Chapter 8 – Comparing Two Populations

In this chapter we consider how to compare two populations using confidence intervals and hypothesis testing.

8.1 – Introduction

1. Many times real-world situations call for *comparing* two (or more) related populations. For instance,

   a. Does Advil fight pain better than Tylenol?
   b. Does Longhorn Steakhouse have a longer wait than Outback Steakhouse?
   c. Do people spend more at Walmart or Target?
   d. Are men older than women when they graduate from college?
   e. Does one fertilizer work better than another?
   f. Does Machine 1 produce stamp automobile hoods faster than Machine 2?
   g. Does one location of a restaurant have less sales than another?

2. The really creative part of statistics is figuring out:

   1. *What* (variables) you need to measure in order to answer the questions above, and
   2. *How* you are going to *measure* each variable.

   For instance, with example *a*, the variable we are interested in is *pain*. But how do we measure *pain*? We could ask people to rate their pain on a scale of 1-10 (ordinal data). Or, we could measure the time until the subjects report feeling no pain, but this could be problematic as peoples' perceptions of pain are different. Perhaps we could measure certain things in the brain or the muscles. For an example such as this, we would consult heavily with an expert in the field.

   In example *e*, we can answer the question in many ways. We could measure the heights of plants, the weights of plants, the weight of yield (amount of fruit or vegetable), the amount of yield (*e.g.* number of potatoes), the diameter of a bloom, the number of blooms, the amount of water, density of water, *etc*. In a real situation, we would probably use many of these ways to compare fertilizer.

   Example *f* has a much more direct answer. We can simply take a sample of the times it takes to produce parts from each machine.

3. One way to compare two populations is to compare the respective means, μ_1 and μ_2. For instance, to compare the sales at two restaurants, we my hypothesize that the average sales at restaurant 1 are more than restaurant 2, $\mu_1 > \mu_2$.

4. Now, notice the following:

   - Start with the inequality: $\mu_1 > \mu_2$
   - Subtract μ_2 from both sides:

$$\begin{array}{rcl} \mu_1 & > & \mu_2 \\ -\mu_2 & & -\mu_2 \\ \hline \mu_1 - \mu_2 & > & 0 \end{array}$$

   Which says, of course that these two are equivalent: $\mu_1 > \mu_2$ and $\mu_1 - \mu_2 > 0$.

5. Statistical tests that compare two means are designed to draw inference on the *difference* in the two means, $\mu_1 - \mu_2$. Thus, the proper way to write the hypothesis above is:

$$H_o: \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_a: \mu_1 - \mu_2 > 0$$

8.2 – Writing Hypotheses

1. Examples:

   a. A claim is made that on average starting salaries for computer science majors are higher than marketing majors: $\mu_{CS} > \mu_{Mktg}$ which can be expressed:

   $$H_o: \mu_{CS} - \mu_{Mktg} = 0 \quad \textit{vs.} \quad H_a: \mu_{CS} - \mu_{Mktg} > 0$$

   b. A claim is made that on average, a gasoline additive has no effect on gas mileage: $\mu_{w/o} = \mu_{with}$ which can be expressed as:

   $$H_o: \mu_{w/o} - \mu_{with} = 0 \quad \textit{vs.} \quad H_a: \mu_{w/o} - \mu_{with} \neq 0$$

   c. A claim is made that on average men spend less time shopping at Walmart than women: $\mu_{Men} < \mu_{Women}$ which can be expressed as:

   $$H_o: \mu_{Men} - \mu_{Women} = 0 \quad \textit{vs.} \quad H_a: \mu_{Men} - \mu_{Women} < 0$$

   Equivalently, we can say that women spend more time shopping than men: $\mu_{Women} > \mu_{Men}$. Thus, the hypothesis above can also be expressed:

   $$H_o: \mu_{Women} - \mu_{Men} = 0 \quad \textit{vs.} \quad H_a: \mu_{Women} - \mu_{Men} > 0$$

   Either set of hypotheses is correct. The key is to focus on (a) the order the parameters are specified (*e.g.* Women first, then Men) and (b) the direction of the alternate hypothesis (*e.g. greater than*). This will be important when we use the calculator or software to compute pvalues.

   d. A claim is made that on average, women spend at least 10 minutes longer shopping than men. We would express the hypothesis this way:

   $$H_o: \mu_{Women} - \mu_{Men} = 10 \quad \text{vs.} \quad H_a: \mu_{Women} - \mu_{Men} > 10$$

   e. A claim is made that college graduates make at least $300 more per month than high school graduates, which can be expressed:

   $$H_o: \mu_{College} - \mu_{HS} = 300 \quad \text{vs.} \quad H_a: \mu_{College} - \mu_{HS} > 300$$

2. When we compare to means, we *hypothesize* about the difference in the two means. In general, hypotheses should be expressed this way:

$$H_o: \mu_1 - \mu_2 = d_o \quad \text{vs.} \quad H_a: \mu_1 - \mu_2 > d_o$$
$$H_o: \mu_1 - \mu_2 = d_o \quad \text{vs.} \quad H_a: \mu_1 - \mu_2 < d_o$$
$$H_o: \mu_1 - \mu_2 = d_o \quad \text{vs.} \quad H_a: \mu_1 - \mu_2 \neq d_o$$

   where d_o is called the *hypothesized difference.* In the last example immediately above, $d_o = 300$.

3. In this chapter, we study statistical inference (confidence intervals and hypothesis tests) about the difference between two means. There are three cases: large samples, small samples, and dependent samples. We also study the procedure for comparing two proportions.

Homework 8.1

1. A study is done to determine the average time that men and women spend in the library studying. Over the course of several weeks, students are randomly selected as they enter the library and observed. Students that have entered the library to study are timed and their sex is recorded. At the end of the study, the researchers wanted to see if there was any significant difference in the average time that men and women study in the library. What is the appropriate set of hypotheses?

2. Mike has noticed that the trip to his job each morning seems to take less time than his return trip home in the evening, even when he travels the same route. In order to determine if there is evidence that this is true, he records the time it takes traveling to work in the mornings for 20 consecutive days and the time it takes for the return trip. (a) What is the appropriate set of hypotheses? (b) What hypothesis should he test if he suspects that it takes at least 5 minutes less time in the morning as compared to the return trip in the evening?

3. A test is conducted to test two pain relievers. Two random samples of adults are selected and administered the standard dose. A blood monitor is attached to each subject to determine the level of the pain reliever in the subject's system. Timing is started when the drug first reaches 60 *ppm* and continues until the first time the level of the drug drops below 60 *ppm*. (a) Researchers would like to show that the new pain reliever, *No Pain Plus* stays in the system at longer than the leading brand, *Quick Relief*. What is the appropriate set of hypotheses? (b) Suppose that researchers wanted to shot that *No Pain Plus* stays in the system at least 4 hours longer than the leading brand, *Quick Relief*. What is the appropriate set of hypotheses?

8.3 – Large Sample Confidence Intervals

1. **Definitions and Assumptions**

   We will calculate a confidence interval about the true mean difference in population means, $\mu_1 - \mu_2$. We make the following assumptions:

   a. The two samples are independent of one another.
   b. Within each sample, data values are independent. (Sometimes we say that we assume independence *between* and *within* samples).
   c. Both samples are large, $n_1 \geq 30,\ n_2 \geq 30$.

   Our goal is to estimate $\mu_1 - \mu_2$. We remember that all confidence intervals are of the form:

$$Estimator \pm (Table\ Statistic) * (Standard\ Error\ of\ Estimator)$$

   - The estimator of $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$.
   - The standard error of the estimator is $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{V[\bar{X}_1 - \bar{X}_2]} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
   - The confidence interval formula is: $\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. Notice that this is the same $z_{\alpha/2}$ that was used for the confidence interval formula for a single mean and for proportions.

2. **Calculator Directions**

On your calculator, choose: Stats/Tests/2SampZInt (#9) to find a confidence interval on the difference between two means.

The screen will show:

2-SampZInt
Inpt: Data Stat

If you choose: *Stat*, the screen will look like this:

σ_1:
σ_2:
$\bar{x}_1$:
n_1:
$\bar{x}_2$:
n_2:
$C - Level$
Calculate

You will type in the appropriate values. When you choose *Calculate*, the confidence interval will be shown.

3. **Example –** A survey was conducted to compare the average salaries of data-processing managers employed in the financial sectors of two areas of the country - the East Central and the Southern regions. Independent random samples of 600 data-processing managers working in the East Central and 250 working in the Southern region were selected and asked to reveal their annual salaries. The results are summarized in the table on the right.

Region	*N*	$\bar{x}$	*s*
East Central	600	\$32,300	\$3200
South	250	\$30,800	\$4700

a. Find a 90% confidence interval for the difference between the true mean salaries in the East Central and South. On the calculator, we type in these values:

Inpt: Data *Stat*
$\sigma_1: 3200$
$\sigma_2: 4700$
$\bar{x}_1: 32300$
$n_1: 600$
$\bar{x}_2: 30800$
$n_2: 250$
$C - Level: 0.9$
Calculate

The resulting confidence interval is: [\$965.92, \$2034.10].

How do we interpret this confidence interval? We say that we are 90% sure that the true mean difference, $\mu_{East\ Central} - \mu_{South}$ is *in* the confidence interval. Thus, this confidence interval means that we are 90% sure that average salaries in the East Central are at least \$965 more than average salaries in the South and that average salaries in the East Central are up to \$2034 more than average salaries in the South. Clearly, salaries are higher on average in the East Central than in South.

b. Evaluate the following hypothesis using the confidence interval found in part *a.*

$$H_o: \mu_{EC} - \mu_S = 0 \quad \text{vs.} \quad H_a: \mu_{EC} - \mu_S > 0$$

Based on the confidence interval, we would reject the null hypothesis. Why? Because we have 90% confidence that the difference is *at least* \$965 and might be as large as \$2034. So, clearly we have strong evidence that the difference is larger than 0. Thus, we reject the null hypothesis.

c. Evaluate the following hypothesis using the confidence interval found in part *a.*

$$H_o: \mu_{EC} - \mu_S = 800 \quad \text{vs.} \quad H_a: \mu_{EC} - \mu_S > 800$$

Based on the confidence interval, we would again, reject the null hypothesis. Why?

d. Evaluate the following hypothesis using the confidence interval found in part *a.*

$$H_o: \mu_{EC} - \mu_S = 1000 \quad \text{vs.} \quad H_a: \mu_{EC} - \mu_S > 1000$$

Based on the confidence interval, [$965, $2034] we would fail to reject the null hypothesis. Why? The alternate hypothesis is asking if the difference is more than $1000? It **might be** because the difference might be as large as $2034, *but* it **might not be** because the difference might be as small as $965. When we have conflicting answers, we are not *certain*, thus we fail to reject the null hypothesis.

4. **Pattern of Signs for confidence interval for $\mu_1 - \mu_2$**

A confidence interval for the difference in two means can be composed of two positive numbers, two negative numbers, or a negative and a positive number as the chart below shows. Based on the just the *signs* of the numbers in the confidence interval, we can conclude the following things:

Signs in Confidence Interval	Interpretation	Conclusion
[+, +]	$\mu_1 > \mu_2$	Strong
[-, +]	$\mu_1 = \mu_2$	Weak, not evidence of a difference between the two means
[-, -]	$\mu_1 < \mu_2$	Strong

5. **Example –** A consumer group would like to see if interest rates on credit cards have an effect on the amount people charge on the cards. The group takes a simple random sample of single people using a credit card with an interest rate of 12% and determines the amount charged over the course of a year. A sample is also taken from single people using a card with 16% interest. The results are shown in the table below.

Interest Rate	*n*	$\bar{x}$	*s*
12%	150	$2056	$392
16%	150	$1987	$413

a. Find a 95% confidence interval on the true mean difference between the amounts charged under the two interest rates. On the calculator, we type in these values:

Inpt: Data *Stat*
$\sigma_1 : 392$
$\sigma_2 : 413$
$\bar{x}_1 : 2056$
$n_1 : 150$
$\bar{x}_2 : 1987$
$n_2 : 150$
$C - Level : 0.95$
Calculate

The resulting confidence interval is: [-$22.12, $160.12].

This confidence interval means that the average amount charged under the 12% card could be $22 less than the average amount charged with the 16% card. On the other hand, the average amount charged under the 12% card could be up to $160 more than the average amount charged with the 16% card. Notice that this is *conflicting* evidence. The first sentence says that people charge more with the 16% card and the second sentence says people charge more with the 12% card. What can we infer from this? There is not evidence to say that the mean amounts charged with the two cards differs. Another way to say this is to note that zero is *in* the confidence interval which means that it is possible that there is no (zero) difference.

b. Use the confidence interval to evaluate this hypothesis:

$$H_o: \mu_{12\%} - \mu_{16\%} = 0 \quad \text{vs.} \quad H_a: \mu_{12\%} - \mu_{16\%} \neq 0$$

We would fail to reject the null hypothesis. Why? We ask ourselves, could there be no difference (0 difference). The answer is, "Yes, 0 is a possible value for the difference. 0 is in the confidence interval." We would conclude by saying that there is not strong evidence that there is a statistical difference between the mean amounts charged by people using the two cards.

6. Theory – We can always swap the order of the inference by swapping the lower and upper limits, multiplying each by -1.

If a confidence interval on $\mu_A - \mu_B$ is $[x, y]$ then a confidence interval on $\mu_B - \mu_A$ is $[-y, -x]$, for example:

Confidence Interval on $\mu_A - \mu_B$	Equivalent Confidence Interval on $\mu_B - \mu_A$
$[-9, -1]$	$[1, 9]$
$[-9,4]$	$[-4, 9]$
$[2, 11]$	$[-11, -2]$

7. **Example –** A survey is done to compare the amounts of money people spend at Flash Foods and Wal-Mart and the results are shown in the table below.

Store	*n*	$\bar{x}$	*s*
Flash Foods	50	\$7.22	\$6.12
Wal-Mart	50	\$28.34	\$24.78

a. Find a 95% confidence interval on the true mean difference between the average amounts spent at Flash Foods and Wal-Mart. On the calculator, we type in these values:

Inpt: Data *Stat*
$\sigma_1: 6.12$
$\sigma_2: 24.78$
$\bar{x}_1: 7.22$
$n_1: 50$
$\bar{x}_2: 28.34$
$n_2: 50$
$C - Level: 0.95$
Calculate

The resulting confidence interval $\mu_{FF} - \mu_{WM}$ is:
[-\$28.20, -\$14.04].

The first thing we note is that we are 95% sure that the mean of the *first* population (Flash Foods) is less than the *second* population (Wal-Mart) because of the two negative signs in the confidence interval. Thus, we have strong statistical evidence that $\mu_{FF} - \mu_{WM} < 0$, or $\mu_{FF} < \mu_{WM}$. Sometimes, it can be confusing to interpret a confidence interval with two negative signs. Part *b* should clarify things.

b. Repeat the problem above by finding a 95% confidence interval on $\mu_{WM} - \mu_{FF}$. Notice that we switch the order when supplying the values for the calculator.

Inpt: Data *Stat*
$\sigma_1: 24.78$
$\sigma_2: 6.12$
$\bar{x}_1: 28.34$
$n_1: 50$
$\bar{x}_2: 7.22$
$n_2: 50$
$C - Level: 0.95$
Calculate

The resulting confidence interval is: [\$14.04, \$28.20].

Notice that this confidence interval is the same as the previous ones except the numbers are positive and have switched places. This relationship is always true. We can reverse the order of inference by taking the negative of each number and swapping their order.

Now we can more easily interpret this confidence interval by saying that we are 95% sure that people spend on average at least \$14.04 more at Wal-Mart, compared to Flash Foods and that people at Wal-Mart may spend up to \$28.20 more on average than at Flash Foods.

c. Can we reach the strong conclusion that people spend on average at least \$10 more at Wal-Mart? The answer is, "Yes." In other words, evaluate this hypothesis:

$$H_o: \mu_{WM} - \mu_{FF} = \$10 \quad \text{vs.} \quad H_a: \mu_{WM} - \mu_{FF} > \$10$$

We reject the null hypothesis.

d. Can we reach the strong conclusion that people spend on average at least \$50 more at Wal-Mart? The answer is, "No." In other words, evaluate this hypothesis:

$$H_o: \mu_{WM} - \mu_{FF} = \$50 \quad \text{vs.} \quad H_a: \mu_{WM} - \mu_{FF} > \$50$$

We would fail to reject the null hypothesis

8. **Example –** For each cell in the table, consider the hypothesis at the top and the confidence interval on the left. Answer, "yes" or "no".

