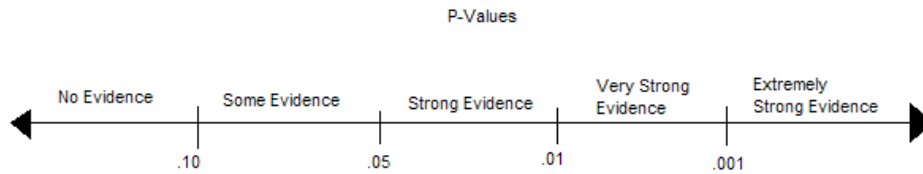


# Stat 3000 Review for Exam 3

In my other class we check “conditions” instead of “assumptions”. So when you see the word “condition” you know it means “assumption”.

	size	mean	standard deviation	variances	proportions
population parameter		$\mu$	$\sigma$	$\sigma^2$	$p$
sample statistic	$n$	$\bar{x}$	$s$	$s^2$	$\hat{p}$



## Paired Sample Test

Just use the one sample T test on the differences.

$$t = \frac{\bar{x}_d}{s_d / \sqrt{n_d}}$$

$$df = n - 1$$

## Two Sample Tests for $\mu_x$ and $\mu_y$

### When to Use:

- We want to compare two population means.
- We have independent samples.

**Assumptions:** The populations are normally distributed or *both* sample sizes are at least 30.

**Null Hypothesis:**  $H_0 : \mu_x = \mu_y$

### Test Statistic:

- **Two Sample z Test:**

If you know the population variances  $\sigma_x^2$  and  $\sigma_y^2$ , use the test statistic

$$z = \frac{(\bar{x} - \bar{y})}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$$

and the standard normal distribution.

- **Two Sample t Test (Equal Variances):**

If you don't know the population variances  $\sigma_x^2$  and  $\sigma_y^2$ , but you think that they are the same, use the test statistic

$$t = \frac{(\bar{x} - \bar{y})}{\sqrt{s_p^2 \left( \frac{1}{n} + \frac{1}{m} \right)}}$$

where the pooled variance is

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

and the t distribution with degrees of freedom  $df = n + m - 2$ .

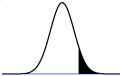
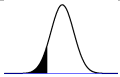

- **Two Sample t Test (Unequal Variances):**

If you don't know the population variances  $\sigma_x^2$  and  $\sigma_y^2$ , but you think that they are not the same, use the test statistic

$$t = \frac{(\bar{x} - \bar{y})}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

and the  $t$  distribution with the Satterthwaite approximation of degrees of freedom

$$df = \frac{\left( \frac{s_x^2}{n} + \frac{s_y^2}{m} \right)^2}{\frac{s_x^4}{n^2(n-1)} + \frac{s_y^4}{m^2(m-1)}}$$

Alternative Hypothesis	P-value
$H_A : \mu_x > \mu_y$	
$H_A : \mu_x < \mu_y$	
$H_A : \mu_x \neq \mu_y$	

## The ANOVA F Test

**When to Use:** We want to compare several population means.

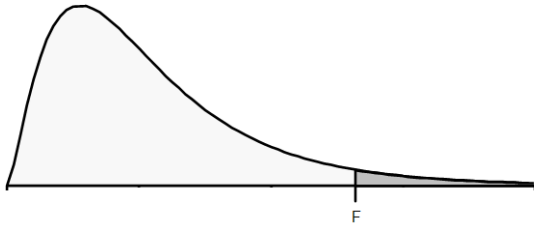
**conditions:** (We won't check these)

**Null Hypothesis:**  $H_0$  : all the population means are equal

**Alternative Hypothesis:**  $H_a$  : at least one of the population means is different

**Test Statistic:**  $F = \frac{MSG}{MSE} = \frac{\text{variation between groups}}{\text{variation within groups}}$

**P-Value:** The p-value is the area under the F curve to the right of the test statistic.



\*\*This is always a right tail test.

Remember that the alternative hypothesis only tells us that at least one of the population means is different. If we decide that at least one of the population means is different, we could go on to do a POST-HOC test to compare the different population means to each other to try to pinpoint which of the means is different from the others.

## Coefficient of Determination

$$R^2 = \frac{SSG}{SST} = \frac{\text{sum square group}}{\text{sum square total}}$$

$R^2$  tells us the percentage of the total variation that is explained by differences between the groups.

The rest of the variation comes from the natural variation from person to person within a group.

# Chi-Square Test for Goodness of Fit

## When to Use:

- We have a proposed distribution and we want to know if it is a good fit for a data. (We want to know if our theoretical probabilities are correct).
- The total sample size is  $n$ .
- The sample items are classified into one of  $k$  groups.

**conditions:** All the expected cell counts are at least 5.

**Null Hypothesis:**  $H_0$  : all our theoretical probabilities are correct **or**

$H_0$  : the proposed distribution is a good fit for our data

**Alternative Hypothesis:**  $H_a$  : at least one of the theoretical probabilities is incorrect **or**

$H_a$  : the proposed distribution is not a good fit for our data

## Expected Count:

$$\text{expected count} = n \cdot p_i$$

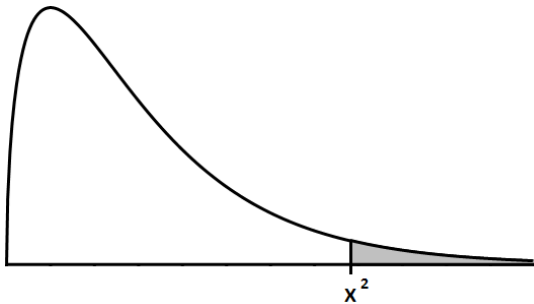
$$\text{expected count} = (\text{total sample size}) \times (\text{theoretical probability for that group})$$

## Test Statistic:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

**Degrees of Freedom:**  $df = k - 1$

**P-value:** The area under the chi-square curve to the right of  $X^2$ .



\*This is a right tail test.

# Chi Square Test of Independence

## When to Use:

- We have a contingency table with two variables.
- We want to see if two variables affect each other.

**conditions:** All the expected cell counts are at least 5.

**Null Hypothesis:**  $H_0$  : the two variables are independent **or**

$H_0$  : there is no relationship between the two variables

**Alternative Hypothesis:**  $H_a$  : the two variables are dependent **or**

$H_a$  : there is a relationship between the two variables

## Expected Count:

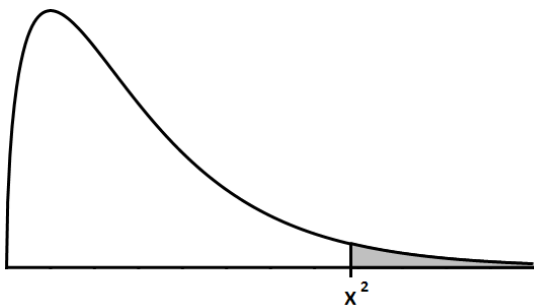
$$\text{expected cell count} = \frac{\text{row total} \times \text{column total}}{\text{total sample size}}$$

## Test Statistic:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

**Degrees of Freedom:**  $df = (r - 1)(c - 1)$

**P-value:** The area under the chi-square curve to the right of  $X^2$ .



\*This is a right tail test.

# Linear Regression

Different textbooks use different symbols. We use  $\hat{\beta}_0$  and  $\hat{\beta}_1$  for our slope and intercept we find from the sample data. But my other class uses  $b_0$  and  $b_1$ .

Symbol/Equation	Meaning
$y = \beta_0 + \beta_1 x + \epsilon$	The equation of the line that fits the population data.
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$	The least squares line. Notice that it has the estimated values of the slope and intercept. We can use this to predict y values which is why we have the symbol, $\hat{y}$ . We often use $\hat{\phantom{x}}$ in statistics to show that it is a predicted value.
$\beta_0$ and $\beta_1$	The true population parameters for y-intercept and slope.
$\hat{\beta}_0$ and $\hat{\beta}_1$	The estimates for the true population parameters. We find these from our sample data.
$r$	Correlation
$r^2$	The coefficient of determination. This tell us the percent of total variation in the y values explained by our line. It is a measure of how well we can predict the y values. $r^2$ is a positive number between 0 and 1. The higher the $r^2$ value, the better our line fits the data.
$t^*$	The critical values for confidence intervals
$SE_{b_1}$	The standard error of the estimate of the slope.
$\left[ \hat{\beta}_1 \pm t^* SE_{\hat{\beta}_1} \right]$	The confidence interval for the true population slope, $\beta_1$ .
$df = n - 2$	degrees of freedom
$t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$	The test statistic to test whether or not $\beta_1 = 0$ .

