

Stat 3000 Review for Exam 3 KEY

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

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 $\alpha = .01$)

Target: $n_T = 36$, $\bar{x}_T = 1.52$, $s_T = .92$

Bidder: $n_B = 36$, $\bar{x}_B = 1.20$, $s_B = .93$

Test: Two sample T Test. I am going to use EQUAL VARIANCES.

condition: Both sample sizes are at least 30. ✓

Level of significance: $\alpha = .01$

Hypotheses: $H_0 : \mu_T = \mu_B$ versus $H_A : \mu_T \neq \mu_B$

Test Statistic:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(36 - 1)(.92)^2 + (36 - 1)(.93)^2}{36 + 36 - 2} = .855$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{1.52 - 1.20}{\sqrt{.855 \left(\frac{1}{36} + \frac{1}{36} \right)}} = \frac{.32}{.2180} = 1.47$$

$$df = n_1 + n_2 - 2 = 36 + 36 - 2 = 70$$

We will have to round down to $df = 60$ to use the table.

P-value: area in both tails. The area in the right tail is between .10 and .05. The total p-value is between .20 and .10.

Conclusion: big p-value, fail to reject H_0

Interpret: we didn't find any evidence that the population means are not equal

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.

Current: $n_{cur} = 31$, $\bar{x}_{cur} = 11$, $s_{cur} = 3.4$

Competing: $n_{com} = 35$, $\bar{x}_{com} = 9.5$, $s_{com} = 1.6$

Test: Two sample T test, Unequal variances.

condition: both sample sizes at least 30. ✓

Level of significance: you pick α . I pick $\alpha = .05$.

Hypotheses: $H_0 : \mu_{cur} = \mu_{com}$ versus $H_A : \mu_{cur} > \mu_{com}$

Test Statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{11 - 9.5}{\sqrt{\frac{3.4^2}{31} + \frac{1.6^2}{35}}} = 2.246$$

Satterhwaite degrees of freedom:

$$df = 41.514 \text{ round to } 41$$

P-value: area in right tail. Using the table we find p-value is between .025 and .01. The computer gives the exact p-value is

$$p - \text{value} = .01504$$

Conclusion: small p-value; reject null hypothesis

Interpret: we have strong evidence that $\mu_{cur} > \mu_{com}$. (i.e. strong evidence that the mean of all the calls with the current provider is greater than the mean of all the calls with 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.

Onkyo: $n_O = 6$, $\bar{x}_O = 189$, $\sigma_O^2 = 144$

JVC: $n_J = 12$, $\bar{x}_J = 145$, $\sigma_J^2 = 100$

Test: Two Sample Z Test

condition: normal \checkmark

Level of significance: I chose $\alpha = .05$.

Hypotheses: $H_0 : \mu_O = \mu_J$ versus $H_A : \mu_O > \mu_J$

Test Statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{189 - 145}{\sqrt{\frac{144}{6} + \frac{100}{12}}} = 7.74$$

P-value: use z table; area in right tail. p-value is essentially 0

Conclusion: small p-value; reject H_0

Interpret: we have extremely strong evidence that $\mu_O > \mu_J$ (i.e. that the population mean cost of Onkyo CD players is higher than the population mean cost of JVC CD players.)

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.

Brand X: $n_1 = 15$, $\bar{x}_1 = 255$, $s_1 = 8.7$

Brand Y: $n_2 = 14$, $\bar{x}_2 = 271$, $s_2 = 9.1$

Test: Two Sample T Test, you could have chosen Equal or Unequal Variances

condition: normal \checkmark

Level of significance: I choose $\alpha = .05$.

Hypotheses: $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 < \mu_2$

- **Unequal Variances:**

- **Test Statistic:**

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{255 - 271}{\sqrt{\frac{8.7^2}{15} + \frac{9.1^2}{14}}} = -4.83$$

satterthwaite degrees of freedom:

$$df = 26.639 \text{ round to } 26$$

****We always round down on degrees of freedom.**

- **P-value:** area in left tail; Table gives us p-value is less than .0005. The exact p-value is

$$2.465 \times 10^{-5}$$

Conclusion: small p-value; reject null hypothesis

Interpret: we have extremely strong evidence that $\mu_1 < \mu_2$ (i.e. that the population mean of brand X is smaller than the population mean of brand Z)

Real Life: on average, brand Z clubs will hit the bar further than Brand X clubs

- **Equal Variances:**

- **Test Statistic:**

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(15 - 1)(8.7)^2 + (14 - 1)(9.1)^2}{15 + 14 - 2} = 79.12$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{255 - 271}{\sqrt{79.12 \left(\frac{1}{15} + \frac{1}{14} \right)}} = -4.84$$

$$df = n_1 + n_2 - 2 = 15 + 14 - 2 = 27$$

- **P-value:** use T table; area in left tail; p-value is less than .0005. (exact p-value is .00002)

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.

1995: $n_1 = 40$, $\bar{x}_1 = 151000$, $s_1 = 5332$

1996: $n_2 = 35$, $\bar{x}_2 = 160000$, $s_2 = 7468$

Test: Two Sample T Test; Unequal Variances

condition: both sample sizes are at least 30 ✓

Level of significance: $\alpha = .05$

Hypotheses: $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 < \mu_2$

Test Statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{151000 - 160000}{\sqrt{\frac{5332^2}{40} + \frac{7468^2}{35}}} = \frac{-9000}{1517.9} = -5.93$$

Satterthwaite degrees of freedom: $df = 60.58$, round down to

$$df = 60$$

P-value: area in left tail;

$$p - value = 7.843 \times 10^{-8}$$

Conclusion: small p-value; reject H_0

Interpret: we have extremely strong evidence that $\mu_1 < \mu_2$ (i.e. that the mean selling price of all homes in 1995 is smaller than the mean selling price of all homes in 1996)

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 $\alpha = .05$.

Men: $n = 34$, $\bar{x} = 25.6$, $\sigma^2 = 17.64$, $\sigma = \sqrt{17.64} = 4.2$

Women: $n = 28$, $\bar{x} = 21.2$, $\sigma^2 = 14.44$, $\sigma = \sqrt{14.44} = 3.8$

Test: Two Sample Z Test

condition: one sample size is large, but the other sample size is less than 30. But we are told to assume normality, so we are good. \checkmark

Level of significance: $\alpha = .05$

Hypotheses: $H_0 : \mu_M = \mu_W$ versus $H_A : \mu_M > \mu_W$

Test Statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{25.6 - 21.2}{\sqrt{\frac{4.2^2}{34} + \frac{3.8^2}{28}}} = \frac{25.6 - 21.2}{\sqrt{\frac{17.64}{34} + \frac{14.44}{28}}} = \frac{4.4}{1.017} = 4.325$$

P-value: use Z table; area in right tail; p-value is essentially 0

$$p - value = 7.6 \times 10^{-6} \text{ or } 0.0000076$$

Conclusion: small p-value, reject H_0

Interpret: we found extremely strong evidence that $\mu_M > \mu_W$ (i.e. evidence that on average, men eat out more than women)

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.

Test: two sample T test; unequal variances

condition: both sample sizes are at least 30 \checkmark

Level of significance: I choose $\alpha = .05$

Hypotheses: $H_0 : \mu_{\text{online}} = \mu_{\text{onsite}}$ versus $H_A : \mu_{\text{online}} \neq \mu_{\text{onsite}}$

Test Statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{76 - 83}{\sqrt{\frac{13.77^2}{47} + \frac{11.18^2}{46}}} = -2.69$$

Satterthwaite degrees of freedom: $df = 88.106$ round to

$$df = 88$$

P-value: area in both tails

$$p - value = .008454$$

Conclusion: small p-value; reject H_0

Interpret: we have very strong evidence that the mean score for all her online students is different from the mean scores for all her onsite students.

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.)

Their City: $n = 12$, $\bar{x} = 217000$, $s = 2200$

Neighboring City: $n = 15$, $\bar{x} = 167607$, $s = 2100$

Test: Two sample T test; Equal Variances

condition: normal populations ✓

Level of significance: I choose $\alpha = .05$.

Hypotheses: $H_0 : \mu_T = \mu_N$ versus $H_A : \mu_T > \mu_N$

Test Statistic:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{11(2200)^2 + 14(2100)^2}{12 + 15 - 2} = 4599200$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = 59.47$$

$$df = n_1 + n_2 - 2 = 25$$

P-value: use T table; area in right tail; p-value is less than .0005.