	Reach a decision on each hypothesis based on the corresponding confidence interval	
95% Confidence Interval for $\mu_M - \mu_B$	$H_a: \mu_M - \mu_B > \$500$	$H_a: \mu_M - \mu_B < \$500$
(\$300, \$400)		
(\$475, \$525)		
(\$550, \$600)		

Homework 8.2

1. Researchers suspect that among men and women that workout on a regular basis, men spend more time in an aerobic state during a workout than women. A simple random sample of 52 men and 56 women was taken. The results showed that men averaged 72.3 minutes in an aerobic state while women averaged 61.7. The standard deviations for men and women were 33.1 minutes and 29.4 minutes, respectively. Find a 90% and 95% confidence intervals on the men difference in aerobic times for men and women and interpret the intervals.

2. A company makes frozen dinners. They have two plants in different parts of the country and thus use different suppliers for the raw ingredients. The company would like to see if there is any difference in the percentage of salt in a dinner between the two pants. The summary statistics for the data are shown below. Find a 95% confidence interval on the difference between the two plants average percentage salt content. Is there evidence that there is a significant difference between the two plants in terms of salt content? What is the appropriate hypothesis? Use the confidence interval to reach a decision.

Plant	*n*	$\overline{x}$	*s*
Moab, UT	40	4.17%	1.4%
Hilo, HI	40	3.95%	1.2%

3. Mike has noticed that the trip to his job each morning seems to take less time than his return trip home in the evening, even when he travels the same route. In order to determine if there is evidence that this is true, he records the time it takes traveling to work in the mornings for 30 consecutive days and the time it takes for the return trip. The average time it takes his to get to work is 23.4 minutes with a standard deviation of 1.3 minutes. The return trip in the evenings takes on average 30.7 minutes with a standard deviation of 3.4 minutes. (a) Find a 95% confidence interval on the difference between average evening and morning times and interpret the interval. (b) Suppose that we want to see if there is strong evidence that it takes longer in the evenings. What is the appropriate hypothesis? Use the confidence interval to reach a decision. (c) Can we make the claim that it takes at least 5 minute more in the evening, on average? What is the appropriate hypothesis? Use the confidence interval to reach a decision?

4. For each cell in the table, consider the hypothesis at the top and the confidence interval on the left. Anwer, "yes" or "no". Explain why in each case.

95% Confidence Interval for $\mu_A - \mu_B$	If $H_a: \mu_A - \mu_B < 0$ would you reject the corresponding null hypothesis based on the confidence interval?	If $H_a: \mu_A - \mu_B > 25$ would you reject the corresponding null hypothesis based on the confidence interval?
(-100, -50)		
(-75, 35)		
(60, 110)		

5. Consider the following 95% confidence interval which shows the average difference between men and women in how long enlisted personnel serve in the armed services: [2.4 years, 3.6 years]. Make a decision for each of the following hypotheses.

a. $H_o: \mu_{Men} - \mu_{Women} = 0$ vs. $H_a: \mu_{Men} - \mu_{Women} \neq 0$
b. $H_o: \mu_{Men} - \mu_{Women} = 0$ vs. $H_a: \mu_{Men} - \mu_{Women} > 0$
c. $H_o: \mu_{Men} - \mu_{Women} = 2.5$ vs. $H_a: \mu_{Men} - \mu_{Women} > 2.5$
d. $H_o: \mu_{Men} - \mu_{Women} = 3$ vs. $H_a: \mu_{Men} - \mu_{Women} < 3$
e. $H_o: \mu_{Men} - \mu_{Women} = 4$ vs. $H_a: \mu_{Men} - \mu_{Women} < 4$
f. $H_o: \mu_{Men} - \mu_{Women} = 2$ vs. $H_a: \mu_{Men} - \mu_{Women} > 2$

8.4 – Large Sample Hypothesis Testing

1. Calculator Details

We can also do hypothesis testing on differences using the p-value approach. We remember that these are the hypotheses we can test:

$$H_o: \mu_1 - \mu_2 = d_o \text{ vs. } H_a: \mu_1 - \mu_2 > d_o$$
$$H_o: \mu_1 - \mu_2 = d_o \text{ vs. } H_a: \mu_1 - \mu_2 < d_o$$
$$H_o: \mu_1 - \mu_2 = d_o \text{ vs. } H_a: \mu_1 - \mu_2 \neq d_o$$

However, the calculator can only consider the case where $d_o = 0$ and the calculator expresses the hypotheses differently:

Your Brain and Calculator		**Statistically**
$H_o: \mu_1 = \mu_2$ vs. $H_a: \mu_1 > \mu_2$	$\Rightarrow$	$H_o: \mu_1 - \mu_2 = 0$ vs. $H_a: \mu_1 - \mu_2 > 0$
$H_o: \mu_1 = \mu_2$ vs. $H_a: \mu_1 < \mu_2$	$\Rightarrow$	$H_o: \mu_1 - \mu_2 = 0$ vs. $H_a: \mu_1 - \mu_2 < 0$
$H_o: \mu_1 = \mu_2$ vs. $H_a: \mu_1 \neq \mu_2$	$\Rightarrow$	$H_o: \mu_1 - \mu_2 = 0$ vs. $H_a: \mu_1 - \mu_2 \neq 0$

The calculator will provide the p-value for the hypothesis test and we will use it and interpret it the same way we have in the past.

2. Calculator Directions

On your calculator, choose: Stats/Tests/2SampZTest (#3). The screen will show:

2-SampZTest
Inpt: Data Stats

If you choose: *Stats*, the screen will look like this:

σ_1:
σ_2:
$\bar{x}_1$:
n_1:
$\bar{x}_2$:
n_2:
$\mu_1 : \neq \mu_2, < \mu_2, > \mu_2$
Calculate

You will type in the appropriate values. When you choose *Calculate*, the p-value will be shown.

3. Example – A survey was conducted to compare the average salaries of data-processing managers employed in the financial sectors of two areas of the country - the South and the East Central regions. Independent random samples of 250 data-processing managers working in the South and 600 working in the East Central area were selected and asked to reveal their annual salaries. The results are summarized in the table shown on the right. Is there evidence at the 10% significance level that on average, salaries in the South are lower than in the East Central? Use the p-value approach.

Region	n	$\bar{x}$	s
South	250	\$30,800	\$4700
East Central	600	\$32,300	\$3200

(1) Hypothesis: $H_o: \mu_S - \mu_{EC} = 0$ vs. $H_a: \mu_S - \mu_{EC} < 0$
(2) p-value calculation:

$\sigma_1: 4700$
$\sigma_2: 3200$
$\bar{x}_1: 30800$
$n_1: 250$
$\bar{x}_2: 32300$
$n_2: 600$
$\mu_1 : \neq \mu_2, < \mu_2, > \mu_2$
Calculate

(3) p-value: $1.92 * 10^{-6} = 0.00000192 \approx 0$
(4) Decision: Since 0 < 0.1, reject the null hypothesis
(5) Conclusion: There is strong evidence that the average salary in the South is less than in the East Central region.

4. A consumer group would like to see if interest rates on credit cards have an effect on the amount people charge on the cards. The group takes a simple random sample of single people using a credit card with an interest rate of 12% and determines the amount charged over the course of a year. A sample is also taken from single people using a card with 16% interest. The results are shown in the table on the right. Is there evidence at the 5% significance level that single people charge more with a lower interest rate? Use the p-value approach.

Interest Rate	n	$\bar{x}$	s
12%	150	\$2056	\$392
16%	150	\$1987	\$413

(1) Hypothesis: $H_o: \mu_{12\%} - \mu_{16\%} = 0$ vs. $H_a: \mu_{12\%} - \mu_{16\%} > 0$

(2) p-value calculation:

$\sigma_1: 392$
$\sigma_2: 413$
$\bar{x}_1: 2056$
$n_1: 150$
$\bar{x}_2: 1987$
$n_2: 150$
$\mu_1: \neq \mu_2, < \mu_2, > \mu_2$
Calculate

(3) p-value = 0.0689
(4) Decision: Since 0.0689 > 0.05, fail to reject the null hypothesis
(5) Conclusion: There is not evidence to say that the average charge with 12% interest is higher than with 16% interest.

5. A production line is designed on the assumption that the difference between mean assembly times for two operations is 5 minutes. Independent tests for the two assembly operations revealed the results shown in the table above. At the 5% significance level, is there evidence that the true mean difference in assembly times differs from 5 minutes? Use the CI approach.

Operation	n	$\bar{x}$	s
A	100	14.8 minutes	0.8 minutes
B	50	10.4 minutes	0.6 minutes

(1) Hypothesis: $H_o: \mu_A - \mu_B = 5$ vs. $H_a: \mu_A - \mu_B \neq 5$
(2) CI calculation:

Inpt: Data *Stat*
$\sigma_1: 0.8$
$\sigma_2: 0.6$
$\bar{x}_1: 14.8$
$n_1: 100$
$\bar{x}_2: 0.6$
$n_2: 50$
$C - Level: 0.95$
Calculate

(3) CI = [4.17 minutes, 4.63 minutes]
(4) Decision: Since 5 minutes is not in the interval, we reject the null hypothesis.
(5) Conclusion: There is sufficient evidence supplied by the data to say that the true mean difference in assembly times is not 5 minutes.

Homework 8.3

1. A study is done to determine if there is evidence that men spend less time in the restroom than women. The data was collected by observing 40 men and 40 women as they entered and exited a bathroom in a shopping mall. The results of the study are shown in the table above. Assume a 1% significance level to evaluate the hypothesis using the p-value approach.

	n	$\bar{x}$	s
Men	40	98 *sec*	47 *sec*
Women	40	143 *sec*	41 *sec*

2. A study is done to determine the average time that men and women spend in the library studying. Over the course of several weeks, students are randomly selected as they enter the library and observed. Students that have entered the library to study are timed and their sex is recorded. At the end of the study, the researchers wanted to see if there was evidence that men study longer than women. Assume a 5% significance level to evaluate the hypothesis and use the p-value approach to evaluate the hypothesis.

	n	$\bar{x}$	s
Men	35	83 *min*	37 *min*
Women	35	72 *min*	42 *min*

3. A study is undertaken to determine if visitors to the Investors.com web site spend more time, on average, than they do at YourMoney.com. The results of the study are shown in the table below.

Web Site	n	$\bar{x}$	s
Investors.com	30	14.6 *min*	10.2 *min*
YourMoney.com	30	12.4 *min*	9.8 *min*

Assume a 10% significance level to evaluate the hypothesis.

4. A study of the number of quarters that it took men and women to graduate with a four-year bachelor's degree was done. The 30 men in the sample averaged 18.33 quarters while the 30 women averaged 16.2 quarters. The standard deviations for men and women were 1.2 and 0.75 quarters, respectively. Is there evidence at the 10% significance level that there is any difference between the average number of quarters for men and women?

8.5 – Small Samples Confidence Intervals and Hypothesis Tests

1. Definitions and Assumptions

   When we have two small samples, we have to make the following assumptions:

   1. The two samples are independent of one another.
   2. Within each sample, data values are independent. (Sometimes we say that we assume independence between and within samples).
   3. Both populations are normally distributed.
   4. Both samples are small, $n_1 < 30,\ n_2 < 30$.
   5. We must decide whether we can assume the variances of the two populations are equal, $\sigma_1^2 = \sigma_2^2$. When we can assume this, we get more precise results. We'll see how to check this out below.

2. Rule-of-thumb for checking for equal variances. If $\frac{\max\{s_1,s_2\}}{\min\{s_1,s_2\}} < 2$ then it is OK to assume equal variances. When we decide that it is OK to assume equal variances, we say that we are going to use a pooled estimate of the common variance. You calculator has a yes/no field labeled: Pooled.

3. Example: A consumer group would like to see if interest rates on credit cards have an effect on the amount people charge on the cards. The group takes a simple random sample of single people using a credit card with an interest rate of 12% and determines the amount charged over the course of a year. A sample is also taken from single people using a card with 16% interest. The standard deviations are $s_{12\%} = \$392$ and $s_{16\%} = \$413$. Thus, the equal variance assumption is satisfied because $\frac{413}{392} < 2$.

4. Calculator Directions

   To calculate a confidence interval on the difference in two means using small samples, on your calculator, choose: Stats/Tests/2SampTInt (#0). The screen will show:

   2-SampTInt
   Inpt: Data Stat

If you choose: *Stat*, the screen will look like this:	If you choose: *Data*, the screen will look like this:
$\bar{x}_1$: S_{x1}: n_1: $\bar{x}_2$: S_{x2}: n_2: $C - Level$: $Pooled$: No Yes	$List1$: $List2$: $Freq1: 1$ $Freq2: 1$ $List1$: $C - Level$: $Pooled$: No Yes

To do hypothesis testing using two small samples, on your calculator, choose: Stats/Tests/2SampTTest(#4). The screens will be similar to above except that the *C*-Level row will be replaced with the specification of the alternate hypothesis: $\mu_1 : \neq \mu_2, < \mu_2, > \mu_2$

5. Example – The speed of cars on a particular road is known to be normally distributed. A sample of 10 cars taken during the morning rush hour revealed a sample average of 47.3 mph with a standard deviation of 3.6 mph. A sample of 12 cars taken during the evening rush hour had a sample average of 43.2 mph with a standard deviation of 4.8 mph. At the 10% level, is there evidence that there is any difference in speeds, on average, during morning and evening rush hour?

   a. Hypothesis: $H_o: \mu_{Morning} - \mu_{Evening} = 0$ vs. $H_a: \mu_{Morning} - \mu_{Evening} \neq 0$
   b. Use pooled variance since 4.8/3.6 < 2.

   c. Put values in calculator

$\bar{x}_1$: 47.3 S_{x1}: 3.6 n_1: 10 $\bar{x}_2$: 43.2 S_{x2}: 4.8 n_2: 12 μ_1: $\neq \mu_2$ $< \mu_2$ $> \mu_2$ *Pooled*: *No* *Yes*	d. p-value: 0.0376 e. Decision: Since 0.0376 < 0.1, reject the null hypothesis f. Conclusion: There is strong evidence that the average speed of cars during morning and evening rush hour differs.

6. The speed of cars on a particular road is known to be normally distributed. A sample of 15 cars taken during the morning rush hour revealed a sample average of 50 mph with a standard deviation of 2 mph. A sample of 20 cars taken during the evening rush hour had a sample average of 52.5 mph with a standard deviation of 6 mph.

   a. At the 5% level, is there evidence that people drive faster on average during the evening rush hour?
      i. Hypothesis: $H_o: \mu_{Morning} - \mu_{Evening} = 0$ vs. $H_a: \mu_{Morning} - \mu_{Evening} < 0$
      ii. Do not use pooled variance since 6/2 > 2.
      iii. Put values in calculator:

$\bar{x}_1$: 50 S_{x1}: 2 n_1: 15 $\bar{x}_2$: 52.5 S_{x2}: 6 n_2: 20 μ_1: $\neq \mu_2$ $< \mu_2$ $> \mu_2$ *Pooled*: *No* *Yes*	iv. p-value: 0.0473 v. Decision: Since 0.0473 < 0.05, reject the null hypothesis vi. Conclusion: There is strong evidence that the average speed of cars during morning and evening rush hour differs.

b. Find a 90% Confidence interval on the average difference between morning and evening rush hour speeds?

i. Put values in calculator:

$\bar{x}_1$: 50 S_{x1}: 2 n_1: 15 $\bar{x}_2$: 52.5 S_{x2}: 6 n_2: 20 $C - Level$: 90 $Pooled$: *No* Yes	ii. CI: [-4.958 mph, -0.0418 mph] iii. Conclusion: Average speeds in the evening are at least very slightly faster than in the mornings and this difference might be up to almost 5 mph.