**Response Variable (Y):** Variable you want to predict

**Explanatory Variable (X):** Variable you use to explain the response variable

**Correlation (r):** Measures the strength of the linear relationship between two variables.

Correlation is between -1 and 1.

Correlation close to -1 or 1 means points are close to line.

Correlation close to 0 means the relationship is very weak and the points are not close to the line.

**Intercept:** If  $x = 0$ , then  $y = \dots$

**Slope:** If  $x$  goes up 1, then  $y$  goes up ...

**Residual:** difference between observed and predicted  $y$  value

$$\epsilon_i = y_i - \hat{y}_i$$

**Extrapolation:** This is when we use a regression line for predictions, but the  $x$  values are far outside the range of the  $x$  values used to obtain the line.

These predictions are often inaccurate. We can't guarantee that the relationship between  $x$  and  $y$  remains the same for  $x$  values outside our original data range.

**Association vs Causation:** We can use our regression line to use  $x$  to predict  $y$ .

But we **can not show** that a change in an explanatory variable **causes** a change in the dependent variable.

## T Test for Slope

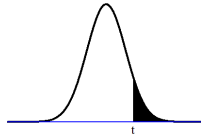
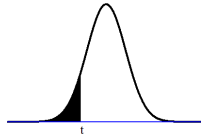
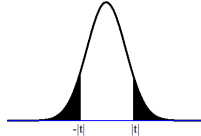
**When to Use:** We want to know the value of the population slope  $\beta_1$ .

**conditions:** Just check the three plots to see if linear regression is appropriate.

**Null Hypothesis:**  $H_0 : \beta_1 = 0$  (There is no linear relationship between the variables.)

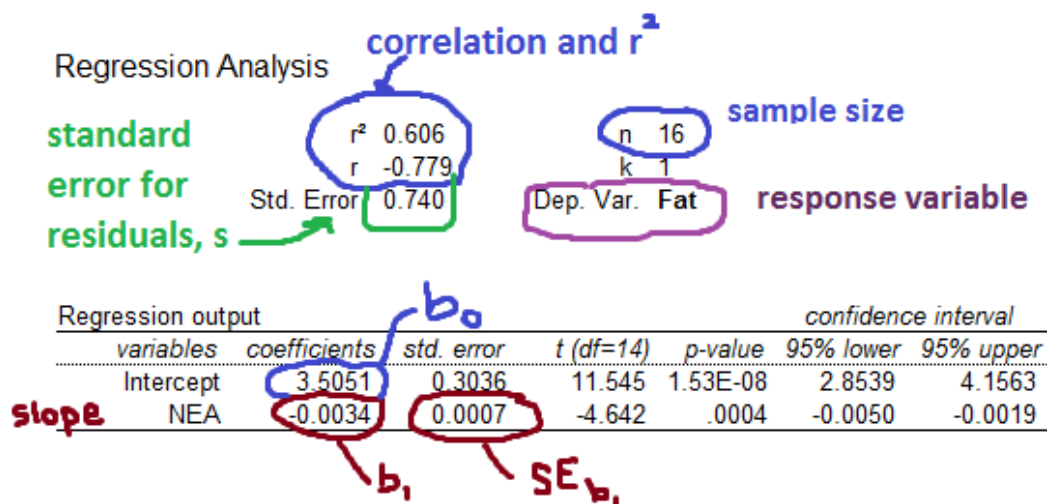
**Test Statistic:**  $t = \frac{b_1}{SE_{b_1}}$

**Degrees of Freedom:**  $n - 2$

Alternative Hypothesis	P-value	
$H_a : \beta_1 > 0$ There is a <i>positive</i> linear relationship between the variables.	$P(T \geq t)$	
$H_a : \beta_1 < 0$ There is a <i>negative</i> linear relationship between the variables.	$P(T \leq t)$	
$H_a : \beta_1 \neq 0$ There is a linear relationship between the two variables.	$2P(T \geq  t )$	

\*Mega Stat automatically does a two sided alternative hypothesis. If you want to test specifically for a positive or negative linear relationship, you have to find the p-value by hand (or just divide it by 2).

## MegaStat Output



## Review for Two Sample Tests

1. John compared the earnings of “target firms” and “bidder firms”.

He took a sample of 36 target firms and found a mean earnings per share of \$1.52 with a standard deviation of \$0.92.

A sample of 36 bidder firms has a mean earnings per share of \$1.20 with a standard deviation of \$0.93.

Test to see if the mean earnings are different between target firms and bidder firms. (Use  $\alpha = .01$ )

2. A library wants to know if they will get faster customer service if they switch to a competing provider. They place 31 calls to the current provider and get a sample mean of 11 minutes with a standard deviation of 3.4 minutes. They place 35 calls to the competing provider and get a sample mean of 9.5 minutes with a standard deviation of 1.6 minutes.

Assume unequal variances to test if the library will spend less time per call, on average, if they switch to the competing provider.

3. A marketing research manager wishes to compare the mean prices charged for two brands of CD players. She conducts a random survey of retail outlets and obtains independent random samples of prices with the following results:

	Onkyo	JVC
sample mean	\$189	\$145
sample size	6	12

Based on previous experiments, she believes that the population variance of Onkyo prices is 144 and the population variance for JVC prices is 100.

Conduct an hypothesis test to see if Onkyo charges more for CD players. Assume normality. You pick an appropriate significance level.



4. After months of working overtime, you have saved some money for a set of new golf clubs and you want to make sure that you are buying the best. You can a really good deal on Brand X clubs, but you are willing to make the sacrifice to buy brand Z clubs if the clubs really improve your game. The salesperson allows you to take the number 3 wood from each brand and use them to hit balls on a driving range. The data in yards is summarized below:

	Brand X	Brand Z
Sample size	15	14
Sample mean	255	271
sample standard deviation	8.7	9.1

Assume normality and test to see if Brand Z improves your golfing game.

5. The Board of Realtors for Greater Bridgeport took random samples of homes sold in 1995 and 1996 and found the sample statistics:

	1995 sales	1996 sales
Sample size	40	35
sample mean	\$151000	\$160000
sample standard deviation	\$5332	\$7468

Conduct an appropriate test with .05 level of significance to determine if the mean selling price of a home was higher in 1996 than in 1995.

6. A study focuses on whether there is a difference in the mean number of times per month that the men and women buy take out food for dinner. They found a sample of 34 men with a sample mean of 25.6 and a sample of 28 women with a sample mean of 21.2 days. Because there is so much historical data, the population variances are known to be 17.64 for men and 14.44 for women.

Conduct a test to see if, on average, men eat out more days a month than women do. Assume normality and use  $\alpha = .05$ .

7. A professor teaches 2 on site courses and 2 online courses. She wants to see which way gives higher exam scores. She keeps track of their grades on an exam and the data is summarized below.

	Online	On Site
sample size	47	46
sample mean	76%	83%
sample standard deviation	13.77%	11.18 %

Can the professor say that, on average, there is a difference in student scores based on which type of class they are in? Conduct a test to find out. Use unequal variances.

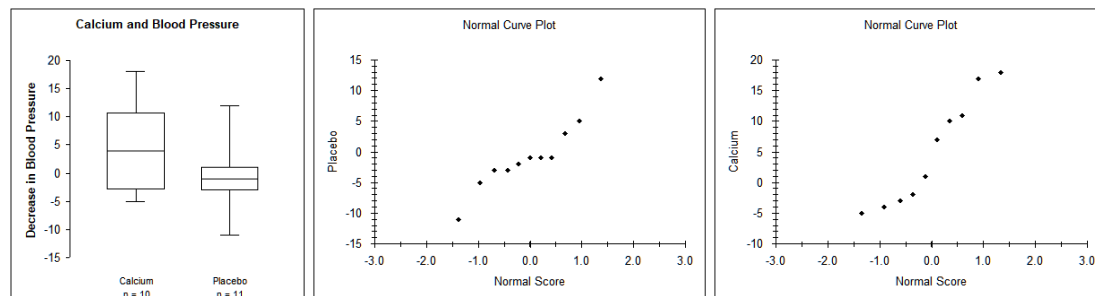
8. The members of the Chamber of Commerce are looking at the amount of vacant office space in their city compared to the space in the neighboring city.

	Their City	Neighboring City
sample size	12	15
sample mean	217,000	167,607
sample standard deviation	2200	2100

Conduct a hypothesis test to see if the mean space available is greater in their city than in the neighboring city. (Assume equal variances and normal populations.)