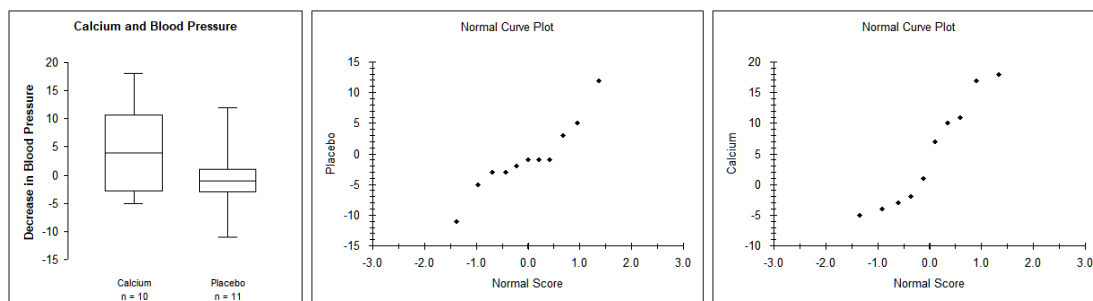
(computer says exact p-value is 9.51×10^{-29})

Conclusion: small p-value; reject $H_0 : \mu_T = \mu_N$

Interpret: we have extremely strong evidence that $\mu_T > \mu_N$ (i.e. extremely strong evidence that the population mean space in their city is greater than the population mean space in the neighboring city.)

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



We start off this problem by deciding that we will use a T test because we know the SAMPLE standard deviations. The boxplots and the sample standard deviations convince me to use UNEQUAL VARIANCES. But when I check the conditions I realize that both sample sizes are less than 30 and both of the normal quantile plots have a definite curve. So I don't think either population is normal. So the conditions aren't met and we can't do this problem.

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

One of the sample sizes is less than 30. We don't have a plot of the data, so we can't tell if it is normal. So the conditions aren't met. We can't do this problem.
(You would have to use non-parametric techniques.)

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.

We subtract Pre-Post. If the attack had a negative impact, then Post should be smaller than Pre and the differences should be positive.

$$n_d = 12, \bar{x}_d = 29.7, s_d = 2.975$$

Test: Paired Differences (Use One Sample T Test)

condition: differences are normally distributed ✓

Level of significance: you can pick α . I choose $\alpha = .05$.

Hypotheses: $H_0 : \mu_{\text{diff}} = 0$ versus $H_A : \mu_{\text{diff}} > 0$

Test Statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{29.7 - 0}{2.975/\sqrt{12}} = 34.58$$

$$df = n - 1 = 12 - 1 = 11$$

P-value: area in right tail. P-value is essentially 0.

Conclusion: small p-value; we reject H_0

Interpret: We found extremely strong evidence that $\mu_{\text{diff}} > 0$. (i.e. we found extremely strong evidence that the population mean of pre-attack mileages is greater than post-attack mileages.)

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)

Difference=Post-Pre

If treatment works, then Post should be smaller than Pre and the differences should be negative.

$n_d = 9$, $\bar{x}_d = -86$, $s_d = 101.83$

Test: Matched Pairs, use one sample T test

condition: normal \checkmark

Level of significance: $\alpha = .05$

Hypotheses: $H_0 : \mu_{\text{diff}} = 0$ versus $H_A : \mu_{\text{diff}} < 0$

Test Statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{-86 - 0}{101.83/\sqrt{9}} = -2.533$$

$$df = n - 1 = 8$$

P-value: use T table; area in left tail; p-value is between .025 and .01.

Conclusion: small p-value; reject H_0

Interpret: We have strong evidence that $\mu_{\text{diff}} < 0$. (i.e. evidence that the mean obstructions are less for post treatment than the mean obstructions for pre treatment.)

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.

<https://youtu.be/9SbJRksAcN8>

Difference=Computability-PC

If PC is cheaper then the differences will be positive.

$n_d = 10$, $\bar{x}_d = 7$, $s_d = 2$

Test: Matched Pairs, use one sample T test

condition: normal \checkmark

Level of significance: $\alpha = .10$

Hypotheses: $H_0 : \mu_{\text{diff}} = 0$ versus $H_A : \mu_{\text{diff}} > 0$

Test Statistic: (just use your calculator) $t = 11.06$

P-value: area in right tail; p-value is 7.68×10^{-7}

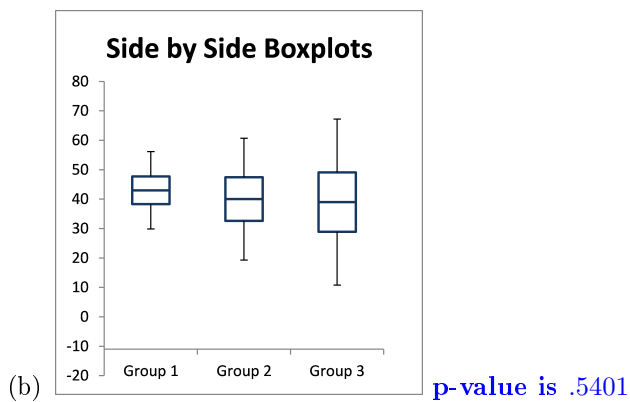
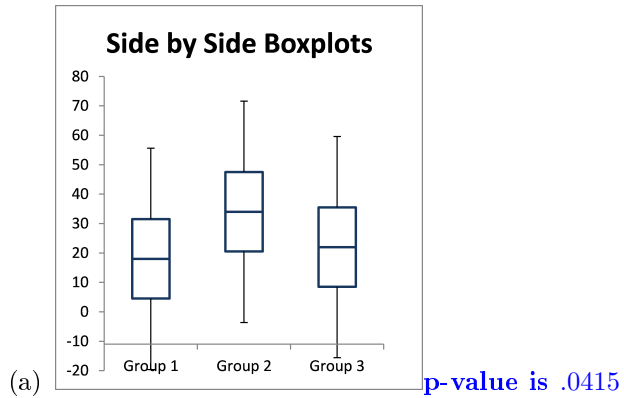
Conclusion: small p-value; reject H_0

Interpret: We have extremely strong evidence that $\mu_{\text{diff}} > 0$. (i.e. evidence that the population mean at PC connection is smaller than the population mean at Computability.)

ANOVA Review

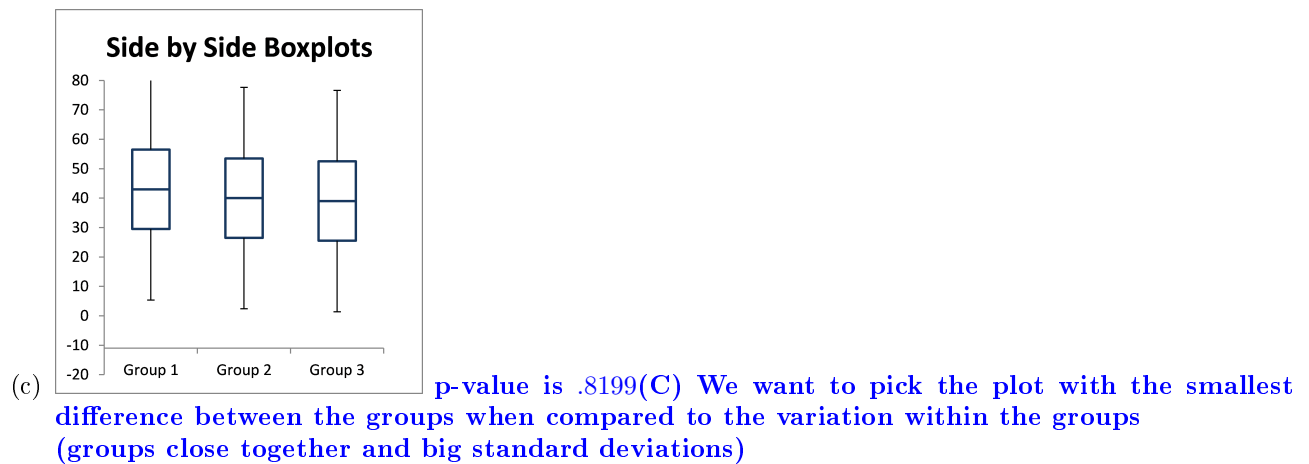
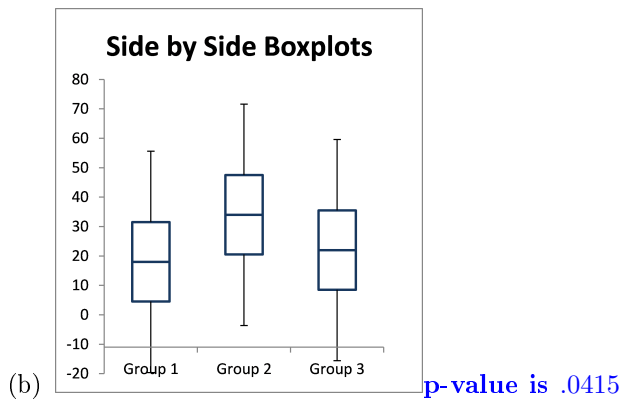
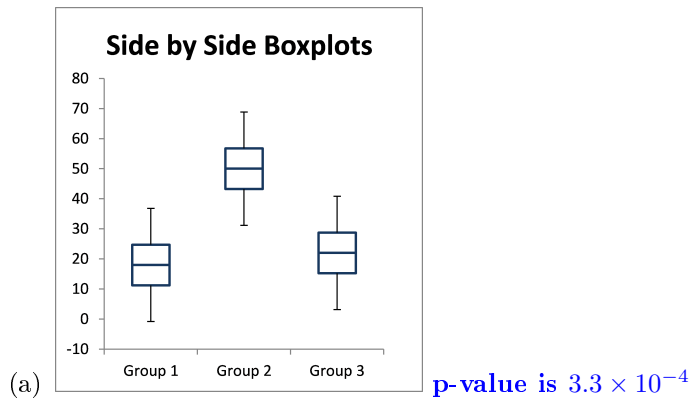
14. For which data set will you be most likely to believe that the population means are different? (The plots are all on the same scale.)

