Unit 4

Other Ideas in Statistics

Unit Objectives

Upon completion of this unit, you should be able to:

1. recognize the chi-square and *F*-distributions;

2. demonstrate chi-square tests to make inferences about the distribution of a variable and to determine if an association of dependence exists between two variables;

3. apply analysis of variance to compare more than two means;

4. construct a linear regression equation and measures for determining the utility of the equation, as well as for determining the strength of the relationship between the two variables; and

5. apply *t*-tests to decide if the regression equation is useful for making predictions and to decide if the variables are linearly correlated, as well as determining the nature of the correlation.

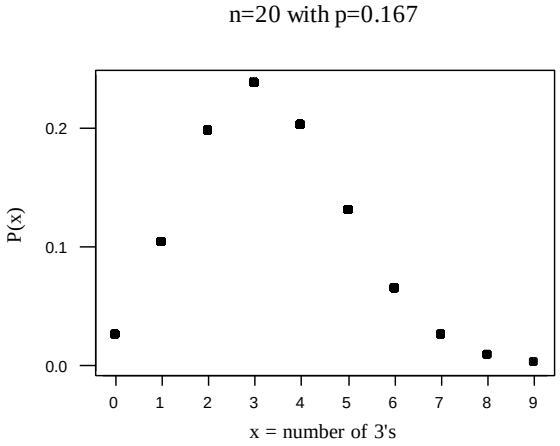
Instructor's Notes

Chi-Square Procedures

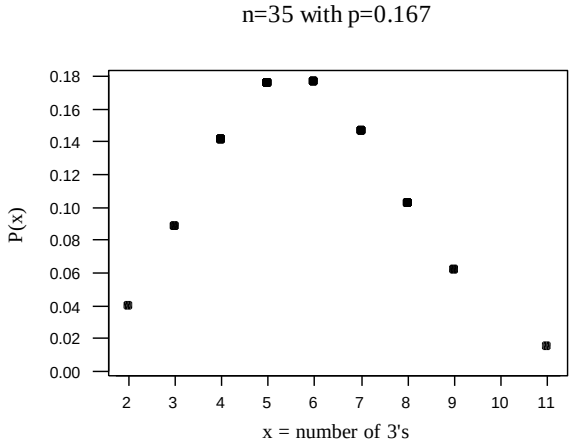
Up to this point we have dealt with performing hypothesis tests on means. In particular, all our generalizations have been based on numerical data. What if we start asking questions about populations characterized by qualitative data. In other words, how do we address questions that are concerned with relationships between characteristics of a population? For a moment reconsider the "first assignment" mentioned at the beginning of Unit 1. Students surveyed 20 men and 20 women and recorded characteristics such as gender and handedness. By the end of this section we will be able to answer the previously posed question: does gender have any effect on handedness?

Recall that the Central Limit Theorem says the distribution of sample **means** will be normally distributed if the sample size is sufficiently large. These new questions about relationships between characteristics do not involve calculating means; therefore, the Central Limit Theorem will not apply. This may mean that we can no longer make use of the normal distribution and the *t*-distribution. As a result, asking such questions will force us to learn about a new probability distribution. What will this new distribution look like? Will we be able to address all questions about such relationships using the same type of distribution? The answer to this last question is yes.

To develop this new distribution, we will consider tossing a fair six-sided die. The list of all possible outcomes is 1, 2, 3, 4, 5, and 6. Since the die is fair, each outcome is equally likely to occur with probability $1/6 \approx 0.167$. For this example, we are only interested in rolling a 3. Thus, P(roll 3) = 0.167, while P(not roll 3) = 0.833. Our experiment will consist of rolling the die 5 times. We may define a variable ***x*** to be the number of times we roll a 3. Using Minitab, we can simulate rolling a die 20 times and recording the number of times a 3 appears, the associated probability distribution for the possible values of ***x*** is given below**.**



Suppose we modify our experiment to consist of rolling a die 35 times and recording the number of 3's. Then the associated graph of the probability distribution would look like:



A quick glance at both of these graphs reveals a non-normal distribution. Each graph is right-skewed. The reason that the distribution is skewed is due to the fact that there is a greater probability of failure (0.833) than of success (0.167).

This type of distribution is called a **Chi-Square** (χ^2) **distribution**. There are infinitely many chi-square curves. Like the *t*-distribution, we differentiate between chi-square curves by their **degrees of freedom**. As the degrees of freedom increases, the chi-square curve looks more symmetric and bell-like.

Section 12.1 introduces the chi-square distribution and its associated probability table. Like the **t**-table, we will use the χ^2 table to find critical values and corresponding probabilities.