9. A study was conducted to see if taking calcium reduces blood pressure. Two independent groups were chosen. One group was given calcium and one group was given a placebo. The decrease in their blood pressure was recorded. Conduct a hypothesis test to see if taking calcium reduced blood pressure. (A positive number represents a decrease and a negative number represents an increase in blood pressure).

Group	Treatment	$n$	$\bar{x}$	$s$
1	Calcium	10	5	8.743
2	Placebo	11	-.273	5.901



10. You want to compare the population means of two different groups. Conduct a test to determine if the means are different.

	sample size	mean	standard deviation
A	10	3.4	2.02
B	32	4.2	1.53

11. We want to know if the September 11 terrorist attack had an effect on U.S. airline demand. We found a sample of 12 airline routes whose passenger miles were tracked for one year before the attack and for one year afterward. We subtracted the post attack mileages from the pre attack mileages for each airline route. The mean of the sample of paired differences was 29.7 million miles and the standard deviation of the sample was 2.975 million miles.

Test to see if the attack had a negative impact on how much passengers fly. Assume normality.

12. Because of skyrocketing health-care costs, many hospital administrators are working to contain costs. They want to know if they can treat OSAS (obstructive sleep apnea syndrome) at home effectively. They take a sample of 9 patients and count the number of obstructions before treatment and after treatment. If the treatment is effective there should be less obstructions after treatment. They find the paired differences (post treatment-pre treatment) and get a sample mean of -86 and standard deviation of 101.83.

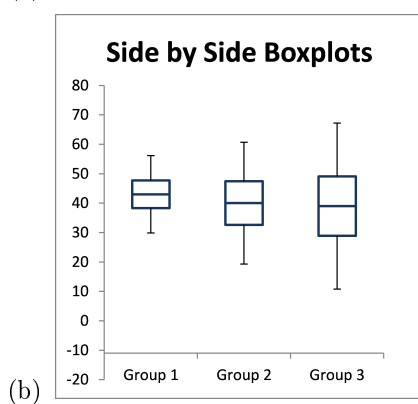
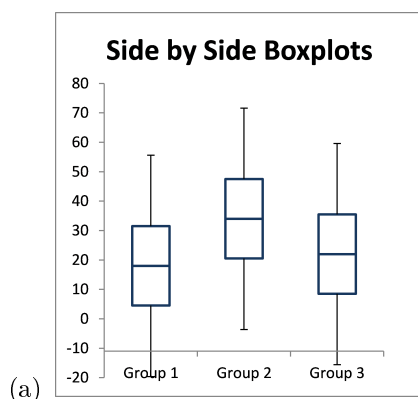
Conduct a hypothesis test with  $\alpha = .05$  to determine if the home treatment was effective. (Assume normality)

13. We want to know which website has better prices for business software packages, Computability or PC Connection. We select 10 business software packages such as Virus Scan and Quick Books and check the price for each package at each website. We find the paired differences of costs at computability and the costs at PC connection. (Computability-PC) The mean of the 10 differences is \$7 with a standard deviation of \$2. Conduct an appropriate test with  $\alpha = .10$  to determine if PC Connection is cheaper. Assume normality.

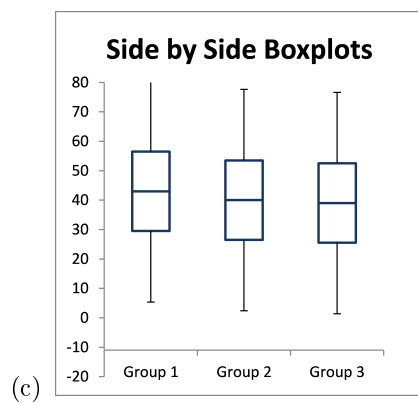
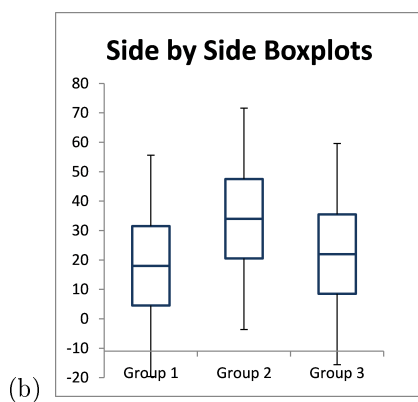
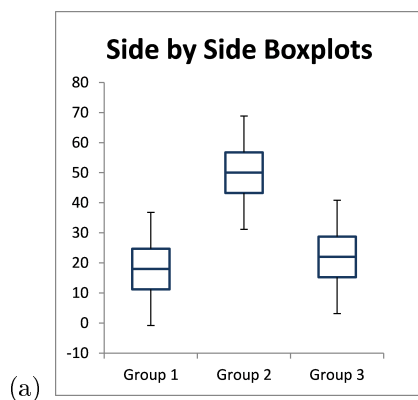
## ANOVA Review

14. For which data set will you be most likely to believe that the population means are different? (The plots are all on the same scale.)

This is just what you can see visually. You aren't doing any computations. I have included the p-values for the ANOVA tests on the key just so you can see how the p-values relate to what you are seeing visually.



15. For which data set will you be most likely to believe that the population means are the same? (The plots are all on the same scale.)



16. A study was conducted to compare five different training programs for improving endurance. Forty subjects were randomly divided into five groups of eight subjects in each group. A different training program was assigned to each group. After two months, the improvement in endurance was recorded for each subject. A one-way ANOVA is used to compare the five training programs, and the resulting p-value is .023. At a significance level of  $\alpha = 0.05$ , what is the appropriate conclusion about mean improvement in endurance?
- (a) The average amount of improvement appears to be the same for all five training programs.
  - (b) The average amount of improvement appears to be different for each of the five training programs.
  - (c) It appears that at least one of the five training programs has a different average amount of improvement.
  - (d) One training program is significantly better than the other four.
17. For an ANOVA test, the p-value is the area to the \_\_\_\_\_.
- (a) left of the test statistic
  - (b) right of the test statistic
  - (c) it depends on the alternative hypothesis
18. The alternative hypothesis for ANOVA is:
- (a) all of the population means are different
  - (b) at least one of the population means is different from the others
  - (c) all of the population means are the same
  - (d) all but one of the population means are the same

19. We want to know if what month a baby is born in affects how early the baby learns to crawl. We wonder if babies who are born during colder months take longer to crawl since they are more likely to be bundled tightly. We kept track of how long in weeks it took babies to crawl. We looked at babies born in January, May, and September.

Group	Mean	Std Dev	Sample Size
January	29.84	7.08	32
May	28.58	8.07	27
September	33.83	6.93	38

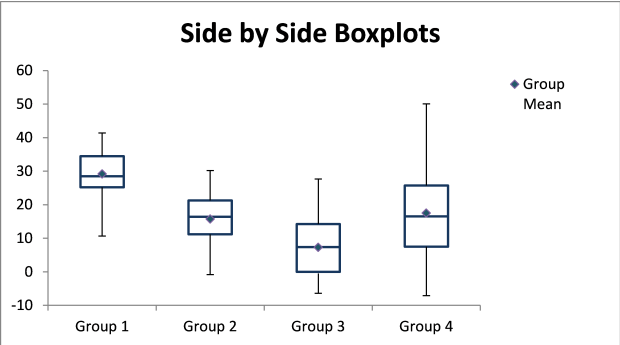
Source	df	SS	MS	F	p-value
Group	2	505.25	252.63	4.73	.011
Error	94	5024.09	53.45		
Total	96	5529.34			

- (a) Let's practice pulling values from the table. Use the ANOVA table from above. Tell me the:
- Group Mean Square
  - Total Sum of Squares
  - Total Degrees of Freedom
  - Sum of Squares for Error
  - Mean Square Error
  - Group Sum of Squares
  - Error Degrees of Freedom
  - F Test Statistic
  - P-value
- (b) Use the ANOVA output to determine if the month that a baby is born affects when the baby is able to crawl.

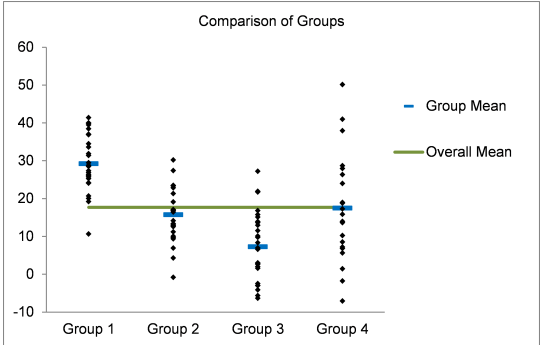
- (c) Find and interpret  $R^2$ .