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.



(A) We want to pick the plot with the biggest difference between the groups when compared to the variation within the groups
(groups far apart and small standard deviations)

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.
- (C) all ANOVA tells us is that at least one of the population means is different from the others**

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
- (B) The p-value for ANOVA is always the area to the right.**

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
- (B) at least one of the population means is different from the others**

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 **252.63**
- Total Sum of Squares **5529.34**
- Total Degrees of Freedom **96**
- Sum of Squares for Error **5024.09**
- Mean Square Error **53.45**
- Group Sum of Squares **505.25**
- Error Degrees of Freedom **94**
- F Test Statistic **4.73**
- P-value **.011**

- (b) Use the ANOVA output to determine if the month that a baby is born affects when the baby is able to crawl.

ANOVA

$\alpha = .05$

H_0 : all the population means are equal

H_A : at least one of the population means is different from the others

$F = 4.73$

$p - value = .011$

Small p-value, reject H_0

We found strong evidence that the mean time to crawl is for at least one of the birth months is different from the others.

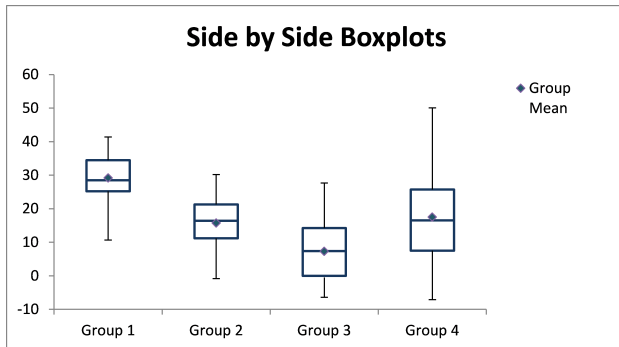
(We don't know which mean is different, or how many means are different. We would need to do a "Post-Hoc" test to determine which mean(s) is different.)

- (c) Find and interpret R^2 .

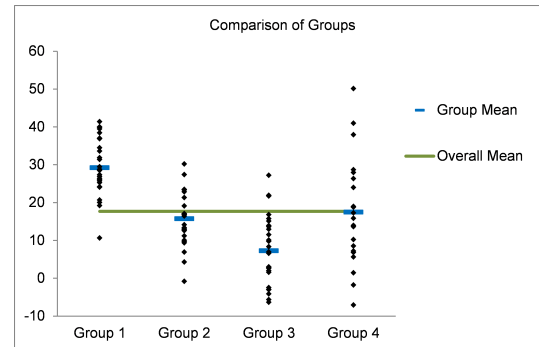
$$R^2 = \frac{SSG}{SST} = \frac{505.25}{5529.34} = .0913$$

So 9% of the variation in how long it takes a baby to crawl is explained by which month the baby was born in. The rest of the variation is natural variation from baby to baby born in the same month.

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
Total	103	17.68	12.53



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

ANOVA

$$\alpha = .05$$

H_0 : all the population means are equal

H_A : at least one of the population means is different from the others

$$F = 24.76$$

$$p\text{-value} = 4.9 \times 10^{-12}$$

Small p-value, reject H_0

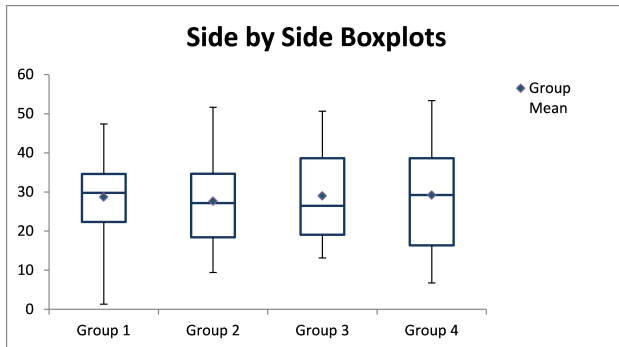
We found extremely strong evidence that at least one of the population means is different from the others.

(We don't know which mean is different, or how many means are different. We would need to do a "Post-Hoc" test to determine which mean(s) is different.)

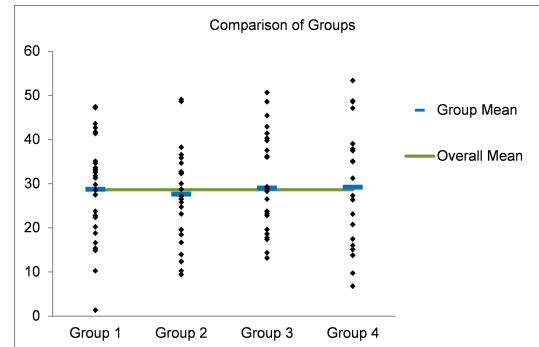
$$R^2 = \frac{SSG}{SST} = \frac{6861.46}{16006.95} = .4286$$

So 42.86% of the variation in the data is explained by which group the observation belongs to.

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
Total	103	28.62	11.87



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			

ANOVA

$$\alpha = .05$$

H_0 : all the population means are equal

H_A : at least one of the population means is different from the others

$$F = .08$$

$$p - value = .9694$$

big p-value, fail to reject H_0

We didn't found any evidence that any of the population means are different from each other.

$$R^2 = \frac{SSG}{SST} = \frac{35.8}{14360.91} = .0025$$

So .25% of the variation in the data is explained by which group the observation belongs to.

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	3.72×10^{-6}
Error	60	5962.05	99.37		
Total	63	9464.12			

ANOVA

$$\alpha = .05$$

H_0 : all four machines produce the same population mean diameter

H_A : at least one of the machines produces a population mean diameter different from the other machines

$$F = 11.75$$

$$p\text{-value} = 3.72 \times 10^{-6}$$

Small p-value, reject H_0

We found extremely strong evidence that at least one of the machines has a different population mean diameter from the others.

$$R^2 = \frac{SSG}{SST} = \frac{3502.07}{9464.12} = .37$$

So 37% of the total variation is explained by which machine produced the brakes. (The rest of the variation is the natural variation from brake to brake from the same machine.)

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			

ANOVA

$$\alpha = .05$$

H_0 : all the population means are equal

H_A : at least one of the population means is different from the others

$$F = 5.696$$

$$p - \text{value} = .005$$

Small p-value, reject H_0

We found very strong evidence that at least one of the population means is different from the others.

(We don't know which mean is different, or how many means are different. We would need to do a "Post-Hoc" test to determine which mean(s) is different.)

- (b) What is the correct conclusion?

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

(iii) at least one population mean is different

- (c) Find and interpret R^2 .

$$R^2 = \frac{SSG}{SST} = \frac{63906.2}{417346.4} = .153$$

So 15.3% of the total variation is explained by which type of college the person is at. The rest of the variation is due to the variation from person to person within a type of college .

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?

We want to know if each of the probabilities are $1/6$.

So we want to use the Chi Square Test for Goodness of Fit.

- (b) Find the expected counts.

The expected counts are $n(p_i)$ where n is the sample size and p_i is the theoretical probability for that group.

Outcome	Observed Count	Theoretical Probability	Expected Count
one	20	$1/6$	$120(1/6) = 20$
two	32	$1/6$	20
three	26	$1/6$	20
four	5	$1/6$	20
five	37	$1/6$	20
six	0	$1/6$	20

- (c) Conduct the test.

condition: all the expected counts are at least 5. \checkmark

I choose $\alpha = .05$.

H_0 : the probabilities are all $1/6$ (the die is fair)

H_A : at least one of the probabilities is not $1/6$ (the die is not fair)

Test statistic:

$$\begin{aligned}
 \chi^2 &= \frac{(obs - exp)^2}{exp} \\
 &= \frac{(20 - 20)^2}{20} + \frac{(32 - 20)^2}{20} + \frac{(26 - 20)^2}{20} + \frac{(5 - 20)^2}{20} + \frac{(37 - 20)^2}{20} + \frac{(0 - 20)^2}{20} \\
 &= 54.7
 \end{aligned}$$

Degrees of Freedom:

$$df = \text{number of categories} - 1 = 6 - 1 = 5$$

P-value: p-value is the area to the right of $\chi^2 = 54.7$. So it is smaller than .0005. Software gives p-value as 1.5×10^{-10} .

small p-value; reject H_0

We found extremely strong evidence that at least one of the probabilities is not $1/6$. (So extremely strong evidence that the die is not fair.)