The first situation involving a random variable with a χ^2 distribution is called a "goodness of fit" test. This occurs when we are interested in determining if a population follows a given distribution. In other words, we will test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution.

Example: The Mars Company claims that its plain M&M candies are distributed with the following color percentages: 30% brown, 20% yellow, 20% red, 10% orange, 10% green, and 10% blue. A sample of 100 plain M&M's yielded the following results: 33 brown, 26 yellow, 21 red, 8 orange, 7 green, and 5 blue. Does this sample fit the distribution described by Mars, Inc.? Use a 0.05 level of significance.

In this problem, we are no longer dealing with means. Instead, we are more interested in proportions (or percentages) of the population. Furthermore, we are interested in many percentages and whether or not each differs from specific values. We need to come up with a statistic that will measure the change in all categories. The goal is to measure the change in all six categories (brown, yellow, red, etc.) and then use a modified version of those values to get **one** statistic.

To measure the change in categories we will be attempting to measure the change in percentages. However, we only have percentages given from Mars, Inc. It will turn out to be easier to measure the change in number not percentages. In other words, assuming that Mars, Inc. is truthful in stating that 30% of plain M&M's should be brown, then in our sample of 100 M&M's we could **expect** that $100 \cdot (0.30) = 30$ would be brown (note that we observed 33 to be brown). We will continue in this manner for each category computing an **expected** value that will be compared to the corresponding **observed** value.

When considering the actual numbers of M&M's, we can see that for every positive change (observed – expected) there will be a negative change since the number of all M&M's in the sample is fixed at 100. If we add up all the changes the net change will equal zero. To get a better sense of the magnitude of the changes we will square the changes so that all differences are positive. Furthermore, it will be useful to modify our values in order to show the measured change relative to the expected value. Finally we will sum all these modified values to get our test statistic.

More formally: when asking questions of this type (i.e., does a population follow a certain distribution?), we are testing the hypotheses:

H_o: the population fits the given distribution
H_a: the population has a different distribution

We will use the χ^2 distribution to test "**goodness of fit**" hypotheses such as this.

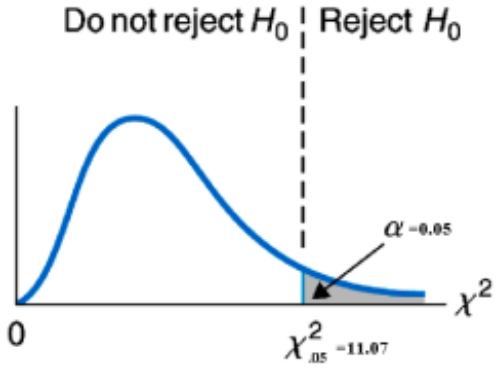
We will compute expected values, **E**, by multiplying the sample size *n* by the appropriate relative frequency (or percent) **p**. Thus, $\boldsymbol{E = np}$. Then for each category we will compute $(\boldsymbol{O - E})$, the observed value minus the expected value. These values will be squared, divided by corresponding expected values, and summed to yield the test statistic.

Therefore, **goodness of fit** tests will use the test statistic $\chi^2 = \sum \frac{(\boldsymbol{O-E})^2}{\boldsymbol{E}}$ with $k - 1$ degrees of freedom, where *k* is the number of categories. In order to apply the chi-square test it is important to make sure that almost all expected frequencies are greater than 5. For our example we get the following:

Color	**Observed O**	**Expected E**	**(O – E)**	**$(O–E)^2/E$**
brown	33	30	3	0.3000
yellow	26	20	6	1.8000
red	21	20	1	0.0500
orange	8	10	-2	0.4000
green	7	10	-3	0.9000
blue	5	10	-5	2.5000
Total:	100			5.9500

The sum of the far right column gives the value of the test statistic $\chi^2 = 5.95$ with 6–1=5 degrees of freedom.

When doing tests of **goodness of fit** using the χ^2 distribution we will always perform a right-tailed test. Therefore, for $\alpha = 0.05$ we have the following rejection region:



The test statistic does not fall within the rejection region. Therefore, we fail to reject the null hypothesis. In other words, there is insufficient evidence to warrant rejection of the

claim that the colors are distributed according to the percentages described by the Mars Company.

The next chi-square procedure that we will employ is a **chi-square test for independence**. Essentially, we will be using this test when we are interested in determining if two variables are statistically related. Keep in mind that we are not required to describe the relationship between the variables, nor are we making predictions about the variables. For this situation our data will be provided (or organized) in the form of a contingency table. Section 12.3 provides many examples of contingency tables and the method by which they are constructed.