20. We want to know if there is a difference in the population means for the four groups. Conduct a hypothesis test and find and interpret  $R^2$ .

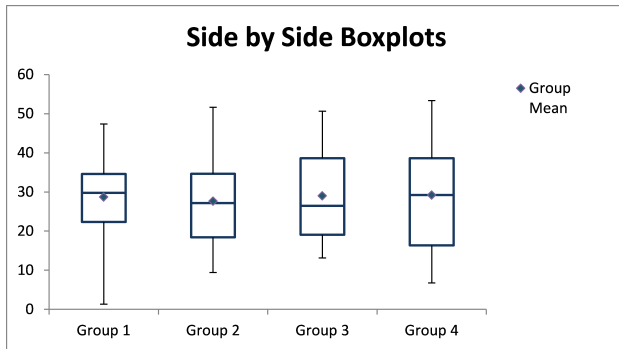


	Size	Mean	Standard Deviation
Group 1	29	29.19	7.30
Group 2	25	15.73	7.36
Group 3	27	7.28	9.20
Group 4	22	17.47	14.05
<b>Total</b>	<b>103</b>	<b>17.68</b>	<b>12.53</b>

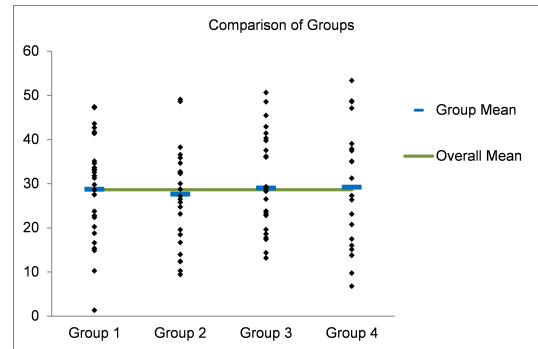


Source	df	SS	MS	F	p-value
Group	3	6861.46	2287.15	24.76	$4.9 \times 10^{-12}$
Error	99	9145.50	92.38		
Total	102	16006.95			

21. We want to know if there is a difference in the population means for the four groups. Conduct a hypothesis test and find and interpret  $R^2$ .



	Size	Mean	Standard Deviation
Group 1	29	28.71	11.19
Group 2	25	27.62	12.00
Group 3	27	29.01	11.28
Group 4	22	29.17	13.89
<b>Total</b>	<b>103</b>	<b>28.62</b>	<b>11.87</b>



Source	df	SS	MS	F	p-value
Group	3	35.80	11.93	0.08	0.9694
Error	99	14325.12	144.70		
Total	102	14360.91			

22. A factory uses four different machines to produce disc brakes. It is important that all the brakes have the same diameter. Conduct an appropriate hypothesis test to determine if there is a difference in the diameters of the brakes between machines. Find and interpret  $R^2$ .

Source	df	SS	MS	F	p-value
Group	3	3502.07	1167.36	11.75	$3.72 \times 10^{-6}$
Error	60	5962.05	99.37		
Total	63	9464.12			

23. A study looked at the SAT Math score for freshman at a random sample of colleges across the nation. The colleges were divided into three groups: Public, Private, or Church. Is there a difference in the SAT Math scores based on the type of college?

(a) Conduct a hypothesis test and find and interpret  $R^2$ .

Source	Sum of squares	DF	Mean Square	F	P-value
Groups	63906.2	2	31953.1	5.696	0.005
Error	353440.2	63	5610.2		
Total	417346.4	65			

(b) What is the correct conclusion?

- i. The average SAT Math scores for freshmen attending colleges with the three different affiliations appear to be the same.
- ii. Each of the three average SAT Math scores for freshmen attending colleges with the three different affiliations appear to be different.
- iii. It appears that freshmen attending at least one of the three different types of college have a different average SAT Math score.
- iv. Freshmen at one type of affiliated college have a significantly better average SAT Math score than the other two.

(c) Find and interpret  $R^2$ .

## Chi Square Tests Review

24. You want to know if a die is fair. You rolled it 120 times. Here are your results.

Outcome	Observed Count
one	20
two	32
three	26
four	5
five	37
six	0

Conduct a hypothesis test to determine if the die is fair.

- (a) Which test should you use?
- (b) Find the expected counts.
- (c) Conduct the test.

25. A study was done with randomly selected students in fourth, fifth, and sixth grade in Georgia. They were asked what their personal goals for school were. Their results are below. Conduct a hypothesis test to determine if their gender affects their goal.

	Boys	Girls
Make good grades	96	295
Be popular	32	45
Be good in sports	94	40

26. We want to know if the favorite size of a pizza affects which topping is preferred. We surveyed 100 randomly selected college students and asked them what size of pizza they prefer and their favorite topping. Conduct a hypothesis test to determine if the favorite size of a pizza affects the favorite topping. (Use the fact that  $\chi^2 = 22.07$ .)

	pepperoni	veggie	cheese	total
small	18	11	6	<b>35</b>
medium	14	12	7	<b>33</b>
large	3	9	20	<b>32</b>
<b>total</b>	<b>35</b>	<b>32</b>	<b>33</b>	<b>100</b>

I've found some of the expected counts for you.

	pepperoni	veggie	cheese	total
small			11.55	
medium			10.89	
large	11.20	10.24	10.56	
<b>total</b>				

27. A study looked at the number of births at a hospital in Switzerland and kept track of which month the babies were born. We want to know if the babies are evenly distributed throughout the year, or if there are certain months that have more babies than other months.

Month	Number of Births
January	66
February	63
March	64
April	48
May	64
June	74
July	70
August	59
September	54
October	51
November	45
December	42

- Which test should you use?
- Find the expected counts.
- Conduct the test. Use  $\chi^2 = 19.73$  so you don't have to calculate it.

28. When police officers respond to a call for help for a spousal abuse situation, they have three basic options: arrest the offender, issue a citation, advise and/or separate the couple. Conduct a hypothesis test to determine if which option the police officer chooses affects the number of subsequent arrests.

Number of subsequent arrests	arrest	citation	advise/separate
zero	175	181	187
one	36	33	24
two	2	7	1
three	1	1	0
four	0	2	0

To save you time, I found the expected counts:

Number of subsequent arrests	arrest	citation	advise/separate
zero	178.77	187.13	177.10
one	30.62	32.05	30.33
two	3.29	3.45	3.26
three	0.66	0.69	0.65
four	0.66	0.69	0.65

Also, the test statistic is  $\chi^2 = 13.64$ .

29. A supermarket claims that its mixed nuts are 30% cashews, 30% hazelnuts, and 40% peanuts (by weight). You bought 20 pounds of nuts and you aren't sure if you believe their proposed distribution. Because you have so much extra time, you divide the nuts and weight each type. You then want to conduct a hypothesis test.

Type of Nut	Weight
cashew	6
hazelnuts	5
peanuts	9

- Which test should you use?
- Find the expected counts.
- Conduct the test.

30. The World Series of baseball is typically the best of seven games. If the two teams are evenly matched, then the probability of the series lasting 4, 5, 6, or 7 games is listed in the table. The actual number of times the series lasted that number of times (up to 2004) is also listed. We want to know if there is evidence that the teams are unevenly matched or if the proposed distribution fits the data.

Games	Theoretical Probability	Actual number of times
four	0.125	17
five	0.25	23
six	0.3125	22
seven	0.3125	35

- (a) Which test should you use?
- (b) Find the expected counts.
- (c) Conduct the test. ( $\chi^2 = 5.03$ )

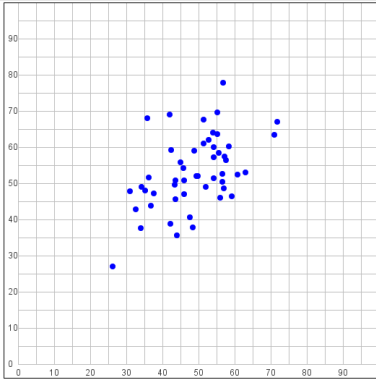
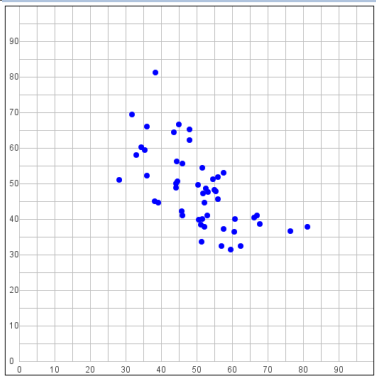
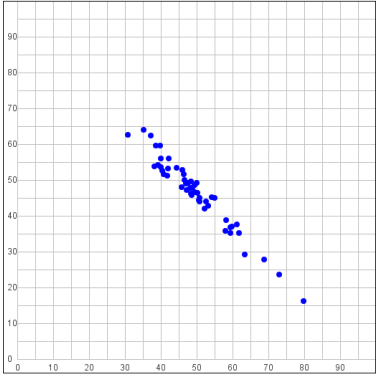
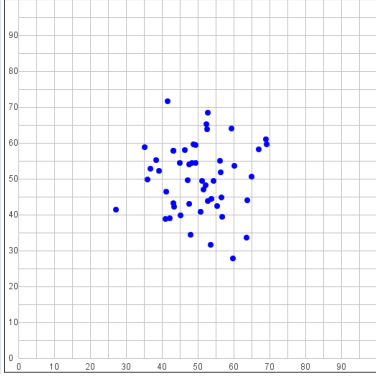
31. A study was conducted to see how many medical patients had supplemental health coverage and how many surgical patients had supplemental coverage. Conduct a test to determine if there is a relationship between whether or not the patient has supplemental health coverage and whether the patient is medical or surgical.