****We don't know how many probabilities are not $1/6$, just that at least one probability is wrong.**

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

- To find the expected counts, we use the formula $\frac{(\text{row total})(\text{column total})}{\text{total sample size}}$.

	Boys	Girls	Total
Make good grades	144.19	246.81	391
Be popular	28.40	48.60	77
Be good in sports	49.42	84.58	134
Total	222	380	602

- condition: all expected counts are at least 5 ✓
- $\alpha = .05$
- H_0 : gender and goal are independent (no relationship; don't affect each other)
- H_A : gender and goal are Dependent (relationship, affect each other)
- Test statistic:

$$\begin{aligned}
 \chi^2 &= \frac{(obs - exp)^2}{exp} \\
 &= \frac{(96 - 144.19)^2}{144.19} + \dots + \frac{(40 - 84.58)^2}{84.58} \\
 &= 89.97
 \end{aligned}$$

- Degrees of Freedom:

$$df = (r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$$

- P-value: area to the right of $\chi^2 = 89.97$. The p-value is less than .0005. (Computer p-value is 2.91×10^{-20})
- small p-value; reject H_0
- We have extremely strong evidence that gender and goal are dependent. (i.e. evidence that they affect each other)

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	35
medium	14	12	7	33
large	3	9	20	32
total	35	32	33	100

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	
total				

- To find the expected counts, we use the formula $\frac{(\text{row total})(\text{column total})}{\text{total sample size}}$.

	pepperoni	veggie	cheese	total
small	12.25	11.20	11.55	35
medium	11.55	10.56	10.89	33
large	11.20	10.24	10.56	32
total	35	32	33	100

- condition: all expected counts are at least 5 ✓
- $\alpha = .05$
- H_0 : size and topping are independent (no relationship; don't affect each other)
- H_A : size and topping are independent are Dependent (relationship, affect each other)
- Test statistic:

$$\begin{aligned}
 \chi^2 &= \frac{(\text{obs} - \text{exp})^2}{\text{exp}} \\
 &= \frac{(18 - 12.25)^2}{12.25} + \dots + \frac{(20 - 10.56)^2}{10.56} \\
 &= 22.07
 \end{aligned}$$

- Degrees of Freedom:

$$df = (r - 1)(c - 1) = (3 - 1)(3 - 1) = 4$$

- P-value: area to the right of $\chi^2 = 22.07$. The p-value is less than .0005. Computer p-value is .0002.
- small p-value; reject H_0
- We have extremely strong evidence that size and topping are dependent. (i.e. evidence that they affect each other)

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

- (a) Which test should you use?

We want to know if each of the probabilities are $1/12$.

So we want to use the Chi Square Test for Goodness of Fit.

- (b) Find the expected counts.

The expected counts are $n(p_i)$ where n is the sample size and p_i is the theoretical probability for that group.

The expected count for each month will be $700(1/12) = 58.33$.

- (c) Conduct the test. Use $\chi^2 = 19.73$ so you don't have to calculate it.

condition: all the expected counts are at least 5. \checkmark

I choose $\alpha = .05$.

H_0 : the probabilities are all $1/12$ (all the months are equally likely)

H_A : at least one of the probabilities is not $1/12$ (not all the months are equally likely)

Test statistic:

$$\begin{aligned}
 \chi^2 &= \frac{(obs - exp)^2}{exp} \\
 &= \frac{(66 - 58.33)^2}{58.33} + \dots + \frac{(42 - 58.33)^2}{58.33} \\
 &= 19.73
 \end{aligned}$$

Degrees of Freedom:

$$df = \text{number of categories} - 1 = 12 - 1 = 11$$

P-value: p-value is the area to the right of $\chi^2 = 19.73$. So it is between .05 and .025. Software gives p-value as .0492.

small p-value; reject H_0

We found strong evidence that at least one of the probabilities is not $1/12$. (So at least one month does not have an equal share of births.)

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 $\chi^2 = 13.64$.

- To find the expected counts, we would use the formula $\frac{(\text{row total})(\text{column total})}{\text{total sample size}}$.
- condition: not all expected counts are at least 5 ✕
- We shouldn't do this test!

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

- (a) Which test should you use?

We want to know if the proposed distribution is correct.

So we want to use the Chi Square Test for Goodness of Fit.

- (b) Find the expected counts.

The expected counts are $n(p_i)$ where n is the sample size and p_i is the theoretical probability for that group.

Type of Nut	Observed Counts	Theoretical Probability	Expected Counts
cashew	6	.30	$20(.3) = 6$
hazelnuts	5	.30	$20(.3) = 6$
peanuts	9	.40	$20(.4) = 8$

(c) Conduct the test.

condition: all the expected counts are at least 5. ✓

I choose $\alpha = .05$.

H_0 : the supermarket's distribution is correct

H_A : the supermarket's distribution is not correct

Test statistic:

$$\begin{aligned}\chi^2 &= \frac{(obs - exp)^2}{exp} \\ &= \frac{(6 - 6)^2}{6} + \frac{(5 - 6)^2}{6} + \frac{(9 - 8)^2}{8} \\ &= .29\end{aligned}$$

Degrees of Freedom:

$$df = \text{number of categories} - 1 = 3 - 1 = 2$$

P-value: p-value is the area to the right of $\chi^2 = .29$. So it is greater than .25. Software gives p-value as .8643.

big p-value; fail to reject H_0

We didn't find any evidence that the supermarket's distribution is not correct.

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?

We want to know if the proposed distribution (the theoretical probabilities) is correct.

So we want to use the Chi Square Test for Goodness of Fit.

- (b) Find the expected counts.

The expected counts are $n(p_i)$ where n is the sample size and p_i is the theoretical probability for that group.

The total sample size is 97.

Games	Theoretical Probability	Observed	Expected Count
four	0.125	17	$97(.125) = 12.125$
five	0.25	23	$97(.25) = 24.25$
six	0.3125	22	$97(.3125) = 30.3125$
seven	0.3125	35	$97(.3125) = 30.3125$

- (c) Conduct the test. ($\chi^2 = 5.03$)

condition: all the expected counts are at least 5. ✓

I choose $\alpha = .05$.

H_0 : the theoretical probabilities are correct (the teams are evenly matched)

H_A : at least one of the theoretical probabilities is incorrect (the teams are not evenly matched)

Test statistic:

$$\begin{aligned}
 \chi^2 &= \frac{(obs - exp)^2}{exp} \\
 &= \frac{(17 - 12.125)^2}{12.125} + \frac{(23 - 24.25)^2}{24.25} + \frac{(22 - 30.3125)^2}{30.3125} + \frac{(35 - 30.3125)^2}{30.3125} \\
 &= 5.03
 \end{aligned}$$

Degrees of Freedom:

$$df = \text{number of categories} - 1 = 4 - 1 = 3$$

P-value: p-value is the area to the right of $\chi^2 = 5.03$. So it is between .15 and .25. Software gives p-value as .1697.

big p-value; fail to reject H_0

We didn't find any evidence that the theoretical probabilities are not correct.

This means that we didn't find any evidence that the teams are not evenly matched.

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

- To find the expected counts, we use the formula $\frac{(\text{row total})(\text{column total})}{\text{total sample size}}$.

	medical patient	surgical patient	total
supplemental health	52.27	39.73	92
no supplemental health	72.73	55.27	128
total	125	95	220

- condition: all expected counts are at least 5 ✓
- $\alpha = .05$
- H_0 : supplemental status and medical/surgical are independent (no relationship; don't affect each other)
- H_A : supplemental status and medical/surgical are Dependent (relationship, affect each other)
- Test statistic:

$$\begin{aligned}
 \chi^2 &= \frac{(obs - exp)^2}{exp} \\
 &= \frac{(56 - 52.27)^2}{52.27} + \frac{(36 - 39.73)^2}{39.73} + \frac{(69 - 72.73)^2}{72.73} + \frac{(59 - 55.27)^2}{55.27} \\
 &= 1.06
 \end{aligned}$$