For tests of independence, the χ^2 test statistic is calculated using the same formula, but the degrees of freedom = $(r-1)(c-1)$, where *r* is the number of rows and *c* is the number of columns. However, the expected values are calculated using $E = \frac{(\text{row total}) \cdot (\text{column to tal})}{\text{sample size}}$. Furthermore, in tests of independence the hypotheses will be defined as

H_o: the variables are independent
H_a: the variables are not independent

Like tests for goodness of fit, tests for independence will always be right-tailed tests.

Example: Suppose that the combined results of the student surveys on men and women yielded the following results: 264 men were right-handed, 27 men were left-handed, and 9 were ambidextrous; while for the women 273 were right-handed, 24 were left-handed, and 3 were ambidextrous. Are handedness and gender independent? Test at the 0.05 level of significance.

For this example we can define the hypotheses as:

H_o: handedness is independent of gender
H_a: handedness is *not* independent of gender

The associated contingency table with expected frequencies in parentheses is given below

		Gender		
		Male	**Female**	**Total**
Handedness	**right-handed**	264 (268.5)	273 (268.5)	537
	left-handed	27 (25.5)	24 (25.5)	51
	ambidextrous	9 (6)	3 (6)	12
	Total	300	300	600

Notice that for this contingency tables the expected values are the same across the rows. This does not happen every time we find expected frequencies. It is occurring in this particular instance because the column totals are identical.

We use the same method from the goodness of fit test to compute the χ^2 test statistic. Doing so yields χ^2 =3.327 with degrees of freedom = (3–1)(2–1) = 2. The critical value associated with $\alpha = 0.05$ is χ^2 = 5.991. Since the test statistic is less than the critical value we fail to reject the null hypothesis. Therefore, we conclude the results of this survey do not provide sufficient evidence to suggest that handedness and gender are related.

Written Assignment

Reminder: these written assignments are for your benefit and are **NOT** to be turned in for a grade.

Do problems 12.2-8, 12.16, 12.37, 12.57, and 12.61

Analysis of Variance (ANOVA) The procedures of this section allow us to compare the means of two populations or the mean responses to two treatments in an experiment. It is not uncommon for studies to compare more than two populations. We will need to have a procedure for comparing any number of means. For the moment consider the following scenario. In 1999 my car died. Considering the volatility of gas prices, I decided it was

Midsize cars		Sport Utility Vehicles		Pickup Trucks	
Model	MPG	Model	MPG	Model	MPG
Acura 3.5RL	25	Acura SLX	19	Chevrolet C1500	20
Audi A6 Quatro	26	Chevrolet Blazer	20	Dodge Dakota	25
BMW 740i	24	Chevrolet Tahoe	19	Dodge Ram	20
Buick Century	29	Chrysler Town & Country	23	Ford F150	21
Cadillac Catera	24	Dodge Durango	17	Ford Ranger	27
Cadillac Seville	26	Ford Expedition	18	Mazda B2000	25
Chevrolet Lumina	29	Ford Explorre	19	Nissan Frontier	24
Chevrolet Malibu	32	Geo Tracker	26	Toyota T100	23
Chrysler Cirrus	30	GMC Jimmy	21		
Ford Taurus	28	Infiniti QX4	19		
Honda Accord	29	Isuzu Rodeo	20		
Hyundai Sonata	27	Isuzu Trooper	19		
Infiniti I30	28	Jeep Grand Cherokee	21		
Infiniti Q45	23	Jeep Wrangler	19		
Jaguar XJ8L	24	Kia Sportage	23		
Lexus GS300	25	Land Rover Discovery	17		
Lexus LS400	25	Lincoln Navigator	16		
Lincoln Mark VIII	26	Mazda MPV	19		
Mazda 626	29	Mercedes ML320	21		
Mercedes-Benz E320	29	Mitsubishi Montero	20		
Mitsubishi Diamante	24	Nissan Pathfinder	19		
Nissan Maxima	28	Range Rover	17		
Oldsmobile Aurora	26	Subaru Forester	27		
Oldsmobile Intrigue	30	Suzuki Sidekick	24		
Plymouth Breeze	33	Toyota RAV4	26		
Saab 900S	25	Toyota 4Runner	22		
Toyota Camry	30				
Volvo S70	25				

necessary to explore the differences between mileage for midsize cars, trucks, and sport

utility vehicles. The following information, on highway gas mileage, was provided by the Environmental Protection Agency's Model Year 1998 Fuel Economy Guide.

From looking at this table, we can see there is a lot of variation in highway mileage. However, it does appear as though midsize cars, on average, have the lowest mileage on the highway.

The problem at hand is to compare the average mileage of three different populations. Are the differences in mean mileage significant enough to warrant choosing one type of automobile over another? Notice, you have not been asked which one has the lowest mileage.

