to the class using clickers. Record the results below and provide an estimate of the population proportion of all UM students that _______________ (that is, report the sample proportion).

# of students responding:	
# of student who responded "yes":	
Sample proportion of student who responded "yes" : (symbol and value)	

3. (a) The estimate in part (2) is a sample proportion. About how far away from the population proportion would you expect such estimates to be, on average? (i.e. report the standard error of the sample proportion).

   (b) There are certain conditions required for this inference procedure to be valid. Are they met here? If not, how would you update this study so the conditions would be met?

   (c) Use your results to construct a 95% confidence interval estimate for the population proportion of all UM students that _______________.

(d) What would happen to the width of your interval if you were to construct a 99% CI instead (using the same survey results)?

(e) Interpret the **interval** you constructed in part (c).

(f) Explain what the corresponding confidence **level** means.

(g) What (minimum) sample size would be needed for a 95% (conservative) confidence interval with a margin of error of 0.06 (or 6%)?

Cool-Down: Interpretations of Confidence Intervals

Consider once again the survey of American families, and recall that the 95% (conservative) confidence interval for the proportion of all such families that ate dinner together is given by (0.56, 0.70).

Below are answers provided by students when asked to now give an interpretation of the actual 95% confidence interval computed in the study. Assess which are correct and which are not correct explanations as to what the interval (0.56, 0.70) means (in the context of this study)?

1. There is a 95% chance the confidence interval of (0.56, 0.70) will contain the true population proportion of all families that ate dinner together last Sunday night.

   Correct   **Not Correct**

2. We estimate with 95% confidence that the true population proportion of all families that ate dinner together last Sunday night will be found in the interval 0.56 to 0.70.

   Correct   **Not Correct**

3. If we repeated this study many times, the true proportion of all families that ate dinner together last Sunday night will be in the interval (0.56, 0.70) about 95% of the time.

   Correct   **Not Correct**

Lab 4: Hypothesis Testing for a Population Proportion

Objective: In this lab, you will learn an important statistical technique that will allow you to answer the question, *"Was our observation due to chance, or is it more significant*?" The objective is to guide you through the ideas behind tests of statistical significance and the statistical language involved. This lab first presents a general overview of testing. Then, the In-Lab Project discusses the large sample Z-test for a population proportion, as well as providing practice in answering the question of whether an event can be attributed to chance.

Application: Lucy is an Ann Arbor resident who is considering the possibility of running for a seat on the City Council. To decide if she should invest time and money in her campaign, she wants to conduct some research to see if there is evidence that she could win the election (i.e. she could receive a majority of all the votes).

Overview:

Hypothesis Tests: A **test of hypotheses** (or **significance test**) is a procedure designed to assess what the evidence provided by the data says about some statement about a population parameter. When we conduct this hypothesis test for Lucy, we are interested in testing if the population proportion of votes that would be cast for Lucy is greater than 0.5, ensuring her victory.

Hypothesis Testing Steps:

1. Determine appropriate null and alternative hypotheses and set $\propto$.
2. Check assumptions for performing the test.
3. Calculate the test statistic and determine the *p*-value.
4. Evaluate the *p*-value and report a conclusion in the context of the problem.

The first step of this hypothesis test is to identify the hypotheses; this step is crucial, as it dictates the procedures for the remainder of the test. The **null hypothesis**, H_0, represents the status quo or statement of no effect. The **alternative hypothesis**, H_a (or sometimes H_1), represents the experimenter's new model, or what the experimenter would like to show.

For our application regarding Lucy, we could define *p* to represent the true proportion of all Ann Arbor residents who would vote for Lucy. We want to test the hypothesis H_0: $p = 0.5$ against the alternative H_a: $p > 0.5$.

Both of the hypotheses are postulated about the same **population** proportion. The alternative hypothesis can take three different forms – it may be a denial of the null hypothesis (uses$\neq$; called a **two-sided test**), or it may specify a direction of interest (uses > or <; called a **one-sided test**). Notice that we <u>never</u> test for equality in H_a.

The purpose of a significance test is to assess whether or not the observed data are consistent with the null hypothesis (within the reasonable bounds of sampling variability). If we collect our sample and compute a population proportion of votes for Lucy that is much higher than the 0.5 we specified in the null hypothesis, we have evidence to reject that null hypothesis.

To help us make this decision, we use a **test statistic**, which represents a summary of the data. When we conduct a test about a population proportion, we are performing a Large Sample Z test. The test statistic for this Large sample z test has the following form:

$$z = \frac{\hat{p}-p_0}{\sqrt{(p_0)(1-p_0)/n}}.$$

This p_0 is the value specified in the null hypothesis. For Lucy's test: $p_0 = 0.5$. The sample proportion of people who will vote for Lucy is our $\hat{p}$. The test statistic tells us how many standard errors the sample proportion, $\hat{p}$, is from the test value, p_0.

The test relies on two key assumptions: that we have a random sample (our sample of responses is representative of the large population of all such responses), and that our sample size is large enough. To achieve a large enough sample size, we need at least 10 hypothesized yes answers and at least 10 hypothesized no answers, or said another way, under the null hypothesis, would we expect to see at least 10 yes answers and at least 10 no answers.

Conditions: np_0>10 and $n(1-p_0)$>10

The Z test statistic has a known probability distribution (under the null hypothesis), and will be examined for evidence in favor of or against H_0. Under the null hypothesis that the true proportion of votes for Lucy is equal to 0.5, the test statistic Z has a standard normal distribution.

In hypothesis testing, another frequently reported value is the ***p*-value**, a number that is used to indicate the degree of significance of the data. *The **p-value** is the probability of getting a **test statistic** as extreme or more extreme than the observed value of the **test statistic**, assuming the null hypothesis is true.* For Lucy's test, the p-value is the probability of getting a z test statistic like we did or greater, assuming that the true proportion of votes for Lucy is equal to 0.5.

We must decide in advance how much evidence against H_0 we will require for rejection. This designated amount of evidence is called the **level of significance**, denoted by α (alpha). Common values of α are 0.01, 0.05, and 0.10. If the *p*-value is less than or equal to α, we make the decision to reject H_0. For Lucy's test, if we encounter a p-value smaller than the significance level, we would decide to reject the hypothesis that the true proportion of votes for Lucy is exactly 0.5.

If we reject the null hypothesis, the results of the test are said to be **statistically significant** at level α. A "significant" result in the statistical sense does not necessarily imply an "important" result in the practical sense. It simply means that such a difference from the null hypothesis is not likely to happen just by chance.

Of course, no procedure is perfect, and as such, there are two types of errors possible during hypothesis testing. If the null hypothesis is true but the decision is to reject H_0, then we say that a **Type I error** has occurred. If we commit a Type I error for Lucy's test, we would reach the decision that Lucy will win the election by receiving more than 50% of the votes, when in fact the true proportion of votes cast for Lucy is 0.5. Lucy would invest her time and money in her campaign when in fact she will not win the election.

A **Type II error** occurs when the alternative hypothesis is true, but we fail to reject H_0. A Type II error could be committed in Lucy's test if we decide Lucy will lose the election by receiving 50% of the votes or less, when in fact Lucy will receive greater than 50% of the votes and win the election. Lucy would decide not to run in the election when she would have won a seat on the City Council.

Each type of error has a probability of occurring. If the null hypothesis is true, the level of significance, α, is also the probability of a Type I error, while the probability of a Type II error is denoted by β.

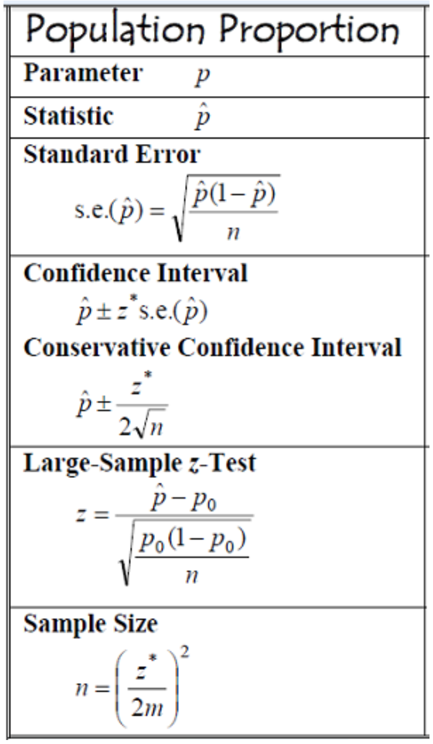
Truth	Decision Made	Result	Associated Probability
H_0 True	Reject H_0	Type I Error	α
	Do Not Reject H_0	Correct Decision	$1 - \alpha$
H_a True	Reject H_0	Correct Decision	$1 - \beta$ = power
	Do Not Reject H_0	Type II Error	β

Another important component of a hypothesis test is its **power**. The **power** of a test measures its ability to detect an alternative hypothesis when it is true. Power of a particular test is calculated as the probability that the test will reject H_0 when the alternative hypothesis is true.

For Lucy's test, the power is calculated as the probability that we conclude Lucy will receive more than ½ the votes in the election when the truth is that she actually will win the election. Since we just learned that β is the probability that we did NOT reject H_0 when the H_a is true, we can see that power is represented by $1 - \beta$.

Formula Card

Corresponding details and formulas as they appear on the Stats 250 formula card are shown at the right.



Population Proportion

Parameter	p
Statistic	$\hat{p}$

Standard Error

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Confidence Interval

$$\hat{p} \pm z^{*}\text{s.e.}(\hat{p})$$

Conservative Confidence Interval

$$\hat{p} \pm \frac{z^{*}}{2\sqrt{n}}$$

Large-Sample z-Test

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Sample Size

$$n = \left(\frac{z^{*}}{2m}\right)^2$$

Warm-Up: Stating the Hypotheses and Defining the Parameter of Interest

For each example, fill in the hypotheses and define the parameter of interest.

1. The Detroit Tigers advertising team believes that about 70% of Ann Arbor residents will attend a Tigers game this season. Suppose the general manager wants to cut back on advertising in the area and speculates that Ann Arbor residents are less likely to attend a Detroit Tigers game this season than previously assumed.

   Let the parameter ____ represent

   H_o:_______________________________ H_a:_________________________________

2. The University of Michigan is interested in examining the proportion of students who graduate in four years. Suppose that a researcher speculates that the University has a majority of students who graduate four years after they enroll.

   Let the parameter ____ represent

   H_o:_______________________________ H_a:_________________________________

ILP: Hypothesis Testing for a Population Proportion

In this ILP, we are going to use a sample proportion to test a theory about the value of the population proportion. Again, we assume the lab section is a representative random sample of the UM student body.

1. Determine a question to investigate by filling in the following question.

   "What proportion of UM students …

   ____?"

2. Once the question has been determined, the instructor will pose the question to the class using clickers. Record the results below and provide an estimate of the population proportion of all UM students that ________ (that is, report the sample proportion).

# of students responding:	
# of student who responded "yes":	
Sample proportion of student who responded "yes" : (symbol and value)	

3. The estimate in part (2) is a sample proportion. About how far away from the population proportion would you expect such estimates to be, on average? (i.e. report the standard error of the sample proportion).

4. We wish to use the data collected to assess:
"Do a … (select one) ***minority*** ***majority*** of UM students_____________?"
Express this as an appropriate null and alternative hypothesis:

H_0:______________________________ H_a:___________________________________

where ***p*** represents:____

5. A large sample Z test can be performed using a normal approximation for the binomial if the sample size is large enough. Provide the checks necessary to see if this approximation may be made.

6. Assume the sample is large enough and calculate the corresponding z test statistic value. Show your work.

What is the distribution of the test statistic under the null hypothesis?

7. Use your test statistic to find the corresponding *p*-value. First make a quick sketch by hand and use Table A.1 to find your *p*-value; then try the new pval() script in R to check your answer.

What would the *p*-value look like if the alternative hypothesis had been that the population proportion did not equal 0.5?

8. What is your decision at the 5% significance level? **Reject H_0** **Fail to Reject H_0**

Also write out your real world conclusion in the context of the problem.

Cool-Down: Errors and Power Computation

Maize and Blue Marbles: Cory is going to play the game 'What is in the Box?' He will win season tickets to the Red Wings if he can correctly identify the contents of a box (which he cannot see). There are two possibilities for the box contents:

H_0: The box contains eight red marbles and four white marbles.

H_a: The box contains two red marbles and ten white marbles.

Cory will select one marble from the box without looking. He must make his decision based on the color of that one selected marble. He has picked the following decision rule to use: **Reject H_0 if the observed color of the selected marble is WHITE.**

It may help to make a quick visual picture of these two boxes.

a. For this situation, what is the probability of incorrectly rejecting H_0?

2/12 4/12 10/12

b. Which of the following terms is appropriate for the probability described in part (a)?

Type I error Type II error Power

c. For this situation, what is the probability of incorrectly failing to reject H_0?

2/12 4/12 10/12

d. Which of the following terms is appropriate for the probability described in part (c)?

Type I error Type II error Power

e. For this situation, what is the power?

2/12 4/12 10/12

Lab 5: Understanding Normal and Random Data

Objective: In this lab, you will use some additional graphical tools to summarize the distribution for a variable or response and check assumptions before performing a statistical test. Graphs you might need to examine include time plots for data collected over time and QQ plots for checking whether a normal (bell-curve) model is a reasonable distribution for a quantitative variable. These techniques can be very useful at the start of data analysis to get a feel for the data.

Application: Brad is the manager of the Detroit Tigers and to prepare for the next round of the playoffs, he would like to run a hypothesis test involving the mean number of runs his players have scored over the last month of play. Brad knows that one assumption required for performing this analysis about the mean is that his data must be considered a random sample (the observations can be viewed as coming from the same parent population). He can examine this assumption by collecting the number of runs scored over the last month and creating a time plot.

Overview: Data on a quantitative variable should be examined graphically. If the data has been collected over time, the first graph to examine is a **time plot**. If the resulting time plot appears to be stable, or if the data was not collected over time, then graphs that can be used to summarize the distribution for a single quantitative variable or response are a **histogram,** a **boxplot**, and perhaps a **QQ plot**. Each graph provides different information about the distribution. The overall shape of the distribution and existence of outliers can generally be used to assess if the data appear to be coming from a relatively homogenous population. If so, then various **numerical summaries** may be used to characterize the center of the distribution and the spread of the distribution.