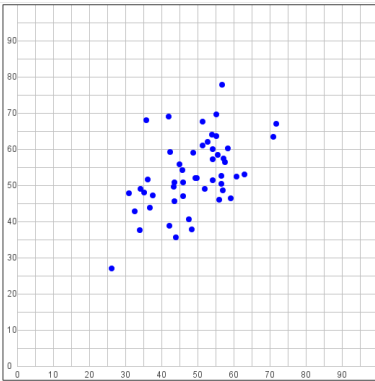
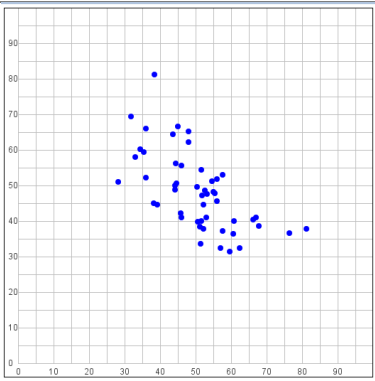
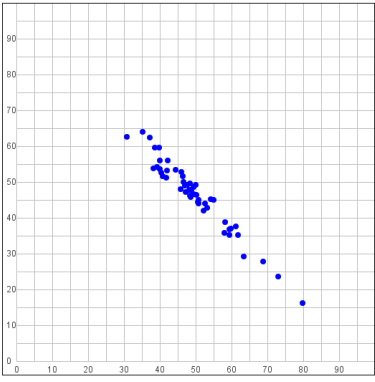
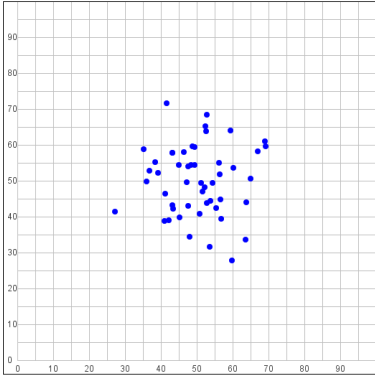
- Degrees of Freedom:

$$df = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$$

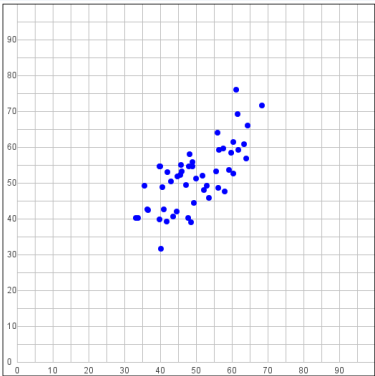
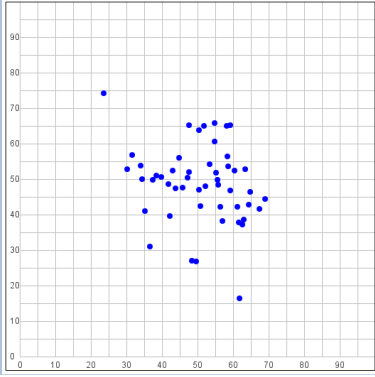
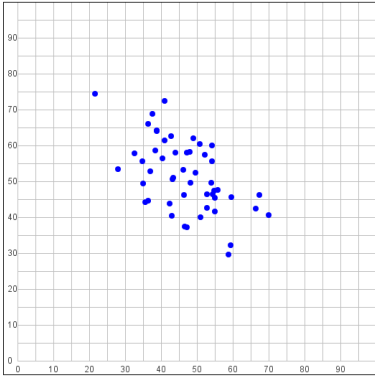
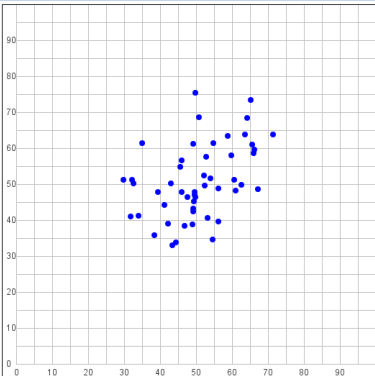
- P-value: area to the right of $\chi^2 = 1.06$. The p-value is bigger than .25. Computer p-value is .3037.
- big p-value; fail reject H_0
- We don't have any evidence that supplemental/no supplemental and medical/surgical are dependent. (i.e. no evidence that they affect each other)

Linear Regression Review-Show Your Work **KEY**

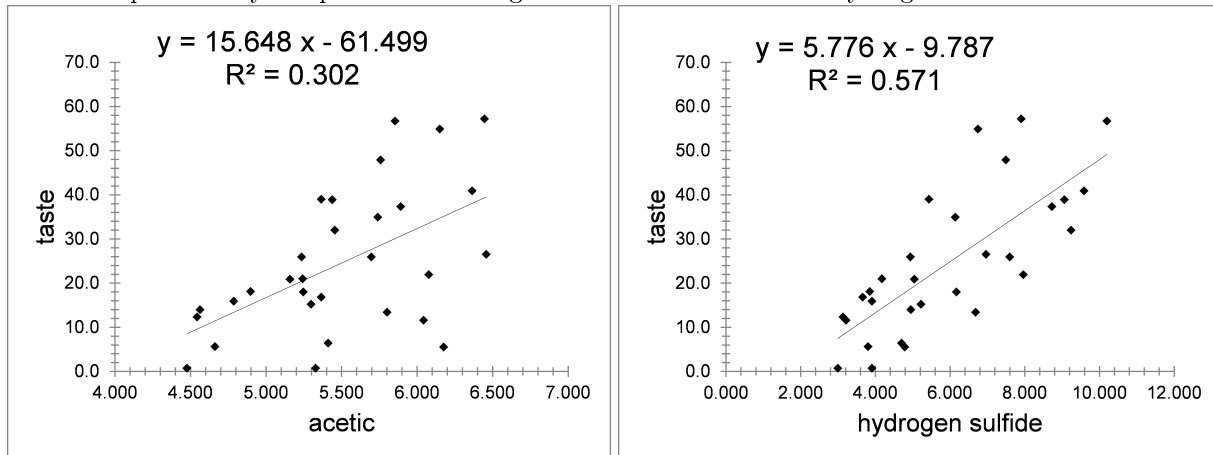
32. Match each plot to its correlation. All the plots are on the same scale. Your choices are $-.98$, $-.63$, $.02$, and $.5$.

Plot	Correlation
	$r = .5$
	$r = -.63$
	$r = -.98$
	$r = .02$

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
	$r = .71$
	$r = -.23$
	$r = -.55$
	$r = .43$

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.

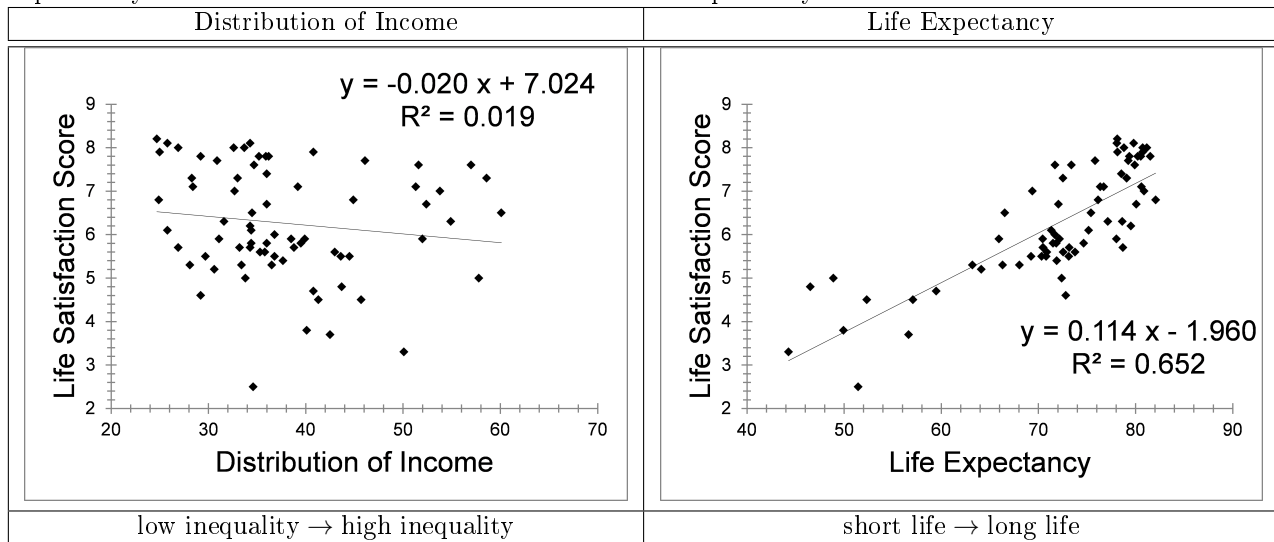


Would you rather use acetic acid or hydrogen sulfide to try to predict the taste rating of cheddar cheese? Why?

The points in the hydrogen sulfide scatterplot seem to be closer to the line. Also the r^2 value is higher. We know that higher r^2 values are better because then the line explains more of the variation in the taste ratings.

So I would use hydrogen sulfide to try to predict the taste rating.

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.

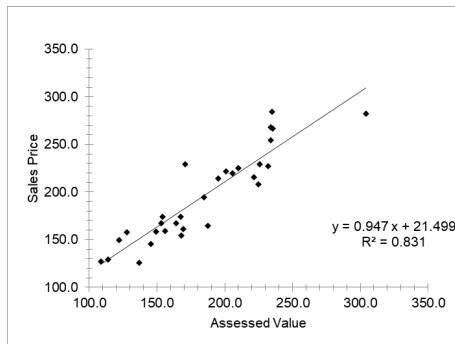


If you had to choose one variable to try to predict the Life Satisfaction Score, would you choose distribution of income or life expectancy, and why?