7. A company assembles large electronic devices. The assembly method they use is rather simple and has been used for a long time. Recently, the company has become aware of a new assembly technique. The company is basically satisfied with their current method of assembly. However, if the company finds that the new method significantly reduces assembly time, they would adopt this new method. The company decides to do a test to compare the two methods over the course of a day. One production line will use the current method and the other production line will try the new method. The data (in minutes) is shown above. Should the company adopt the new method? Assume the assembly times for both methods are normally distributed. Use the p-value approach with a significance level of 10%.

Current	47.4	57.3	39.8	60.4	46.7	36.4
New	37.3	55.9	45.4	52.2	39.1	34.4

i. Hypothesis: $H_o: \mu_{Current} - \mu_{New} = 0$ *vs.* $H_a: \mu_{Current} - \mu_{New} > 0$

ii. Put the data into two lists in you calculator.

iii. Find the numerical descriptive statistics for each data set. (Stats/Calc/1/L1, then Stats/Calc/1/L2). Write down the standard deviations: 9.424 and 8.624.

iv. Check to see if it is OK to assume equal variances: $\frac{9.42}{8.62}$=1.1<2. Since the ratio is less than 2, it is OK to assume equal variances.

v. Put values into calculator:

$List1: L_1$ $List2: L_2$ $Freq1: 1$ $Freq2: 1$ $List1: L_1$ μ_1: $\neq \mu_2$ $< \mu_2$ $> \mu_2$ $Pooled$: No Yes	vi. p-value: 0.233 vii. Decision: Since 0.233 > 0.1, fail to reject the null hypothesis viii. Conclusion: The company should not adopt the new method. There is not statistical evidence that the new method reduces the average assembly time.

8. A survey of recent graduates with degrees in marketing and business revealed the following starting salaries (in thousands of dollars).

Marketing	24.6	27.6	25.8	20.8	23.1	25.2	24.7	26.1	22.6	23.8					
Business	26.7	24.9	27.1	23.1	27.5	27.4	24.9	26.3	28.9	26.4	28.1	29.7	25.5	24.9	28.5

a. Is there evidence at the 5% significance level that Business majors have higher average starting salaries than Marketing majors (assume salaries are normally distributed? Use the p-value approach.

The descriptive statistics obtained from these samples are shown below.

Variable	N	Mean	Median	StDev
Marketing	10	24.43	24.65	1.946
Business	15	26.66	26.7	1.786

Check to see if it is *OK* to assume equal variances: $\frac{1.946}{1.786} = 1.1 < 2$. Since the ratio is less than 2, it is OK to assume equal variances.

i. Hypothesis: $H_o : \mu_{Business} - \mu_{Marketing} = 0$ *vs.* $H_a : \mu_{Business} - \mu_{Marketing} > 0$

ii. p-value: 0.00358

iii. Decision: Since 0.00358 < 0.05, reject the null hypothesis

iv. Conclusion: There is strong evidence that the average starting salary for Business majors is higher than Marketing majors.

b. Use Minitab to evaluate the hypothesis and confidence interval.

Two-sample T for Business vs Marketing

	N	Mean	StDev	SE Mean
Business	15	26.66	1.79	0.46
Marketing	10	24.43	1.95	0.62

```
Difference = mu (Business) - mu (Marketing)
Estimate for difference:  2.230
95% lower bound for difference:  0.935
T-Test of difference = 0 (vs >): T-Value = 2.95  P-Value = 0.004
DF = 23
Both use Pooled StDev = 1.8507
```

We see that the p-value, 0.004 is less than the significance level, 5%, so we reject the null hypothesis. We also see that we are 95% certain that the true value of μ is at least 0.935 ($935). Since $\mu_o = 0$ is not in the confidence interval, this also leads us to reject the null hypothesis.

c. Using the Minitab results above, can we make the claim that the true mean difference exceeds $500? $1000? Yes and No.

Homework 8.4

1. The spending habits of residents of Valdosta at the Outlet Mall is being compared to people who are from out-of-state that stopped in to shop. It is hypothesized that there is no difference between the two groups of consumers. The results of the study are shown in the table below. (a) Use the p-value approach at the 5% significance level to evaluate the hypothesis. (b) Find a 95% confidence interval and interpret it.

Consumer	n	$\bar{x}$	s
Valdosta resident	18	\$24.27	\$13.26
Out-of-state	12	\$32.25	\$33.71

2. A study is done to determine if professors live further away from campus than students. A simple random sample of 5 professors and 5 students is selected and each subject is asked how long it takes them (in minutes) to drive to school. The results are shown in the table below. (a) Use the p-value approach at the 10% significance level to evaluate the hypothesis. (b) Find a 90% confidence interval and interpret it.

Professors	15	22	16	14	15
Students	17	8	6	12	15

3. A new flight route has been proposed for flights from La Guardia International Airport in New York City to Heathrow International Airport in London, England. It is hoped that this new route will reduce the overall flight time by at least 30 minutes. To determine if there is evidence that the new route does reduce the flight time, an analysis of the old route and the new route is done. The statistical results are shown below (the units of measurement are hours). Evaluate these results as completely as possible.

Two-Sample T-Test and CI

```
Sample  N   Mean  StDev  SE Mean
Old     7  8.071  0.368     0.14
New     6  7.283  0.471     0.19


Difference = mu(Old) - mu(New)
Estimate for difference:  0.788
90% lower bound for difference:  0.471
T-Test of difference = 0 (vs >): T-Value = 3.39  P-Value = 0.003  DF = 11
Both use Pooled StDev = 0.4180
```

8.6 – Dependent Samples Inference

8.5.1. Introduction

Sometimes a situation arises where we have two ***dependent*** samples that we want to use to draw inference on the difference in population means. Since the samples are dependent, we can't use the methods of previous sections. First, what does *dependent* mean? It means that knowledge of one value in a sample tells you something about a value in the other sample.

Example: Our goal is to determine if an SAT preparation course increases SAT scores. Suppose we choose 30 people at random from the population (*e.g.* all 9th graders who have not taken the SAT nor any preparation courses) and give them the SAT. Next, we administer the SAT preparation course over the next six weeks to the *same* students. Following this, we give the SAT again, to the same students. This is called a *matched pairs design* and an example is shown below.

Person	*Before* Score	*After* Score	*Difference* After-Before
Dave	630	720	90
Sue	1180	1240	60
Delia	1020	1070	50
⋮	⋮	⋮	⋮

This is an example of the case where the two samples are dependent. The data come in *matched pairs*, a *before* score and a *after* score for each person. For instance, note that Sue's score is relatively high in the *before* sample. Thus, we would expect Sue's score to be relatively high in the *after* sample. This is dependence. Similarly, if Dave is at the bottom of the *before* sample, he will probably still be at the bottom of the *after* sample.

The technique we use to do a hypothesis test with *matched pairs (dependent)* data is called a *paired t-test*. This technique is preferable to the techniques considered earlier when we have independent samples because the precision is increased.

Example: Suppose, again, that our goal is to determine if an SAT preparation course increases scores. However, this time we choose two *independent* samples of 30 people each from the same population. The first group is administered the SAT with no preparation course. The second group is administered a preparation course and then the SAT as shown in the table below.

Before Prep. Course		*After Prep. Course*	
Person	**Score**	**Person**	**Score**
Dave	940	Sharon	1260
Sue	1180	Mike	820
Delia	1020	Steve	970
⋮	⋮	⋮	⋮

So, effectively, we now have two populations, those who have not had the preparation course and those who have. Now, consider comparing these two populations statistically. What influences the *precision* of results in a confidence interval or hypothesis test? One thing is the *variability*; more variability means less precise results. For this example, what are the *sources of variability*? What factors contribute to the total variability, σ_{Total} that we observe? These can be grouped into three categories:

1. **Controllable Factors**: variability due to the individual people involved. Each person is different and will score at different levels on the test. It may seem strange that we call this a *controllable* factor, but we will see shortly that it is.
2. **Study Factor**: variability due to the effectiveness of the preparation course. This is what we are trying to study.

3. **Uncontrollable/Environmental Factors**: (a) day the test is administered, (b) time test is administered, (c) location of testing facility, *etc*.

All these factors contribute to the total variability we see in scores. In some sense:

$$\sigma_{Total} = \sigma_{Controllable} + \sigma_{Study\ factor} + \sigma_{Uncontrollable}$$

These factors *add* to the error in the study by increasing the *total variability*, making inference less precise. With a *matched pairs design* and analysis with a *paired t-test*, we can eliminate, to some degree, the effect of the controllable factor.

When we compute the *difference* in the *after* and *before* scores (see table on previous page), something very important happens. When we subtract the *before* scores from the *after* scores, we are literally (in some sense) subtracting out the *variability due to the individual people*. This reduces the *total variability* which in turn makes the inference more precise.

We now have *one* sample, the sample of differences and this sample contains only variability due to the preparation course itself (and the *uncontrollable* factors) but not variability due to the individuals (to some degree).

So, why is *dependence* such a big deal? Often, this can lead to a smaller sampling error, which leads to more precision in statistical inferences we make. In other words, we can reduce the *total variability* by using a *matched pairs design.*

If we want to compare two populations, we should always try to arrange things so that we have a match pairs design. This is, of course, not always possible.

Other examples. What source of variability has been removed or at least controlled?

1. Which brand of rubber lasts longer on tennis shoes, A or B?
2. Which variety of tomato plant produces more tomatoes?
3. How fast does a forest regenerate itself after a fire?
4. Is there any difference between two different brands of caliper?
5. Does a fuel additive increase gas mileage?

8.5.2. Data, Definitions and the Paired t-test

We will use the terminology, *before* and *after,* though as you have seen from some of the examples this isn't always the scenario.

The *after* sample: $A_1, A_2, \cdots, A_n$

The *before* sample: $B_1, B_2, \cdots, B_n$

The sample of differences: $\{d_1, d_2, \cdots, d_n\} = \{A_1 - B_1, A_2 - B_2, \cdots, A_n - B_n\}$

Once we compute the differences, we have *one* sample, a sample of differences. Thus, we make statistical inferences about μ_d, the true mean difference. The way we do this is exactly the same way we handled inference for one population earlier.

The estimator is: $\bar{d} = \frac{1}{n}\sum_{i=1}^{n}(A_i - B_i) = \frac{1}{n}\sum_{i=1}^{n} d_i$, which is just the average of the differences.

The *standard error* of the estimator - We use s_d to represent the standard deviation of the sample of differences. This is just the usual standard deviation but it is computed from the sample of differences, not the actual data. So the standard error of the estimator is: $s_{\bar{d}} = s_d/\sqrt{n}$ which is exactly the same as earlier when we talked about inference on a single mean, and as shown below, all hypothesis testing is the same.

$$H_o: \mu_d = d_o \text{ vs. } H_a: \mu_d > d_o$$
$$H_o: \mu_d = d_o \text{ vs. } H_a: \mu_d < d_o$$
$$H_o: \mu_d = d_o \text{ vs. } H_a: \mu_d \neq d_o$$

Use Calculator:

CI: Stats/Tests/ #8 TInterval
HT: Stats/Tests/ #2 TTest

Note, use $\bar{d}$ for $\bar{x}$ and s_d for Sx.

8.5.3. Example

1. A group of 10 people is selected to evaluate a well know product. Each person was asked to rate on a scale of 1-10 how likely they are to buy the product where higher scores mean a stronger likelihood of purchasing the product. The people were then shown a commercial for the product and then asked again to rate how likely they were to buy the product. The results are shown in the table below. If there is evidence that the mean purchase rating has increased after viewing the commercial, then the commercial will be aired on TV.

Person	1	2	3	4	5	6	7	8	9	10
Before	4	4	3	7	3	4	8	5	4	3
After	5	6	6	8	5	7	7	9	7	8

   a. Use the p-value approach to determine if the commercial should be aired. Use 10% significance.

      i. Define difference to be *After - Before*.

      ii. Hypothesis: $H_o: \mu_d = 0$ vs. $H_a: \mu_d > 0$

      iii. Compute sample of differences:

Differences	1	2	3	1	2	3	-1	4	3	5

      iv. Put values in a list in your calculator.

      v. Choose Stats/Tests/T-Test(#2)

      vi. p-value = 0.001

      vii. Decision: Since 0.001 < 0.1, reject the null hypothesis

      viii. Conclusion: The commercial should be aired. There is evidence that the mean purchase rating

increased after viewing the commercial.

b. Find a point estimate for the mean difference in purchase potential rating.

This is just $\overline{d}$. On your calculator, choose: Stats/Calc/1-var stats. The result is $\overline{d} = \overline{x} = 2.3$.

c. Find a 90% confidence interval on the mean difference in purchase potential rating.

i. Choose Stats/Tests/TInterval(#8)
ii. Confidence interval: [1.3, 3.3]

Homework 8.5

1. A new home weight scale has been developed that is much cheaper than a model currently being sold. A statistical test is conducted to determine if there is any difference between the new scale and the existing scale. 5 women are selected at random and weighed with each of the two scales and the results are shown in the table below. (a) What factor is being controlled for by using a matched pairs design. (b) Use the p-value approach at the 10% level to evaluate the hypothesis. (c) Find a 90% confidence interval on the true mean difference between the weights reported by the new scale and the existing scale.

Person	1	2	3	4	5
New Scale	124	103	92	141	118
Existing Scale	126	100	96	143	117

2. A study is undertaken to determine which of two varieties of tomato plants produces more tomatoes. 20 locations are selected from around the country. At each location, a *Ruby Red* variety tomato plant is planted and an *Oppulent Orange* variety plant is planted. At the end of the growing season, the total number of tomatoes is tallied for each plant in each location. The average difference (Oppulent Orange - Ruby Red) was 4.85 tomatoes and the standard deviation of the differences was 3.167. (a) What factor is being controlled for when we use this experimental design? (b) Is there evidence that the Oppulent Orange variety produces more tomatoes, on average than the Ruby Red variety using 5% significance? (c) Find and interpret a 95% confidence interval on the true mean difference

8.7 – 2-Sample Proportion Inference

Here we consider inference on the difference in two (true) proportions: $p_1 - p_2$. On the calculator, we

Confidence Interval: Stat/Test/2-PropZInt

Hypothesis Test: Stat/Test/2-PropZTest

Example: Students were surveyed to determine the percentage that went to the beach for spring break. 587 students were sampled from colleges in the Mid West where it was found that 212 were going to the beach. 634 students were sampled from colleges in the South East and 242 were going to the beach.

a. Can we claim that there a larger proportion of students in the south-east go to the beach for spring break as compared to the proportion from the mid-west?

$$H_o: p_{South\ East} - p_{Mid\ West} = 0 \quad \text{vs.} \quad H_a: p_{South\ East} - p_{Mid\ West} > 0$$

$$p - value = 0.2290$$

Thus, fail to reject the null hypothesis and conclude that there is not enough evidence to say that the true proportion of students in the South East who go to spring break at the beach is larger than for students in the Mid West.