We can see that comparing means for more than two populations is more complex than what we have previously encountered. Using the techniques of this section we could merely test all possible pairings of the three means, but this would be time consuming. Furthermore, comparing pairs limits us to talking about the differences in respective pairs. The ideal situation would be if we could compare the three populations simultaneously.
Asking if there is a difference between the three populations of automobiles means that we are really interested in determining if there is a great deal of variation in the results of the survey. In particular, if we were to compute the means for the individual samples from each population would there be a significant variation in the three means? Keep in mind that the term **variation** carries the same meaning that it did in Unit 1. In analyzing the data from the survey, we might think about applying the same ideas used to find the standard deviation of a distribution. This is the principle behind the **analysis of variance**.

The basic idea is to measure the variation in the data and to report the sum total of the variation. If the variation is large then we can claim that the means are indeed different. If the variation is small then we can say the means are essentially the same.

Because the procedures associated with performing analysis of variance require complicated calculations we will not emphasize a great deal of the associated statistical theory, but instead we will focus more on the use and interpretation of computer software such as Minitab. However, you will be expected to do the calculations for small samples. Keep in mind that we will be using a procedure to test the claim that two or more means are equal. Although the procedure is complicated, when we make use of computer software our conclusion will be based on a *p*-value. If the *p*-value is small (0.05 or lower) then we will reject the null hypothesis and conclude that a difference in means exists. Otherwise, we will fail to reject the equality of the means.

As with tests for independence and goodness of fit, analysis of variance procedures rely on a different distribution, in particular, the ***F*-distribution**. The *F*-distribution is a right-skewed curve that can be identified by **two** numbers of degrees of freedom. The chapter details the methods by which the degrees of freedom are determined and the appropriate way to read an *F*-table.

The hypotheses associated with analysis of variance are typically defined as

H_o: there is **no** difference among means
H_a: there **is** a difference among means

It is important to note that there are a few assumptions that must be satisfied before performing an analysis of variance hypothesis test. We will require that the samples be independent, for each population the variable under consideration must be normally distributed, and we will assume the standard deviations are the same for all populations of interest.

Return to the automobile example. The hypotheses for this problem may be restated as

H_o: the mean mileage for midsize cars, pickup trucks, and SUVs are the same
H_a: there **is** a difference in mean mileage for midsize cars, pickup trucks, and SUVs

Using Minitab to employ an analysis of variance, yields the following printout:

Minitab -- One-way Analysis of Variance

```
    Analysis of Variance for MPG
(1) Source     DF        SS        MS         F        P
(2) Vehicles    2    606.37    303.19    40.12    0.000
(3) Error      59    445.90      7.56
(4) Total      61   1052.27
                                        Individual 95% CIs For Mean
                                        Based on Pooled StDev
(5) Level       N      Mean    StDev  ---+---------+---------+---------+---
(6) Midsize    28    27.107    2.629                            (---*----)
(7) Pickups     8    23.125    2.588          (-------*------)
(8) SUV        26    20.423    2.914  (----*---)
                                       ---+---------+---------+---------+---
(9) Pooled StDev =    2.749            20.0      22.5      25.0      27.5
```

Glancing over the output reveals a few things. In lines 6 – 8 we can see that the mean highway mileage for midsize cars does appear to be higher than for pickups and SUVs. Or that perhaps mean mileage does decrease as we move from cars to pickups to SUVs. Line 2 provides the *F* test statistic, $F = 40.12$, and its associated *p*-value = 0.000. This means that the *p*-value is zero to three decimal places, or $p < 0.0005$. Such a small *p*-value indicates that we should reject the null hypothesis. We may conclude that there is strong evidence to suggest that the three types of vehicles do not have the same mean highway mileage.

Be aware that the *F*-test does not indicate which of the three means are significantly different. Looking at the means it appears that perhaps all three may differ. The computer output reveals additional information regarding degrees of freedom and other values used in producing the F-statistic.

We can see that the computer output also provides confidence intervals (lines 6 – 8) that support the same conclusion. The confidence interval for midsize cars does not overlap

either of the other confidence intervals. Remark: these are 95% confidence intervals for each mean separately. This **does not** imply that we are 95% confident that **all three** intervals cover the three means.

Written Assignment

Do problems 13.5-9, 13.15-17, 13.29, 13.33, and 13.35

Descriptive Methods in Regression and Correlation

We learned methods for determining if relationships exist between two variables. In particular, when applying the chi-square tests, we dealt with either two quantitative variables or two qualitative variables. Recall that we only allowed ourselves to ask if a relationship (dependence) existed, we did not address questions that could describe the type of relationship.

In this chapter when answering such questions we will be restricting our analysis to pairs of quantitative variables. We will consider questions such as

- How are the variables related?
- Does an increase in the value of one cause a decrease in the other?
- Does an increase in the value of one cause an increase in the other?
- Can we develop a formula to describe this relationship?
- If so, can we use this formula to make predictions about our variables?
- How strong is this relationship? How useful is the formula?