	medical patient	surgical patient
supplemental health	56	36
no supplemental health	69	59

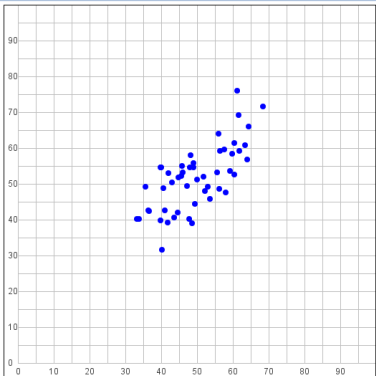
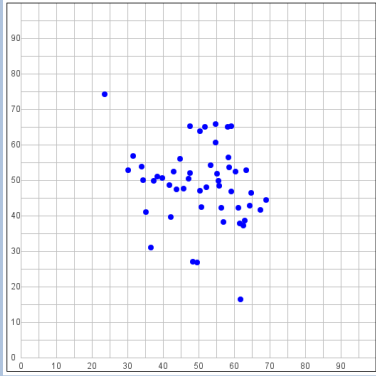
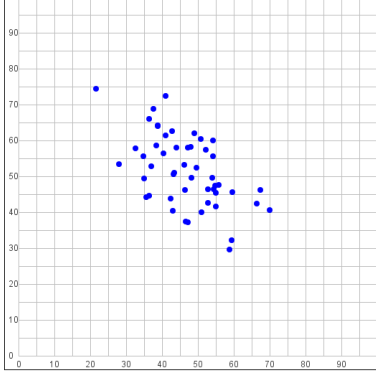
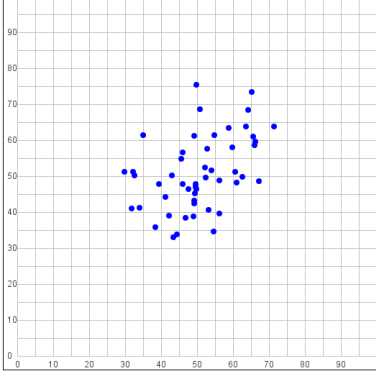


Linear Regression Review-Show Your Work

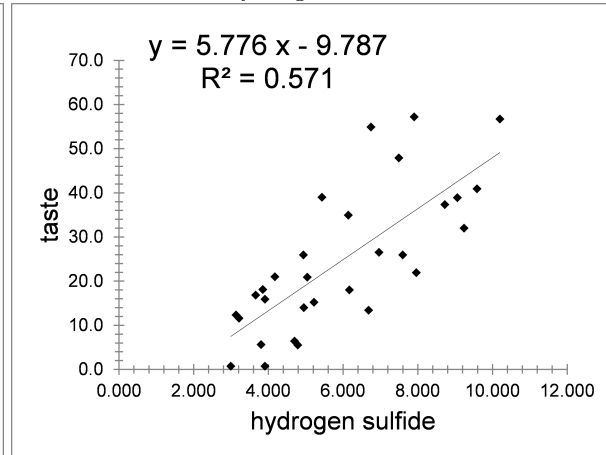
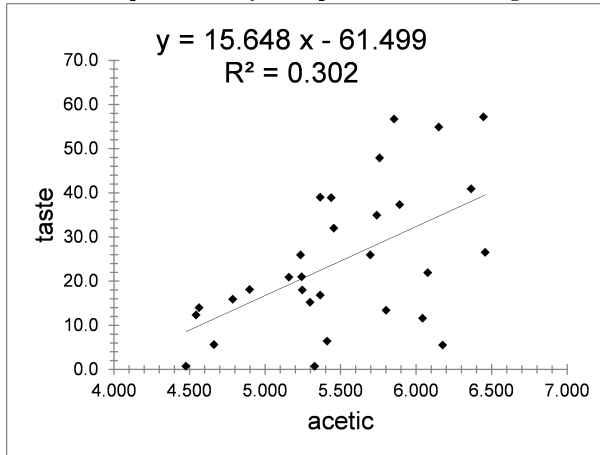
32. Match each plot to its correlation. All the plots are on the same scale. Your choices are  $-.98$ ,  $-.63$ ,  $.02$ , and  $.5$ .

Plot	Correlation
	
	
	
	

33. Match each plot to its correlation. All the plots are on the same scale. Your choices are  $-.55$ ,  $-.23$ ,  $.43$ , and  $.71$ .

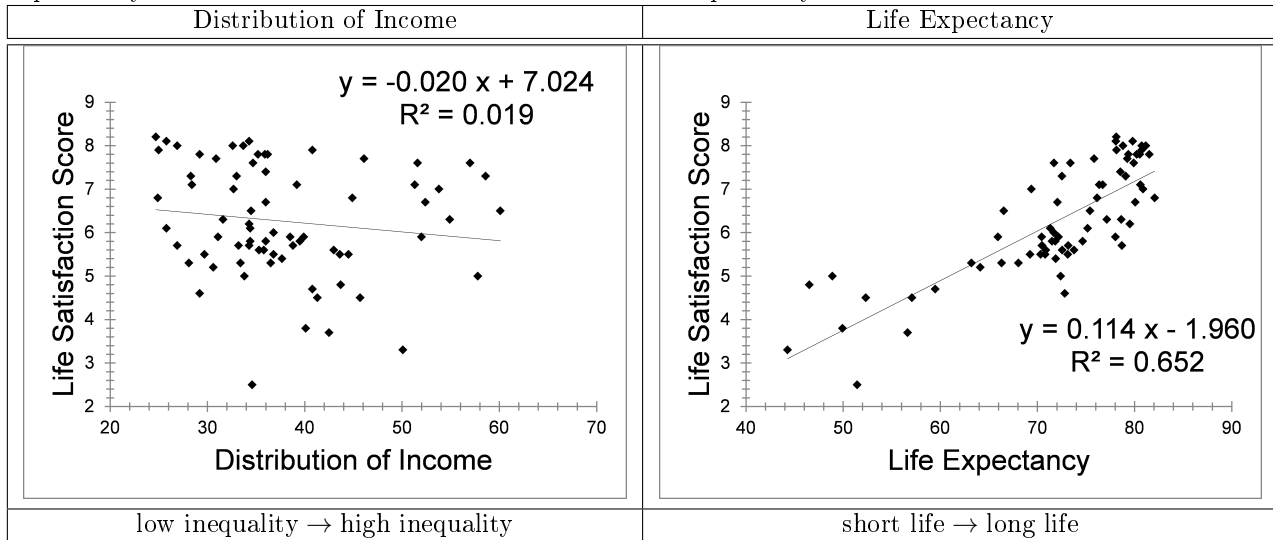
Plot	Correlation
	
	
	
	

34. The taste of cheddar cheese depends on the concentrations of several chemicals. You want to be able to predict the taste rating of cheese based on either acetic acid or hydrogen sulfide. You did a preliminary sample and linear regression for acetic acid and hydrogen sulfide.



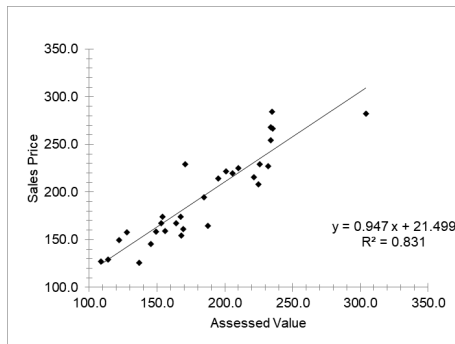
Would you rather use acetic acid or hydrogen sulfide to try to predict the taste rating of cheddar cheese? Why?