Note that some graphical tools are introduced solely in lab, not in lecture,
so it will benefit you to read this overview thoroughly

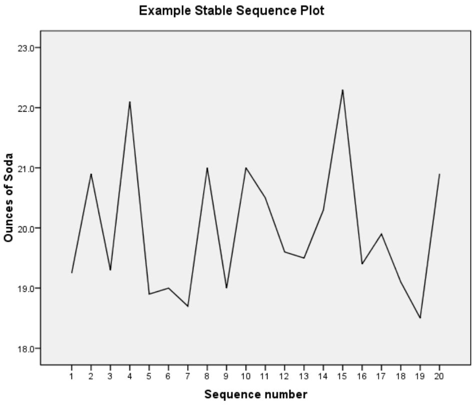
Sequence (Time) Plots: Data might be gathered over time. Employment rate, stock prices, and sales figures are just a few examples. When data is gathered over time, such as the number of runs scored over a one month by the Tigers, it is generally wise to examine the data plotted against time. Plots against time can reveal the main features of a time series, overall patterns and striking deviation from those patterns. Some overall patterns that may arise are:

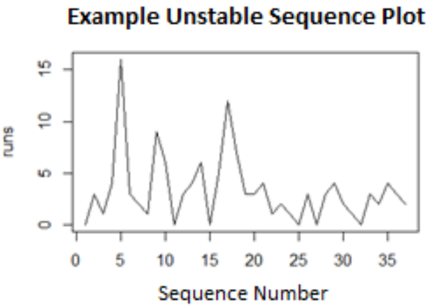
- ❑ A persistent, long-term rise or fall called a **trend** (either increasing or decreasing).
- ❑ A pattern that repeats itself at regular intervals of time called **seasonal variation**.
- ❑ A persistent, long-term increase or decrease in the **variation** of the observations called a **pattern in variation**.

If data is collected over time, a sequence plot can be used to check the assumption of a random sample, often needed for inference procedures. A random sample consists of *independent* and *identically distributed* (i.i.d.) observations. This means the observations can be considered as all coming from the same parent population (with the same or *identical* distribution) and are *independent* of one other.

With a sequence plot, you can check the *identically distributed* aspect of a random sample by looking for evidence of **stability** in the plot. Stability is supported when both the mean of the observations and the amount of variation among observations appear to be constant over time and there does not appear to be any *pattern* in the resulting plot.

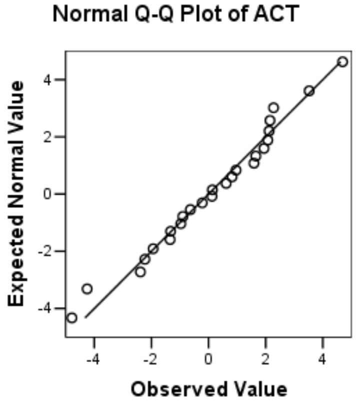
Below are two sequence plots; in the first plot the observations appear to support that the underlying process that generated the observations is stable, but that is not the case for the observations in the second plot on the right. In this case, there appears to be an increasing trend, thus the underlying process does not appear to be stable; the observations should not be considered a random sample.



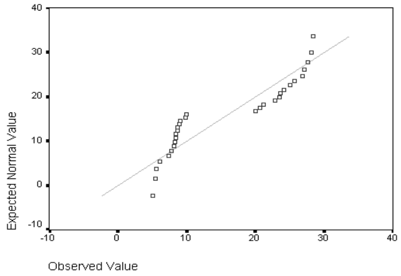


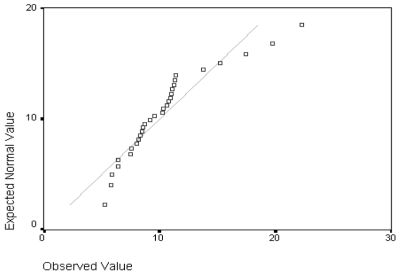
Going back to the earlier application, here is the time plot Brad created of the number of runs the Detroit Tigers scored between September 1, 2013 and October 19, 2013. This is another example of an unstable time plot. Brad notes that there is a large variation in the number of runs scored by the Tigers early in September but less variation in their run total around game 21. The underlying process for the number of runs scored does not appear to stable, so Brad should not consider this a random sample.

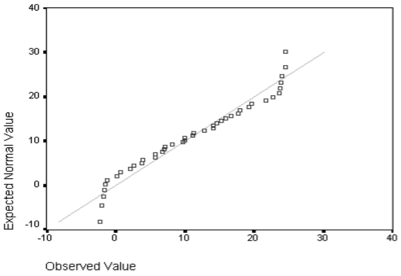
QQ Plots: Sometimes the **assumption of a normal model for a population of responses will be needed in order to perform certain inference procedures**. A histogram can be used to get an idea of the shape of a distribution. However, there are more sensitive tools for checking whether the shape is *close* to a normal (bell-curve) model. The *best* plot that can be used to check for normality is called a QQ Plot, which plots the percentiles (quantiles) of a standard normal distribution against those of the observed data. If the observations follow an approximately normal distribution, the resulting points should follow roughly a straight line with a positive slope. Strong deviations would indicate possible departures from a normal distribution. At the right is an example of a QQ Plot showing data that does seem to come from a population with an approximately normal distribution.



The three graphs below are examples for which a normal model for the response does not seem reasonable. The QQ plot on the far left indicates the existence of two clusters of observations. The QQ plot in the center shows an example where the shape of the distribution appears to be skewed right. The QQ plot on the far right shows evidence of an underlying distribution that has shorter tails compared to those of a normal distribution.

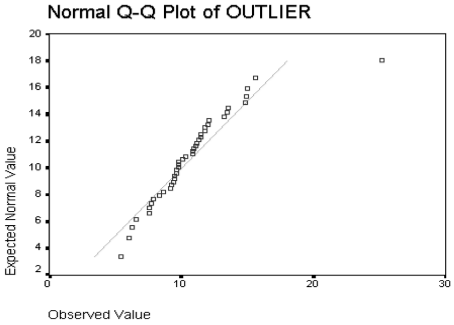






Note: Many inference procedures, including some you will use later in the semester, require the assumption of normally distributed population(s). Most of these procedures are robust which means we would need to see strong evidence of a departure from normality to conclude this assumption is not met. Some mild departures from normality would still allow us to conclude the underlying model for the response is reasonably normally distributed.

Finally, consider the QQ plot at the right. In this case, we would say the QQ plot shows evidence of an underlying distribution which is approximately normal except for one large outlier that should be further investigated. Outliers could appear in either the upper or lower tail.



Warm-Up: Matching

Match the graph or descriptive statistic to one of its primary uses.

____ 1. Histogram A. Measure of center, not sensitive to outliers

____ 2. Bar Chart B. Compare distributions (but not their shapes)

____ 3. Mean C. Examine distribution of a categorical variable

____ 4. Median D. Helps assess if can treat data as a random sample

____ 5. Side-by-side Boxplots E. Measure of spread

____ 6. IQR F. Examine distribution of a quantitative variable

____ 7. Time Plot G. Helps assess if underlying distribution is bell-shaped.

____ 8. QQ Plot H. Measure of center, sensitive to outliers

ILP: More Visualizing and Exploring Quantitative Data

In this In-Lab Project, you will create additional useful graphs and obtain more descriptive statistics for quantitative data using R.

Task 1: **QQ Plots.**

1. Recall the data set **employee.Rdata** contains information on employees at a company, and that a histogram of the current salary data indicated a strongly skewed to the right distribution. Let's examine a **QQ plot** (otherwise known as a Quantile-comparison plot) to see that it supports this non-normal feature. To create a QQ plot, go to **Graphs > Quantile-comparison plot** and select the **SALARY** variable.

   Provide a sketch and explain how the QQ plot supports that the underlying distribution of salary does not appear to be normal (or bell-shaped)

2. We can also consider making QQ plots for different populations. Recall that the distribution of salary for minorities was far less skewed than the whole distribution. Let's try making a QQ plot of salary for minorities. To do this, we will need to make a new dataset that only contains minorities in it. We do this by going to **Data > Active data set > Subset active data set**. The subset expression should be formatted like VARIABLE=="value" – here, the variable we are splitting by is MINORITY, and we want all observations with the value "Yes" for MINORITY; so type in **MINORITY=='Yes'**. Give the data set a new name and click OK – if done correctly, the new data set will appear in blue text at the top.

   Now, make a QQ plot of the **SALARY** variable as we did before with the new data set. How does this distribution of minorities compare to the entire distribution from the previous question? Would you consider this a normal distribution?

3. Let's try analyzing the normality of IQ scores given in the dataset **iq.Rdata,** which contains information from high school students collected for the purpose of examining the relationship between IQ scores and GPA. Before creating a QQ plot, we can get an idea of the distribution of IQ scores by simply making a histogram. Make a sketch of this histogram below. Do you think the population of IQ scores might follow a normal distribution?

4. Let's see if our suspicions from the previous problem are correct, and make a QQ plot of the IQ scores. You will need code similar to what we used in problem 1 – recall that variable names can be viewed by clicking on **View data set**.

   Draw the resulting QQ plot below. Is a normal distribution a reasonable model for IQ scores in the population based on this QQ plot?

Task 2: **Time Plots.**

1. The data set **oldfaithful.Rdata** contains the date and duration of eruptions (in minutes) of the Old Faithful geyser. The data was collected several times per day over 23 consecutive days. To make a time plot, we will have to use some custom code, as there is no R Commander function for a time plot. To make a time plot of the **DURATION** variable from the **oldfaithful** data set, type the following code into the **R Script** box at the top of your R Commander window (note the type is the letter L but must be lowercase):

   ```
   plot(oldfaithful$DURATION, type=“l”, main="Time Plot of variable
   by name")
   ```

   Highlight this code, and then click the **Submit** button. Make a sketch of this time plot.

2. Does the plot appear to show any trends or changing variation? Would you consider it to be stable?

3. Would it be reasonable to conclude these data are a random sample of eruptions?

4. Would it make sense to go on to make a histogram of these data? A QQ plot of these data? Descriptive summaries? Why or why not?

5. Let's examine another time plot – download the data set **chemical.Rdata**, which contains the concentrations of a certain chemical in 15 consecutively produced batches of solution. Create a time plot of the **concentration** variable and make a quick sketch here, using code similar to problem 1.

6. Does the plot appear to show any trends or changing variation? Would you consider it to be stable?

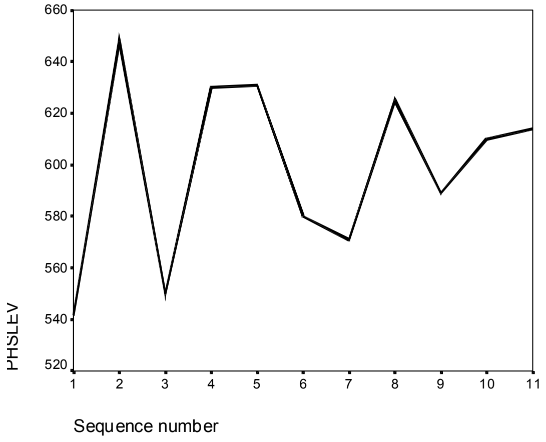
7. Would it be reasonable to conclude these data are a random sample of eruptions?

8. Would it make sense to go on to make a histogram of these data? A QQ plot of these data? Descriptive summaries? Why or why not?

Cool-Down: Check Your Understanding About Time Plots

A new method of measuring phosphorus levels in soil is under consideration. A sample of 11 soil specimens is analyzed using the new method. The time series (sequence plot) for the 11 observations is presented.

Comment on the overall stability of these data based on this plot.



Would it be appropriate to go on to make a histogram of these data? Explain.

Optional Review of Time Plot and QQ Plot Examples

If you would like to look at more examples of QQ plots (which help assess whether the model for the underlying population of responses seems to be normal) and examples of time plots (which help assess whether our underlying process appears to be stable and if we can consider the data to be a random sample) … try out these tasks using some simulations with R (R scripts).

Task 1: Go to the **Extra Review** link on your course site and click on the **QQ Plots in R** link in the list. Review the background about the simulator and download the **qqplot** script file (which will open up the R program).

1. Begin the program by entering the following command. **qqplot ()**
2. Select your sample size by entering a number between 1 and 10000.
3. Select if you want a QQ plot from a normal or non-normal distribution.
4. If a non-normal distribution, select the type of distribution you would like to see.
5. Once your QQ plots and the corresponding histograms have been created, you will be asked if you wish to save the plot. You will then be prompted to select a sample size. Create QQ plots for many different samples.
6. Sketch the QQ plot and Histogram for a sample from a skewed right distribution.
   QQ plot: Histogram:

Task 2: Go to the **Extra Review** link on your course site and click on the **Time Plots in R** link in the list. Review the background about the simulator and download the **timeseries** script file (which will open up the R program).

1. Begin the program by entering the following command: **timeseries ()**
2. Select your sample size by entering a number between 1 and 10000.
3. Select if you would like an example of a stable or unstable time plot.
4. If unstable time plots are selected, select the type of pattern you want to see.
5. Once your time plots have been created, you will be asked if you want to save the plot. You will then be prompted to select a sample size. Try out the various options and explore the various patterns of time plots.
6. Create a time plot for a sample of 1000 data points taken from a process with an increasing mean and decreasing variance. Sketch the time plot below.

Lab 6: Learning about a Population Mean

Objective: In this lab, you will learn an important statistical technique that will allow you to answer the question, ***"Was our observation due to chance, or is it more significant?"*** The objective is to guide you through the ideas behind tests of statistical significance and the statistical language involved. This In-Lab Project discusses the one-sample *t*-test for a population mean, as well as providing practice in answering the question of whether an event can be attributed to chance.

Application: Derek is considering attending the University of Michigan and he wants to consider the cost of rent in Ann Arbor to help make his decision. Ann Arbor's website claims that the average price for apartments in Ann Arbor is $800 a month. However, Derek's friends lead him to believe that the average price for all apartments in Ann Arbor is actually greater than $800 a month. Derek would like to conduct a hypothesis test to see if his friends' statement is supported.

Overview:
Hypothesis Tests: A **one-sample t test** is a procedure designed to assess what the evidence provided by the data says about some statement about the population mean. The key elements of the one-sample t test include: a null and alternative hypothesis, assumptions that must be checked, a test statistic t and its *p*-value, a decision, and a conclusion.

Hypothesis Testing Steps:

1. Determine appropriate null and alternative hypotheses about μ, and set $\propto$.
2. Check assumptions for performing the test.
3. Calculate the test statistic and determine the *p*-value.
4. Evaluate the p-value and report a conclusion in context of the problem.

The first step of a hypothesis test is to identify the hypotheses; this step is crucial, as it dictates the procedures for the remainder of the test. A one-sample *t*-test is used to test whether the mean of a quantitative variable is significantly different from some value. This value, the **test value** or μ_0, is given in the null hypothesis (H_0: $\mu = \mu_0$), and is often taken to be zero (i.e., H_0: $\mu = 0$). In Derek's test, his null hypothesis is H_0: $\mu = 800$.

The **alternative hypothesis**, H_a (or sometimes H_1), represents the experimenter's new model, or what the experimenter would like to show. The alternative hypothesis can take three different forms – it may be a denial of the null hypothesis (uses $\neq$; called a **two-sided test**), or it may specify a direction of interest (uses > or <; called a **one-sided test**). Notice that we never test for equality in H_a. Derek's alternative hypothesis will be H_a: $\mu > 800$.