The methods of **regression** and **correlation** will help us answer these types of questions.

Section 4.1 reviews the algebra associated with straight lines and their equations. Be sure you understand how to graph, use, and interpret the **slope-intercept** form of a line, $y = b_0 + b_1 x$.

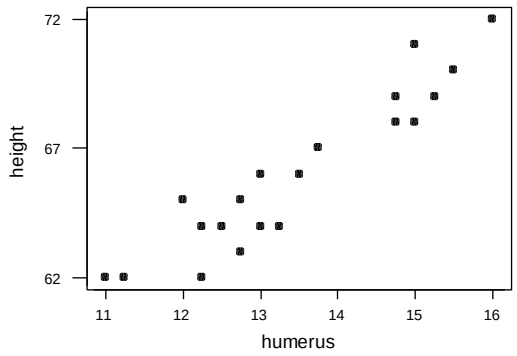
Example: Archaeologists and forensic scientists have found a strong linear relationship between the length of the humerus bone and height. In fact, they have collected enough data to develop a linear equation that can predict the height of a person given the length of the humerus bone. (Note: the humerus is the bone between the elbow to the shoulder.) Since we endeavored to collect measurements of height and humerus length in the "first assignment" we will try to generate our own equation to predict height when the length of the humerus is known. The results of one student's survey of 20 women are given below.

humerus	11	11.25	12	12.25	12.25	12.5	12.75	12.75	13	13
height	62	62	65	62	64	64	63	65	64	66

humerus	13.25	13.5	13.75	14.75	14.75	15	15	15.25	15.5	16
height	64	66	67	68	69	68	71	69	70	72

The humerus lengths are recorded in increasing order. Looking at the table, there appears to be a relationship between humerus length and height. In general, larger values for humerus length coincide with larger values for height. Since we are amateur forensic scientists, we will follow the lead of the experts and try to use the length of the humerus to predict height.

Thinking back on the algebra we learned, we might recall that we can graphically display these results on a coordinate system. Since humerus length will predict height we will view humerus length as the independent (or **predictor**) variable along the horizontal axis. Likewise, the height will be considered the dependent (or **response**) variable along the vertical axis. In particular, each ordered pairing of (humerus, height) may be graphed as described in section 4.1. Using Minitab, we can produce the following **scatterplot**.



The scattering of points does not appear to be random. In fact, the points seem to **cluster** around an imaginary line. We could easily take a pencil and draw a line through the cluster that would **approximate** the pattern of dots. Looking at this graph seems to verify our suspicion that height increases as the length of the humerus increases. Since the points seem to follow a straight line it is reasonable for us to try to construct an equation to represent that line.

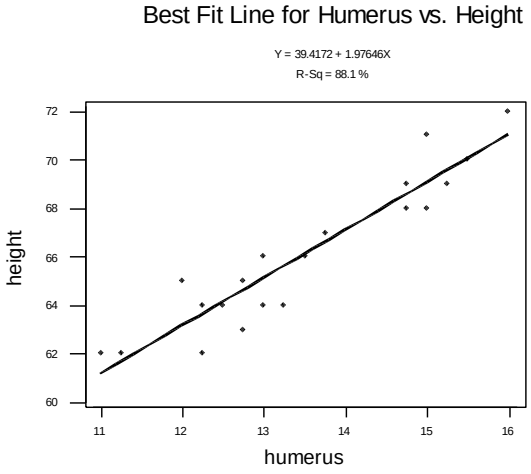
Clearly there are many lines that we could sketch on the graph that would provide a close approximation to the points. So the question is: how do we find the best one? For each line that we draw there will be a vertical distance between each point and the corresponding point on our line. Depending on how we draw our line, these vertical distances will be great or small. Each of these distances represents the error between the observed value and the predicted value given by the line. When finding the best line we want the sum of the squares of these distances to be as small as possible. The method of **least squares** is a mathematical procedure by which we construct an equation for the line of **best fit**. We will call this line the **regression line** and its associated equation will be the **regression equation**.

The least squares procedure involves many computations. The notation and formulas are provided in section 4.2. We may once again invoke the use of Minitab to determine the regression equation and to plot the regression line over the scatterplot. The output is given below.