35. In 2007-2008, a study looked at the average happiness or life satisfaction score for 72 nations. Two of the explanatory variables were distribution of income and life expectancy.



If you had to choose one variable to try to predict the Life Satisfaction Score, would you choose distribution of income or life expectancy, and why?

36. A study looked at 28 randomly selected homes in a Midwestern city that recently sold. Every year, real estate is assessed for property tax purposes, but sales prices are usually different from the assessed price. We want to try to use the assessed values to predict the sales price. Both assessed value and sales price are measured in thousands of dollars.



### Regression Analysis

$r^2$  0.831                      n 30  
 $r$  0.912                        k 1  
 Std. Error 19.728            Dep. Var. **Sales Price**

Regression output					<i>confidence interval</i>	
<i>variables</i>	<i>coefficients</i>	<i>std. error</i>	<i>t (df=28)</i>	<i>p-value</i>	<i>90% lower</i>	<i>90% upper</i>
Intercept	21.4992	15.2794	1.407	.1704	-4.4930	47.4914
Assessed Value	0.9468	0.0806	11.741	2.49E-12	0.8096	1.0840

- Which variable should be the explanatory variable and which is the response variable?
- Conduct a hypothesis test to determine if there is a significant linear relationship between the assessed value and sales price.
- Do you have evidence that the increasing the assessed value causes a change in the sales price?
- What is the equation for the linear regression line? (Read it off the results)
- What is the y-intercept? Interpret. Is it a logical/practical interpretation?
- What is the slope? Interpret.

(g) Predict the mean sales price for all homes that were assessed at \$150,000.

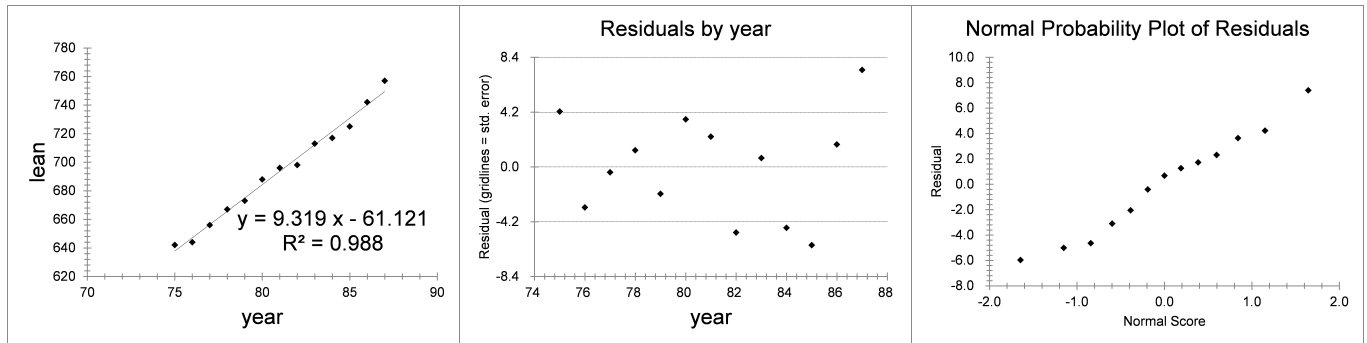
(h) Predict the sales price for John's house that was assessed at \$150,000.

(i) What is the correlation? Interpret.

(j) What is  $r^2$ ? Interpret.

(k) Find a 90% confidence interval for the slope  $\beta_1$ . Interpret.

37. The Leaning Tower of Pisa seems to lean more every year. Engineers measured the lean of the tower each year from 1975 to 1987. The lean is measured in tenths of millimeters past 2.9 meters. The year 1974 is coded as 74 and 1985 is coded as 85, etc.



### Regression Analysis

$r^2$  0.988                      n 13  
 $r$  0.994                          k 1  
 Std. Error 4.181              Dep. Var. lean

Regression output					confidence interval	
variables	coefficients	std. error	t (df=11)	p-value	95% lower	95% upper
Intercept	-61.1209	25.1298	-2.432	.0333	-116.4312	-5.8105
year	9.3187	0.3099	30.069	6.50E-12	8.6366	10.0008

- Which variable should be the explanatory variable and which is the response variable?
- Conduct a hypothesis test to determine if there is a significant positive linear relationship between the year and the lean.
- Do you have evidence that the increasing the year causes the tower to lean further?
- What is the equation for the linear regression line? (Read it off the results)
- What is the y-intercept? Interpret. Is it a logical/practical interpretation?
- What is the slope? Interpret.

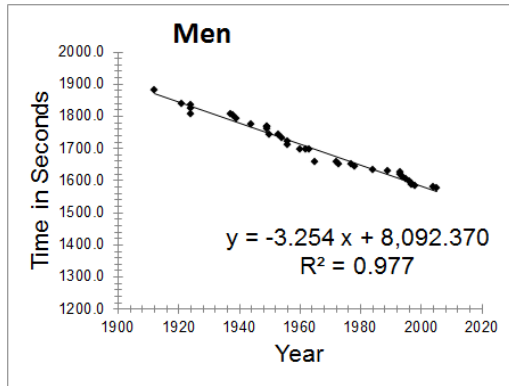
(g) Predict the lean in the year 1988.

(h) What is the correlation? Interpret.

(i) What is  $r^2$ ? Interpret.

(j) Find a 95% confidence interval for the slope  $\beta_1$ . Interpret.

38. Let's look at the world record times (in seconds) for the 10,000 meter run for just the men. The year variable is the actual year when the record was made. The year 1975 is coded as 1975, etc.



### Regression Analysis

$r^2$  0.977                      n 37  
 $r$  -0.988                        k 1  
 Std. Error 13.648              Dep. Var. **record**

### Regression output

variables	coefficients	std. error	t (df=35)	p-value	confidence interval	
					90% lower	90% upper
Intercept	8,092.3704	167.0960	48.429	.3241	7809.7	8,374.3000
year	-3.2544	0.0851	-38.240	1.90E-30	-3.4000	-3.1000

- Which variable should be the explanatory variable and which is the response variable?
- Conduct a hypothesis test to determine if there is a significant linear relationship between the year and the record time for men. Assume the conditions are met.
- Do you have evidence that the increasing the year causes male athletes to run faster?
- What is the equation for the linear regression line? (Read it off the results)
- What is the y-intercept? Interpret. Is it a logical/practical interpretation?
- What is the slope? Interpret.



(g) Predict the world record time for the year 1997.

(h) Predict the world record time for the year 2500.

(i) What is the correlation? Interpret.

(j) What is  $r^2$ ? Interpret.

(k) Find a 90% confidence interval for the slope  $\beta_1$ . Interpret.

## Multiple Choice

39. The dependent variable is the variable that is being described or predicted.
- (a) true
  - (b) false
40. The residual is the difference between the observed value of the dependent variable and the predicted value of the dependent variable.
- (a) true
  - (b) false
41.  $r^2$  is the percentage of variation in the  $y$  values that is explained by our least squares line.
- (a) true
  - (b) false
42. When using simple regression analysis, if there is a strong correlation between the independent and dependent variable, then we can conclude that an increase in the value of the independent variable *causes* an increase in the value of the dependent variable.
- (a) true
  - (b) false
43. If  $r = -1$ , then we can conclude that there is a perfect linear relationship between X and Y.
- (a) true
  - (b) false
44. A significant positive correlation between X and Y implies that changes in X cause Y to change.
- (a) true
  - (b) false
45. In a simple linear regression analysis, the correlation coefficient and the slope \_\_\_\_\_ have the same sign.
- (a) always
  - (b) sometimes
  - (c) never
46. \_\_\_\_\_ measures the strength of the linear relationship between the dependent and the independent variable.
- (a) Distance value
  - (b) Y Intercept
  - (c) Correlation coefficient
  - (d) Residual

47. The least squares regression line minimizes the sum of the
- (a) Differences between actual and predicted Y values
  - (b) Absolute deviations between actual and predicted Y values
  - (c) Absolute deviations between actual and predicted X values
  - (d) Squared differences between actual and predicted Y values
  - (e) Squared differences between actual and predicted X values
48. The \_\_\_\_\_ the  $r^2$ , the stronger the relationship between the dependent variable and the independent variable.
- (a) lower
  - (b) higher
49. In simple regression analysis the quantity that gives the amount by which Y (dependent variable) changes for a unit change in X (independent variable) is called the
- (a) Coefficient of determination
  - (b) Slope of the regression line
  - (c) Y intercept of the regression line
  - (d) Correlation coefficient
  - (e) Standard error
50. The correlation coefficient may assume any value between
- (a) 0 and 1
  - (b)  $-\infty$  and  $\infty$
  - (c) 0 and 8
  - (d) -1 and 1
  - (e) -1 and 0
51. If the correlation is positive, then the slope is
- (a) negative
  - (b) positive
  - (c) zero
  - (d) it could be negative or positive