The test relies on two key assumptions: (1) the data can be considered a **random sample** from the larger population of interest and (2) the data are observations from a **normally distributed population**. The Central Limit Theorem allows the assumption of normality to be more relaxed if our sample size, *n*, is large. (Generally, "large" means more than 30 observations, although it also depends somewhat on how serious the data depart from normality.) The sampling distribution applet that you worked with demonstrated the CLT using a *large* sample size of 25. Thus we will accept *at least 25* as large enough for inference about means.

The purpose of a significance test is to assess whether or not the observed data are consistent with the null hypothesis (within reasonable bounds of sampling variability). If the data seem unlikely to occur if the null hypothesis is assumed true, then we would reject the null hypothesis statement.

To help us make this decision, we use a **test statistic**, which represents a summary of the data. Going back to the application, Derek samples 40 students at the University by asking about their monthly rent. From the sample he computes a sample mean rent of $900 as well as a standard error for this sample mean. This test statistic takes the form: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$.

The **test statistic tells us how many standard errors the sample mean, $\bar{x}$, is from the test value, μ_0** . Derek calculated a test statistic of t = 1.2. His interpretation would state: The sample mean rent of $900 per month is 1.2 standard error away from the hypothesized mean rent of $800.

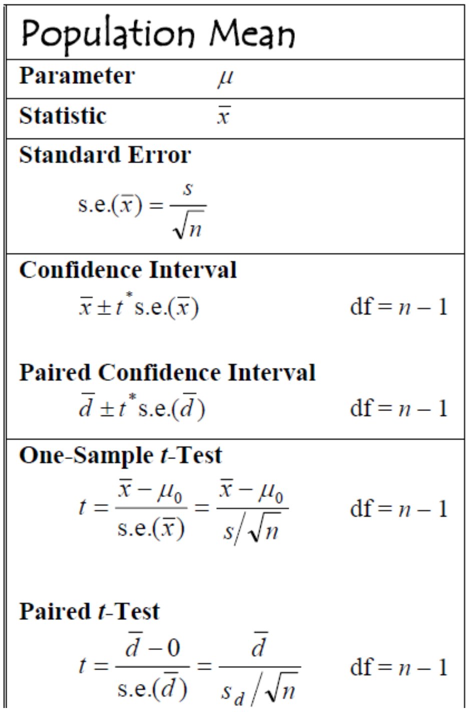
In hypothesis testing, another frequently reported value is the ***p*-value**, a number that is used to indicate the degree of significance of the data. The *p*-value is the probability of getting a test statistic as extreme or more extreme than the observed value of the test statistic, assuming the null hypothesis is true.

In Derek's case, the *p*-value is the probability of observing a test statistic of 1.2 or larger, if the population mean rent in Ann Arbor really was 800 dollars more month. Note that more "extreme" means in the direction of Ha. His test statistic, t, has a t-distribution with n-1 degrees of freedom (df), where n represents the sample size. In Derek's test, the t statistic has a t (39) distribution.

We must decide in advance how much evidence against H_0 we will require for rejection. This designated amount of evidence is called the level of significance, denoted by α (alpha). Common values of α are 0.01, 0.05, and 0.10. If the *p*-value is less than or equal to α, Derek would make the decision to reject H_0. If Derek makes the decision to reject the null hypothesis, the results of the test are said to be statistically significant at level α.

A "significant" result in the statistical sense does not necessarily imply an "important" result in the practical sense. It simply means that such a difference from the null hypothesis is not likely to happen just by chance.

Formula Card



Population Mean

Parameter μ

Statistic $\bar{x}$

Standard Error

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}}$$

Confidence Interval

$$\bar{x} \pm t^{*}\text{s.e.}(\bar{x}) \qquad \text{df} = n-1$$

Paired Confidence Interval

$$\bar{d} \pm t^{*}\text{s.e.}(\bar{d}) \qquad \text{df} = n-1$$

One-Sample *t*-Test

$$t = \frac{\bar{x}-\mu_0}{\text{s.e.}(\bar{x})} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \qquad \text{df} = n-1$$

Paired *t*-Test

$$t = \frac{\bar{d}-0}{\text{s.e.}(\bar{d})} = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad \text{df} = n-1$$

Warm-Up: Check Your Understanding

A researcher wants to determine if a filling machine at a cereal plant fills the boxes well on average or if the machine needs to be replaced. The machine is intended to fill the packages with 28 grams of cereal. He wants to test the machine to see if the average weight of cereal is less than the intended 28 grams. He takes a sample of 30 cereal boxes and measures the weight of each box. The histogram of the weights indicates some skewness.

Which result allows him to conduct the one sample t-test even though his histogram shows some skewness?

He performs a one sample t test and finds a statistically significant *p*-value.
If the significance level was 5%, what are possible options for the *p*-value?

0.0067 0.043 0.14 0.22

ILP: How Fast is your Stats 250 Lab at the Tangrams game?

Background: The Tangram is a dissection puzzle that is claimed to have originated in China, and players must try to make the shapes out of seven pieces of varying shapes. The most common goal of the game is to make a square out of the seven pieces, but the goal can be changed to a varying number of shapes. In today's lab, you will be completing the "House of Tangrams" puzzle, with your time to complete the puzzle recorded. A previous Stats 250 GSI claims that her lab completed the puzzle in 120 seconds, on average – can your lab beat this?

Task: Perform a test to assess if the average time to complete the "House of Tangrams" puzzle is significantly faster than 120 seconds.

Complete the "House of Tangrams" puzzle.

Your GSI will provide you with a link to access the Tangrams puzzle online. Some tips before you begin:

- Use an alias when playing so you can find the row corresponding to you in the dataset. It does not need to be your real name, but it can be if you'd like.
- Please, only play the game once, as repeated entries will be added to the dataset and will cause our results to be incorrect!
- To rotate a piece clockwise, double-click it (and hold alt to rotate counter-clockwise). If you are having trouble rotating multiple times, it may help to slightly move the cursor after each click.
- If you are attending remotely, please make a quick post in the backchannel saying that you are starting the tangrams puzzle. Reply to that same post when you are done so I can keep track of who still needs to complete the puzzle.

We will need to wait until all the results are entered before we download the data set, so review the overview section of this lab (before the Warm-Up) and when done, you can start working on the next page while you wait for your GSI to announce that the data is ready to download.

Procedure: The appropriate inference procedure for this scenario is the one-sample t test and the specific parameter of interest is ____________________ .

Hypothesis Test: You can complete the steps for conducting a test of hypotheses.

1. **State the Hypotheses:** H_0: ___________ = ___________ and H_a: ___________ ___________ ,

   where __________ represents:

   Determine Alpha: We were told the significance level was 5%.

   > **Remember:** Your hypotheses and parameter definition should always be a statement about the **population(s)** under study.

2. **Checking the Assumptions**

   a. For this scenario, we need to assume that the data are a random sample. To check this assumption, we would make a ________ plot (if there was time order) of the observations, and look for ____________________ .

   b. We also need to assume that the responses come from a normally distributed ____________________ . To check this assumption, we would make a __________ plot.

   c. Create the qqplot of the data on time to complete the "House of Tangrams" puzzle and comment.

   d. Based on your sample size, is the assumption of normality for the population of all times to complete the "House of Tangrams" puzzle? Why or why not?

3. **Importing Tangrams data into R Commander**

   When the GSI tells you that everyone has completed the tangrams game, you can then download the data. To do this, click on "Get Game Data" on the page from the link that your GSI gave you earlier. The GSI will give you a group name for your class. Make sure that you select "CSV" as the type to return.

   CSV files are "Comma Separated Values," so each item in a row is separated by a column. Most Microsoft Excel spreadsheets have the option to export data as CSV files, so you can use the following importing method to import your own data for analysis into R Commander.

   To import this CSV data in R Commander, go to **Data > Import data > From text file**. This gives some options on how to import text -- you can give the dataset a name like "Tangrams" or whatever suits you. For our file, the **Field Separator** will be **commas**, as this is a comma-separated-values file. After that, you can click **OK**, and find the file that you downloaded from the tangrams website.

4. Compute the Test-Statistic and Calculate the *p*-value

Test-Statistic

a. Before conducting the test, we will compute some summary statistics used in computing a test statistic. Go to **Statistics > Summaries > Numerical summaries** to generate these, and fill out the table below. Be sure to check the standard error option under the **Statistics** tab.

Summary Statistics			
Mean	**Std. Dev (s)**	**Sample Size (n)**	**Std. Error**

b. Generate the *t*-test output using **Statistics > Means > Single-Sample T Test**. Enter a test value of __________ and specify the correct direction for the alternative hypothesis. (This should match your answer in question 1) Use this to fill out the following output table for this test. Note that when conducting a one-sided test, R will provide a confidence *bound* (instead of a confidence interval), which is not taught in this course.

One Sample T Results		
t	***df***	***p-value***

c. What distribution does *the test statistic follow* if the null hypothesis is true?

Note: This will be the model used to find the *p*-value.
This is not the same as the distribution *of the population that the data were drawn from*.

Calculate the *p*-value:

d. Draw a picture of the *p*-value, with labels for the distribution and x-axis.

4. Evaluate the *p*-value and Conclusion

Evaluate the *p*-value:

What is your decision at a 5% significance level? **Reject H_0** **Fail to Reject H_0**

Remember: Reject H_0 ⇔ Results statistically significant
Fail to Reject H_0 ⇔ Results not statistically significant

Conclusion:

What is your conclusion *in the context of the problem*?

Note: Conclusions should always include a reference to the population parameter of interest.
Conclusions should not be too strong; you can say that you have sufficient evidence,
but do NOT say that we have proven anything true or false.

5. Providing a Confidence Interval Estimate

> **Note:** In order to get a 95% confidence interval with R Commander,
> you will need to run a two-sided test (regardless of what test you are conducting).

a. Generate the two-sided test output and report the 95% confidence interval (include units):

The 95% confidence interval estimate for the population mean time to complete the

"House of Tangrams" puzzle was found to go from _________________ to ____________________.

b. How would the 99% confidence interval compare to the 95% interval in part (a)?

Cool-Down: Check Your Understanding

Provide an **interpretation of the test statistic value** that you found in lab to test if the average time to complete the "House of Tangrams" puzzle is significantly faster than 120 seconds.

How would you explain the meaning of the **95% confidence level** used in making the confidence interval for the population mean time to complete the "House of Tangrams" puzzle above?

Lab 7: Paired Data Analysis

Objective: In this lab, you will learn how to perform a hypothesis test in the case when we have two quantitative variables collected in pairs, called a **paired *t*-test**. You will make a confidence interval for and test hypotheses about the population mean difference, μ_{d}. Using these, you will be able to provide a statement about how confident you are regarding your interval estimate or in your decision.

Application: Mackenzie believes that college students can run a mile faster in the afternoon then they can in the morning. She has ten of her friends run a mile early in the morning and also late in the afternoon and she records their time in seconds. For each of her ten friends Mackenzie computes a difference: time to run the morning mile (in seconds) – time to run the afternoon mile (in seconds).

Runner	AM	PM	Difference (AM – PM)
Runner 1	633.8	618.9	14.9
Runner 2	588.9	569.6	19.3
Runner 3	619.4	630.9	-11.5
Runner 4	640.9	628.5	12.4
Runner 5	613.9	574.2	39.7
Runner 6	590.0	627.3	-37.3
Runner 7	613.4	603.3	10.1
Runner 8	568.6	593.1	-24.5
Runner 9	637.0	596.2	40.8
Runner 10	648.6	613.4	35.2

Overview:

Matched or paired data results from a deliberate experimental design scheme. Mackenzie's scenario is one example of a paired design. Another example of a paired design is an experiment where rats are matched by weight, where one rat in each match receives a new diet and the other rat in the match receives a control diet.

These types of design are called **paired data designs**. Note that paired designs can occur when you have two measurements on the same individual OR when you have two individuals that have been matched or paired prior to administering a treatment.

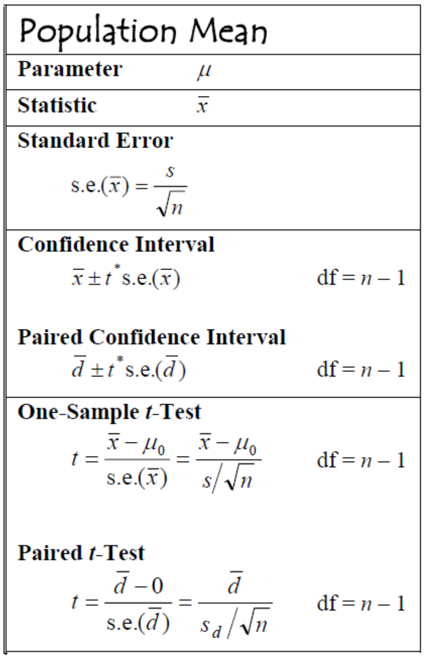
The inference procedures for a paired data design are based on the one-sample *t*-test procedures from the previous lab. The change is that we want to estimate or test hypotheses about the population mean difference μ_{d}, which is generally compared to a hypothesized value of zero, indicating no difference on average.

The assumptions are similar to those assumptions made for the one sample t-test. Going back to the application, Mackenzie now must assume the sample of the ten differences in running time is a random sample. This can be verified by creating a time plot of those ten sample differences and checking for stability. She must also assume that the population of differences is normally distributed. Mackenzie can check this assumption by creating a QQ plot of those ten sample differences.

The test statistic t is created by taking the sample mean difference and dividing by the standard error of the sample mean difference, $t = \frac{\bar{d}}{se(\bar{d})}$.

If the null hypothesis is true, this test statistic has a t distribution with *n*-1 degrees of freedom, where *n* is the number of pairs. For Mackenzie's 10 runners, the degrees of freedom will be 10 – 1 = 9.

Formula Card



Population Mean

Parameter	μ
Statistic	$\bar{x}$

Standard Error

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}}$$

Confidence Interval

$$\bar{x} \pm t^{*}\text{s.e.}(\bar{x}) \qquad \text{df} = n-1$$

Paired Confidence Interval

$$\bar{d} \pm t^{*}\text{s.e.}(\bar{d}) \qquad \text{df} = n-1$$

One-Sample *t*-Test

$$t = \frac{\bar{x} - \mu_0}{\text{s.e.}(\bar{x})} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad \text{df} = n-1$$

Paired *t*-Test

$$t = \frac{\bar{d} - 0}{\text{s.e.}(\bar{d})} = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad \text{df} = n-1$$

Warm-Up: Check Your Understanding

Eight cars were run to determine their mileage, in miles per gallon (MPG). Then each car was given a tune-up, and run again to measure the mileage a second time. The difference in mileage was computed as difference = After MPG minus Before MPG. Assume that the selection of 8 cars represents a random sample of cars.