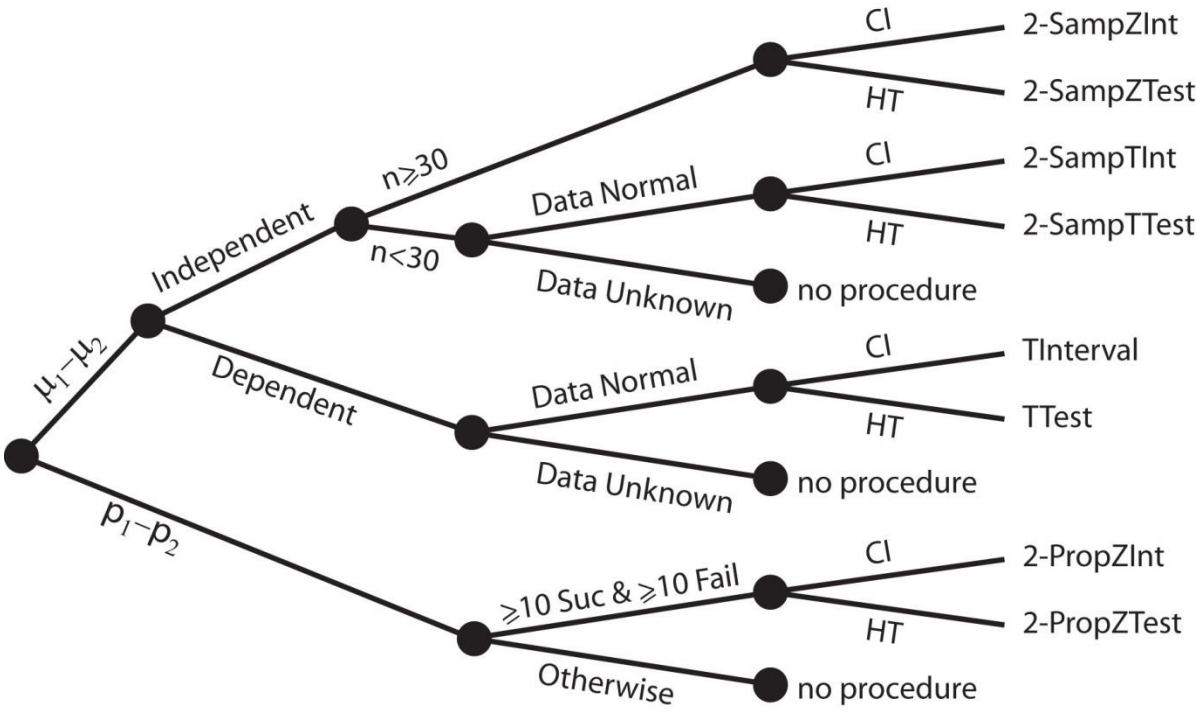
b. Use the confidence interval approach. A 90% confidence interval on $p_{South\ East} - p_{Mid\ West}$ is: (-0.025, 0.066) = (-2.5%, 6.6%). Since zero is in this confidence interval we conclude that there is no evidence of a difference between the two areas of the country.

c. Find the margin of error for confidence interval above. The margin of error is

$$E = \frac{0.066 - (-0.025)}{2} = 0.0455 = 4.55\%$$

8.8 – 2-Sample Inference Summary

The figure below shows a summary of the types of 2-sample inference we made and the tool to use on the calculator.



Chapter 9 – Correlation and Regression

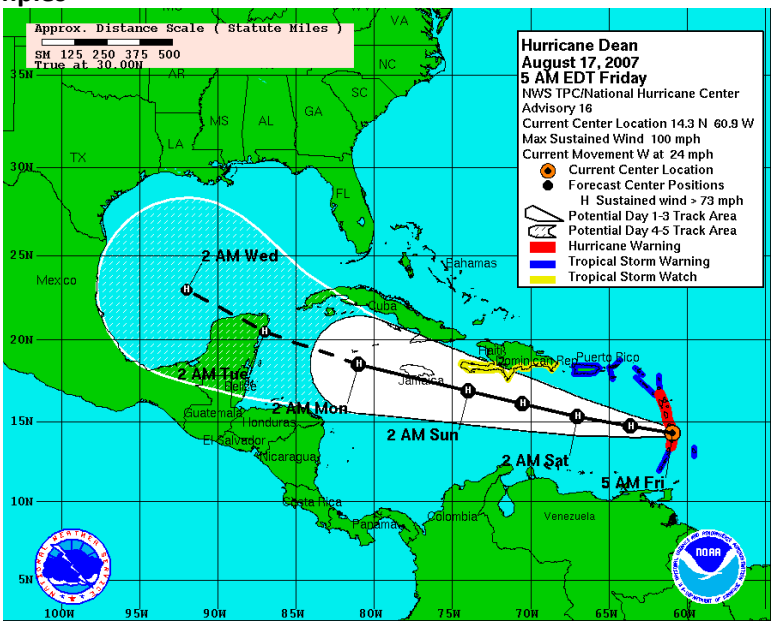
In this chapter, we begin to investigate the situation where we have two variables (or more) and there is a relationship between them.

9.1 – Introduction

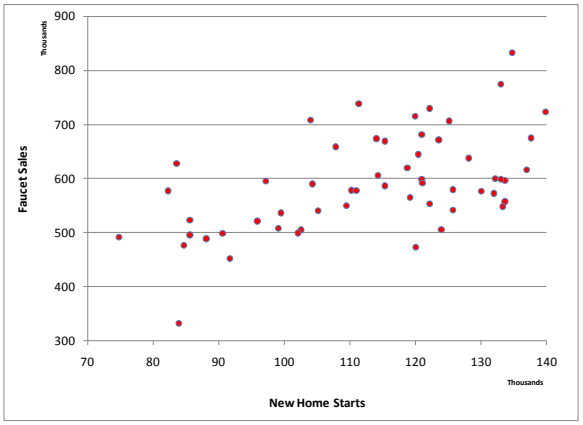
In this chapter, we begin to investigate the case where we have two variables (or more) and there is a relationship between them. Often, we are interested in questions such as, "given the value of one variable, can we predict the value of the other?" In general, we refer to this situation as *regression*. For example, can we predict:

1. ...power usage given the outside temperature?
2. ...the stopping distance given the speed of a car?
3. ...sales given television advertising expenditures?
4. ...the number of customers an ad will attract given the size of the ad?
5. ...someone's starting salary given their college GPA?
6. ...someone's college GPA given their high school GPA?
7. ...the width of flood plain given rainfall?
8. ...storm surge given the amount of rainfall?
9. ...the path of a hurricane given certain environmental conditions?
10. ...the number of mortgage foreclosures at a bank in the next month given the number of foreclosures this month? given interest rates?

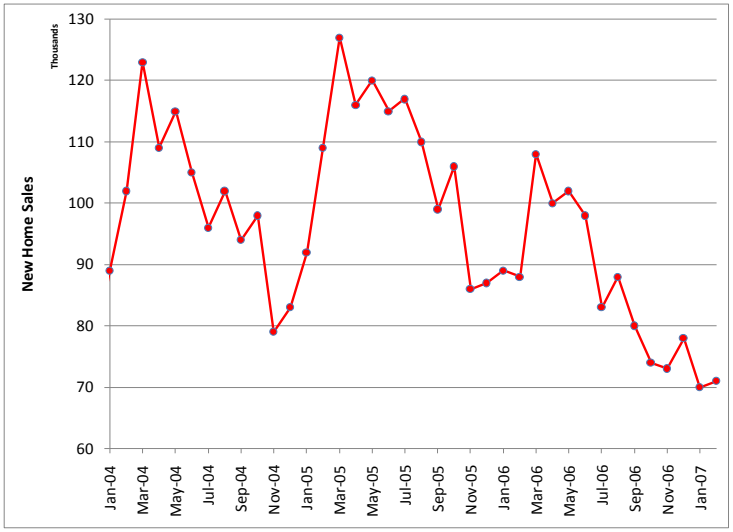
1. Examples



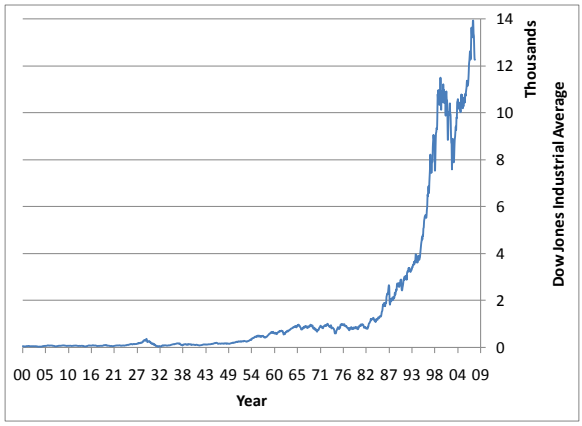
Bivariate data involves two variables that have a relationship between them. For instance, the scatter plot below shows how many homes that were started under construction and the resulting number of faucets sold in a month. This is real data from a large manufacturer of faucets. This data shows that, generally the more homes started, the more faucets sold. But, could we use this information to predict how many faucets would be sold in a month if we knew how many new homes were going to be started?

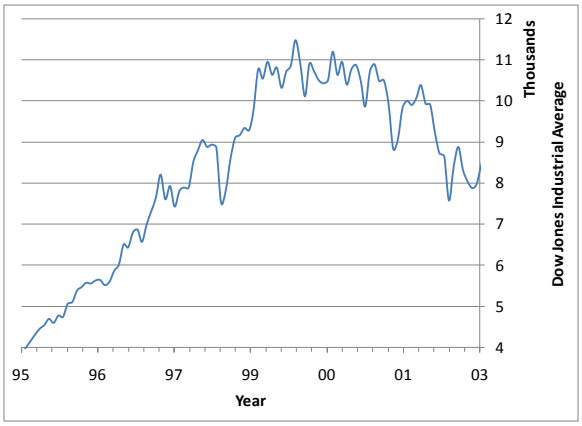


The data shown in the scatter plot below is also bivariate. However, when time is involved, we usually refer to this as a special type of bivariate data called a *time-series*. When we look at a time series, we look for *trend* and *oscillation*. In the scatter plot, we have added a line to connect the points which helps with identifying trend and oscillation. We will not study time-series' any further.



The data below shows the Dow Jones Industrial Average (DJIA) since its inception and the years of the *Internet Bubble*. These are also times-series graphs.



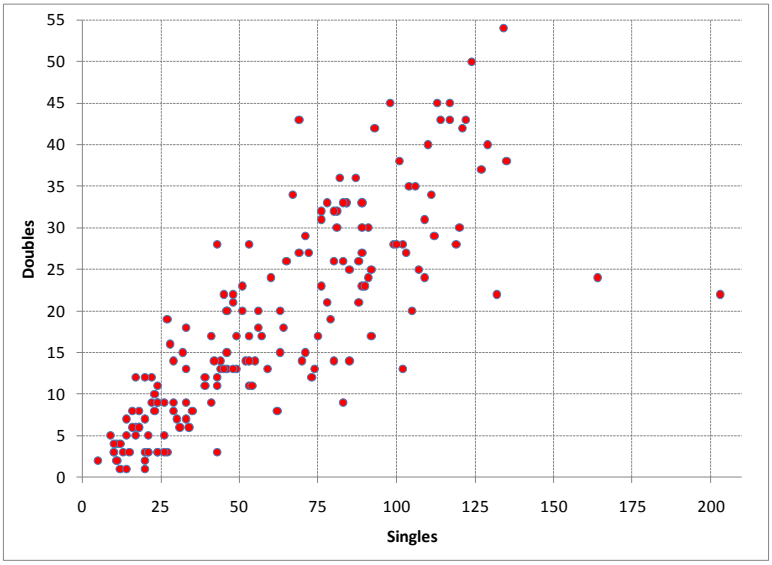


For our purposes, *regression* refers to *the process of using past data to come up with the equation of the best line that fits the data*. We will refer to this *best line* as the *regression equation*. We will refine this definition as we go along.

Suppose that we have recorded the number of singles and the number of doubles for the 173 outfielders in MLB who had at least 50 at-bats for the 2007 season.

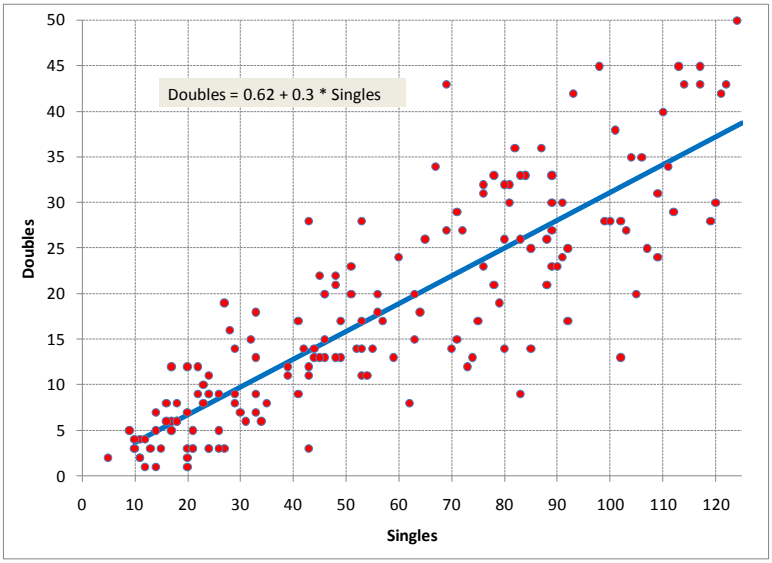
Player	Singles	Doubles
Ichiro Suzuki	203	22
Juan Pierre	164	24
Delmon Young	135	38
Alex Rios	117	43
Jeff Francoeur	129	40
Nick Markakis	122	43
Matt Holliday	124	50
Grady Sizemore	111	34
...	...	...

When we have data like this, we often make a *scatterplot* as shown in the figure below.



Suppose someone tells you that a player has 75 singles, can you predict how many doubles they have? Of course for a specific player with 75 singles, you could find the answer on the internet, but the question here is could we *predict* the number of doubles for a player that has 75 singles? Using the scatter plot above you could come up with some type of guess.

However, analyzing this data using *regression* yields the blue, *regression line* in the figure below. (Note: the two extreme points in the scatterplot above, 164 singles and 203 singles have been removed. More on this later).



Thus, the regression equation says that a player with 75 singles will have about:

$$\begin{aligned} Doubles &= 0.62 + 0.3 * Singles \\ &= 0.62 + 0.3 * 75 \\ &= 23.12 \approx 23 \end{aligned}$$

Best Line – An interesting question is what does, "...best line" mean? In other words, how exactly do we define *best*? Looking at the figure above, you probably have an intuitive feel for what *best* means. However, to use mathematics to figure out the equation of the line, we must explicitly define what *best* means. There are many choices. However, we choose one that is very convenient and useful mathematically, and seems to fit our world rather nicely, from an empirical point of view. *Best* is frequently defined as *the line which minimizes the sums of squared distances between each data point and the line*. So, you can imagine placing the regression line on the scatter plot. Then, drop a straight line from each data point to the regression line and measure the distance for each. Now, pivot the regression line around until the each distance, squared, then summed is as small as possible.

Terminology

The *X* variable in regression is referred to as the *independent variable*, the *explanatory variable*, the *predictor* or *predictor variable.*

The *Y* variable in regression is referred to as the *dependent variable* or the *response*.

For instance, if we wanted to predict starting salaries (*Y*) given a student's GPA (*X*), we could say, salary *depends* on GPA, GPA *explains* salary, GPA is a *predictor* of salary, salaries *respond* to GPAs.

9.2 Correlation

Positive Correlation - when one variable is large (small) the other variable *tends* to be large (small). The variables *move in the same direction*.

Examples: height and weight, income and mortgage payment, flow rate and amount of eroded soil, outside temperature and cost to cool house, distance between two cities and time to drive between them.

Negative Correlation - when one variable is large (small) the other variable *tends* to be small (large). The variables

move in opposite directions.

Examples: hours worked and hours spent with family, SAT scores and the percentage who take the SAT, alcohol consumption and ability to drive a car, speed of car (*mph*) and fuel efficiency (*mpg*).

No Correlation - when two variables do not influence one another. Changes in one variable do not effect changes in the other variable.

Examples: height and salary, heart rate and money spent at grocery store.

Sometimes combinations of these correlations are at work over the range of the independent variable.

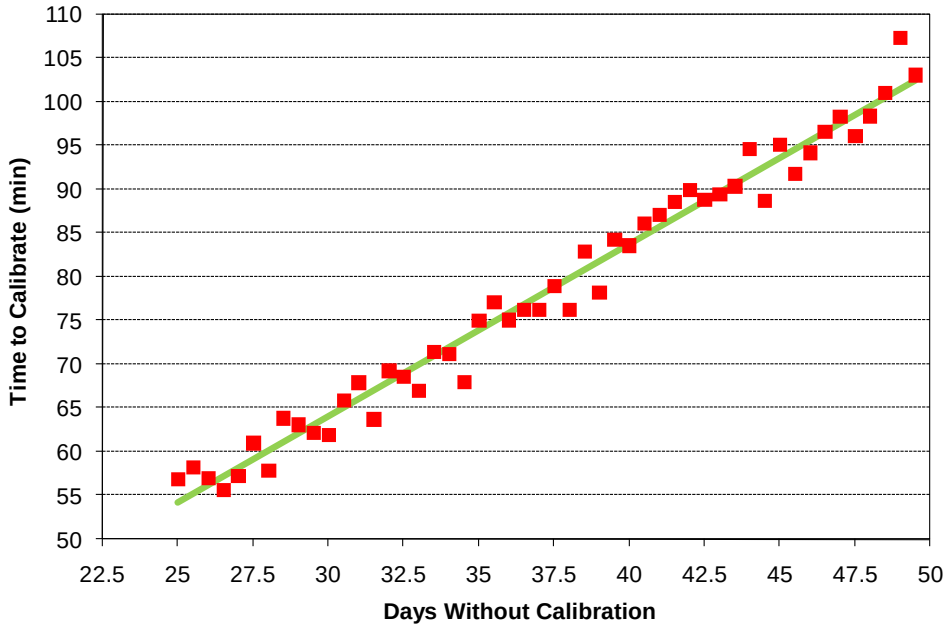
Examples: The relationship between the amount of water and the number of tomatoes that result exhibits both positive and negative correlation. The more water the more tomatoes, but only to a point (positive correlation). At some point, the more water you give tomatoes, the fewer tomatoes you get (negative correlation).