Minitab – Regression Analysis

```
(1)  The regression equation is
(2)  height = 39.4 + 1.98 humerus

(3)  Predictor        Coef       StDev          T        P
(4)  Constant       39.417       2.319      17.00    0.000
(5)  humerus        1.9765      0.1711      11.55    0.000

(6)  S = 1.087       R-Sq = 88.1%     R-Sq(adj) = 87.4%

(7)  Analysis of Variance

(8)  Source            DF          SS          MS          F        P
(9)  Regression         1      157.67      157.67     133.38    0.000
(10) Residual Error    18       21.28        1.18
(11) Total             19      178.95
```

Lines 1 and 2 of the Minitab output tell us that the regression equation is given by: **height** = 39.4 + 1.98*__humerus__. We can use this equation to make a prediction. I measure my humerus to be 12.5 inches long. I can predict my height by substituting 12.5 into the formula to get $\textbf{\textit{height}} = 39.4 + 1.98(12.5) = 64.15$ inches. That is fairly accurate considering that, in reality, I am slightly taller than 5'4". The graph of the best fit line is shown below. We can see that points seem to cluster around this line.



For now, we will ignore lines 3 – 5 of the output.

Now that we have an equation that provides us with a model for making predictions we need to determine if it is in fact a good model. We have established that a relationship exists between the length of the humerus and height in women. How strong is this relationship? Basically, we are interested in knowing how much of the variation in the observed values of the response variable may be explained by the predictor variable (or the regression line).

The **coefficient of determination**, r^2, is a number that can be computed by looking at the measure of total variation in the observed values of the response variable (**SST**) and the measure of variation in the observed values of the of the response variable that can be explained by the regression (**SSR**). The coefficient of determination will be a number between 0 and 1. Values close to 0 imply that the regression is not overly useful. While values close to 1 will imply that the regression equation is very useful.

Line 6 of the Minitab output for our regression example shows the value of r^2:

```
(6)  S = 1.087       R-Sq = 88.1%     R-Sq(adj) = 87.4%
```

We can see that r^2=88.1%. This tells us that, for our sample, approximately 88% of the variation in height can be attributed to (or explained by) the regression. In other words, our regression equation is very useful and we can feel comfortable in using it to make predictions. Lines 9 and 11 of the Minitab output provide us with the actual values for SSR and SST, respectively. These values are found under the SS column. (We can check the math: SSR/SST=157.67/178.95=.881=88.1%.)

Often, we hear statements about the correlation between two variables. For example, "alcohol consumption and automobile accidents are positively correlated" and "there is no correlation between shoe size and mathematical ability." These are statements that attempt to describe the relationship, or lack thereof, between two variables.

The **correlation coefficient**, ***r***, is a number ranging from –1 to 1 that we will use to describe the strength and type of relationship between two variables. It turns out that *r* is just the square root of the coefficient of determination.

The correlation between two variables is considered to be negative if one variable decreases when the other variable increases, this will coincide with negative values for *r*. Similarly, the correlation between two variables is considered to be positive if one variable increases when the other variable increases, this will coincide with positive values for *r*. The sign (+ or –) of *r* will be consistent with the sign of the slope of the regression line, i.e., if the slope of the regression line is negative, then *r* will also be negative.

We will say that there is a strong negative correlation between to variables when the value of *r* is close to –1. Likewise, we say that there is a strong positive correlation

between to variables when the value of *r* is close to 1. For values of *r* that are close to 0 we will say that there is a weak correlation between the variables.

The Minitab output does not provide us with the value of the correlation coefficient. However, both r^2 and the slope of the regression equation are provided so we can quickly compute r. For our example, $r^2 = .881$ (the percent has been converted to a decimal) and the slope is positive, this tells us that *r* is the positive square root of 0.881. In particular, $r = \sqrt{0.881} = .939$, which implies a strong positive correlation between the length of the humerus and height.

Suppose we work as crime scene investigators for the local police department. As a result, we encounter many gruesome crime scenes. At one particular site, a small humerus bone is found measuring 5 inches. It is suspected that the humerus comes from a female. Despite all of our work, we cannot use our regression equation to predict the height of the female. The reason is because a length of 5 inches is not within the range of the observed values of our predictor variable – humerus lengths ranging from 11 to 16 inches. Similarly, if the bone found was 16 inches and suspected to be from a male we still could not use our equation. Our equation was constructed from a sample of women (over the age of 18) and therefore should only be applied to situations that satisfy the same criteria.

Written Assignment

Do problems 4.5, 4.7, 4.13, 4.15, 4.32, 4.35, 4.40, 4.41, 4.62, 4.63, 4.84, and 4.85

(Note: for 40, 62, and 84 you should get $y = 1.75 + 0.25x, r^2 = 0.2$, and $r = 0.447$.)