1. Select the appropriate alternative hypothesis to assess if on average mileage significantly improves after a tune up.

   a. H_a: $\mu_d > 0$

   b. H_a: $\mu_d < 0$

   c. H_a: $\mu_d \neq 0$

2. The researcher of course is hoping that the results of the experiment are statistically significant. What type of *p*-value would the researcher want to obtain?

   a. A large *p*-value.

   b. A small *p*-value.

   c. The magnitude of a *p*-value has no impact on statistical significance

ILP: Do Books Purchased from Barnes and Noble (In-store) Cost More on Average than If Purchased at Amazon.com (Online)?

Background: The popularity of purchasing books online has increased dramatically, and the conventional bookstore no longer dominates the sales of books. The most influential factor that sways customers into purchasing online is lower prices. A group of statistics students decided to perform a comparison of the Amazon.com prices versus local Barnes and Noble bookstore prices based on a sample of 40 books, selected from a wide range of categories. For Amazon, a standard ground shipping of $4.29 and local state tax were included in the cost. The corresponding costs are available in the R data set called **books.Rdata** (Source: Statistics group project). Do the data provide sufficient evidence to conclude that, on average, *Barnes and Noble (in-store) books are more expensive than Amazon.com books?*

Task: Perform a test regarding the mean difference in book price, μ_D, where the differences are computed as "*Barnes and Noble less Amazon*" (i.e. 'price at Barnes and Noble' minus 'price on Amazon').

Procedure: The appropriate inference procedure for this scenario is a paired t-test, and the specific parameter of interest is ____________________ . Think about why this is this a paired procedure.

Hypothesis Test:

1. **State the Hypotheses:** H_0: _____________________ and H_a: ______________________ ,

   where the parameter___________ represents:

   Determine Alpha: We were told the significance level was 5%.

   Remember: Your hypotheses and parameter definition
   should always be a statement about the **population(s)** under study.

2. **Checking the Assumptions**
   a. For this scenario, we need to assume that the sampled differences are a random sample. Since the data was collected at one time point, a time plot is not appropriate to make. However we would want to learn more about how the books were selected.

   b. We also need to assume that the population of differences is normally distributed. To check this assumption, we would make a __________ plot of the ____________________. (What you would hope to see?)

3. **Compute the test statistic and calculate the *p*-value**
   Test statistic
   a. Let's create a variable for the *differences* so we can get summary statistics needed to compute the test statistic. First, we need to compute a difference for each pair. We can do this in R fairly easily by going to **Data > Manage variables in active data set > Compute new variable**. Call the new variable **difference**, and enter the difference of the two variable names in the **Expression to compute** box. You can double click on a variable name to insert it into the expression box.

b. Now, we will compute those summary statistics. Go to **Statistics > Summaries > Numerical summaries** to generate these, making sure you select your difference variable, and fill out the table below. Be sure to check the standard error option under the **Statistics** tab.

Summary Statistics			
Mean diff ($\bar{d}$)	**Std. Dev (s_d)**	**Sample size (n)**	**Std. Error**

c. Generate the *t*-test output using **Statistics > Means > Paired T Test**. Select the two variables, and specify the correct direction for the alternative hypothesis. (This should match your answer in question 1.) Use the output to fill out the following output table for this test.

Paired T Results		
t	***df***	***p-value***

d. Our sample mean difference is __________ standard errors below the hypothesized mean difference of zero.

e. What is the distribution *of the test statistic* if the null hypothesis is true?

Note: This is not the same as the distribution *of the population that the data were drawn from*, and will be the model used to find the *p*-value.

f. Now, try conducting a one mean test on the difference variable we created earlier, using the same hypotheses from problem 1. This can be found by going to **Statistics > Means > One Mean T Test**. Are the results similar to those you found in part c? Why is this the case?

One Sample T Results for the Differences		
t	***df***	***p-value***

Visualize the *p*-value:

g. Draw a picture of the *p*-value, with labels for the distribution and x-axis.

4. Evaluate the *p*-value and Conclusion

Evaluate the *p*-value:
What is your decision at a 5% significance level? **Reject H_0** **Fail to Reject H_0**

Remember:	Reject H_0	⇔ Results statistically significant
	Fail to Reject H_0	⇔ Results not statistically significant

Conclusion:
What is your conclusion *in the context of the problem*?

Note: Conclusions should always include a reference to the population parameter of interest.
Conclusions should not be too strong; you can say that you have sufficient evidence,
but do NOT say that we have *proven* anything true or false.

Cool-Down: Setting Up Hypotheses and Writing Conclusions

The Major League Baseball Organization would like to assess whether or not players have the same batting averages during the nighttime games as during the daytime games using a 5% significance level. Eight players are selected at random and their nighttime and daytime batting averages are collected for a given period of time.

Let μ_d represent the population mean difference between the nighttime minus the daytime batting average. Write out the corresponding null and alternative hypothesis using the appropriate notation

Hypothesis Test: **H_0**:_________________________ **H_a**:_________________________

Why is this a paired t-test?

The *p*-value for this paired t-test was 0.028. Make a decision for this test and write a conclusion in context of the problem.

Lab 8: Comparing Two Means

Objective:

In this lab, you will learn an important statistical technique that will allow you to compare two populations with respect to their means by looking at μ_1-μ_2. The aim is to help you understand the ideas behind confidence intervals, tests of significance, and the statistical language involved in the comparison of two population means.

Application:

Jim believes that the mean GPA for all English majors is different from the mean GPA for all Math majors. He takes a random sample of students in the Math and English buildings and he ends up with 32 English majors and 35 Math majors in his sample. He has more Math majors, which leaves Jim with no method to pair an English major with Math major and no way to perform the paired t test.

Overview:

The two independent samples *t* procedures (sometimes called Student's *t* procedures) are used when you want to compare the means of two populations that are not related or matched in any way. The idea is that we use the two sample means to estimate the corresponding population means.

We may want to construct a confidence interval to estimate the difference between the two population means. Or, we may wish to test if the difference in the two population means equals a value specified in a null hypothesis; generally, the hypothesized test value is 0.

Going back to the application, we will assume that we have a random sample from each of the two populations. This means Jim will assume that he has a random sample of English majors and a random sample of Math majors. To check this assumption, he will create a time plot with the GPAs for the English majors and assess the stability as well as creating a separate time plot with the GPAs for the Math majors.

The second assumption needed to perform the two independent sample t-test is that the variable being measured has a normal model for each population, although possibly with different means. To assess this assumption, Jim will create two separate QQ plots; one of the GPAs for English majors and one of the GPAs for Math majors.

In addition, the two random samples are assumed independent of each other. However, there are two versions of independent samples *t* procedures: pooled and unpooled (also known as general). In a pooled *t* procedure, we assume equal population variances for the two populations of responses.

Whether or not we can assume the two populations have equal variances will dictate which of the two procedures we use. Details on determining which version to use are provided next.

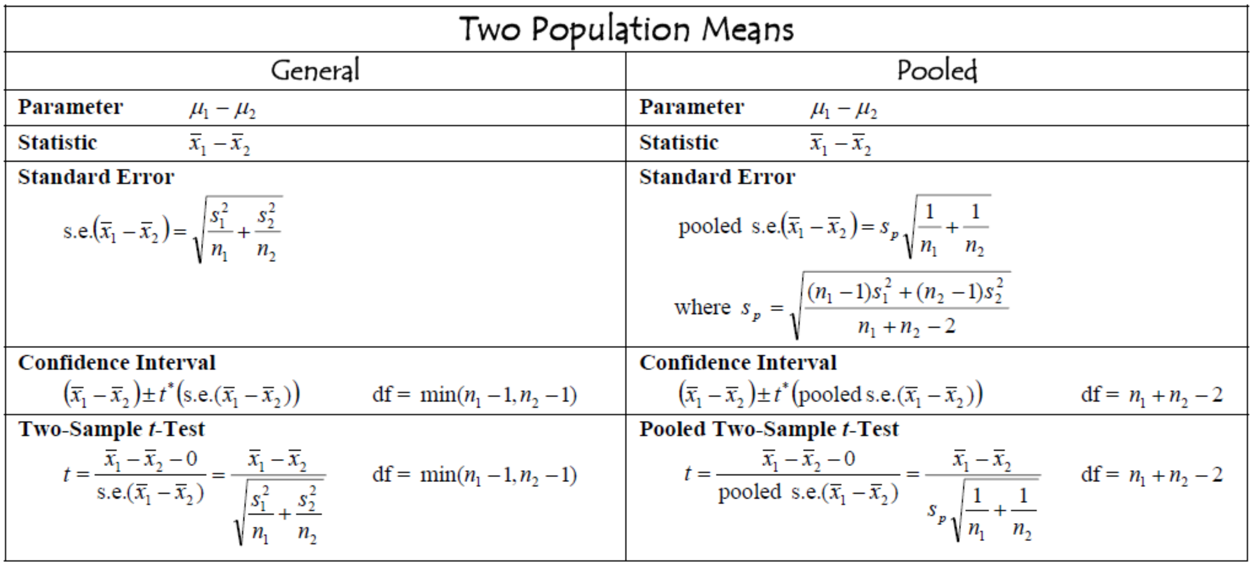
Checking the Equal Population Variance Assumption: There are several ways to check the equal population variances assumption required for a pooled *t*-test. Which one you will use will often depend on the information provided.

- **Side-by-side Boxplots:** Examine the IQRs of the sample data. If they are similar, then the assumption is valid. (NOTE: "Similar" in this sense does not mean that the boxes need to line up right next to each other. It means that the lengths or sizes of the boxes should be similar.) If one IQR is twice as large as the other, the assumption of equal population variances is not valid.
- **Sample Standard Deviations:** Since variance is standard deviation squared, if the sample standard deviations are similar, then the assumption is valid. If one standard deviation is twice as large as the other, the assumption of equal population variances doesn't hold.
- **Levene's Test for Homogeneity of Variances:** Levene's test is a hypothesis test, but it is a test about population variances rather than means. In Jim's case the null hypothesis is that the populations of English and Math majors do have equal variances (H_0: $\sigma_1^2 = \sigma_2^2$) and the alternative is that these populations have equal variances (H_a: $\sigma_1^2 \neq \sigma_2^2$). When the null hypothesis from the Levene's test is **NOT** rejected, it would be reasonable for Jim to perform the pooled *t*-test for the two population mean GPAs. If the null hypothesis from Levene's test **IS** rejected, then the assumption of equal population variances is unreasonable, and Jim should use the unpooled (general) version instead of the pooled. To make this determination, we will use an alpha of 10% for Levene's tests.

Common Population Standard Deviation: If Jim concludes that the populations of Math and English majors have the same variances, he is also stating that these populations have the same standard deviation. In order to run the pooled version of this test, he must estimate the common population standard deviation with the common sample standard deviation. This common standard deviation is written as s_p and Jim will calculate it by taking a weighted average of the standard deviation of English majors' GPAs and the standard deviation of Math majors' GPAs.

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

Formula Card



Two Population Means	
General	Pooled
Parameter $\mu_1 - \mu_2$	**Parameter** $\mu_1 - \mu_2$
Statistic $\bar{x}_1 - \bar{x}_2$	**Statistic** $\bar{x}_1 - \bar{x}_2$
Standard Error $\text{s.e.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	**Standard Error** $\text{pooled s.e.}(\bar{x}_1 - \bar{x}_2) = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ $\text{where } s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$
Confidence Interval $(\bar{x}_1 - \bar{x}_2) \pm t^*(\text{s.e.}(\bar{x}_1 - \bar{x}_2))$ $\text{df} = \min(n_1-1, n_2-1)$	**Confidence Interval** $(\bar{x}_1 - \bar{x}_2) \pm t^*(\text{pooled s.e.}(\bar{x}_1 - \bar{x}_2))$ $\text{df} = n_1+n_2-2$
Two-Sample *t*-Test $t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\text{s.e.}(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ $\text{df} = \min(n_1-1, n_2-1)$	**Pooled Two-Sample *t*-Test** $t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\text{pooled s.e.}(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $\text{df} = n_1+n_2-2$

Warm-Up: Check Your Understanding

A student researcher, Jackie, is interested in learning about how much time students spend studying per week. She is curious to find out if study time is higher for female students (group 1) versus male students (group 2). Data from a survey of 47 female and 45 male undergraduate students is collected. One question asked students to report the number of hours that they study in a typical week. Assume that all the needed conditions are met to perform a pooled two independent samples t-Test.

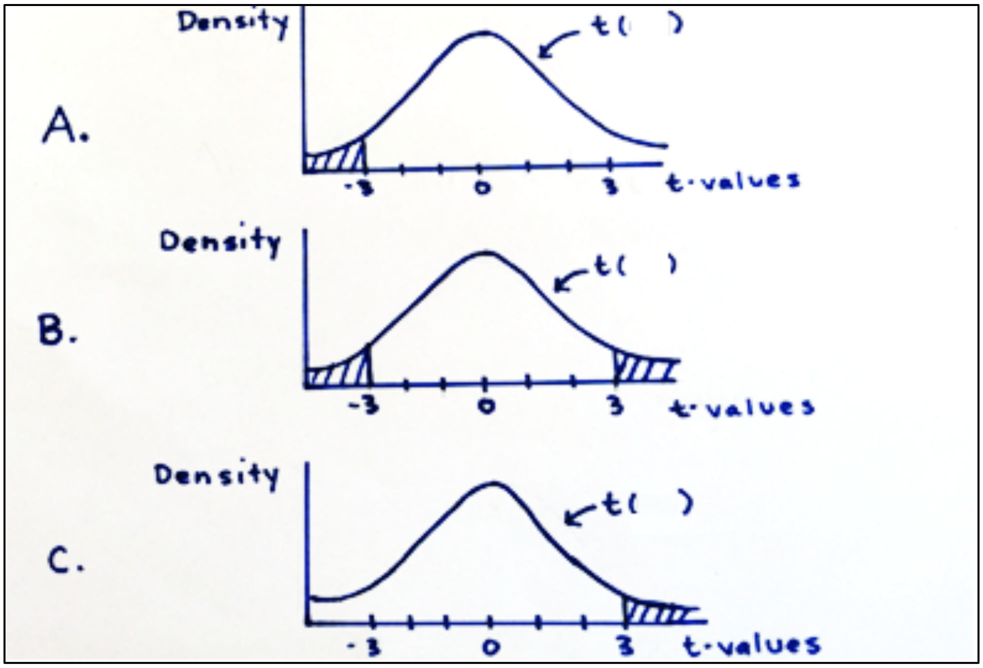
State the hypotheses to be tested and provide the test statistic value.

H_0: ______________________ versus H_a: ______________________

The test was run and the difference in the two sample means, $\bar{x}_1 - \bar{x}_2$ has been computed. This difference turned out to be 3 standard errors *above* the hypothesized difference for $\mu_1 - \mu_2$ of 0.