The points in the life expectancy scatterplot seem to be closer to the line. Also the r^2 value is higher. We know that higher r^2 values are better because then the line explains more of the variation in the life expectancy scores.

So I would use life expectancy to try to predict the life satisfaction score.

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

- (a) Which variable should be the explanatory variable and which is the response variable?

Explanatory: assessed value

Response Variable: sales price

- (b) Conduct a hypothesis test to determine if there is a significant linear relationship between the assessed value and sales price.

- **condition: the plots look okay.**
- **I choose $\alpha = .05$.**
- **$H_0 : \beta_1 = 0$ (the slope is zero, there is no linear relationship between the variables)**
- **$H_A : \beta_1 \neq 0$ (the slope is not zero, there is a linear relationship between the variables.)**
- **Test statistic:**

$$t = 11.741$$

- **Degrees of freedom**

$$df = 28$$

- **The p-value will be the area in the both tails.**

$$p - value = 2.49 \times 10^{-12}$$

- **small p-value; reject H_0**
- **We have extremely strong evidence that the population slope is not zero.**
- **We have extremely strong evidence that there is a significant linear relationship between the assessed value and sales price.**

- (c) Do you have evidence that the increasing the assessed value causes a change in the sales price?
No! All we know from the statistics is that the two variables are associated. We can't show cause and effect with linear regression.

- (d) What is the equation for the linear regression line? (Read it off the results)
 $\hat{y} = .9468x + 21.4992$

- (e) What is the y-intercept? Interpret. Is it a logical/practical interpretation?
 $b_0 = 21.4992$
If the assessed price is \$0, the predicted sales price is \$21,499.

(Practical? I'm not sure if a city would assess a price to be \$0, but I don't know.)

- (f) What is the slope? Interpret.
 $b_1 = .9468$
If the assessed value increases by \$1000, we predict the sales price will increase by \$946.80.

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

We just need to plug $x = 150$ into the equation.

$$\begin{aligned}\hat{y} &= .9468x + 21.4992 \\ &= .9468(150) + 21.4992 \\ &= 163.5192\end{aligned}$$

So the predicted mean sales price for all the homes is \$163,519.

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

We just need to plug $x = 150$ into the equation.

$$\begin{aligned}\hat{y} &= .9468x + 21.4992 \\ &= .9468(150) + 21.4992 \\ &= 163.5192\end{aligned}$$

So the predicted sales price is \$163,519 for John's house.

- (i) What is the correlation? Interpret.

$$r = .912$$

So assessed value and sales price have a positive linear relationship.

Since $r = .912$ is close to 1, the points will be very close to the line.

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

$$r^2 = .831$$

So we can explain 83.1% of the variation in the sales prices with our line.

- (k) Find a 90% confidence interval for the slope β_1 . Interpret.

We know the formula is

$$b_1 \pm t^* SE_{b_1}$$

where t^* is the critical value from the t curve with $n - 2$ degrees of freedom.

Reading off the output, the confidence interval is

$$.8096 \text{ to } 1.0840$$

So we are 90% confident that the population slope β_1 is between .81 and 1.08

So we are 90% confident that as the assessed value increases by \$1000, the sales price increases by between \$810 and \$1080. If you want to do it by hand, then remember that t^* is the critical value to find a confidence interval. This is exactly the same as what we did when we found T confidence intervals for the population mean μ . (Except for slope we use $df = n - 2$).

So to review t^* go back to page 142 in the course reader and work through a couple problems. You can also watch the video <https://youtu.be/aljivrK8oy8>

So for this problem, we want .90 area in the middle. That means .05 area in each tail. Use the T table and

$$df = n - 2 = 30 - 2 = 28$$

and .05 area in the right tail to find $t^* = 1.701$

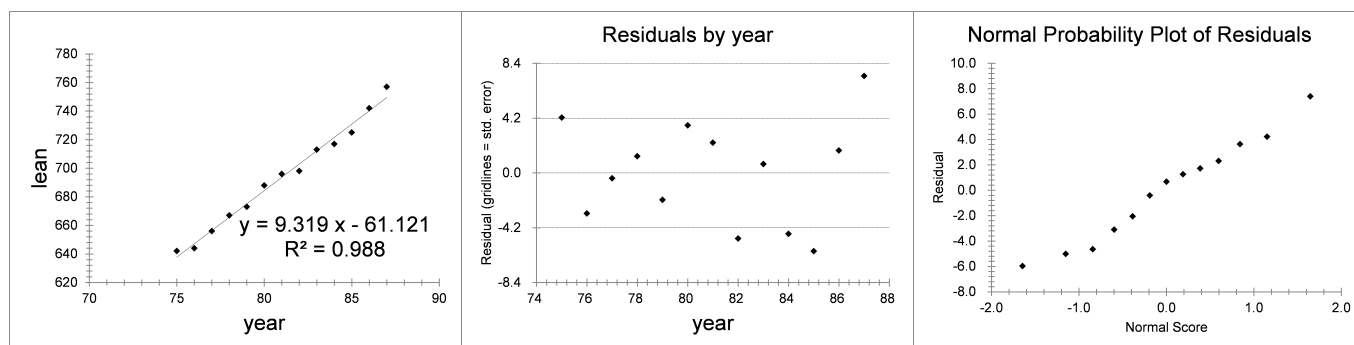


$$b_1 \pm t^* SE_{b_1}$$

$$.9468 \pm 1.701 (.0806)$$

$$.8096 \text{ to } 1.084$$

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

variables	coefficients	std. error	t (df=11)	p-value	confidence interval	
					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

- (a) Which variable should be the explanatory variable and which is the response variable?

Explanatory: year

Response Variable: lean

- (b) Conduct a hypothesis test to determine if there is a significant positive linear relationship between the year and the lean.

- I choose $\alpha = .05$.
- $H_0 : \beta_1 = 0$ (the slope is zero, there is no linear relationship between the variables)
- $H_A : \beta_1 > 0$ (the slope is greater than zero, there is a positive linear relationship between the variables.✓)
- Test statistic:

$$t = \frac{b_1}{SE_{b_1}} = \frac{9.3187}{.3099} = 30.069$$

- Degrees of freedom are $n - 2 = 13 - 2 = 11$.
- The p-value will be the area in the right tail. Megastat gives the area in both tails, so just divide it by 2. (All the computer output in this class for linear regression is from MegaStat).

$$p - value = \frac{6.50 \times 10^{-12}}{2} = 3.25 \times 10^{-12}$$

- small p-value; reject H_0
- We have extremely strong evidence that the population slope is greater than zero.
- We have extremely strong evidence that there is a significant positive relationship between the year and the lean of the tower.

- (c) Do you have evidence that the increasing the year causes the tower to lean further?

No! All we know from the statistics is that the two variables are associated. We can't show cause and effect with linear regression.

- (d) What is the equation for the linear regression line? (Read it off the results)

$$\hat{y} = 9.3187x - 61.1209$$

- (e) What is the y-intercept? Interpret. Is it a logical/practical interpretation?

$$b_0 = -61.1209$$

In the year 0 (i.e. 1900 since they coded 1974 as 74 and 1985 as 85, etc), we predict the lean was -61.1209.

This next part is some extra logic.

In this problem the lean is measured as tenths of millimeters past 2.9 meters. And the intercept is specifically -61.1209 tenths of a millimeter from the 2.9 meter lean.

So to see what the total lean would have been I subtracted that 61.1209 from 2.9 (because 2.9 is the baseline for this problem). But because 2.9 is in meters and 61.1209 is in tenths of millimeters I have to divide it by 10,000.

So the tower was only leaning $2.9 - \frac{61.1290}{10,000} = 2.89388$ meters in 1900 A.D. This could be logical.

There is nothing to make me think it is unlogical.

- (f) What is the slope? Interpret.

$$b_1 = 9.3187$$

For each additional year, the predicted lean increases by 9.3187 tenths of a millimeter.

- (g) Predict the lean in the year 1988.

We just need to plug $x = 88$ into the equation.

$$\begin{aligned}\hat{y} &= 9.3187x - 61.1209 \\ &= 9.3187(88) - 61.1209 \\ &= 758.92\end{aligned}$$

So we predict that in the year 1988, the tower will have leaned 758.92 tenths of a millimeter (or .0758 meters) past the 2.9 meter lean.

(1988 is slightly past our data range of 1975 to 1987, but not much, so we should be okay)

- (h) What is the correlation? Interpret.

$$r = .994$$

So time and lean have a positive linear relationship.

Since $r = .994$ is so close to 1, we know that the points will be very close to the line.