52. The following results were obtained from a simple regression analysis:

$$\hat{y} = 37.2895 - 1.2024x$$

$$r^2 = .6744$$

$$SE_{b_1} = .2934$$

For each unit change in X (independent variable), the estimated change in Y (dependent variable) is equal to:

- (a) -1.2024
- (b) .6774
- (c) 37.2895
- (d) .2934

53. The following results were obtained from a simple regression analysis:

$$\hat{y} = 37.2895 - 1.2024x$$

$$r^2 = .6744$$

$$SE_{b_1} = .2934$$

If X is equal to zero, the estimated value of Y is:

- (a) -1.2024
- (b) .6774
- (c) 37.2895
- (d) .2934

54. The following results were obtained from a simple regression analysis:

$$\hat{y} = 37.2895 - 1.2024x$$

$$r^2 = .6744$$

$$SE_{b_1} = .2934$$

The proportion of variation in the y values that can be explained by our line is:

- (a) -1.2024
- (b) .6774
- (c) 37.2895
- (d) .2934

55. The strength of the relationship between two quantitative variables can be measured by:

- (a) The slope of a simple linear regression equation
- (b) The Y intercept of the simple linear regression equation
- (c) The coefficient of correlation
- (d) The standard error

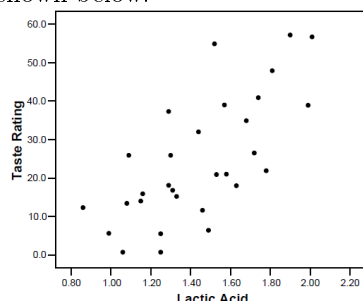
56. The local grocery store wants to predict the daily sales in dollars. The manager believes that the amount of newspaper advertising significantly affects the store sales. He randomly selects 7 days of data consisting of daily grocery store sales (in thousands of dollars) and advertising expenditures (in thousands of dollars). The Excel/Mega-Stat output given below summarizes the results of the regression model.

Regression Analysis						
	$r^2$	0.762	$n$	7		
	R	0.873	$k$	1		
	Std. Error	11.547	Dep. Var.	Sales		
Regression output						
<i>Variables</i>	<i>Coefficients</i>	<i>std. error</i>	<i>t (df=5)</i>	<i>p-value</i>	<i>Confidence interval</i>	
					95% lower	95% upper
Intercept	63.3333	7.9682	7.948	.0005	42.8505	83.8162
Advertising	6.6667	1.6667	4.000	.0103		

- (a) What is the estimated simple linear regression equation?
- $\hat{y} = 7.9682 + 1.667x$
  - $\hat{y} = 63.333 + 6.667x$
  - $\hat{y} = 7.948 + 4.000x$
  - $\hat{y} = 11.547 + 1.667x$
  - $\hat{y} = 6.667 + 63.333x$
- (b) If the manager decides to spend \$3000 on advertising, based on the simple linear regression results given above, the estimated sales are:
- \$68,333
  - \$20,063.33
  - \$83,333
  - \$20,064,333
  - \$70,000
- (c) At a significance level of  $\alpha = .05$ , test the significance of the slope (is the slope not zero) and state your conclusion. (use the p-value on the computer output)
- We reject  $H_0$  and conclude there is sufficient evidence that dollars spent on advertising is a useful linear predictor of the grocery store sales.
  - We failed to reject  $H_0$  and conclude there is not sufficient evidence that dollars spent on advertising is a useful linear predictor of the grocery store sales.
  - We failed to reject  $H_0$  and conclude there is sufficient evidence that dollars spent on advertising is a useful linear predictor of the grocery store sales.
  - We reject  $H_0$  and conclude that there is sufficient evidence that grocery store sales in dollars is a useful linear predictor of the dollars spent on advertising.
  - We reject  $H_0$  and conclude that there is not sufficient evidence that dollars spent on advertising is a useful linear predictor of the grocery store sales.

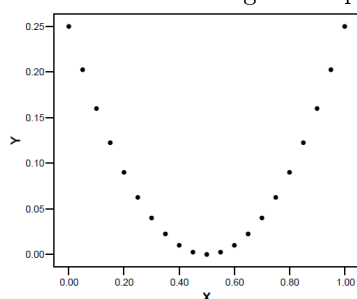
57. When creating a scatterplot, one should use the \_\_\_\_\_ axis for the explanatory variable.
- (a) x axis
  - (b) y axis
58. A study is conducted to determine if one can predict the yield of a crop based on the amount of yearly rainfall. The variable \_\_\_\_\_ is the response variable in this study.
- (a) yield of crop
  - (b) amount of rainfall
59. Negative linear relationships are represented by values of the correlation,  $r$ , that are \_\_\_\_\_.
- (a) greater than zero
  - (b) less than zero
  - (c) zero
  - (d) equal to 1 or -1
60. The lack of a linear relationship between two quantitative variables is represented by the correlation,  $r$ , with values \_\_\_\_\_.
- (a) greater than zero.
  - (b) less than zero.
  - (c) equal to zero.
  - (d) equal to 1 or -1.
61. A college newspaper interviews a psychologist about a proposed system for rating the teaching ability of faculty members. The psychologist says, "The evidence indicates that the correlation between a faculty member's research productivity and teaching rating is close to zero." What would be a correct interpretation of this statement?
- (a) Good researchers tend to be poor teachers and vice versa.
  - (b) Good teachers tend to be poor researchers and vice versa.
  - (c) Good researchers are just as likely to be good teachers as they are bad teachers. Likewise for poor researchers.
  - (d) Good research and good teaching go together.

62. As Swiss cheese matures, a variety of chemical processes take place. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study of cheese in a certain region of Switzerland, samples of cheese were analyzed for lactic acid concentration and were subjected to taste tests. The numerical taste scores were obtained by combining the scores from several tasters. A scatterplot of the observed data is shown below:



What is a plausible value for the correlation between lactic acid concentration and taste rating?

- (a) 0.999
  - (b) 0.7
  - (c) 0.07
  - (d) -0.7
  - (e) -0.999
63. Consider the following scatterplot of two variables  $x$  and  $y$ :



What can we conclude from this graph?

- (a) The correlation between  $x$  and  $y$  must be close to 1 because there is nearly a perfect relationship between them.
  - (b) The correlation between  $x$  and  $y$  must be close to -1 because there is nearly a perfect relationship between them, but it is not a straight-line relation.
  - (c) The correlation between  $x$  and  $y$  could be any number between -1 and +1. Without knowing the actual values, we can say nothing more.
  - (d) We shouldn't use correlation at all because it isn't a linear relationship.
64. A company computed the correlation between the length of their products ( $x$ ) and the height of their products ( $y$ ). It is  $r = .827$ . What would the correlation be if they computed the correlation between the height of their products ( $x$ ) and the length of their products ( $y$ )? (switch  $x$  and  $y$  variables)
- (a) -0.827
  - (b) .827
  - (c) 0
  - (d) we need to know the actual values

65. Which of the following best describes correlation?

- (a) Correlation measures the strength of the relationship between two quantitative variables whether or not the relationship is linear.
- (b) Correlation measures how much a change in the explanatory variable causes a change in the response variable.
- (c) Correlation measures the strength of the relationship between any two variables.
- (d) Correlation measures the strength of the linear relationship between two quantitative variables.
- (e) Correlation measures the strength of the linear association between two categorical variables.

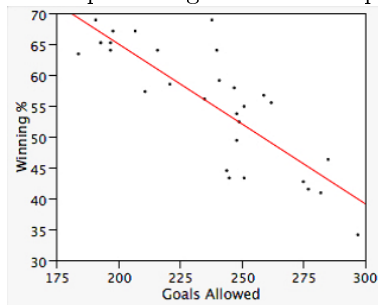
66. In a study of 1991 model cars, a researcher computed the least-squares regression line of price (in dollars) on horsepower. He obtained the following equation for this line.

$$\hat{price} = -6677 + 175(\text{horsepower})$$

Based on the least-squares regression line, what would we predict the cost to be of a 1991 model car with horsepower equal to 200?

- (a) \$41,677
- (b) \$35,000
- (c) \$28,323
- (d) \$13,354
- (e) We don't have enough information. We need to know the correlation.