9.3 Scatter Plots

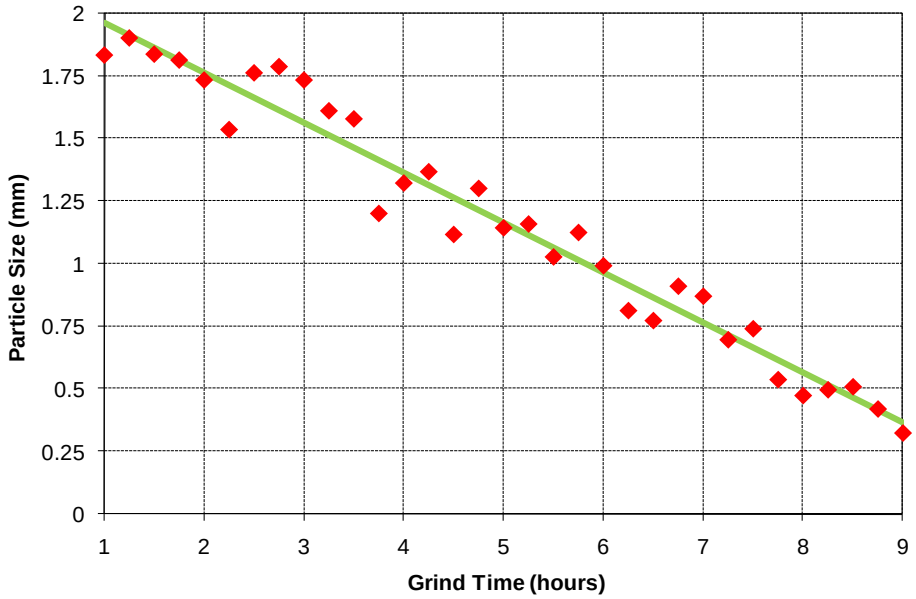
Scatter Plot – The first step in investigating bivariate data is to make a scatter plot of the data. Do this by plotting the independent variable along the *x*-axis and the dependent variable along the *y*-axis.

Examples

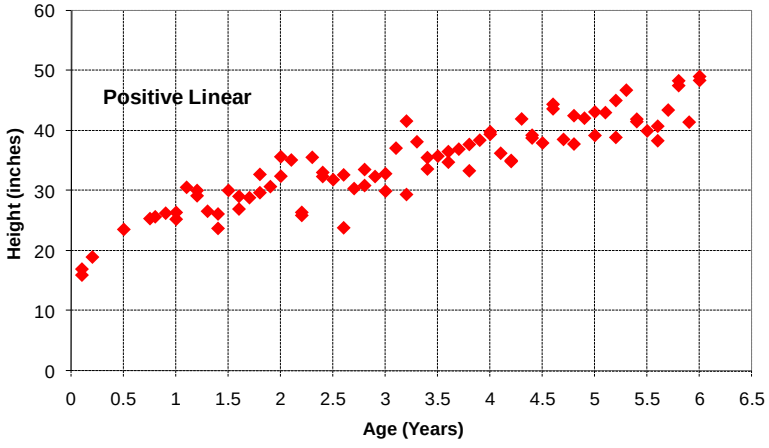
A. Consider the situation where you have a machine in a shop that performs an important function. Every so often, the machine must be taken off-line and calibrated. Plant managers have noticed that the longer it has been since the last calibration (in days), the longer it takes to calibrate the machine (in minutes). A scatter plot of the collected data is shown below.



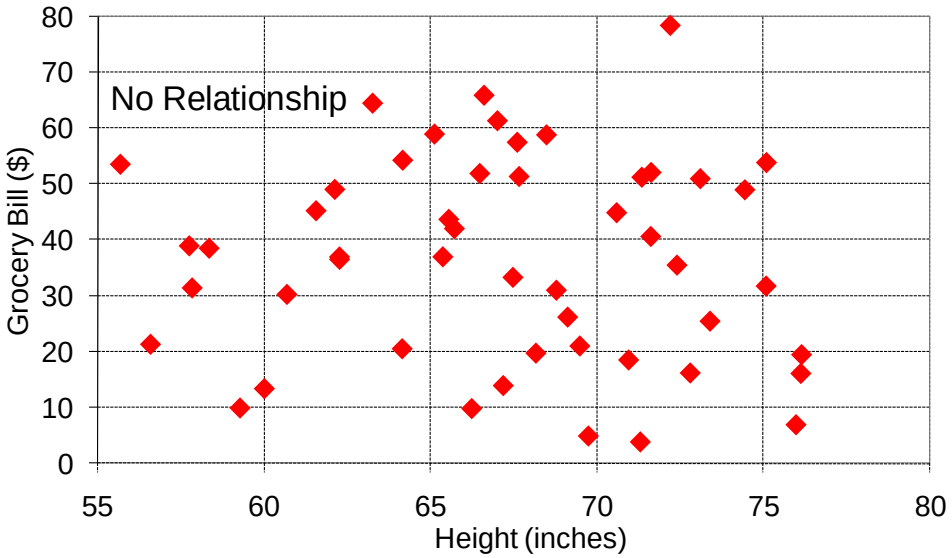
B. To make fertilizer and other agricultural chemicals for farming operations, several chemicals are mixed together in a huge vat and ground together. If the chemicals are ground to finely, then they will tend to stick together and clog up the sprayer on the back of a tractor. Similarly, if the particles are not ground fine enough, the resulting mixture will have particles that are too thick to be sprayed through a nozzle. The company would like to be able to describe the relationship between grind time (in hours) and particle size (in millimeters). A scatter plot of the collected data is shown at the right.



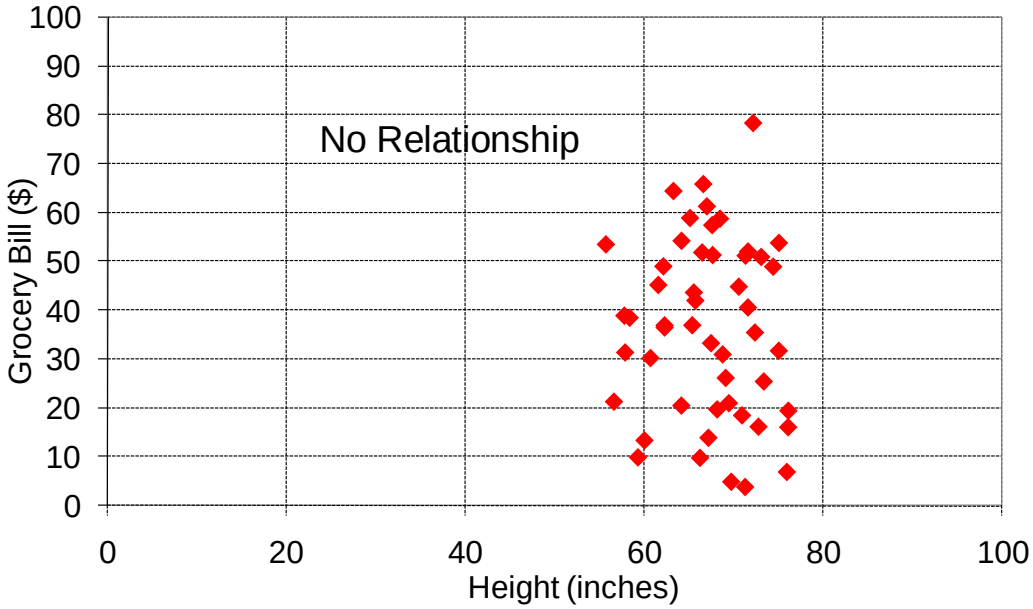
C. What is the relationship between age and height? Obviously, as a child gets older, they get taller and at some point they stop growing. What would this graph look like over the life of a person? Over a certain range of years, it has been shown that children's growth is fairly linear (this means that they are growing at a constant rate). We probably know from experience that when children are very young, the first 6 months or so, they tend to grow very fast. Both of these observations can be seen in the graph shown below.



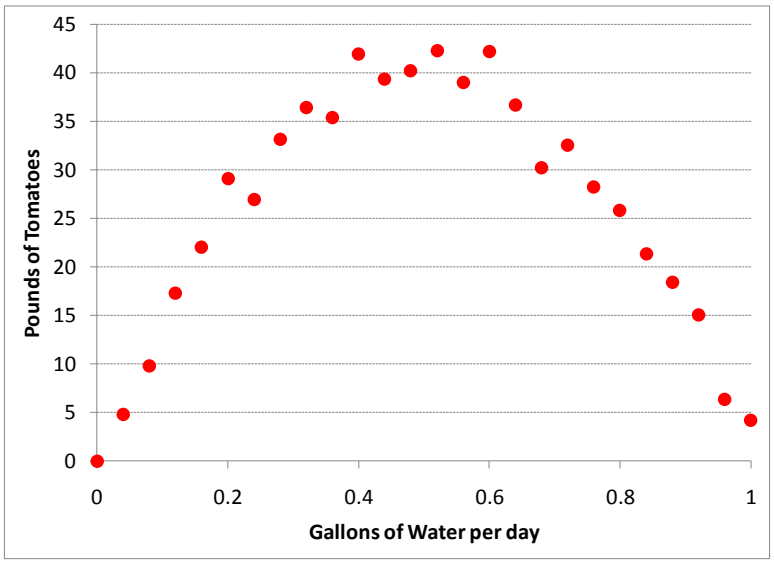
D. Single people shopping in a grocery store were asked their height (in inches) and the total amount of their grocery bill (in dollars). As we might expect, a scatter, as shown on the right does not reveal any relationship. Regression is not appropriate in this situation. Note, however, that we are making this judgment visually. In other words, we could be tricked. In situations that are more important than this, we would use more sophisticated techniques to determine if there is a relationship.



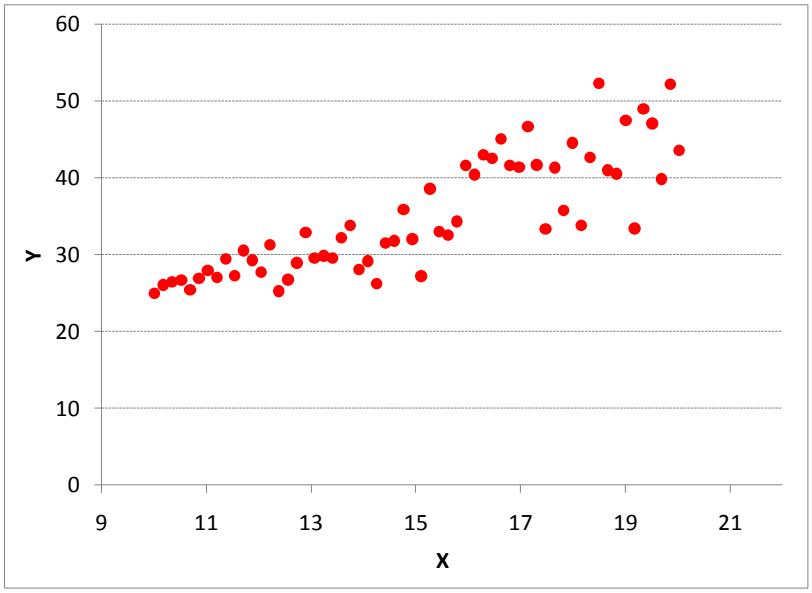
E. This graph is the same data as the previous graph and again, we are seeing the effect of changes in scale. Regardless, we still see no relationship in the data. Can you explain why?



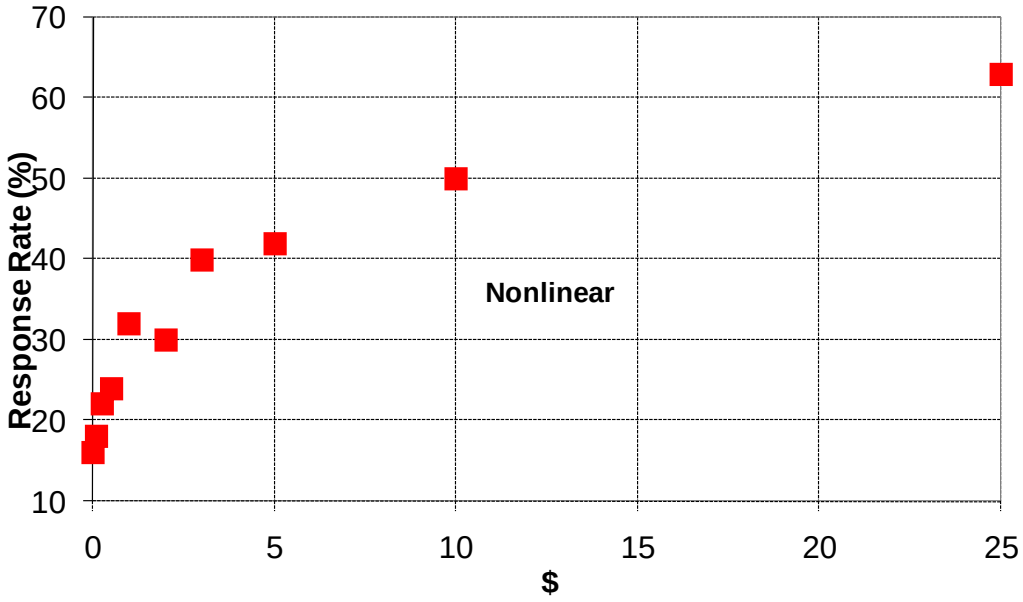
F. The number of tomatoes on a plant is certainly influenced by the amount of rainfall. The interesting thing about this relationship is that there is some *optimal* amount of rainfall that will produce the most tomatoes. In other words, too little water or too much water will reduce the number of tomatoes. The scatter plot at the right is a hypothetical example of this relationship. Note that industry is very interested in these types of situations.



G. The hypothetical data shown in the scatter plot on the right shows a situation that occurs in real life some times: *unequal variance*. What we see is that the larger the *x* value is, the more variable the *y* value is. This is a very important thing to observe and note when present. Special techniques must be used in this situation.



H. The data shown on the right is from a famous experiment done by psychologists in France. People were sent a meaningless questionnaire and asked to fill it out. The response rate (measured in percents) was calculated and recorded. Some people were also sent money as an incentive to fill out the questionnaire and again, response rates were recorded. What are some of the implications of this graph?



9.4 TI-83 Directions, Scatter Plot

Enter Data

1. Put X data in L1 and Y data in L2. To start entering data, choose Stat/Edit and then Enter.

Scatter Plot

1. Choose 2nd + Stat Plot.
2. Select one of the plot numbers, for instance the first, 1.
3. Select: *On*. Press Enter.
4. Choose the Type: select **first** graph from the six shown. Press Enter.
5. Enter the Xlist (2nd + L1, for instance) – This is the explanatory variable.
6. Enter the Ylist (2nd + L2, for instance) – This is the response variable.
7. Choose a Mark
8. Choose Zoom/9
9. Trace (use left/right arrow keys)

Homework 9.1

4. Consider the four data sets shown below. Notice that the *x* values are the same for each data. Thus, you only have to enter that list one time in your calculator. (a) Make a scatter-plot of each data set. (b) comment on the correlation you see in each scatter-plot.