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

$$r^2 = .988$$

So we can explain 98.8% of the variation in the amount of lean with our line.

(So we can use the year to predict the amount of lean very well.)

- (j) Find a 95% confidence interval for the slope β_1 . Interpret.

We know the formula is

$$b_1 \pm t^* SE_{b_1}$$

where t^* is the critical value from the t curve with $n - 2$ degrees of freedom.

So degrees of freedom are $n - 2 = 13 - 2 = 11$.

The critical value is $t^* = 2.201$.

From the MegaStat results we see that $b_1 = 9.3187$ and $SE_{b_1} = .3099$.

$$b_1 \pm t^* SE_{b_1}$$

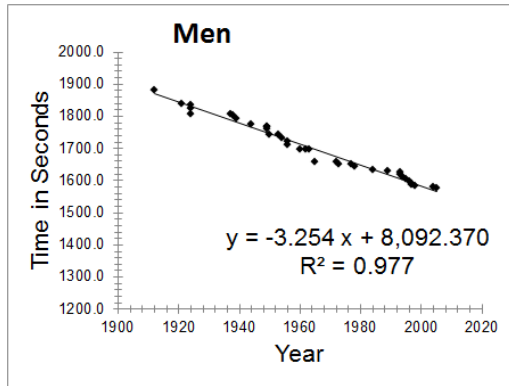
$$9.3187 \pm 2.210 (.3099)$$

$$(8.63, 10)$$

So we are 95% confident that the population slope β_1 is between 8.63 and 10.

So we are 95% confident that as for each additional year, the amount of lean increases between 8.63 and 10 tenths of a millimeter.

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

- (a) Which variable should be the explanatory variable and which is the response variable?

Explanatory: year

Response Variable: world record time

- (b) 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.

- **I choose $\alpha = .05$.**
- $H_0 : \beta_1 = 0$ (the slope is zero, there is no linear relationship between the variables)
- $H_A : \beta_1 \neq 0$ (the slope is not zero, there is a linear relationship between the variables.✓)
- **Test statistic:**

$$t = \frac{b_1}{SE_{b_1}} = \frac{-3.2544}{.0851} = -38.24$$

- **Degrees of freedom are $n - 2 = 37 - 2 = 35$.**
- **The p-value will be the area in the both tails. (computer p-value is 1.9×10^{-30})**
- **small p-value; reject H_0**
- **We have extremely strong evidence that the population slope is less than zero.**
- **We have extremely strong evidence that there is a significant negative linear relationship between the year and the world record time for the 10k for the men.**

- (c) Do you have evidence that the increasing the year causes male athletes to run faster?

No! All we know from the statistics is that the two variables are associated. We can't show cause and effect with linear regression.

- (d) What is the equation for the linear regression line? (Read it off the results)

$$\hat{y} = -3.2544x + 8092$$

- (e) What is the y-intercept? Interpret. Is it a logical/practical interpretation?

$$b_0 = 8092$$

**In the year 0, we predict the world record to run the 10k was 8092 seconds (2.247 hours).
(I don't think this is practical. First of all, we don't have records for 10k runs for 0 A.D.. I don't think they ran the 10k then. But even if they did, I don't think the fastest person in the world would be that slow.)**

- (f) What is the slope? Interpret.

$$b_1 = -3.2544$$

For each additional year, the predicted record time decreased by 3.2544 seconds.

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

We just need to plug $x = 1997$ into the equation.

$$\begin{aligned}\hat{y} &= -3.2544x + 8092 \\ &= -3.2544(1997) + 8092 \\ &= 1592.9\end{aligned}$$

(1997 is inside our data range of 1912-2005, so we don't have to worry about extrapolating)

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

We just need to plug $x = 2500$ into the equation.

But we shouldn't do this problem because it is extrapolation.

Who knows if we have already pushed our bodies as fast as they can go and we can't improve more? Or perhaps we will have technological advancements that make us run twice as fast. We can't guarantee that the linear trend continues.

Also, if you accidentally do this problem you will get

$$\begin{aligned}\hat{y} &= -3.2544x + 8092 \\ &= -3.2544(2500) + 8092 \\ &= -44\end{aligned}$$

which is complete nonsense. So you would realize that something is wrong.

- (i) What is the correlation? Interpret.

$$r = -.988$$

So time and length have a negative linear relationship.

Since $r = -.988$ is so close to 1, we know that the points will be very close to the line.

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

$$r^2 = .977$$

So we can explain 97.7% of the variation in the world record times with our line.

(So we can use the year to predict the world record times very well.)

- (k) Find a 90% confidence interval for the slope β_1 . Interpret.

We know the formula is

$$b_1 \pm t^* SE_{b_1}$$

where t^* is the critical value from the t curve with $n - 2$ degrees of freedom.

So degrees of freedom are $n - 2 = 37 - 2 = 35$. (round down to $df = 30$)

The critical value is $t^* = 1.697$.

From the MegaStat results we see that $b_1 = -3.2544$ and $SE_{b_1} = .0851$.

$$\begin{aligned}b_1 \pm t^* SE_{b_1} \\ -3.2544 \pm 1.697 (.0851) \\ (-3.4, -3.1)\end{aligned}$$

So we are 90% confident that the population slope β_1 is between -3.4 and -3.1.

So we are 90% confident that as the year increases, the world record time decreases by between 3.4 and 3.1 seconds.

Multiple Choice

39. The dependent variable is the variable that is being described or predicted.
- (a) true
 - (b) false
 - true**
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
 - true**
41. r^2 is the percentage of variation in the y values that is explained by our least squares line.
- (a) true
 - (b) false
 - true**
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
 - false, linear regression never lets us show cause and effect**
43. If $r = -1$, then we can conclude that there is a perfect linear relationship between X and Y.
- (a) true
 - (b) false
 - true, all points are exactly on the line**
44. A significant positive correlation between X and Y implies that changes in X cause Y to change.
- (a) true
 - (b) false
 - false, linear regression never lets us show cause and effect**
45. In a simple linear regression analysis, the correlation coefficient and the slope _____ have the same sign.
- (a) always
 - (b) sometimes
 - (c) never
 - (a) always**
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
 - (C) Correlation coefficient**

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
- (D) Squared differences between actual and predicted Y values**
48. The _____ the r^2 , the stronger the relationship between the dependent variable and the independent variable.
- (a) lower
 - (b) higher
- (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
- (b) slope**
50. The correlation coefficient may assume any value between
- (a) 0 and 1
 - (b) $-\infty$ and ∞
 - (c) 0 and 8
 - (d) -1 and 1
 - (e) -1 and 0
- (d) negative one and one**
51. If the correlation is positive, then the slope is
- (a) negative
 - (b) positive
 - (c) zero
 - (d) it could be negative or positive
- (b) slope has to be positive if correlation is 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
- (a) slope is $b_1 = -1.2024$**

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
- (c) y intercept is $b_0 = 37.2895$**

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
- (b) $r^2 = .6744$**

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
- (c) correlation, r**

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					Confidence interval	
Variables	Coefficients	std. error	t (df=5)	p-value	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?
- $\hat{y} = 7.9682 + 1.667x$
 - $\hat{y} = 63.333 + 6.667x$
 - $\hat{y} = 7.948 + 4.000x$
 - $\hat{y} = 11.547 + 1.667x$
 - $\hat{y} = 6.667 + 63.333x$
- (ii) $\hat{y} = 63.333 + 6.667x$**
- (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
- (iii) Remember that both sales and advertising are in thousands of dollars. We plug \$3000 or $x = 3$ into our equation.**

$$\begin{aligned}
 \hat{y} &= 63.333 + 6.667x \\
 &= 63.333 + 6.667(3) \\
 &= 83.334
 \end{aligned}$$

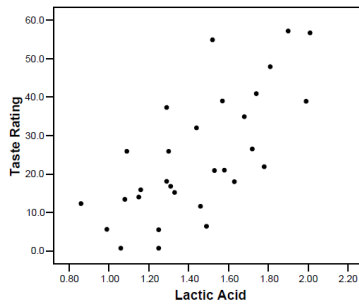
So for \$3000 in advertising, we expect about \$83,333 in sales.

- (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.

(i) the computer p-value for slope is .0103, we reject H_0 and we would usually say that we have evidence that there is a significant linear relationship between advertising and sales. This is equivalent to saying that the amount of advertising is useful to predict sales.

57. When creating a scatterplot, one should use the _____ axis for the explanatory variable.
- (a) x axis
 - (b) y axis
 - (a) x-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
 - (a) yield of crop**
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
 - (b) less than zero**
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.
 - (c) equal to zero**
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.
 - (c) zero correlation means that how good a teacher is tells us nothing about how good a researcher they are, and vice versa.**

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:



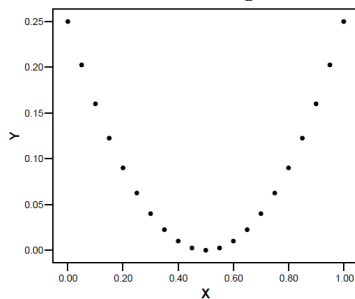
What is a plausible value for the correlation between lactic acid concentration and taste rating?