Test Statistic Value: ____________________

Which of the following is a correct graph of the p-value?



If the *p*-value for this test was 0.076,

are the results significant at a 5% level? **Yes** **No**

What are the appropriate degrees of freedom for the pooled t test? **91** **43** **90**

ILP: Do Men and Women Differ in Their SSHA Scores?

Background: A total of 38 college freshmen at a private college were administered the Survey of Study Habits and Attitudes (SSHA), a psychological test designed to measure motivation and attitude towards study habits in college students. The sampled students were a simple random sample consisting of 18 females and 20 males. Scores on the test range from a low of 0 to a high of 200 and it is known that they may explain collegiate success. School administrators are interested in whether or not there is a difference between the mean scores for males and females. The scores for females (Group 1) and males (Group 2) are listed in the **SSHA.Rdata** data set. (Source: Moore and McCabe (1999), pg 563).

Task: Perform a test to assess if there is a difference between the mean score for women on the SSHA and the mean score for men on the SSHA.

Procedure: Since the SSHA score is a quantitative response and we have two sets of scores to compare, the appropriate inference procedure for this scenario is the two independent samples t-test, and the specific parameter of interest is ___________. Think about why is this not a paired procedure.

Hypothesis Test

1. **State the Hypotheses:** H_0: ______________________ versus H_a: ________________________ ,

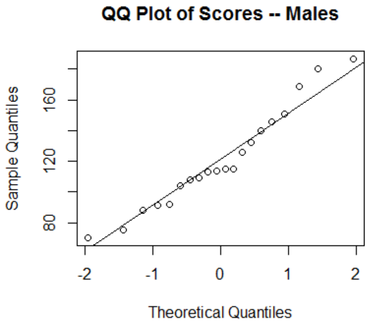
   where μ_1 represents____

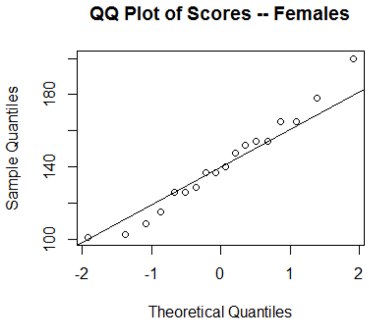
   and μ_2 represents____

   Note: In order to state the appropriate direction in the alternative hypothesis, it is important to know which population is being referred to as population 1 and which is population 2.
   Determine Alpha: We were told the significance level was 5%.

   Remember: Your hypotheses and parameter definition
   should always be a statement about the **population(s)** under study.

2. **Check Assumptions (Exploratory Data Analysis)**
   a. Checking Normality
      i. Based on the information provided in the background of the problem, we will assume that we have independent, random samples of SSHA scores. One remaining assumption that needs to be checked is that each sample comes from a normally distributed population. The QQ plots for each sample are provided next.





      ii. Does it appear that the assumption that each sample comes from a normally distributed population is met? Why?

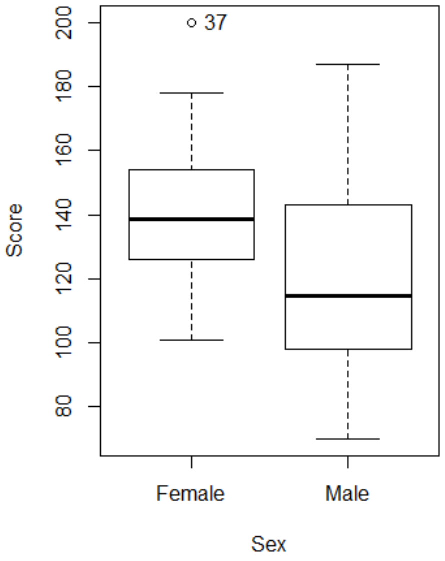
      iii. What remaining assumption needs to be checked before we can conduct the two independent samples t-test?

b. Now, we must determine if we can assume equal population variances or not. We will utilize all three methods discussed earlier to do this.

i. *Side-by-side Boxplots:*
The side-by-side boxplots show that the IQRs are:

similar **not similar**

ii. *Sample Standard Deviations:* Remember that we can generate summary statistics by using **Statistics > Summaries > Numerical Summaries,** and be sure to **summarize by groups** to get results for both sexes.



Summary Statistics			
Group	**Mean**	**Std. Dev**	**Sample Size**
Male			
Female			

These sample standard deviations are: **similar** **not simi**

iii. *Levene's Test:* Write the appropriate hypotheses for Levene's test using correct notation:

H_0: ___________________________ H_a: ________________________

We can perform Levine's test in RCommander using **Statistics > Variances > Levene's Test** (be sure to **use MEANS** and not MEDIANS).

The Levene's test statistic is F = _____________ , with a *p*-value of _______________.

Therefore, we **can** **cannot** reject the hypothesis that the population variances are equal.

Based on these results, we can say:
the assumption of equal population variances **is** **is not** valid.

Thus, the procedure that we will use for this test is the **pooled** **unpooled** procedure.

The symbol for the estimate of the common population standard deviation is ____.

Calculate its value:

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

Next, use the estimate s_p to compute the pooled standard error:

$$\text{pooled s.e.}(\bar{x}_1-\bar{x}_2) = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

3. Compute the Test-Statistic and Calculate the *p*-value

Test Statistic and P-value

a. Generate the *t*-test output using **Statistics > Means > Independent-Samples T Test**. Make sure to choose the appropriate direction for this test. Recall, we want to assess **if there is a difference** between the mean score for women on the SSHA and the mean score for men on the SSHA. Go to the **Options** tab, and you will find an **"Allow equal variances?"** setting. We have already looked at the data in various ways to make our decision of a **Yes** (pooled) setting. Using this output, fill out the table below.

Two Sample T Results					
	t	***df***	***p-value***	***95% CI lower***	***95% CI upper***
Pooled					

b. What is the distribution *of the test statistic* if the null hypothesis is true?

Note: This is not the same as the distribution *of the population that the data were drawn from*, and will be the model used to find the *p*-value.

Visualize the *p*-value:

c. Draw a picture of the *p*-value,
with labels for the distribution and x-axis.

d. Provide an interpretation of the *p*-value.

e. Think about it … how would you report the *p*-value for this test if you were trying to determine if males' mean SSHA score is larger than the females SSHA score on average? (Hint: You would need to be careful on the order of the subtraction!)

4. Evaluate the *p*-value and Conclusion

Evaluate the *p*-value:

What is your decision at a 5% significance level? **Reject H_0** **Fail to Reject H_0**

Remember: Reject H_0 ⇔ Results statistically significant
Fail to Reject H_0 ⇔ Results not statistically significant

Conclusion:

What is your conclusion *in the context of the problem*?

Note: Conclusions should always include a reference to the population parameter of interest.
Conclusions should not be too strong; you can say that you have sufficient evidence,
but do NOT say that we have <u>*proven*</u> anything true or false.

5. Confidence Interval (CI):

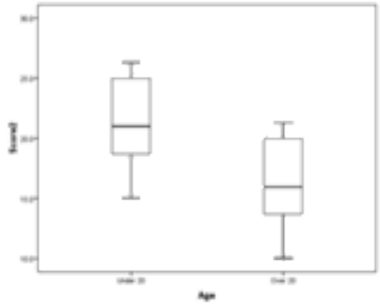
a. Provide the corresponding 95% confidence interval from the output for the difference in the two population mean scores.

a. Based on the confidence interval, would you reject the null hypothesis of no difference in population means at a 5% significance level?
Circle one: **Yes** **No**
Explain.

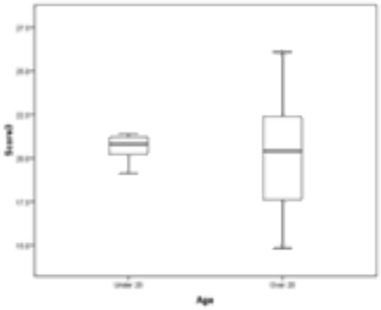
Did your conclusion here match the one you made in the previous part 4? **Yes** **No**

Cool-Down: Check Your Understanding

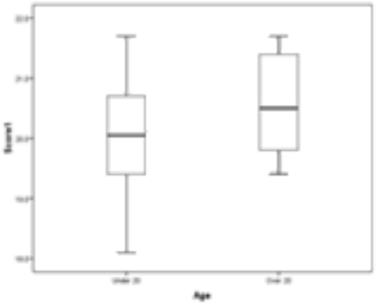
Consider the following sets of boxplots of scores between two age groups.



Set 1



Set 2



Set 3

a. Which set(s) that indicate a pooled test is appropriate: **Set 1** **Set 2** **Set 3**

b. For which set(s) are you most likely to reject the null hypothesis that the population mean scores are equal? **Set 1** **Set 2** **Set 3**

Lab 9: One-Way Analysis of Variance (ANOVA)

Objective: In this lab you will perform a **one-way analysis of variance**, often abbreviated **ANOVA**. We have already seen that the two independent samples *t*-test can be used to compare the means of two populations (when the samples are independent). However, when we want to compare the means of *three or more populations*, we turn to ANOVA. You can think of ANOVA as an extension of the two independent sample pooled *t*-test since it compares several population means and **requires** the assumption that the populations have equal variances.

Application: Andy is conducting a test for UM Hospital to explore the effects of a new antibiotic drug. They wish to explore if this drug has the same effects on the mean white blood cell count for the four different age populations***:*** 1 = 0 to 19 years, 2 = 20 to 29 years, 3 = 31 to 40 years, and 4 = 40 years and older. They need a test that would allow them to compare the mean white blood cell count for these four populations. In this example, the number of populations under study is *k*=4, and the total sample size is 24 (6 people in each age group).

Overview: ANOVA is a statistical tool for analyzing how the mean value of a quantitative response (or dependent) variable is affected by one or more categorical variables, known as treatment variables or factors. While ANOVA allows us to compare the means of more than two populations, it can only tell us whether differences appear to exist, not specifically which population means are different. Consequently, the appropriate hypotheses for ANOVA are H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ (that the population mean responses are equal, where k is the number of populations or treatment groups) and H_a: at least one of the population mean responses, μ_i, is different.

Going back to the application, for Andy's test we have:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_a: at least one of the population mean white blood cell counts, μ_i, is different.

As with our other hypothesis tests, several assumptions are required for ANOVA. Andy will need to assume that the white blood cell count measurements in each of the four age populations follows a normal model. To check this assumption, he can create four QQ plots, one for each sample of measurements.

He will also need to assume that the four populations of white blood cell counts have equal variance. The equal population variance assumption can be checked as done in the two independent samples t-test – side-by-side boxplots, comparing sample standard deviations, and Levene's test. Further, the data are assumed to consist of independent random samples.

The analysis of variance that Andy will run involves decomposing the total variation of the white blood cell count into two parts: (1) that due to variation among the four sample white blood cell counts (*between groups* variation) and (2) that due to the natural variation of the white blood cell count in each of the four age groups (variation due to error, or *within groups* variation): **SS Total = SS Groups + SS Error**.

If Andy finds that the sum of squares between groups (SS Groups) is large relative to the sum of squares within groups (SS Error), it implies that the model of different white blood cell means explains a significant portion of the observed variability.

In order to determine what is "relatively large," the sum of squares values are divided by their respective degrees of freedom, creating what are called mean square terms. The degrees of freedom for SS Groups is the number of treatment groups, *k*, minus one (*k* - 1).

$$\text{MS Groups} = \text{SS Groups}/(k - 1)$$

$$\text{MS Error} = \text{SS Error}/(N - k)$$

In Andy's test, the degrees of freedom for SS Groups is equal to 3. For SS Error, the degrees of freedom is the <u>total</u> number of observations, *N*, minus the number of treatment groups (*N* - *k*). For Andy's ANOVA procedure, the degrees of freedom for the SS Error is equal to 20.

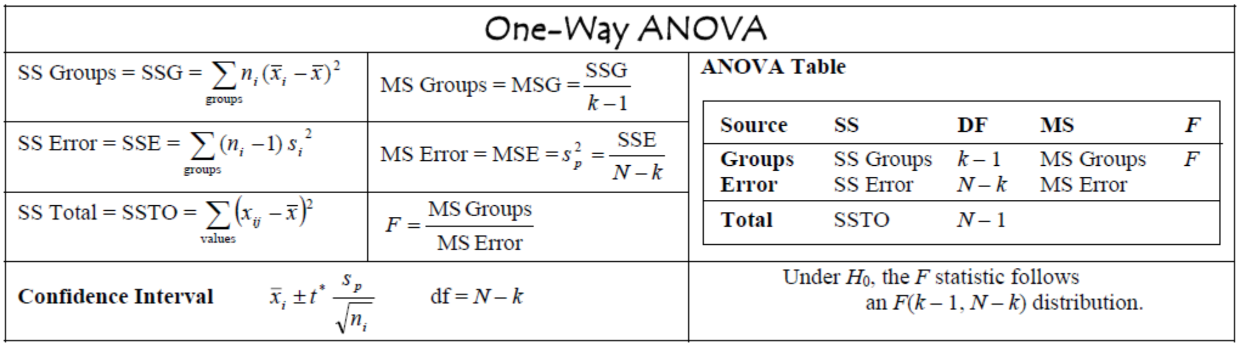
The ratio of these two mean squares forms the ***F*-statistic**, which has numerator degrees of freedom (*k* - 1) and denominator degrees of freedom (*N* - *k*). Note in the equation that MSE stands for MS Error.

$$F = \frac{\text{Variation among sample means}}{\text{Natural variation within groups}} = \frac{\text{MS Groups}}{\text{MSE}}$$

We can view this ***F*-statistic** as the ratio of two estimators of the common population variance, σ^2, where the denominator (MSE) is a good (unbiased) estimator, and the numerator (MS Groups) is only good when the H_0 is true (otherwise, it tends to overestimate σ^2). If Andy's data results in a large *F* value then there is some evidence against the null hypothesis of equal population white blood cell count means.

If Andy rejects the null hypothesis, indicating that at least one of the population mean white blood cell count is different, then we can turn to a **multiple comparisons procedure** for determining which population mean(s) appear to be different and how they differ. The most common method to analyze this is by looking at the set of all *pairwise comparisons*. Two equivalent techniques can be used by Andy for each pair of means: either perform a hypothesis test to see if two population mean white blood cell counts are significantly different, or construct a confidence interval for the difference in population mean white blood cell counts to see whether the value of 0 is contained in the interval. Specifically, a multiple comparisons procedure called **Tukey's procedure**, which is available in most computer packages, controls for the overall Type I error rate (overall significance level) or the overall confidence level.