Data Set 1		Data Set 2		Data Set 3		Data Set 4	
X	Y	X	Y	X	Y	X	Y
5	11	5	11	5	29	5	32
6	15	6	8	6	29	6	28
7	18	7	29	7	30	7	18
8	22	8	21	8	23	8	11
9	23	9	14	9	45	9	20
10	25	10	36	10	31	10	20
11	29	11	19	11	24	11	15
12	33	12	17	12	18	12	10
13	35	13	12	13	23	13	10
14	40	14	17	14	19	14	18
15	41	15	23	15	36	15	21
16	45	16	37	16	14	16	20
17	47	17	15	17	21	17	32
18	50	18	21	18	29	18	49
19	53	19	19	19	8	19	47
20	57	20	19	20	11	20	69

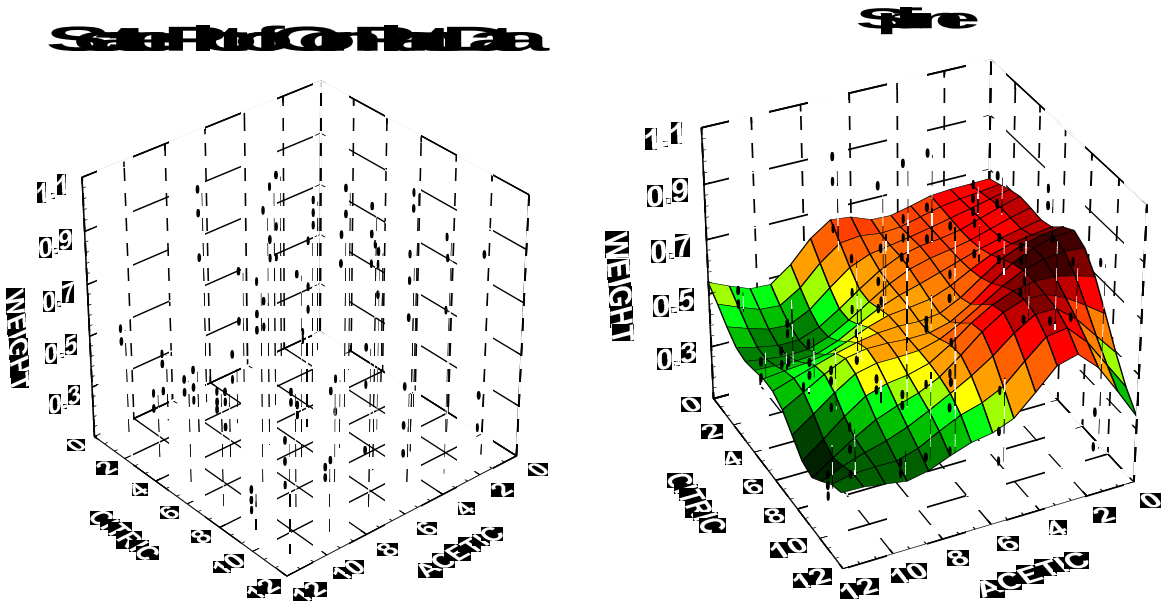
9.5 Regression

Simple Linear Regression – We will study *simple linear regression* in the remainder of this chapter. *Simple* means that we have *one* independent variable. *Linear* means that we are going to find the *best straight line* that fits through the data. *Regression* refers to the process of *finding* that line, the mathematical technique used. These are the steps that we will follow to do simple linear regression:

1. Make a scatter plot of the data
2. Check to see if a straight line may "fit" the data. If so, proceed to the next step; otherwise, stop. We will not consider data with any type of curvature.
3. Build regression line/equation.
4. Check to see if line is statistically significant. If so, proceed to the next step; otherwise, stop.
5. Check to see how good the line is at predicting things. If it is good, then we may use the regression line with confidence. Otherwise, we may not use the regression line, even though it is statistically significant.
6. Investigate line further.

Multiple Regression – The term *multiple regression* refers to the situation where we have more than one independent variable, as shown in the corn plant example below. There, we consider testing to see how effective a weed killer is. The active ingredients in the weed killer are citric acid and acetic acid. There are two independent variables in this situation: the *percent acetic acid* and the *percent citric acid*. The response is the *weight* of a plant. Consider spraying a plant with a combination of citric acid, acetic acid, and water. This will kill individual leaves where it is sprayed, eventually killing the plant, depending on the concentrations of the two acids. To measure the effectiveness of different concentrations of the acids for killing a plant, we weigh the plants after they have been sprayed and left for a few days. Low weights mean the plant has been killed (there is very little moisture left in the plant). High weights indicate that the plant has not suffered much damage and is still living.

Example: The scatter plot on the left is very hard to look at. In real life situations we almost always have 2 or more (up to 10 or so) independent variables. Thus, in higher dimensions it is very hard to visualize the data and get an intuitive feel for the relationship among the variables. The graph on the right shows a *surface* that has been fit to the data using a regression technique.



9.6 Simple Linear Regression

1. Model

To do regression, we must first *assume* an underlying model that we think may describe the data (the nature of the relationship). Then, we *fit* the data to the model. For simple linear regression, we assume the true, unknown relationship in the data is: $y = \beta_o + \beta_1 x$. In other words, a straight line whose *y*-intercept (β_o) and slope (β) are unknown. We call this the *model* for simple linear regression.

Note that this is exactly the equation of a straight line that you learned in an algebra course: $y = mx + b$. We are simply placing the terms in a different order and giving different names to the y-intercept, $\beta_o = b = y\ intercept$ and slope, $\beta_1 = m = slope$.

Fitting the data to the model, in our case reduces to using the data to *estimate* the slope and y-intercept β_o and β_1. Remembering that in statistics, whenever we estimate something, we put some sort of symbol on top of it, we will denote the estimated y-intercept as $\hat{\beta}_0$ and the estimated slope as $\hat{\beta}_1$. Thus, we denote the actual *regression line* as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The hat on the *y* indicates that when we plug a value into the regression equation, the value we get is an estimate. Sometimes, we refer to the regression equation as "y-hat."

	One Population	Comparing Two Populations	Relationship between Two Populations
Model	μ	$\mu_1 - \mu_2$	$y = \beta_o + \beta_1 x$
Estimator	$\bar{x}$	$\bar{x}_1 - \bar{x}_2$	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

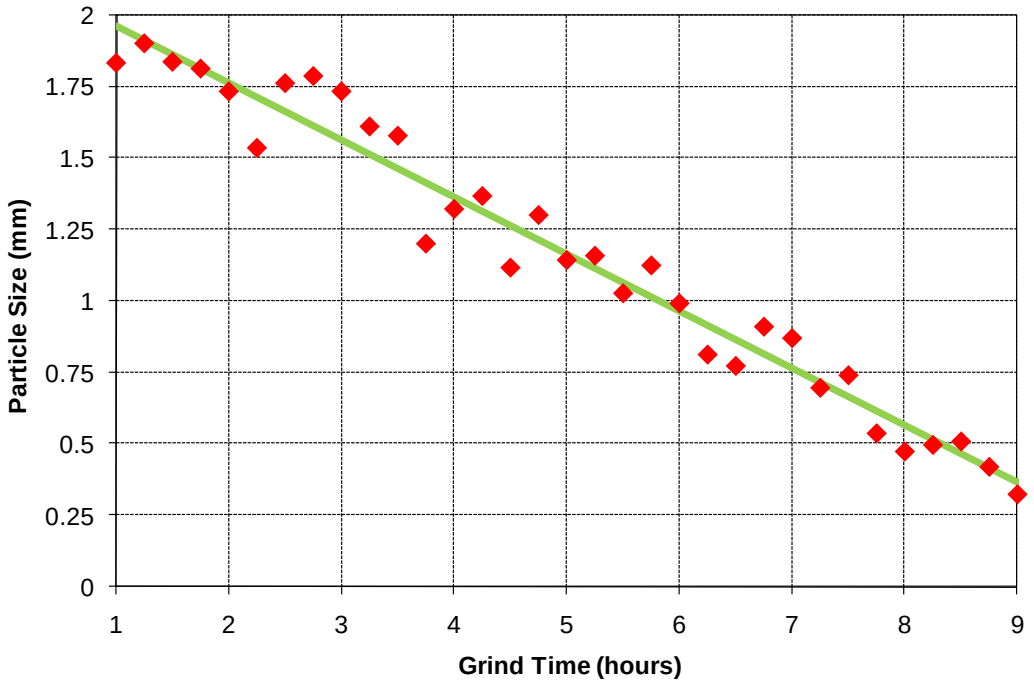
2. Estimating the Regression Equation

We generally use a computer or a calculator to determine the values for β_o and β_1, and thus the regression equation.

3. Example

From the agricultural chemical (grind time) example previously considered (Section 9.3, 2.B). 33 data values were collected from past records. The scatter plot of this data shown below indicates that a straight line relationship appears to be appropriate. Thus, we can proceed with simple linear regression. The 33 data values were put into *Minitab* and analyzed. The results:

Size = 2.16 - 0.200* Time



Remember, that before we use the regression line, we need to determine if it is statistically significant.

9.7 TI-83 Directions, Regression Line

First Time

1. Choose 2nd + 0 (Catalog)
2. Scroll down to DiagnosticOn, Press Enter twice and it will display, "Done"

Enter Data

1. Put X data in L1 and Y data in L2. To start entering data, choose Stat/Edit and then Enter.

Regression Line

1. Choose Stat/Tests/LinRegTTest(alpha+E) and press Enter
2. Supply the appropriate Lists for Xlist and Ylist
3. Make sure Freq is 1
4. Choose the $\neq 0$ alternative
5. Ignore *RegEQ*
6. Select Calculate and press Enter.

   The resulting display will show (for example):

$y = a + bx$
$\beta \neq 0 \text{ and } \rho \neq 0$
$t = 4.24$
$p = 0.002$
$df = 9$
$a = 3.00$
$b = 0.50$
$s = 1.24$
$r^2 = 0.666$
$r = 0.816$

In the example results shown above, the estimated y-intercept is *a=3.00* and the estimated slope is *b=0.50.* Thus, the regression is: $y = 3 + 0.5x$. We will learn what some of the other things in the display mean shortly.

Homework 9.2

1. Consider the four data sets from Homework 9.1. (a) Find the regression line for each one where simple linear regression is appropriate. (b) Which regression line is the steepest? (*i.e.* find the line with the largest absolute value of its slope) (c) Which regression line has slope closest to zero?

9.8 Significance of Regression Line

Overview - Just because we can use a computer or calculator to make a regression line doesn't mean that it is any good. We need to assess how good our regression line is. The first thing we need to do is to determine if the regression line is *statistically significant*.

Remember that when we do simple linear regression, our model is $y = \beta_o + \beta_1 x$. What would happen if our model was wrong and there was no slope ($\beta_1 = 0$). What would the model look like then? Thus, if there is not a linear relationship between *x* and *y* then the slope of the regression line would be zero which results in a horizontal line. This would mean that as *x* changed, *y* didn't change.

So, it makes sense to do a hypothesis test to see if the true (unknown) slope is zero or not. In other words, when we do a regression, we may come up with an estimated slope of $\hat{\beta}_1 = 3.42$, for example. So, the question becomes, can we differentiate this value from zero. In other words, was it simply *chance* that produced the estimate of 3.42, or is the true value definitely different than zero?

Hypothesis Test on β_1: $H_o: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

We are testing to see if the *slope is significant*, to see whether the slope is different than zero.

Method for Conducting Hypothesis Test – We generally use a computer or calculator to determine the *p*-value for the test. If the *p*-value is small (less than α **)** then reject the null hypothesis.

Interpretation of Decision:

If we ***reject*** H_o, then we conclude that the regression is significant, that the slope is significant, that a significant (linear) relationship exists between the independent and dependent variables, that *the regression is good*, that we have a statistically significant regression. When we do regression, we want to reject the null hypothesis.

If we ***fail to reject*** H_o, then we conclude that the data does not fit the model well, that the regression is not statistically significant. This does not mean that *no* relationship exists, it means that there is *not* evidence that the linear model is appropriate. Maybe a quadratic model will work, or some other model with curvature. Maybe we need more *explanatory* (independent) variables.

1. Example

This is *Minitab* output from the grind time example considered previously.

```
The regression equation is
Size(mm) = 2.16 - 0.200 Time(hrs)

Predictor         Coef       StDev          T        P
Constant       2.16303     0.03995      54.14    0.000
Time(hrs     -0.199764    0.007214     -27.69    0.000

S = 0.09865     R-Sq = 96.1%     R-Sq(adj) = 96.0%
```

Suppose we choose a significance level of 5%. Then we see that we reject the null hypothesis $H_o: \beta_1 = 0$, because the *p*-value = 0.000 < 0.05. Thus, we conclude that the slope is definitely not zero, *e.g.* that the slope and the regression line are statistically significant.

Note that the *p*-value is always (for our class) the second value listed under the column titled *P*.

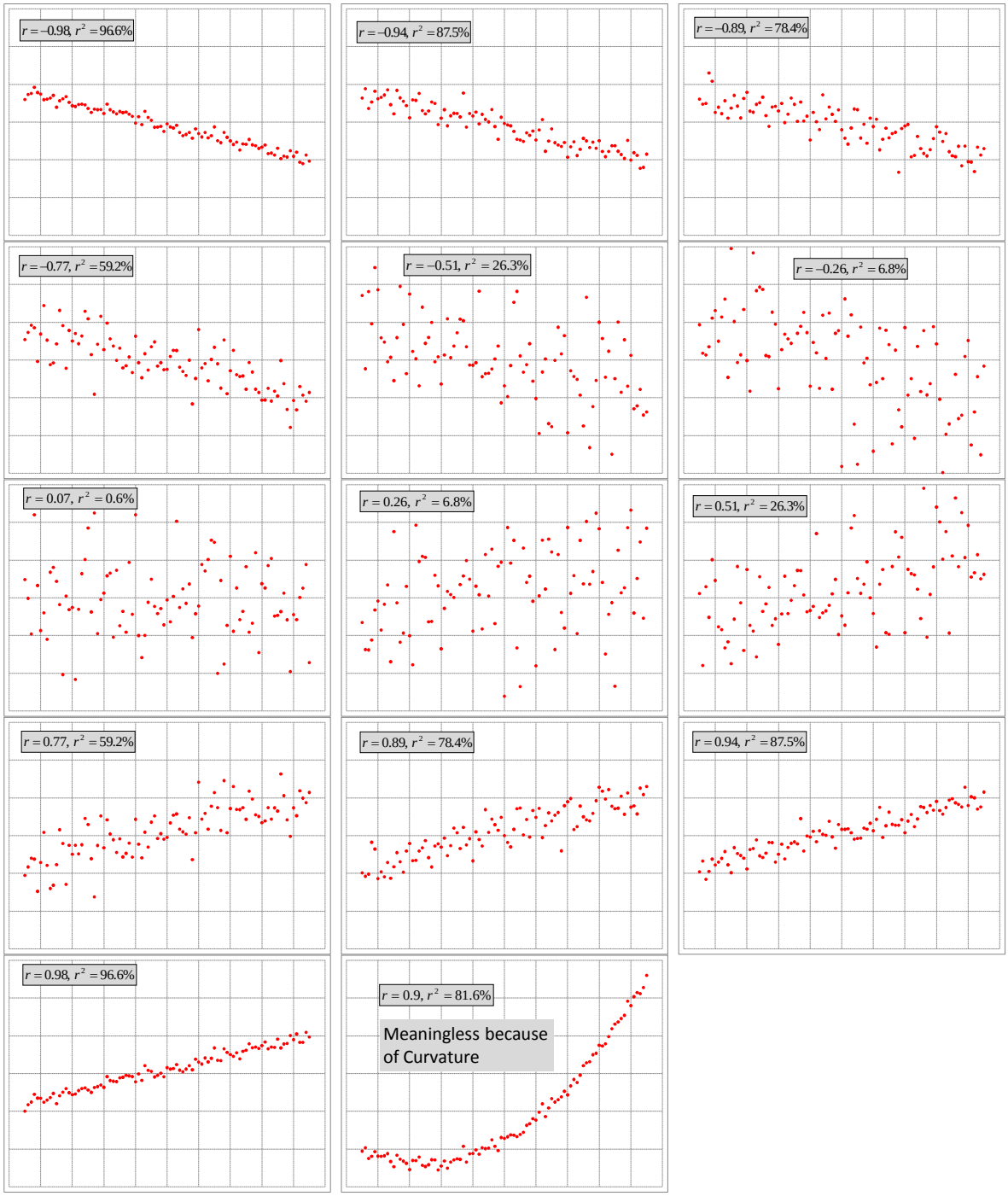
9.9 Sample Correlation Coefficient

Definition - The sample correlation coefficient measures the strength of the linear association between two quantitative variables. The symbol, *r* is used to denote sample correlation. Thus, *r* is only meaningful *after* we have determined that we have a statistically significant regression line. We will use a computer or calculator to calculate the value of *r*.