- (a) 0.999
- (b) 0.7
- (c) 0.07
- (d) -0.7
- (e) -.999

(b) .7

the points are kind a close to a line and the slope is positive.

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.

(d) can't use correlation at all

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) -.827
- (b) .827
- (c) 0
- (d) we need to know the actual values

(b) when you find correlation, it doesn't matter which variable is response and which is explanatory.

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.

(D) strength of linear relationship (quantitative)

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.

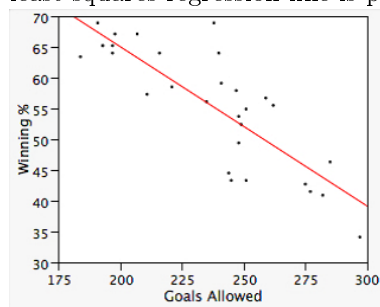
$$\hat{price} = -6677 + 175(\text{horsepower})$$

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.

(C) \$28,323

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%.

(E) If x goes up by one, y goes down by .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.

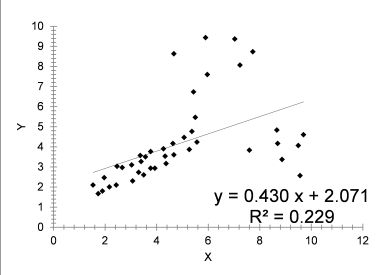
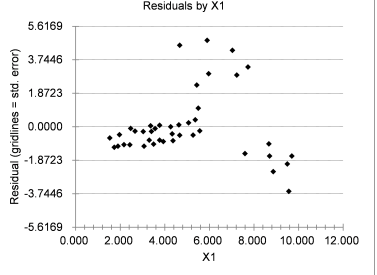
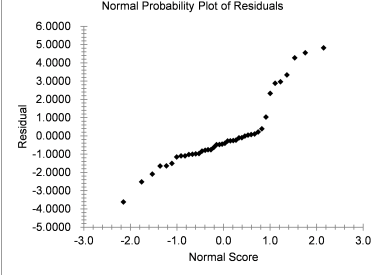
(b) is the only correct answer

Doing the test only tells us that there is a negative linear relationship.

It doesn't tell us anything about how close the points are to the line. The correlation is what tells us how close the points are to the line.

We can never show cause and effect with linear regression. We can only show that there is a linear relationship.

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)
		

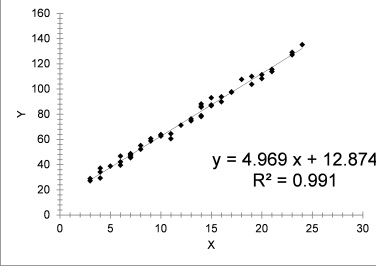
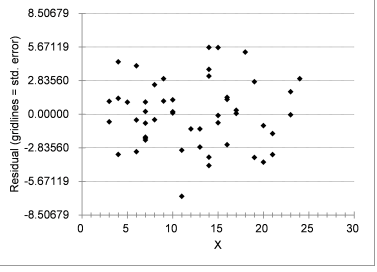
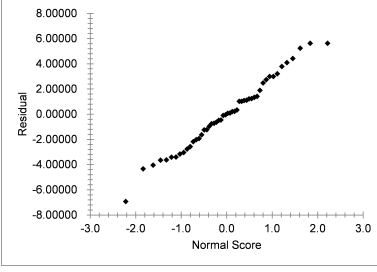
No!

Actual data doesn't look linear.

There is a pattern in the residual plot.

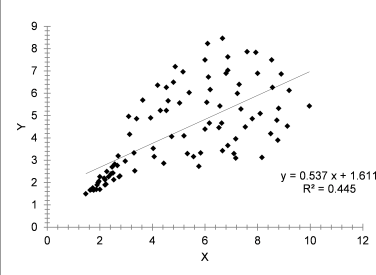
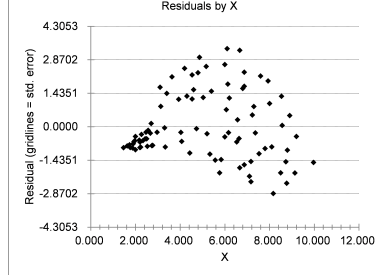
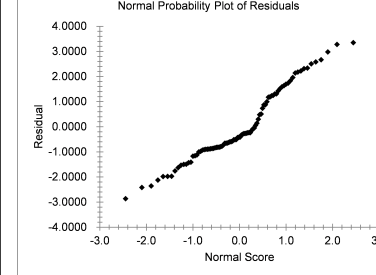
The normality plot looks like it has a curve, so the residuals aren't normally distributed.

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)
		

Yes, the plots look okay.

71. Is linear regression appropriate for the data set?

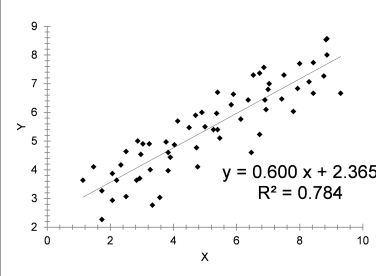
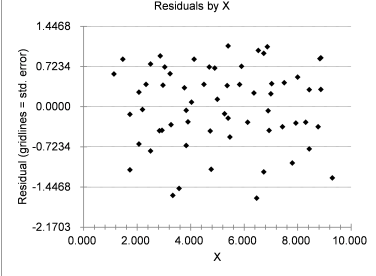
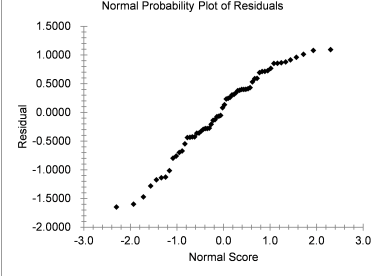
scatterplot of actual data	plot of residuals against explanatory variable	normality plot for residuals
(check to see if the data seems to have a linear relationship)	(there shouldn't be any pattern, the points should be randomly distributed)	(residuals are supposed to be normal, so the normality plot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

No!

Actual data is maybe linear, but it is kind of strange the way it get's further from the line. There is a pattern in the residual plot. It fans out. The points don't seem to be randomly scattered.

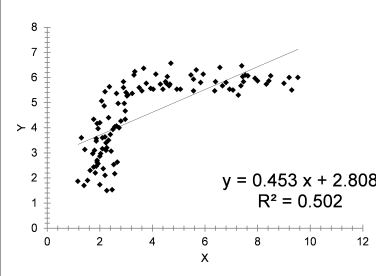
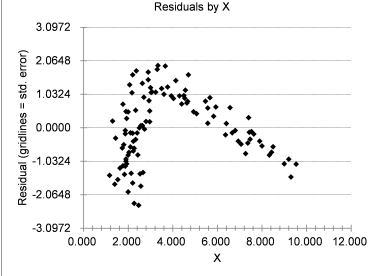
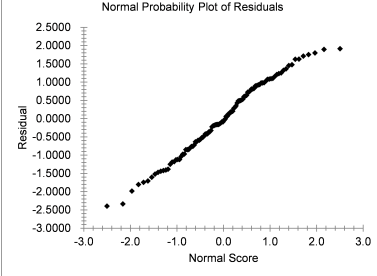
The normality plot looks like it has a curve, so the residuals aren't normally distributed. (But that isn't as noticeable as the plot of the residuals against the x values.)

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)
		

Yes, the plots look okay.

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)
		

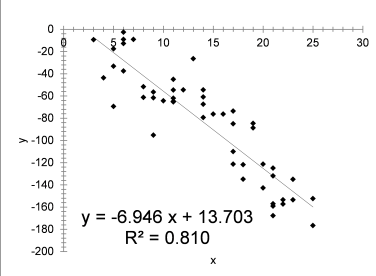
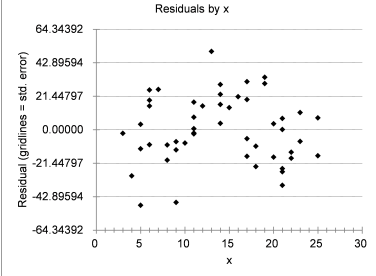
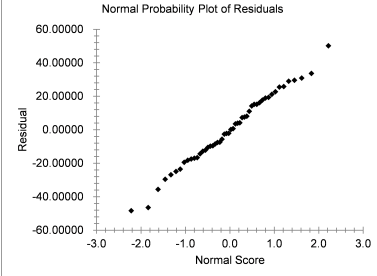
No!

The scatterplot doesn't look like there is a linear relationship. ✗

The residual plot has a definite pattern. ✗

The normality plot looks fine. ✓

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)
		

Yes, the plots look okay.