Formula Card



One-Way ANOVA

$$\text{SS Groups} = \text{SSG} = \sum_{\text{groups}} n_i(\bar{x}_i - \bar{x})^2$$

$$\text{MS Groups} = \text{MSG} = \frac{\text{SSG}}{k-1}$$

$$\text{SS Error} = \text{SSE} = \sum_{\text{groups}} (n_i - 1)\, s_i^{\,2}$$

$$\text{MS Error} = \text{MSE} = s_p^2 = \frac{\text{SSE}}{N-k}$$

$$\text{SS Total} = \text{SSTO} = \sum_{\text{values}} \left(x_{ij} - \bar{x}\right)^2$$

$$F = \frac{\text{MS Groups}}{\text{MS Error}}$$

ANOVA Table

Source	**SS**	**DF**	**MS**	***F***
Groups	SS Groups	$k - 1$	MS Groups	*F*
Error	SS Error	$N - k$	MS Error	
Total	SSTO	$N - 1$		

Confidence Interval $\bar{x}_i \pm t^* \frac{s_p}{\sqrt{n_i}}$ $\text{df} = N - k$

Under H_0, the *F* statistic follows an $F(k - 1, N - k)$ distribution.

Warm-Up: Content of TV Shows

A study examined whether the content of TV shows influence the ability of viewers to recall brand names of items featured in the commercials. The researchers randomly assigned 90 adults to watch one of three programs (30 to each). One program had violent content, another sexual content, and the third neutral content. Each program contained the same nine commercials. After the shows ended, the subjects were asked to recall the brands of products that were advertised. The ANOVA table based on the data is provided below.

```
ANOVA
            Df Sum Sq Mean Sq F value Pr(>F)
Content      2   11.8    5.90    7.47  0.001
Residuals   87   68.6    0.79
```

a. The researcher would like to assess if there is any effect of the program type on the average number of brands recalled, namely, test H_0: $\mu_1 = \mu_2 = \mu_3$. Provide the appropriate alternative hypothesis in the context of the problem.

H_a: ____

b. One assumption is the model for number of brands recalled for each population is normal. What graph(s) would you would make to assess this assumption?

c. At the 5% level, was there a significant effect of the program type on the average number of brands recalled? Explain your answer.

ILP: Is There a Difference Among the Mean Freshman GPAs for Three Different Socioeconomic Classes?

Background: Sociologists often conduct experiments to investigate the relationship between socioeconomic status and college performance. Socioeconomic status is generally partitioned into three classes: lower, middle, and upper. Consider the problem of comparing the mean grade point average (GPA) of college freshmen across the three socioeconomic populations. The GPAs for random samples of seven college freshmen from each of the three socioeconomic classes were selected from a university's files at the end of the first academic year. The data are in the **gpa.Rdata** data set. (Source: Mendenhall and Sincich, 1996, page 589)

Task: Perform a test to assess whether the population mean freshman GPAs among the three socioeconomic classes differ. If there is sufficient evidence to indicate significant differences, determine which groups differ and how.

Procedure: We want to compare three populations with respect to a quantitative response (GPA). The appropriate inference procedure for this scenario is ANOVA and the value of *k* for this problem is _____.

Hypothesis Test:

1. **State the Hypotheses:** H_0: ________________________________

   versus H_a: ____ ,

   Clearly define the one of the parameters in the null hypothesis in context:

   μ_1 represents ____

   ____

   Determine Alpha: We were told the significance level was 5%.

   > **Remember:** Your hypotheses and parameter definition should always be a statement about the **population(s)** under study.

2. **Checking the Assumptions**

   We need to assume:

   - The *k* samples are _______________________ from each other.

   - Each sample is a random sample. To check this assumption, we would make a _________________ plot (if there was time order) for each sample.

   - Each sample needs to come from a normally distributed population. To check this assumption, we would make a ________ plot for each sample.

   - All k populations have equal ____ .

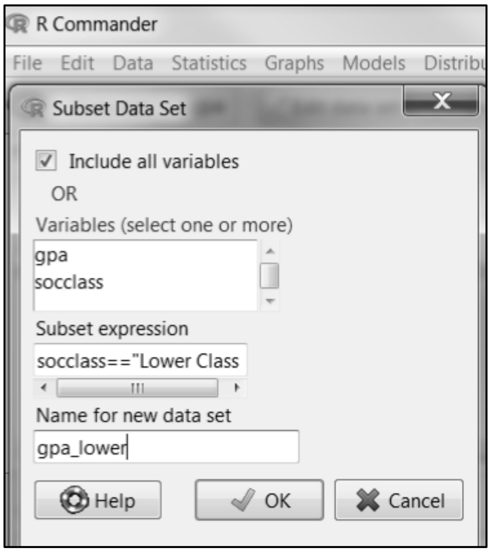
   a. Based on the description about how these samples were collected, can we assume we have random and independent samples?

Note that there is no time order for this data. ***If there were***, since you need EACH sample to be a random sample, how many time plots would you need to make to check this assumption?

Answer = Need to make ______ time plot(s)

b. Construct the QQ plots necessary to check the assumption about normally distributed populations. To do this, we have to subset the file for the three subgroups (or classes).

To do this, go to **Data -> Active data set -> Subset active data set**



Subset Data Set

☑ Include all variables

OR

Variables (select one or more)

gpa
socclass

Subset expression

socclass=="Lower Class

Name for new data set

gpa_lower

Enter **socclass=="Lower"** as the subset expression, give an appropriate new data set name **gpa_lower** and click OK. Note that this corresponds to the line command:

```
gpa_lower <- subset(gpa, subset=socclass=="Lower")
```

This should result in the message: NOTE: The dataset gpa_lower has 7 rows and 2 columns.

With this new data set as the active data set, **create the qqplot** by going to **Graphs > Quantile-comparison plot** and select the **gpa** variable.

Alternatively you could type in, then highlight, and submit these commands:

```
qqnorm(gpa_lower$gpa,main="Normal QQ Plot by yournamehere")
qqline(gpa_lower$gpa)
```

Based on your qqplot, does it appear that the population of GPAs for all students in the lower socioeconomic class is (approximately) normal? Why or why not?

Note: you would need to repeat the above steps for the “Middle Class” and “Upper Class” to complete the normal model for each population assumption checking. You will not need to make the other qqplots here, but when you do repeat these steps, you would need to be sure you go back and select the original gpa data set each time. The resulting three commands lines would look like:

```
gpa_lower <- subset(gpa, subset=socclass=="Lower")
```

```
gpa_middle <- subset(gpa, subset=socclass=="Middle")
gpa_upper <- subset(gpa, subset=socclass=="Upper")
```

3. Compute the Test-Statistic and Calculate the *p*-value

Test-Statistic

a. First, test to see if the variances are equal using Levene's test, using **Statistics > Variances > Levene's test…** making sure to select "mean" as the measure of center, not "median."
If the p-value > 0.10, then we fail to reject the null hypothesis that the variances are equal.

The assumption of equal population variances seems: **valid** **not valid**

b. Next, generate the ANOVA output using **Statistics > Means > One-way ANOVA**. (Make sure you are using the full **gpa** dataset, not one of the subsets)

The symbol for the estimate of the common population standard deviation is ______,

which for this problem is found to be __________________________________ .

c. What is the value of the test statistic? ____ = ____________

d. What is the distribution *of the test statistic* if the null hypothesis is true?

Note: This is not the same as the distribution *of the population that the data were drawn from*, and will be the model used to find the *p*-value.

Calculate the *p*-Value:

e. What is the reported *p*-value? ____________________

f. Draw a picture of the *p*-value, with labels for the distribution and x-axis.
Use the *pval()* function in R to check your work.

4. Evaluate the *p*-value and Conclusion

Evaluate the *p*-value:

What is your decision at a 5% significance level? **Reject H_0** **Fail to Reject H_0**

Remember: Reject H_0 ⇔ Results statistically significant
Fail to Reject H_0 ⇔ Results not statistically significant

Conclusion:

What is your conclusion *in the context of the problem*?

Note: Conclusions should always include a reference to the population parameter(s) of interest.
They should not be too strong; you can say that you have sufficient evidence,
but do NOT say that we have *proven* anything true or false.

5. **Follow-up Analyses:** If ANOVA has indicated that there appears to be significant differences between two or more of groups, we can use a multiple comparison test to tell us which groups appear to be different and by how much.

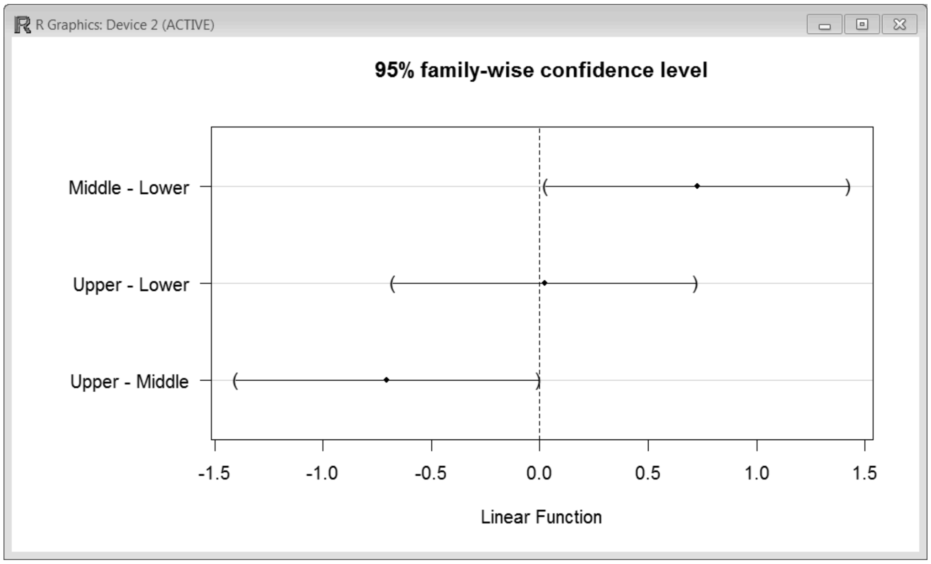
   a. Obtain the multiple comparisons output using using **Statistics > Means > One-way ANOVA**. This time, click on the "Pairwise comparison of mean" box, which has a default significance level of 0.05. The multiple comparisons output contains both *p*-values and confidence intervals for every possible pairwise comparison of groups; either can be used to determine where differences exist. The *p*-values that are less than or equal to 0.05 or confidence intervals that do NOT contain 0 indicate a difference between those two population means.

```
95% family-wise confidence level

Linear Hypotheses:
                    Estimate  lwr       upr
Middle - Lower == 0  0.727143  0.029443  1.424843
Upper - Lower == 0   0.021429 -0.676271  0.719128
Upper - Middle == 0 -0.705714 -1.403414 -0.008014
```

The "estimate" is the estimated difference of means, and the "lwr" and "upr" are the lower and upper bounds of the confidence intervals. If these CI's contain zero, we fail to reject the null hypothesis that the two means are equal.

You also can see this chart, which illustrates the same information:



   b. Summarize the findings about the differences in population means for the GPAs of freshmen in the different socioeconomic classes. Which pairs are significantly different at the 5% level?

c. Calculate a 95% confidence interval for the mean GPA for the middle class group. The sample mean GPA for the 7 subjects in the group was 3.25.

$$\textbf{Confidence Interval} \qquad \bar{x}_i \pm t^* \frac{s_p}{\sqrt{n_i}} \qquad \text{df} = N - k$$

Cool-Down: Check Your Understanding

Complete the following sentences by circling words or filling in blanks as necessary.

- The *p*-value of 0.025 from this activity implies that if this study were repeated many times,

  we would see an *F* test statistic of 4.579 or **greater** **less**

  in about _______ % of repetitions if the population means were really all equal.

- ANOVA procedures can be thought of as

  an extension of the two independent samples **pooled** **unpooled** *t*-test,

  and hence requires the assumption of equal **population** **sample** variances.

- One way to check this assumption is to use Levene's test and see if

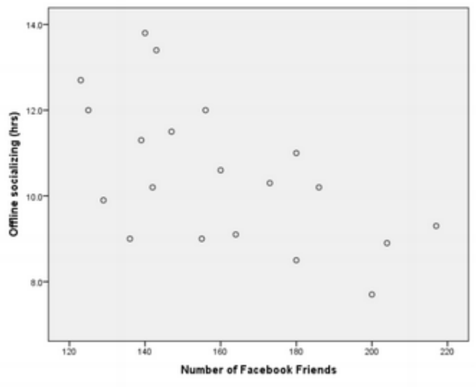
  the *p*-value is **greater than** **less than or equal to** 0.10.

Lab 10: Exploring Linear Regression

Objective: In this lab, you will examine relationships between two quantitative variables using a graphical tool called a scatterplot. You will interpret scatterplots in terms of form, direction, and strength of the relationship, and use it to assess the appropriateness of using a linear regression model to describe the relationship between the two variables. If appropriate, you can then perform a linear regression analysis to produce an estimated model that can be used to predict the value of the response *y* for a given value of the predictor *x*.

Application: Pam believes that the number of Facebook friends a person has could interfere with the number of hours they spends socializing offline. She wants to be able to predict the number of hours a person spends socializing offline for a given number of Facebook friends. She will collect data from her friends and examine the relationship between these two variables. Pam will be able to fit a linear regression model if the relationship between these variables is linear.

Overview: A regression model describes how the mean of one variable is thought to depend on the value of one or more other variables. If we think the number of Facebook friends may explain changes in the amount of time spent socializing online, we call the number of Facebook friends an **explanatory** variable (or **predictor** variable or **independent** variable) and the amount of time spent socializing offline is called the **response** variable (or **dependent** variable).



To start, we use a scatterplot to display the relationship between two quantitative variables, plotting the number of Facebook friends on the *x*-axis and the amount of time spend socializing offline on the *y*-axis. In examining the relationship, Pam looks at the overall pattern showing the **form**, which appears to be linear. She notes that the **direction** of the form is negative and **strength** of the relationship is moderate. She notes that there are no apparent **outliers**.

One of the many misconceptions about regression arises from the concept of association. Scatterplots can show the association between variables, but Pam should remember that *correlation does not imply causation*. For example: weekly flu medication sales and weekly sweater sales for an area with extreme seasons would exhibit a positive association because both tend to go up in winter and down in summer. However, neither causes the other. The observed association between two variables is sometimes due to other factors, such as confounding variables.