1. Rules for interpreting *r*:

A. The value of *r* always falls between -1 and 1. A positive value of *r* indicates positive correlation and a negative value of *r* indicates negative correlation.
B. The closer *r* is to 1, the stronger the positive correlation and the closer *r* is to -1, the stronger the negative correlation. Values of *r* close to zero indicate no linear association (*i.e. random*).
C. The larger the absolute value of *r*, the stronger the relationship between the two variables.
D. *r* is dimensionless
E. *r* measures only the strength of *linear* association between two variables. *r* is meaningless for nonlinear relationships.

2. Examples


$r = -0.98,\ r^2 = 96.6\%$
$r = -0.94,\ r^2 = 87.5\%$
$r = -0.89,\ r^2 = 78.4\%$
$r = -0.77,\ r^2 = 59.2\%$
$r = -0.51,\ r^2 = 26.3\%$
$r = -0.26,\ r^2 = 6.8\%$
$r = 0.07,\ r^2 = 0.6\%$
$r = 0.26,\ r^2 = 6.8\%$
$r = 0.51,\ r^2 = 26.3\%$
$r = 0.77,\ r^2 = 59.2\%$
$r = 0.89,\ r^2 = 78.4\%$
$r = 0.94,\ r^2 = 87.5\%$
$r = 0.98,\ r^2 = 96.6\%$
$r = 0.9,\ r^2 = 81.6\%$
Meaningless because of Curvature

9.10 Coefficient of Determination

Definitions – We use the symbol, r^2 as a measure of how good the regression line is. This value is the *square* of the correlation coefficient and is provided by *Minitab* or a calculator. r^2 is defined to be *the fraction of total variability that is explained by the regression line*. r^2 is referred to as the *coefficient of determination*.

Rules for Interpreting r^2:

A. The value of r^2 always falls between 0 and 1.
B. The closer r^2 is to 1, the better the regression line fits the data. The closer r^2 is to 0, the worse the regression line fits the data.
C. r^2 is dimensionless
D. r^2 is meaningless for nonlinear relationships.
E. Guidelines:

1. social sciences: r^2 as low as 0.25 is sometimes useful
2. physical, medical, and engineering sciences: r^2 as low as 0.6 is sometimes useful
3. sometimes r^2 will be 0.9 or greater

Example

This is *Minitab* output from the grind time example considered previously.

The regression equation is
Size(mm) = 2.16 - 0.200 Time(hrs)

Predictor	Coef	StDev	T	P
Constant	2.16303	0.03995	54.14	0.000
Time(hrs	-0.199764	0.007214	-27.69	0.000

S = 0.09865 R-Sq = 96.1% R-Sq(adj) = 96.0%

Thus, we say that 96.1% of the total variability in the particle size is explained by the regression line. So, not only is this regression line statistically significant, it also does a good job at prediction.

Determining *r* from r^2 - *r* is found from r^2 by taking the square root of r^2 and the *sign* of the slope of the regression line. Thus, in the example above and the correlation is $r = -\sqrt{0.961} = -0.98$ noting that the minus sign is because the slope is negative.

Homework 9.3

1. Consider the regression lines from Homework 9.2. Evaluate each regression lines as completely as possible. Justify your answer.
2. How do we tell if a regression line is statistically significant?
3. How do we tell how much variability is explained by a regression line?
4. What is the first step when considering using simple linear regression?

5. A regression line, $y = 4 - 5x$ has $r^2 = 77\%$. What is the correlation coefficient?
6. A regression line $y = -3.4 + 7.8x$ has $r^2 = 0.25$. What is the correlation coefficient?

9.11 Estimation and Prediction

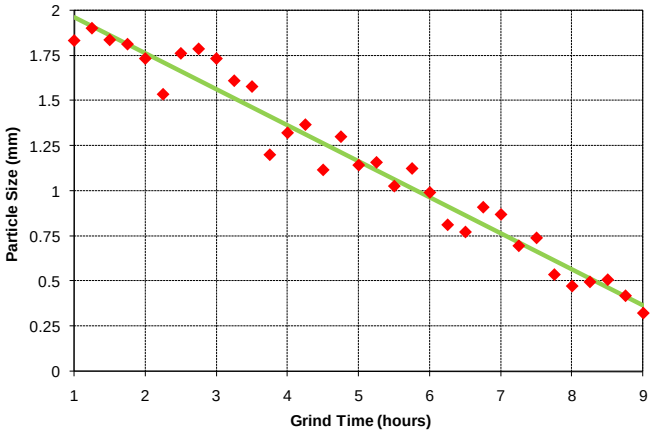
1. Using the Regression Line for Prediction

If we are satisfied with a regression line: small p-value and large r^2, then we can use it for prediction. To use the regression line for prediction, we simply plug a value of the independent variable into the regression equation and calculate the result. A prediction is actually a prediction of the average response for a particular value of the independent variable. Note that predictions are statistically valid only over the range of the independent variable.

2. Example

Consider the grind time example considered previously,

Size = 2.16 - 0.200* Time



We note that the range of the independent variable is from approximately 1 hour to 9 hours. Thus, predictions in this range are valid.

Suppose the particles have been grinding for about 6.5 hours, what size do we predict the particles will be?

$$size = 2.16 - 0.2(6.5) = 0.86\ mm \approx 0.9\ mm$$

Suppose we stop the grinding machine after 6.5 hours and take a sample of the particles and determine their size to be 1.5 *mm*. What does this suggest?

3. Confidence Interval

A *confidence interval* is an interval estimate about the *mean (average) level of a response* (prediction).

For example, consider the case where we are trying to predict starting salaries from GPA's. Suppose that the placement center at VSU would like a 95% confidence interval on the *average starting salary for all people with a GPA of 3.0*.

4. Prediction Interval

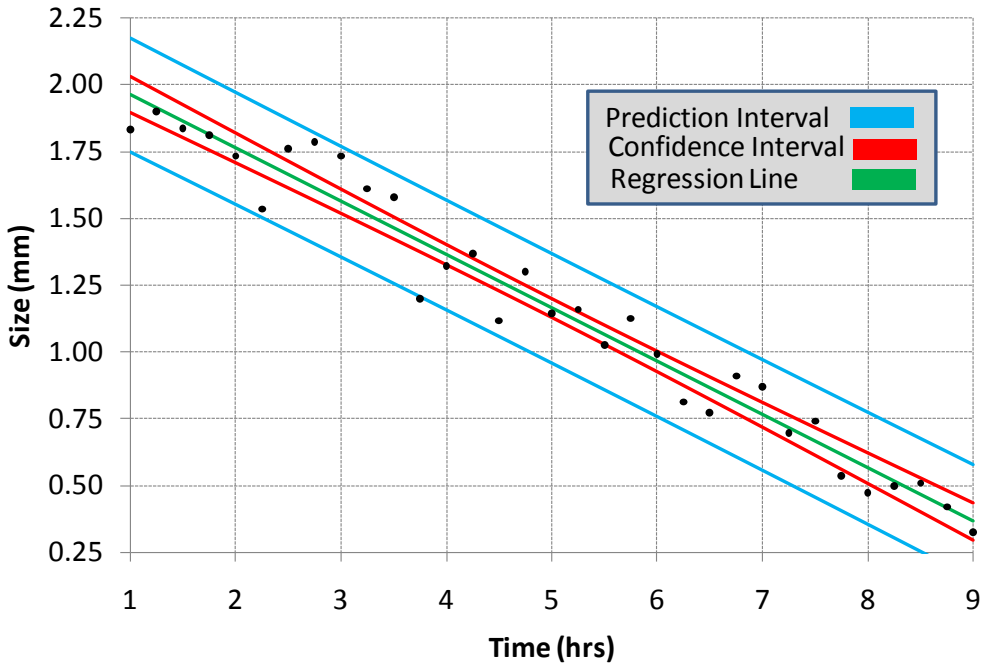
A *prediction interval* is an interval estimate about a particular, individual data value.

For example, again, consider the case where we are trying to predict starting salaries from GPA's. Suppose that a particular student, Paul, is a graduating senior with a GPA of 3.0. What range of starting salaries can he expect? The answer is that Paul requires a *prediction interval*.

Note that a prediction interval is *always* wider than a confidence interval. This should make sense intuitively because the confidence interval is for the average response, and averages have less variability than an individual observations, which is what the prediction interval is about.

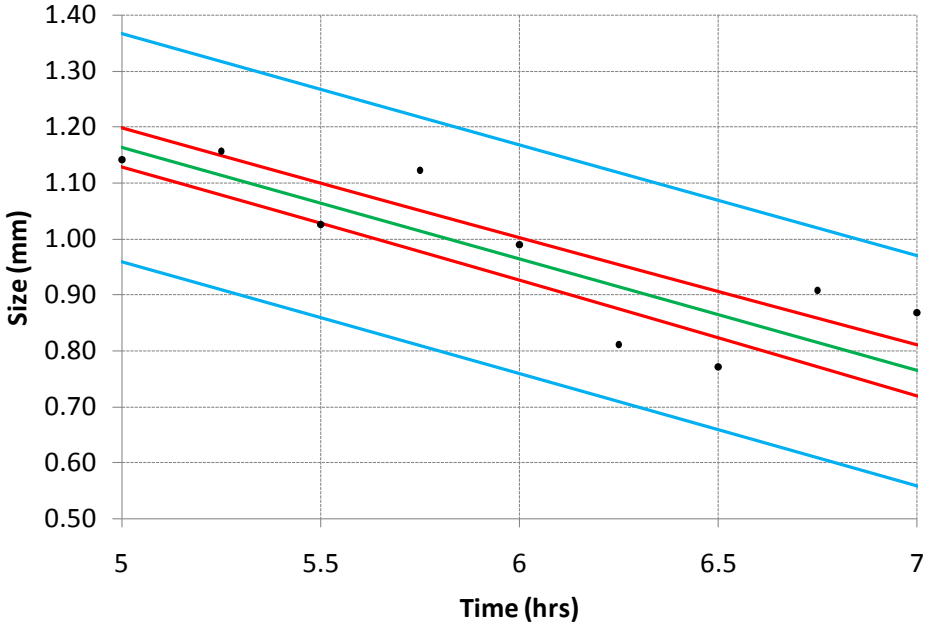
5. Example

Consider the grind time example and the graphs below which shows the confidence and prediction intervals across the entire range of times.



Now, suppose we want to know a confidence interval for the average particle size for all the times when the particles have been ground for 6 hours. Using the graph below, we can see that this confidence interval is about (0.93 *mm*, 1.00 *mm*).

Next, suppose an operator has been monitoring a grind for 6 hours. What range of particle sizes would he expect? Using the graph below, we can see that the prediction interval for 6 hours is about (0.76 *mm*, 1.17 *mm*).



Now, consider confidence intervals and prediction intervals for 2 hrs, 5 hrs, and 8 hrs.

Time(hrs)	Prediction	95% Confidence Interval	ME	95% Prediction Interval	ME
2.00	1.76	(1.71, 1.82)	0.06	(1.55, 1.97)	0.21
5.00	1.16	(1.13, 1.20)	0.04	(0.96, 1.37)	0.20
8.00	0.56	(0.51, 0.62)	0.06	(0.36,0.77)	0.21

In general, we note that inference is more precise towards the middle of the data. We can see this by looking at the margin of errors in the table above.

Homework 9.4

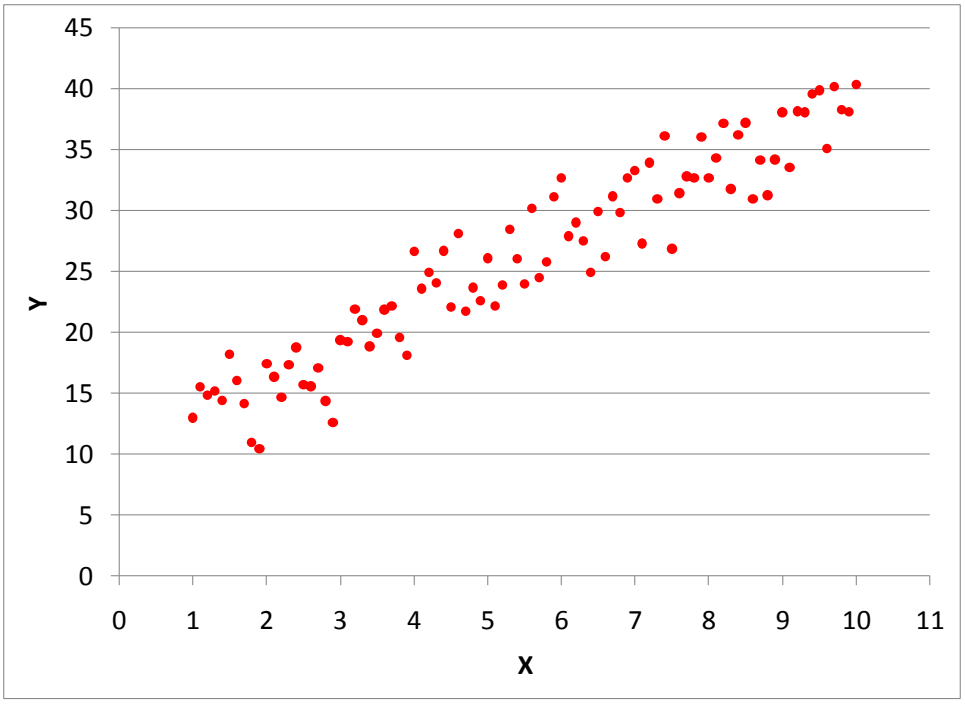
1. Define a *confidence interval* and *prediction interval* in the context of regression.

9.12 Examples

For each of the examples below:

a. Is simple linear regression appropriate? Why or why not?
b. What type of correlation is present?
c. What is the regression equation?
d. What is the estimated slope?
e. What is the estimate y-intercept?
f. Is the regression line statistically significant? Why or why not?
g. What fraction of total variability is explained by the regression line?