67. In the National Hockey League a good predictor of the percentage of games won by a team is the number of goals the team allows during the season. Data were gathered for all 30 teams in the NHL and the scatterplot of their **Winning Percentage** against the number of **Goals Allowed** in the 2006/2007 season with a fitted least-squares regression line is provided:



The results are  $\hat{y} = 116.95 - .26x$  and  $r^2 = .69$ .

Which of the following provides the best interpretation of the slope of the regression line?

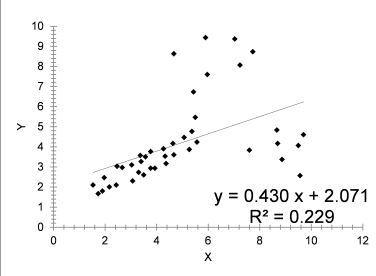
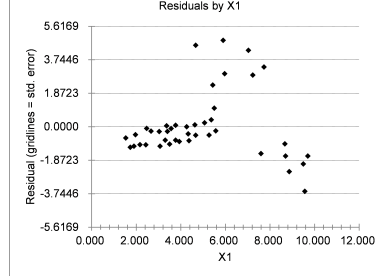
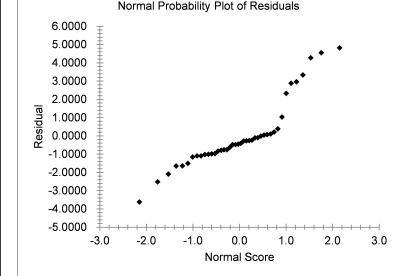
- (a) If the Winning Percent increases by 1%, then the number of Goals Allowed decreases by 0.26.
- (b) If a team were to allow 100 goals during the season, their Winning % would be 90.95%.
- (c) If Goals Allowed increases by one goal, the Winning % increases by 0.26%.
- (d) If the Winning % increases by 1%, then the number of Goals Allowed increases by 0.26.
- (e) If Goals Allowed increases by one goal, the Winning % decreases by 0.26%.

68. Sean conducted a hypothesis test for  $H_0 : \beta_1 = 0$  versus  $H_A : \beta_1 < 0$  for his study. He found a p-value of .00000000324. Which conclusion(s) is correct?

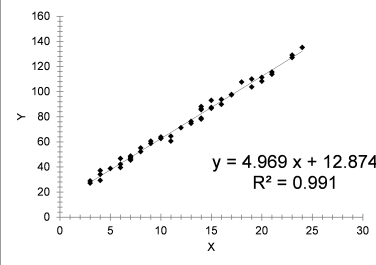
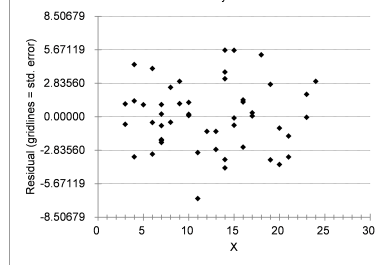
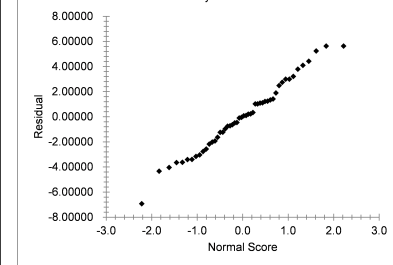
- (a) The p-value is so small that we are very sure that the points are very close to the line.
- (b) The p-value is so small that we are very sure that there is a negative linear relationship.
- (c) The p-value is so small that we can say that his x variable causes the change in his y variable.



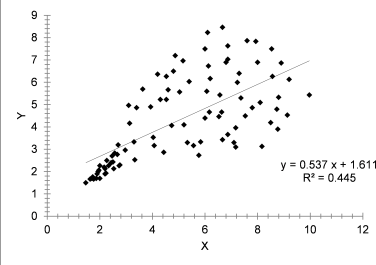
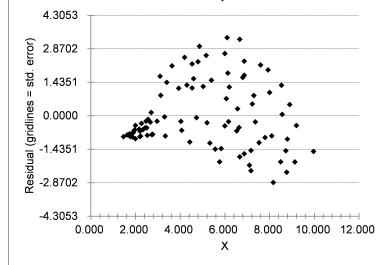
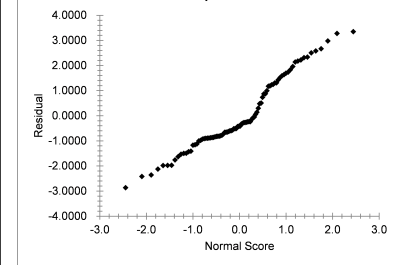
69. Is linear regression appropriate for the data set?

scatterplot of actual data	plot of residuals against explanatory variable	normality plot for residuals
(check to see if the data seems to have a linear relationship)	(there shouldn't be any pattern, the points should be randomly distributed)	(residuals are supposed to be normal, so the normality plot should be a straight line)
		

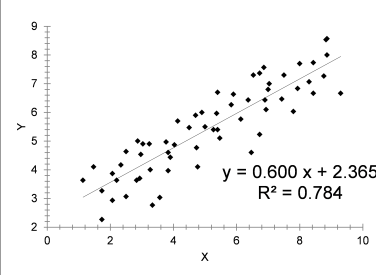
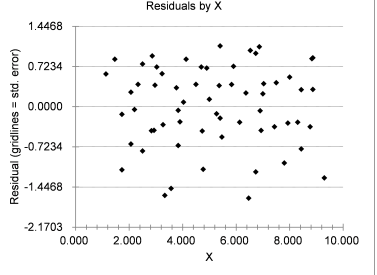
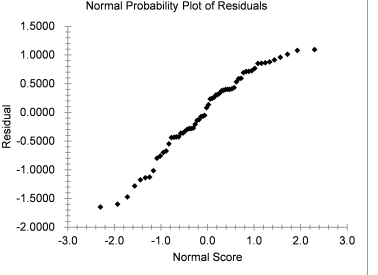
70. Is linear regression appropriate for the data set?

scatterplot of actual data	plot of residuals against explanatory variable	normality plot for residuals
(check to see if the data seems to have a linear relationship)	(there shouldn't be any pattern, the points should be randomly distributed)	(residuals are supposed to be normal, so the normality plot should be a straight line)
		

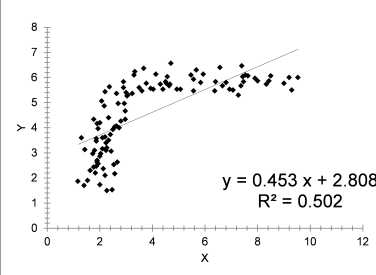
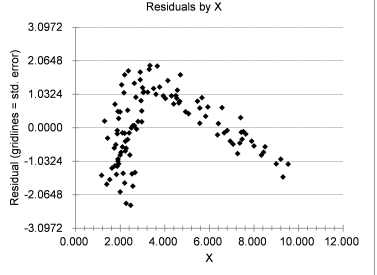
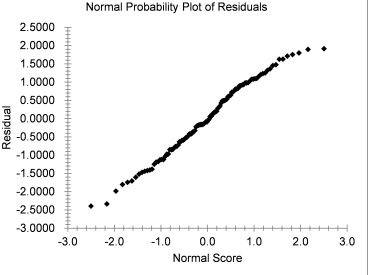
71. Is linear regression appropriate for the data set?

scatterplot of actual data	plot of residuals against explanatory variable	normality plot for residuals
(check to see if the data seems to have a linear relationship)	(there shouldn't be any pattern, the points should be randomly distributed)	(residuals are supposed to be normal, so the normality plot should be a straight line)
		

72. Is linear regression appropriate for the data set?

scatterplot of actual data	plot of residuals against explanatory variable	normality plot for residuals
(check to see if the data seems to have a linear relationship)	(there shouldn't be any pattern, the points should be randomly distributed)	(residuals are supposed to be normal, so the normality plot should be a straight line)
		

73. Is linear regression appropriate for the data set?

scatterplot of actual data	plot of residuals against explanatory variable	normality plot for residuals
(check to see if the data seems to have a linear relationship)	(there shouldn't be any pattern, the points should be randomly distributed)	(residuals are supposed to be normal, so the normality plot should be a straight line)
		

74. Is linear regression appropriate for the data set?

scatterplot of actual data	plot of residuals against explanatory variable	normality plot for residuals
(check to see if the data seems to have a linear relationship)	(there shouldn't be any pattern, the points should be randomly distributed)	(residuals are supposed to be normal, so the normality plot should be a straight line)