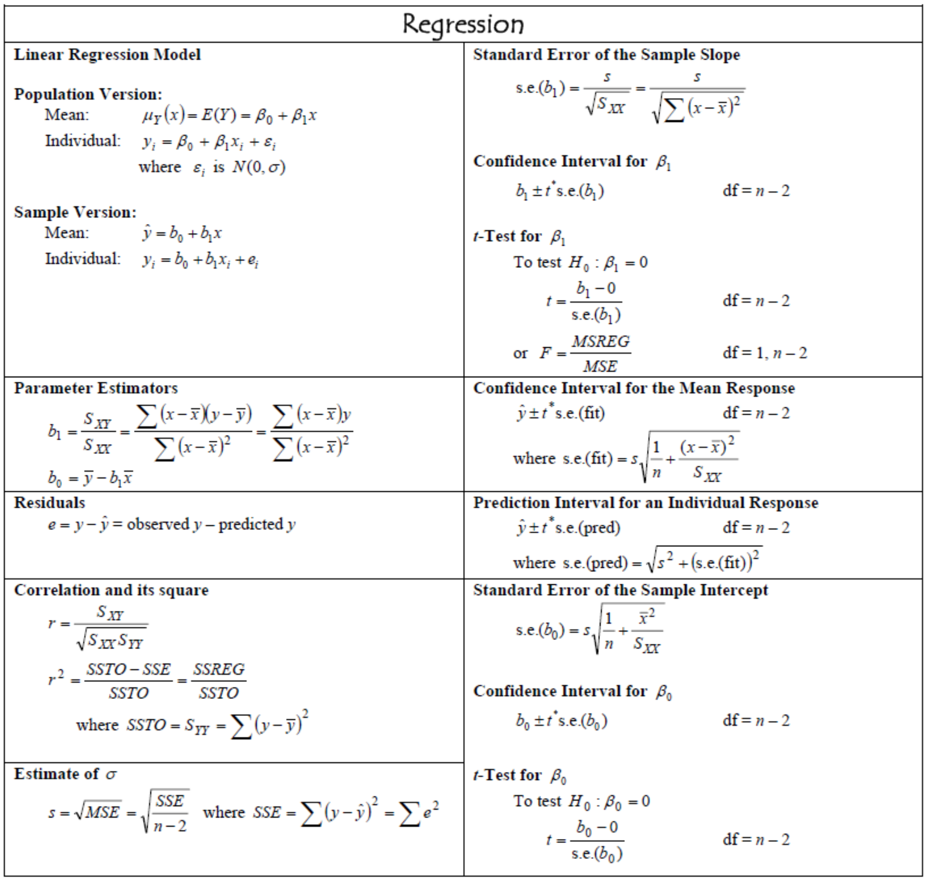
This correlation value, **r**, explains the strength of the linear relationship between x and y. The correlation can take on values between -1 and 1. The sign of the correlation also describes the direction of the linear relationship. The correlation between number of Facebook friends and time spent socializing offline is reported to be -0.59, which confirms what Pam observed in her scatterplot. The square of this correlation is known as the coefficient of determination and also has an important interpretation. Pam calculates this coefficient to be 0.348 and she can interpret this value by stating that 34.8% of the variation in the amount of time spent socializing online can be explained by the linear relationship between number of Facebook friends and time spent socializing online.

Since Pam's scatterplot suggested that the dependence of amount of time spent socializing online on the number of Facebook friends can be summarized by a straight line, the **least squares regression line** can be calculated. The least squares regression line is the line that minimizes the sum of the squared vertical distances of the data points to the line – hence the name **least squares**. This fitted line can be used to describe the linear relationship between the amount of time spent socializing offline and the number of Facebook friends and to predict the amount of time spent socializing online for a given number of Facebook friends.

The distances from the observed amount of time spend socializing offline to the predicted amount of time spent socializing offline are known as the **residuals**. These residuals are estimates of the true error terms associated with Pam's model.

Pam has fit a **linear regression model,** which is what will also be fit during the lab. Pam's regression is based on a linear model relating the number of Facebook friends to the mean amount of time spend socializing offline as follows: E(*time spent socializing offline*) = β_0+ β_1(*number of Facebook friends)*. Here β_0 and β_1 are parameters – fixed but unknown constants. Specifically, β_0 is the population *y*-intercept (the amount of time spent socializing offline when the number of Facebook friends is zero) and β_1 is the population slope (the change in the mean time spent socializing offline for every additional Facebook friend). These two values are unknown, but can be estimated using the least squares criterion. The resulting estimated regression line is generally written as: $\hat{y} = b_0 + b_1(x)$. The estimates, b_0 and b_1 are referred to as the **least squares estimates** of β_0 and β_1.

Formula Card:



Regression

Linear Regression Model

Population Version:

Mean: $\mu_Y(x) = E(Y) = \beta_0 + \beta_1 x$

Individual: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where ε_i is $N(0, \sigma)$

Sample Version:

Mean: $\hat{y} = b_0 + b_1 x$

Individual: $y_i = b_0 + b_1 x_i + e_i$

Standard Error of the Sample Slope

$$\text{s.e.}(b_1) = \frac{s}{\sqrt{S_{XX}}} = \frac{s}{\sqrt{\sum(x-\bar{x})^2}}$$

Confidence Interval for β_1

$b_1 \pm t^* \text{s.e.}(b_1)$ df $= n-2$

***t*-Test for β_1**

To test $H_0 : \beta_1 = 0$

$t = \dfrac{b_1 - 0}{\text{s.e.}(b_1)}$ df $= n-2$

or $F = \dfrac{MSREG}{MSE}$ df $= 1, n-2$

Parameter Estimators

$$b_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum(x-\bar{x})y}{\sum(x-\bar{x})^2}$$

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

Confidence Interval for the Mean Response

$\hat{y} \pm t^* \text{s.e.(fit)}$ df $= n-2$

where $\text{s.e.(fit)} = s\sqrt{\dfrac{1}{n} + \dfrac{(x-\bar{x})^2}{S_{XX}}}$

Residuals

$e = y - \hat{y} =$ observed y – predicted y

Prediction Interval for an Individual Response

$\hat{y} \pm t^* \text{s.e.(pred)}$ df $= n-2$

where $\text{s.e.(pred)} = \sqrt{s^2 + (\text{s.e.(fit)})^2}$

Correlation and its square

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSREG}{SSTO}$$

where $SSTO = S_{YY} = \sum(y-\bar{y})^2$

Standard Error of the Sample Intercept

$$\text{s.e.}(b_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}$$

Confidence Interval for β_0

$b_0 \pm t^* \text{s.e.}(b_0)$ df $= n-2$

Estimate of σ

$s = \sqrt{MSE} = \sqrt{\dfrac{SSE}{n-2}}$ where $SSE = \sum(y-\hat{y})^2 = \sum e^2$

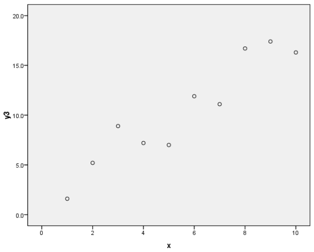
***t*-Test for β_0**

To test $H_0 : \beta_0 = 0$

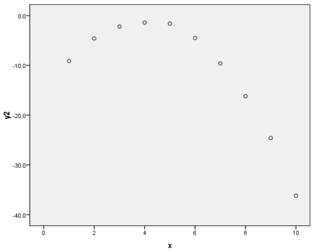
$t = \dfrac{b_0 - 0}{\text{s.e.}(b_0)}$ df $= n-2$

Warm-Up: Check Your Understanding

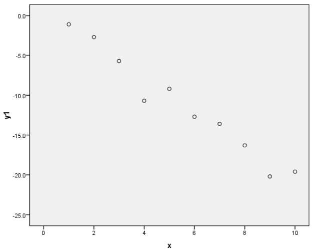
For which scatterplot(s) does a linear regression analysis seem appropriate?



Plot 1



Plot 2



Plot 3

ILP: Using a Scatterplot and Correlation to Analyze the Relationship

Background: As we approach the final exam, it may be of interest to you to know if there is a relationship between your existing exam and homework scores and your final exam score. To analyze this, we have a sample of exam and homework scores from Stats 250 students in a previous term, found in the **Stats250Data.Rdata** file.

***Task*:** First, we would like to determine what measure predicts Final Exam scores best – Homework, Exam 1, or Exam 2? To do this, we will examine scatterplots and correlations to ensure that a linear relationship is appropriate, and determine which model is strongest.

Procedure: We will have **two variables** being measured – Final Exam scores (which is **quantitative** and playing the role of the **response**) and your chosen predictor variable (which is also **quantitative** and plays the role of the **explanatory** variable). The goal is to assess the **relationship** between these two quantitative variables. The appropriate inference procedure for this scenario is **linear regression**.

1. Open the data set and produce a scatterplots for each of the potential predictor variables (AvgHW, Exam1, Exam2) against the Final Exam score, using **Graphs > Scatterplot**. Under **Options**, you will want to uncheck **Show Spread** and **Smooth Line**. Write a few sentences to describe each scatterplot in terms of:
   A. form (does there appear to be a linear relationship?)
   B. direction (positive or negative)?
   C. strength (generally weak, moderate, or strong?)
   D. outliers (any unusual observations or outliers present?)

   Interpret the Scatterplot for Final Exam versus AvgHW

Interpret the Scatterplot for Final Exam versus Exam 1

Interpret the Scatterplot for Final Exam versus Exam 2

2. Let's now assess the strength of the relationship via the correlation, or R (not the software, but the value!). To get R values for each pair of variables, select **Statistics > Summaries > Correlation Matrix**. This will display an array of values, where each value is the correlation of the variables given in the row and column headers. Report the following correlation values:

AvgHW and Final Exam:_______ Exam 1 and Final Exam:________ Exam 2 and Final Exam:________

3. Based on the scatterplots and correlation values, which variable will you use to predict Final Exam scores? Explain.

   AvgHW Exam 1 Exam 2

ILP: Describing a Linear Relationship with a Regression Line

Now that we have chosen an appropriate linear model to predict final exam score, we are ready to estimate that model and use it to predict Final Exam scores.

Task**:** Fit a linear model to the data. If you have questions about the regression output after the activity, refer to Supplement 6 in this workbook for more details.

1. Obtain the linear regression output first by creating a model using **Statistics > Fit Models > Linear Regression**, making sure to enter the appropriate response and explanatory variables. Give your model an appropriate name, and click OK. A summary of the model will appear in the output.
   Report the estimated regression line (the predicting equation or least squares regression line):

   ____

2. Interpret the estimated slope b_1 in terms of the change in your explanatory variable.

3. Report the coefficient of determination, r^2, and interpret it: r^2 = _____________
   Interpretation:

Variable	Value
Exam 1	58
Exam 2	50
Avg HW	25.6

4. Use your regression line to predict the Final Exam score for a student with score given in the table for your predicting (explanatory) variable (include units).

5. The student above actually had a final exam score of 92 out of 100 points. What is this student's residual (include units)?

Cool-Down: Check Your Understanding

Could you use this regression line to predict your final exam score in this class?

What are some potential issues with the prediction this line gives you?

Lab 11: Regression Inference

Objective: Last lab, we looked at a linear regression model for Stats 250 scores, and started to examine if it was an appropriate to use a linear model. In this lab, you will learn how to perform hypothesis tests and compute confidence intervals in regression and assess if the linear relationship is significant, as well as learn how to check the assumptions needed for these inference procedures to be valid.

Application: Recall our example from last lab -- Pam believes that the number of Facebook friends a person has could interfere with the number of hours they spends socializing offline. She wants to be able to predict the number of hours a person spends socializing offline for a given number of Facebook friends. She will collect data from her friends and examine the relationship between these two variables. Pam will be able to fit a simple linear regression model if the relationship between these variables is linear.

Overview: To do this, Pam has fit a **linear regression model,** which is what will also be fit during the lab. Pam's regression is based on a linear model relating the number of Facebook friends to the mean amount of time spend socializing offline as follows: E(*time spent socializing offline*) = $\beta_0 + \beta_1$(*number of Facebook friends)*. Here β_0 and β_1 are parameters – fixed but unknown constants. Specifically, β_0 is the population *y*-intercept (the amount of time spent socializing offline when the number of Facebook friends is zero) and β_1 is the population slope (the change in the mean time spent socializing offline for every additional Facebook friend). These two values are unknown, but can be estimated using the least squares criterion. The resulting estimated regression line is generally written as $\hat{y} = b_0 + b_1(x)$. The estimates, b_0 and b_1 are referred to as the **least squares estimates** of β_0 and β_1.

There are also several assumptions that Pam must check in order for inferences to be valid. First, the time spent socializing offline must be normal with a mean that varies linearly with the number of Facebook friends and a standard deviation that does not depend on the predicted socializing time. To check her assumption Pam would create a residuals vs fitted plot: a plot with residuals on the y-axis and the predicted social time on the x-axis. For this assumption to hold, the residuals should be randomly scattered.

Pam must also assume that the error terms are normally distributed and are identically distributed. To check this assumption about the normal distribution she will create a QQ plot of the residuals. To check if the errors are identically distributed, she will create a time plot of the residuals and look for stability.

Warm-Up: Check Your Understanding

A linear regression model was built to assess the relationship between the weight (pounds) and the height (inches) of men aged 18 to 29. The estimate of the least squares regression line is given to be:

Predicted weight = -250 + 6*(Height)

a. What is the predicted weight for a male with a height of 70 inches?

250 ***170*** ***420***

b. We know that the average height of a male aged 18 to 29 is 72 inches. If we were to construct a 95% confidence interval for the weight of a male that is 72 inches tall, how would the width of that interval compare to a 95% confidence interval for the weight of a male that is 70 inches tall?

NARROWER ***WIDER*** ***STAYS THE SAME***

ILP: Is There a Significant Linear Relationship Between Final Exam Score and Our Selected Predictor?

Recall our previous lab on the Stats 250 data. Is the explanatory variable that we chose a useful linear predictor for the Final Exam score? That is, do we observe a significant, non-zero linear relationship between the Final Exam Score and our selected predictor variable?

Remember that another way to make inferences about the significance of the linear relationship is through a confidence interval for the population slope. Further, recall the basic form of a confidence interval: point estimate ± (a few) standard errors.

Most standard computer regression output provides the slope estimate and its standard error, and the "few" will correspond to a t^* value for the corresponding confidence level with degrees of freedom for regression of $n - 2$. Since a confidence interval provides a range of reasonable values for the parameter, it can be used to perform two-sided hypothesis tests by seeing whether the hypothesized value falls in the interval or not.

Task: Assess if the chosen explanatory variable is significant in the linear model.

> **NOTE:** You may have to re-run some of the output that we ran in our last class, specifically creating the regression model. Refer to the last lab for instructions on how to do this.

Hypothesis Test:

1. State the Hypotheses: H_0: _______________ and H_a: ___________________ ,

where __________ represents:

Determine Alpha: We were told the significance level was 5%.

> **Remember:** Your hypotheses and parameter definition should always be a statement about the **population(s)** under study.

2. Checking the Assumptions - **Covered in the next activity.**

3. Compute the Test-Statistic and calculate the *p*-value:
Test-Statistic

a. Using the computer output generated by the regression model, which two test statistics could you use to test these hypotheses? Give the value for each test statistic.

t = ______________________ F = _____________________

Check it out: What happens if you square your t statistic value? Think about the shape of the t distribution and if you squared all the t values, it would look like an F distribution.