1. Example 1



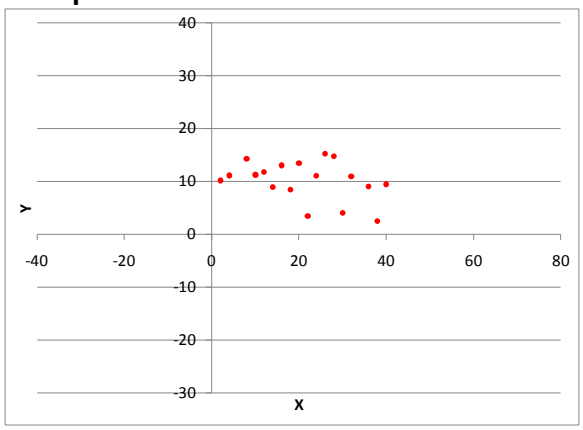
Regression Analysis

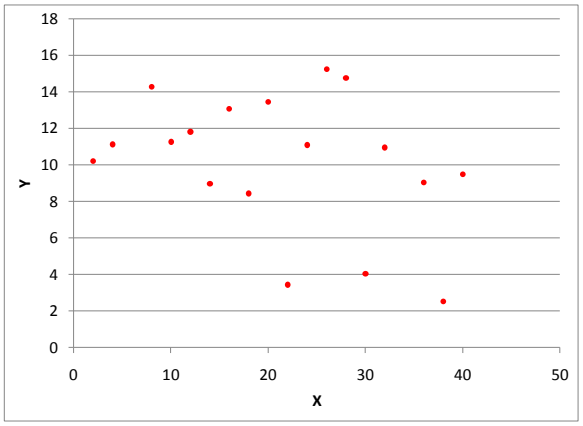
The regression equation is
Y = 10.1 + 2.92 X

Predictor	Coef	StDev	T	P
Constant	10.1331	0.6069	16.70	0.000
X	2.92077	0.09958	29.33	0.000

S = 2.495 R-Sq = 90.6% R-Sq(adj) = 90.5%

2. Example 2

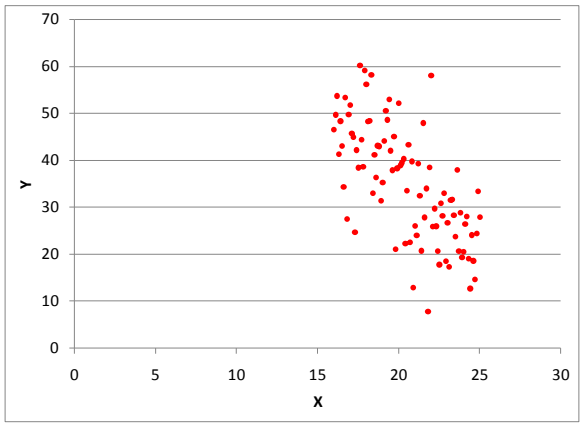


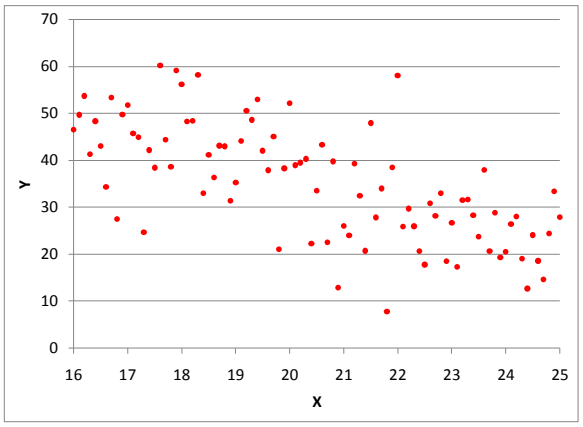


Regression Analysis

```
The regression equation is
Y = 12.6 - 0.112 X

Predictor         Coef       StDev          T        P
Constant        12.551       1.812       6.93    0.000
X            -0.11203     0.07579      -1.48    0.159

S = 3.608       R-Sq = 12.0%     R-Sq(adj) = 6.5%
```

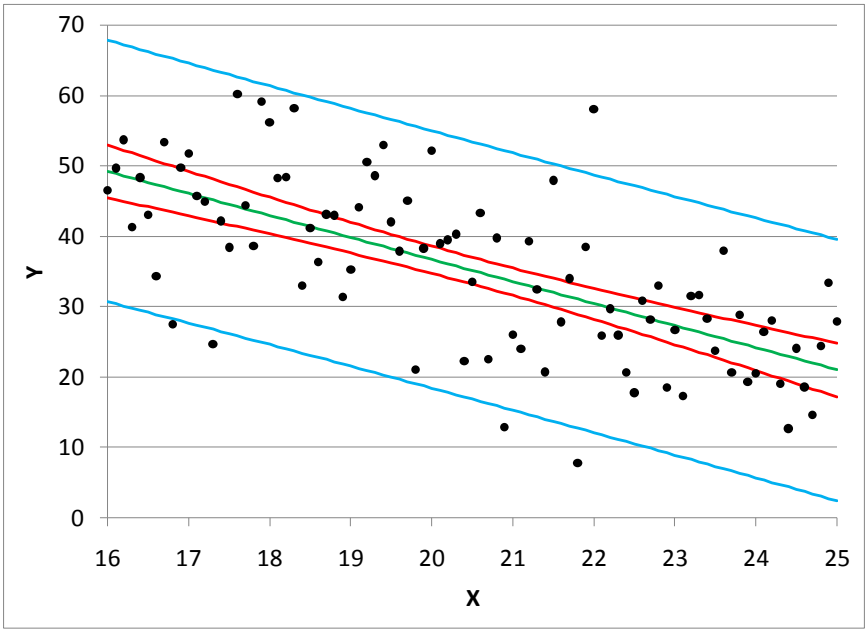
3. Example 3



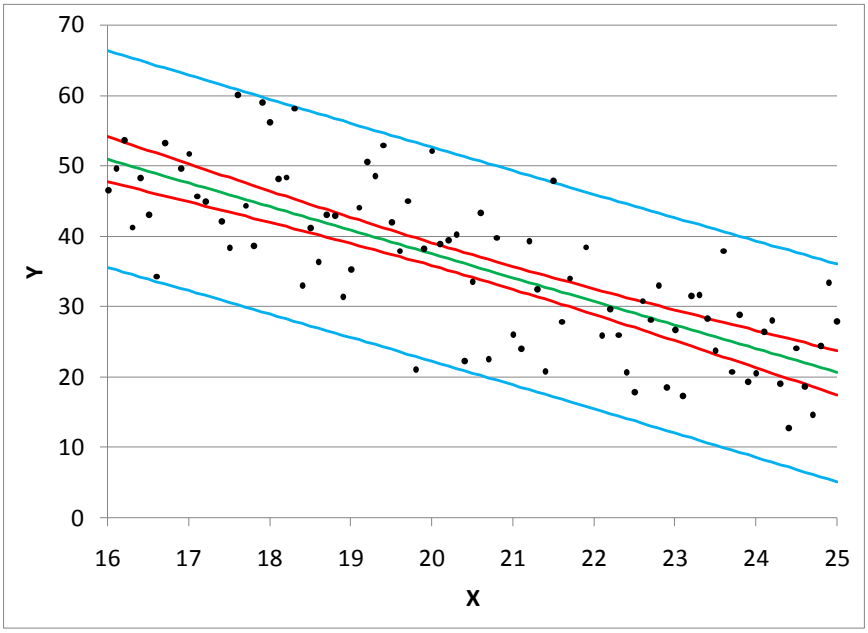


Regression Analysis

```
The regression equation is
C9 = 99.5 - 3.14 C8

Predictor         Coef       StDev          T        P
Constant        99.516       7.550      13.18    0.000
C8            -3.1381      0.3653      -8.59    0.000

S = 9.154       R-Sq = 45.3%     R-Sq(adj) = 44.7%
```



Remove 5 outliers



Regression Analysis

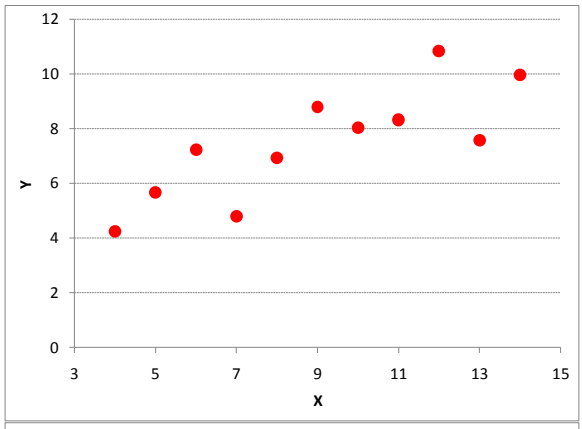
```
The regression equation is
Y = 105 - 3.38 X

Predictor        Coef       StDev          T        P
Constant      105.046       6.424      16.35    0.000
X             -3.3754      0.3102     -10.88    0.000

S = 7.596       R-Sq = 58.5%     R-Sq(adj) = 58.0%
```

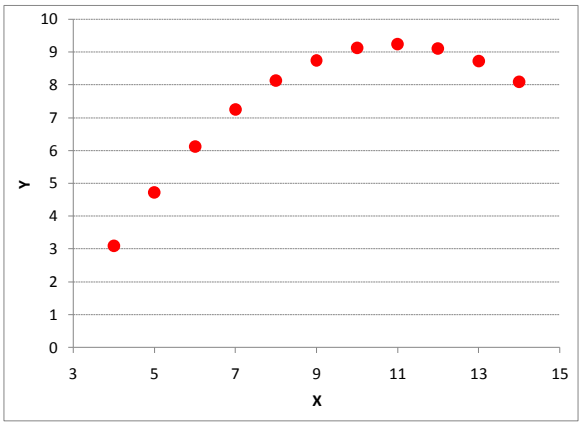
4. Three Data Sets

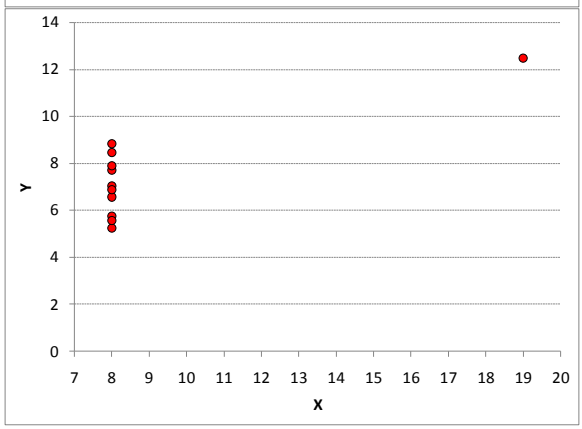
Consider these three data sets. Summarize each as completely as possible.

	Data Set 1		Data Set 2		Data Set 3	
Data Value	**X**	**Y**	**X**	**Y**	**X**	**Y**
1	10	8.04	10	9.14	8	6.58
2	8	6.95	8	8.14	8	5.76
3	13	7.58	13	8.74	8	7.71
4	9	8.81	9	8.77	8	8.84
5	11	8.33	11	9.26	8	8.47
6	14	9.96	14	8.10	8	7.04
7	6	7.24	6	6.13	8	5.25
8	4	4.26	4	3.10	8	5.56
9	12	10.84	12	9.13	8	7.91
10	7	4.82	7	7.26	8	6.89
11	5	5.68	5	4.74	19	12.50

Corresponding Scatter Plots







Corresponding Minitab Output

A. The regression equation is

```
y = 3.00 + 0.500 x

Predictor       Coef       Stdev    t-ratio        p
Constant       3.000       1.125       2.67    0.026
x             0.5001      0.1179       4.24    0.002

s = 1.237       R-sq = 66.7%     R-sq(adj) = 62.9%
```

B. The regression equation is

```
y2 = 3.00 + 0.500 x2

Predictor       Coef       Stdev    t-ratio        p
Constant       3.001       1.125       2.67    0.026
x2            0.5000      0.1180       4.24    0.002

s = 1.237       R-sq = 66.6%     R-sq(adj) = 62.9%
```

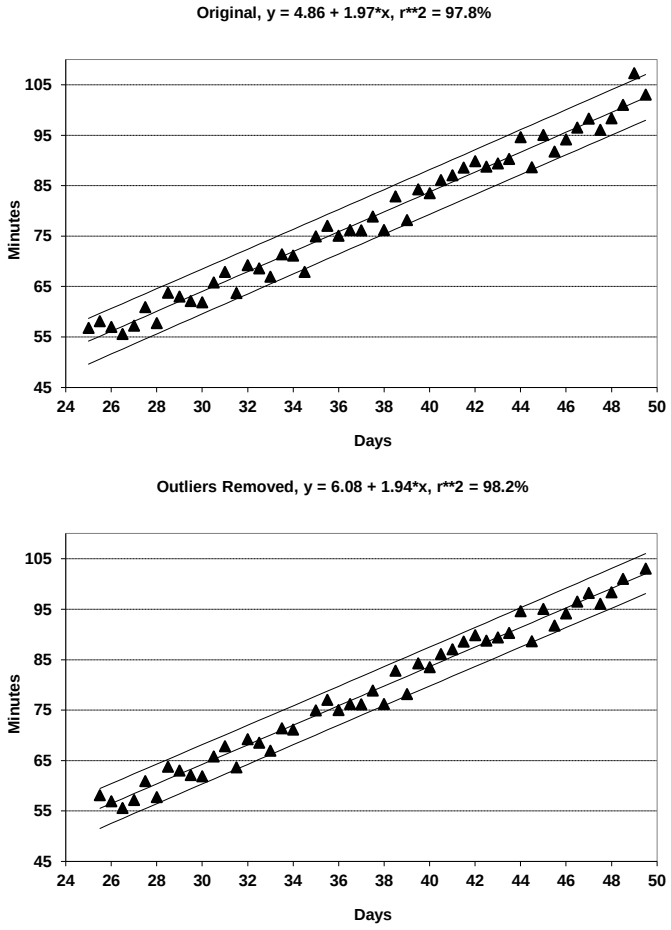
C. The regression equation is

```
y3 = 3.00 + 0.500 x3

Predictor       Coef       Stdev    t-ratio        p
Constant       3.002       1.124       2.67    0.026
x3            0.4999      0.1178       4.24    0.002

s = 1.236       R-sq = 66.7%     R-sq(adj) = 63.0%
```

9.13 Outliers and Influential Observations

1. Outlier

A *residual* is difference between an actual data value the corresponding predicted value. In other words, the vertical distance between a data value and the regression line. A data value with a large residual is called an *outlier*.

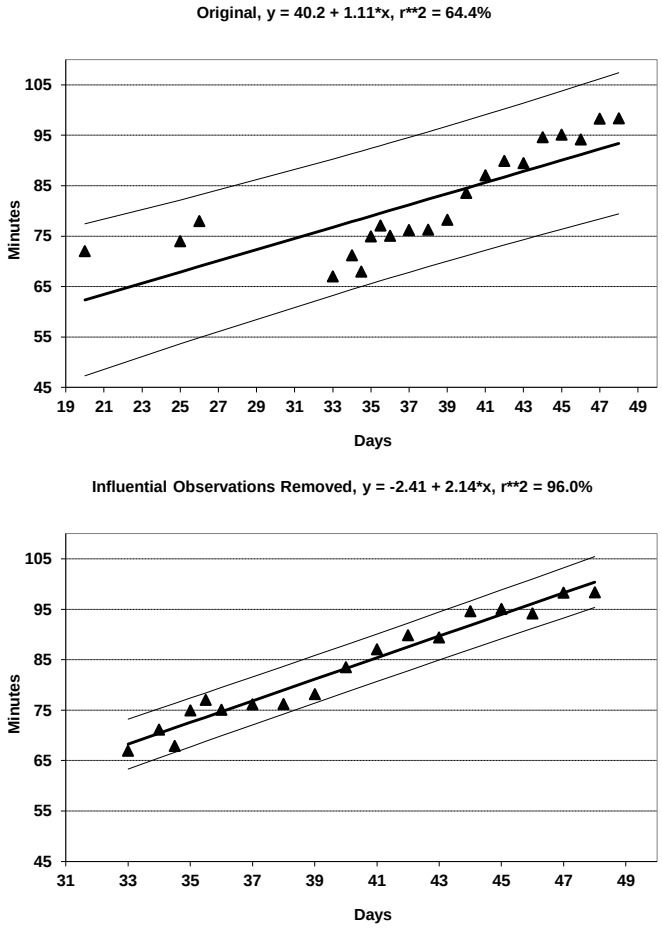


In the example shown above, the original data plotted on the left graph, along with the 95% prediction interval. There are two data values that fall outside the prediction interval: 34.5 days and 49 days. The graph on the right shows the regression line and data (with the possible outliers removed) and the new prediction interval. Notice that the y-intercept of the regression line changes, but the slope does not change very much and neither does r^2.

2. Influential Observation

A data value whose removal would cause a marked change in the position of the regression line is called an *influential observation*. Points that are separated in the *x* direction from the other observations are often influential.

Note that an outlier is not necessarily an influential observation, as we saw in the example above. Similarly, an influential observation is not necessarily an outlier, as shown below.



Note that in this example, the three left most data points (in the top graph) are not outliers; however, they have a strong impact on the regression line. The majority of the data (from 33 days to 49 days) seems to have a different slope than the regression line. When we remove those three influential observations (bottom graph), we note a drastic change in the regression line and in r^2.

Homework 9.5

1. Consider Data Set 3. Remove the three outliers at (9,45), (15,36), and (18,29). Rerun the regression and analyze the situation.
2. Consider Data Set 3. Suppose the independent variable is time as measured in minutes that a mixture has been boiling. The dependent variable is the grams of fat per serving. Suppose we boil a mixture for 15 minutes, what level of grams of fat would we predict?
3. Consider Problem 2. Each additional minute the mixture is boiled, reduces the fat grams by how much?
4. Consider Problem 2. Over what range of values is the regression line valid?