Inferential Methods in Regression and Correlation

The last section dealt with constructing equations to describe the relationship between two variables, producing a number to measure the strength of the relationship, and using the equation to make predictions. We used subjective language to suggest that relationships existed (whether strong or weak). Furthermore, from glancing at a graph of the regression line overlayed onto the scatterplot and the computation of r^2 we determined if a regression equation provided useful predictions. In the study of statistics it is preferable to take a more rigorous approach – in other words, we will use hypothesis tests to provide support to such conclusions. After performing hypothesis tests we may rightfully extend our results to statements (or inferences) about the **general population** instead of merely making inferences about our sample.

Section 14.2 shows how we can use an hypothesis test to decide whether a regression equation is useful for making predictions. In other words, the results of the hypothesis test will allow us to make conclusions about whether we may use one characteristic of a population to predict another. Keep in mind that the regression equation has two important parts: the **slope** and the **intercept**. If a regression line is to be useful then it is imperative that the slope be **non-zero**. Therefore, for all hypothesis tests of this type the null and alternative hypotheses will be defined as follows:

H_o: the slope = 0 (or *x* is **not** useful for predicting **y**)
H_a: the slope ≠ 0 (or *x* is useful for predicting **y**)

Ideally, we would like to reject the null hypothesis in this situation so that we can conclude that our regression equation is useful for making predictions.

Returning to our example on humerus vs. height we can now explain lines 3 – 5 of the Minitab output:

```
(3)  Predictor        Coef       StDev          T        P
(4)  Constant       39.417       2.319      17.00    0.000
(5)  humerus        1.9765      0.1711      11.55    0.000
```

The T value in line 5 is the value of the *t*-statistic for this hypothesis test. The associated *p*-value of 0.000 (i.e., 0 to three decimal places) is enough information to allow us to conclude that we should reject the null hypothesis in favor of concluding that the regression equation is useful in predicting height from humerus length.

Line 4 is associated with performing a test on the intercept—we will not concern ourselves with this type of test.

Section 14.4 shows how we can use a hypothesis test to decide if two variables are linearly correlated, and if so, the nature of the correlation – positive or negative. Keep in mind we have stated that for values of *r* close to zero there is a weak correlation between the two variables. In order to establish a strong linear correlation we would like to demonstrate that the population correlation coefficient, ρ, is either **non-zero**, **positive**, or **negative**. Therefore, for all hypothesis tests of this type the null and (options for the) alternative hypotheses will be defined as follows:

H_o: $\rho = 0$ (or *no* correlation exists)

$$H_a: \begin{cases} \rho \neq 0 & \text{(a correlation exists)} \\ \text{or} & \\ \rho < 0 & \text{(a negative correlation exists)} \\ \text{or} & \\ \rho > 0 & \text{(a positive correlation exists)} \end{cases}$$

Once again, we would like to reject the null hypothesis in favor of concluding a linear correlation exists.

For our example about height vs. humerus the Minitab output does not provide information on a hypothesis test concerning correlation. An additional command is required. Using the **correlation** command yields the value of *r* and the *p*-value for the two-sided test only:

```
Correlation of height and humerus = 0.939, P-Value = 0.000
```

This output only allows us to reject the null hypothesis (due to small *p*-value) in favor of concluding that a correlation exists. However, using Minitab alone we cannot say whether the correlation is positive or negative. To verify that the correlation is indeed positive we would need to do this test the old fashioned way – by hand.

Formulas for the test statistics can be found in the text.

Written Assignment

Do problems 14.31, 14.32, 14.59, and 14.60

In order to do the above hypothesis tests, you will first need to compute the regression equations and either the correlation coefficient or the **error sum of squares** (SSE). Partial Minitab outputs are provided below in order for you to verify that you have correctly computed these values before moving on to the requisite hypothesis tests.

Corvette Prices:

```
The regression equation is y = 372 - 27.9*x

S = 14.25       R-Sq = 93.7%     R-Sq(adj) = 92.9%

Analysis of Variance
Source            DF          SS          MS         F         P
Regression         1       24058       24058    118.53    0.000
Residual Error     8        1624         203
Total              9       25682
```

Custom Homes:

```
The regression equation is y = - 141 + 15.9*x

S = 59.62       R-Sq = 68.7%     R-Sq(adj) = 64.2%

Analysis of Variance
Source            DF          SS          MS         F         P
Regression         1       54562       54562     15.35    0.006
Residual Error     7       24882        3555
Total              8       79445
```

You are now ready to take the Unit 4 test. Please send in the test, with the attached assignment cover, as your assignment for Unit 4.

Unit 4 Test

Use only hand calculators on this exam; Minitab or any other computer software is **not** allowed. The textbook (but no other notes) may be used during the exam. Be sure to submit all your work in order to receive partial credit. For example, when performing hypothesis tests it is necessary that you define the hypotheses, sketch the rejection region, and show computations for the test statistic.