Calculate the *p*-Value:

b. The *p*-value for both test statistics is the same. What is that *p*-value?___________________

Note: this *p*-value is for testing the two-sided alternative H_a: $\beta_1 \neq 0$. If we want to conduct a one-sided version of the alternative hypothesis we could only use the t test statistic and would need to work with the two-sided *p*-value in the output to find the appropriate one-sided *p*-value. Drawing a quick sketch of what that one-sided *p*-value looks like can help.

c. We could also get the F-test statistic value from running an ANOVA table to summarize the variance in our regression model. To do this, go to **Models > Hypothesis Tests > ANOVA**. To ensure that we see the table in a familiar format, select **Sequential "Type I"** as our option and click OK. Use this ANOVA table to calculate the r^2 value, and verify that this is the same as last lab. Remember that SSTotal = SSRegressionModel + SSError(or residuals).

$$r^2 = \frac{SSRegression\ model}{SSTotal} =$$

4. Evaluate the *p*-value and Conclusion

Evaluate the *p*-value:

What is your decision at a 5% significance level? **Reject H_0** **Fail to Reject H_0**

Remember: Reject H_0	⇔	Results statistically significant
Fail to Reject H_0	⇔	Results not statistically significant

Conclusion:

What is your conclusion *in the context of the problem*?

Note: Conclusions should always include a reference to the population parameter(s) of interest.
They should not be too strong; you can say that you have sufficient evidence,
but do NOT say that we have proven anything true or false.

5. Confidence Intervals (CI):

a. Generate the confidence interval for estimating the population slope using **Models > Confidence Intervals**. Give the 95% confidence interval for the population slope.

b. Provide an interpretation of the resulting interval in context.

c. Based on the confidence interval, would you reject the null hypothesis at a 5% significance level? Circle one: **Yes** **No**
Explain.
Did your conclusion here match the one you made in part 4?

ILP: Is the Linear Model Appropriate? Are Assumptions Met for Inference?

You will produce and examine the residuals from the regression line as well as create some plots to assess the fit of the linear model. This will also serve to evaluate the validity of the testing and confidence intervals performed in the earlier activities.

Regression assumptions may be stated in terms of the response variable or in terms of the error terms. The statistical model for simple linear regression assumes that for each value of *x*, the observed values of the response are normally distributed with some mean (that may depend on *x* in a linear way) and a standard deviation σ that does not depend on *x*. **For each *x*, *Y is* N(E(*Y*), σ), where E(*Y*) = $\beta_0 + \beta_1$x**.

Thinking about the error terms, we can say the **true error terms** (those that we do not observe) are the difference of the response and the true mean (for a given *x*). These errors are to have a normal distribution with a mean of 0 and a standard deviation of σ (that does not depend on *x*).

***Task*:** Create diagnostic plots for regression and check that assumptions are met.
To create diagnostic plots for regression, go to **Models > Graphs > Basic diagnostic plots**.

1. The graph generated in the upper left shows a scatterplot of residuals on the y-axis against fitted y values on the x-axis. This is called a residuals vs fitted plot. Sketch the general pattern of the plot.

   What assumption of the error terms does this plot help assess?

   What conclusion can you draw from this plot?

2. The graph generated in the upper right shows a QQ plot of residuals.
   Sketch the general pattern of the plot.

   What assumption of the error terms does this plot help assess?

   Based on the plot what is your conclusion about this assumption?

Cool Down: Check Your Understanding

1. Which is the only test statistic you can use if you want to test if the Final Exam score is a significant positive linear predictor of your chosen explanatory variable?

2. Some other Stats 250 students were interested in predicting their final exam score, so you calculated some prediction intervals for your curious friends. Some of your friends Exam 2 scores are given below. Which score will have the narrowest prediction interval? (Hint: You may need to run some summary statistics to do this)

   52 ***55*** ***58*** ***61*** ***64*** ***67***

Lab 12: Chi-Square Tests

Objective: In this lab, you will learn how to perform three Chi-square tests (the test of goodness of fit, the test of independence, and the test of homogeneity) that are used to analyze categorical responses.

Overview: You will see three **Chi-Square tests**: the tests of **goodness of fit**, **independence**, and **homogeneity**. For all three tests, the data are generally presented in the form of a **contingency table** (a rectangular array of numbers in cells). All three tests are based on the **Chi-Square statistic:**

$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$, where O_i is the observed count and

E_i is the expected count under the corresponding null hypothesis.

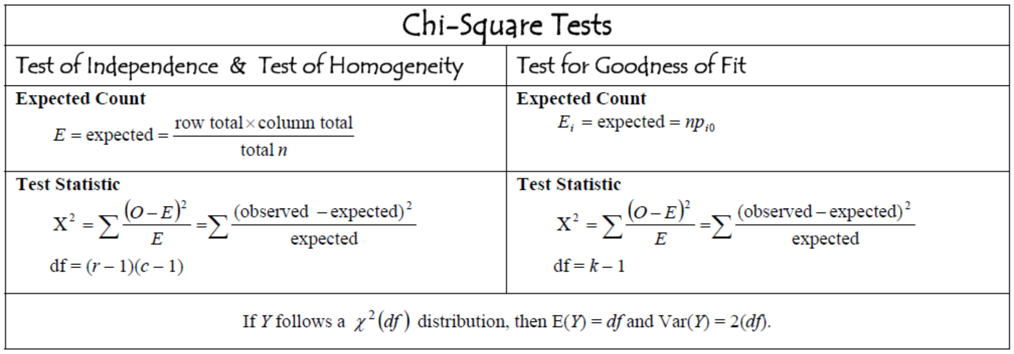
Goodness of Fit Test: This test answers the question, "Do the data fit well compared to a specified distribution?" It considers one categorical response, and assesses whether the proportion of sampled observations falling into each category matches well to a specified distribution. The null hypothesis specifies this distribution which describes the population proportion of observations in each category.

Test of Homogeneity: This test answers the question, "Do two or more populations have the same distribution for one categorical variable?" It considers one categorical response, and assesses whether the model for this response is the same in two (or more) populations. The null hypothesis is that the distribution of the categorical variable is the same for the two (or more) populations.

Test of Independence: This test answers the question, "Are two factors (or variables) independent for a population under study?" It considers two categorical variables (sometimes one is a response and the other is explanatory), and assesses whether there appears to be a relationship between these two variables for a single population. The null hypothesis is that the two categorical variables are independent (not related) for the population of interest.

There are a few properties of the Chi-square distribution that you might find useful. The expected value of a Chi-square distribution is its degrees of freedom (mean = $\mu = df$), and its variance is 2 times its degrees of freedom. Thus, its standard deviation is the square root of 2 times the degrees of freedom ($\sigma^2 = 2*df$ so $\sigma = \sqrt{2*df}$). This frame of reference can help assess if our observed statistic is unusual under the null hypothesis or somewhat consistent with the null hypothesis.

Formula Card



Chi-Square Tests	
Test of Independence & Test of Homogeneity	Test for Goodness of Fit
Expected Count $E = \text{expected} = \frac{\text{row total} \times \text{column total}}{\text{total } n}$	**Expected Count** $E_i = \text{expected} = np_{i0}$
Test Statistic $X^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ $\text{df} = (r-1)(c-1)$	**Test Statistic** $X^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ $\text{df} = k-1$
If Y follows a $\chi^2(df)$ distribution, then $\text{E}(Y) = df$ and $\text{Var}(Y) = 2(df)$.	

Warm-Up: Check Your Understanding

Fill in the blank with the name of the most appropriate Chi-square test to address the research question.

1. A researcher wants to determine if scoring high or low on an artistic ability test depends on being right or left-handed.

   Answer: Chi-Square test of ______ _______________ ________

2. A national organization wants to compare the distribution of level of highest education completed (high school, college, masters, doctoral) for Republicans versus Democrats.

   Answer: Chi-Square test of ______ _______________ ________

3. A preservation society has the percentages of five main types of fish in the river from 10 years ago. After noticing an imbalance recently, they add some fish from hatcheries to the river. How can they determine if they restored the ecosystem from a new sample of fish?

   Answer: Chi-Square test of ______ _______________ ________

ILP: Is There a Different Pattern in the Distribution of Accidental Deaths in a Certain Region Compared to the Pattern in the Entire United States?

Background: According to the records of the National Safety Council, accidental deaths in the United States during 2002 had the following distribution according to the principal types of accidents.

Motor Vehicle	Falls	Drowning	Fire	Poison	Other
45%	15%	4%	3%	16%	17%

Suppose that an accidental death data set from a particular geographical region yielded the following frequency distribution for the principal types of accidents:

Motor Vehicle	Falls	Drowning	Fire	Poison	Other
442	161	42	33	162	150

Do these data show a significantly different pattern in the distribution of accidental deaths in the particular region compared to the pattern in the entire United States? Use a 5% significance level. (Source: National Safety Council Website, 2005)

Task: Perform a Chi-square goodness of fit test to assess whether the data fit well with the model specified in the null hypothesis.

1. **State the null hypothesis:** H_0: ____

   Determine Alpha: We were told to use a significance level of 5%.

Remember: Your hypotheses and parameter definition should always be a statement about the **population(s)** under study.

2. **Checking the Assumptions**

   a. Find the expected counts and fill them in the table below.

	MotorVehicle	Falls	Drowning	Fire	Poison	Other	Total
Null %	45%	15%	4%	3%	16%	17%	100%
Observed	442	161	42	33	162	150	990
Expected							

   b. Do all cells have expected counts greater than 5? **Yes** **No**

3. **Compute the Test Statistic and Calculate the *p*-value**

   Test-Statistic

   a. Complete the calculation of the test statistic based on your table above by calculating the contribution of "other" accidental deaths to the total test statistic.

$$X^2 = \frac{(442-445.5)^2}{445.5} + \frac{(161-148.5)^2}{148.5} + \frac{(42-39.6)^2}{39.6}$$
$$+\frac{(33-29.7)^2}{29.7} + \frac{(162-158.4)^2}{158.4} + \frac{(..................)^2}{} = 3.663$$

   Calculate the *p*-Value:

   b. Based on the output below, answer the following:

```
        Chi-squared test for given probabilities

                  data:  .Table
 X-squared = 3.6635, df = 5, p-value = 0.598805
```

   i. The *p*-value is _____________ .

   ii. The expected value of the test statistic assuming H_0 is true is ____________

   iii. The large *p*-value we obtained is consistent with the fact that our observed test statistic value is

   greater than **less than** the expected test statistic value (under the null hypothesis).

4. **Evaluate the *p*-value and Conclusion:**

   Evaluate the *p*-value:

   What is your decision at a 5% significance level? **Reject H_0** **Fail to Reject H_0**

Remember: Reject H_0 ⇔ Results statistically significant
Fail to Reject H_0 ⇔ Results not statistically significant

   Conclusion:

   What is your conclusion *in the context of the problem*?

ILP: Comparison of the Distribution of Academic Degrees: Males Versus Females

Background: How do women and men compare in the pursuit of academic degrees? The table presents counts (in thousands) from the Statistical Abstract of degrees earned in 1996 categorized by the level of the degree and the sex of the recipient.

	Bachelor	Master	Professional	Doctorate
Female	642	227	32	18
Male	522	179	45	27

Task: Perform a Chi-square test of homogeneity. Use a 1% significance level.

Hypothesis Test:

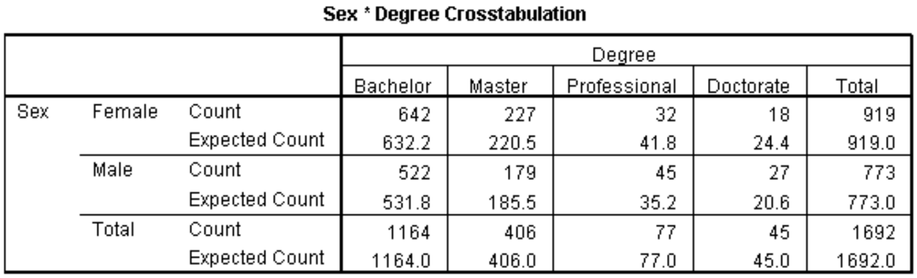
1. State the null hypothesis:

H_0: ____

____

Determine Alpha: We were told to use a significance level of 1%.

2. Checking the Assumptions



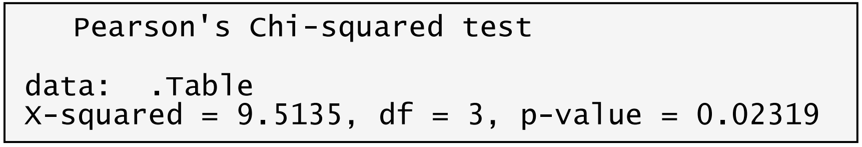
Sex * Degree Crosstabulation

			Degree				
			Bachelor	Master	Professional	Doctorate	Total
Sex	Female	Count	642	227	32	18	919
		Expected Count	632.2	220.5	41.8	24.4	919.0
	Male	Count	522	179	45	27	773
		Expected Count	531.8	185.5	35.2	20.6	773.0
	Total	Count	1164	406	77	45	1692
		Expected Count	1164.0	406.0	77.0	45.0	1692.0

a. Show how the expected count 531.8 (first cell for males) was computed.

b. Based on the table, do the assumptions appear to be met to perform the test? (Are all expected counts greater than 5?) **Yes** **No**

3. Compute the Test-Statistic and Calculate the *p*-value



```
   Pearson's Chi-squared test

data:  .Table
X-squared = 9.5135, df = 3, p-value = 0.02319
```

Test-Statistic

a. Based on the output above, the test-statistic is ____ = ____________

Calculate the *p*-value

b. Based on the output above, the *p*-value is _____________ .

4. **Evaluate the *p*-value and Conclusion**
   Evaluate the *p*-value:
   What is your decision at a 1% significance level? **Reject H_0** **Fail to Reject H_0**

   Conclusion:
   What is your conclusion at a 1% significance level *in context of the problem*?

5. Would your decision and conclusion change if the significance level was:

   - 5% instead of 1%?
   - 3% instead of 1%?
   - 2.3% instead of 1%?
   - 2% instead of 1%?

   Based on your answers, the *p*-value represents the ________________ significance level at which the results would be statistically significant.

Cool-Down: Name That Scenario

Match each research question with the appropriate Chi-Square test that should be used to answer the question

1. Is student status (in-state versus out-of-state) associated with one's eventual graduation outcome (graduating versus not graduating)?

   Answer: Chi-Square test of ______ _______________ ________

2. To test a theory that people have no preference among four different outdoor activities, you ask 100 people to select among jogging, bicycling, hiking, or swimming.

   Answer: Chi-Square test of ______ _______________ ________

3. A biostatistician would like to determine if the ratio of the blood type in the storage for transfusions should be different in Hawaii from the main land. She collected a sample of blood types of 10,000 people in Hawaii and that of 100,000 people in the mainland. She wishes to see if the breakdown of blood types (A, B, AB and 0) is the same for both populations.

   Answer: Chi-Square test of ______ _______________ ________