1. Find the critical value $\chi^2_{0.99}$ associated with a χ^2 curve with 10 degrees of freedom.

2. Find the critical value having an area of 0.05 to its right for an *F*-curve with 8 numerator degrees of freedom and 15 denominator degrees of freedom.

3. A random sample of 100 adults was recently gathered in order to explore a possible relationship between alcohol consumption and high blood pressure. The results are provided in the contingency table below. Your job is to determine if drinking status is independent of high blood pressure.

		Absent	Present	Total
Drinking Status	**non drinker**	26	11	37
	light to moderate	16	29	45
	heavy drinker	13	5	18
	Total	55	45	100

   a. Define the null and alternative hypotheses associated with the χ^2 test for independence.
   b. Complete the contingency table by finding expected values.
   c. Let $\alpha = 0.10$. Find and sketch the rejection region associated with this hypothesis test.
   d. Compute the χ^2 test statistic and state your conclusion.

4. The types of raw material used to construct stone stools found at the archaeological site Casa del Rito are shown below. (Bandelier Archaeological Excavation Project edited by Kohler and Root). A random sample of 1486 stone tools are obtained from a current excavation site.

Raw Material	**Regional Percent of Stone Tools**	**Observed Number of Tools at Current Excavation Site**
Basalt	61.30%	906
Obsidian	10.60%	162
Welded tuff	11.40%	168
Pedernal chert	13.10%	197
Other	3.60%	53
Total	100%	n = 1486

a. Test the claim that the regional distribution of raw materials fits the distribution at the current evaluation site. State null and alternative hypothesis.
b. Let $\alpha = 0.10$. Find and sketch the rejection region associated with this hypothesis test.
c. Compute the χ 2 test statistic and state your conclusion.

5. The presence of harmful insects in farm fields is detected by erecting boards covered with a sticky substance and then examining the insects trapped on the board. To investigate which colors are most attractive to cereal leaf beetles, researchers placed six boards of each of four colors in a field of oats. The table below gives data on the number of cereal leaf beetles trapped.

Color	Insects Trapped					
Yellow	45	59	48	46	38	47
White	21	12	14	17	13	17
Green	37	32	15	25	39	41
Blue	16	11	20	21	14	7

a. Compare the means for the four colors.
b. Test the hypothesis that the color of the board has no effect in attracting cereal leaf beetles at $\alpha = 0.05$, assuming that the number of insects trapped follows a normal distribution and that the standard deviations are the same for all colors.
c. Do these assumptions seem reasonable? Why or why not?

6. Manatees are large sea creatures that live in the shallow water along the coast of Florida. Many manatees are injured or killed each year by powerboats. Here are the data on manatees killed and powerboat registration (in thousands of boats) in Florida for the period 1984 to 1990.

Year	Powerboats registration x	Manatees killed y	x^2	y^2	xy
1984	559	34	312481	1156	19006
1985	585	33	342225		
1986	614	33		1089	20262
1987	645	39	416025	1521	
1988	675	43			29025
1989	711	50	505521	2500	35550
1990	719	47		2209	
SUMS	4508		2925834		182096

a. Complete the table.

b. Using the table, find the following values: $\bar{x}$, $\bar{y}$, S_{xx}, S_{yy}, and S_{xy}.

c. Using $b_1 = \dfrac{S_{xy}}{S_{xx}}$ *and* $b_0 = \bar{y} - b_1\bar{x}$ find the regression equation $\hat{y} = b_0 + b_1 x$.

d. What is the coefficient of determination, r^2?

e. If Florida were to limit powerboat registration to a maximum of 700,000 boats ($x = 700$) how many manatees could we expect to be killed?

7. Applicants for a particular job that involves extensive travel in Spanish speaking countries, must take a proficiency test in Spanish. The sample data below was obtained in a study of the relationship between the numbers of years applicants have studied Spanish and their score on the test.

Number of Years (x)	3	4	4	2	5	3	4	5	3	2
Score (y)	57	78	72	58	89	63	73	84	75	48

Partial Minitab output is provided below.

Mintab -- Regression Analysis

```
The regression equation is score = 31.5 + 10.9*years

S = 5.651       R-Sq = 83.0%     R-Sq(adj) = 80.9%

Analysis of Variance
Source            DF          SS          MS         F        P
Regression         1      1248.6      1248.6     39.09    0.000
Residual Error     8       255.5        31.9
Total              9      1504.1
```

a. What is the correlation coefficient, ***r***?
b. At the 5% level of significance, do the data provide sufficient evidence to conclude that the slope of the population regression line is not zero and hence that the number of years of study is useful as a predictor of score on the test?
c. Do the data provide sufficient evidence to conclude that the number of years of study and test score are linearly positively correlated? Use $\alpha = 0.01$.