Solutions A

Final - STAT 301
Spring 2013

45

Name:

UIN:

Signature:

Version A:

1. Do not open this test until told to do so.
2. This is a closed book examination, However you may use four single-sided sheet of formulas that you have brought with you and the tables. You should have no other printed or written material with you on the exam. But scrap paper is allowed.
3. You have 2 hours to work on this exam. There are 25 multiple choice questions.
4. On the scantron please state the version of exam that you have.
5. You may use a calculator in the exam.
6. If there is no correct answer or if multiple answers are correct, select the **best** answer.
7. If you are unsure of what a question is asking for, do not hesitate to ask the instructor or course assistant for clarification.
8. Please only give one answer per question (the one that is closest to the solution).
9. No wearing hats that can cover ones eyes.
10. Good Luck!!!

(1-3) [ABCDE]

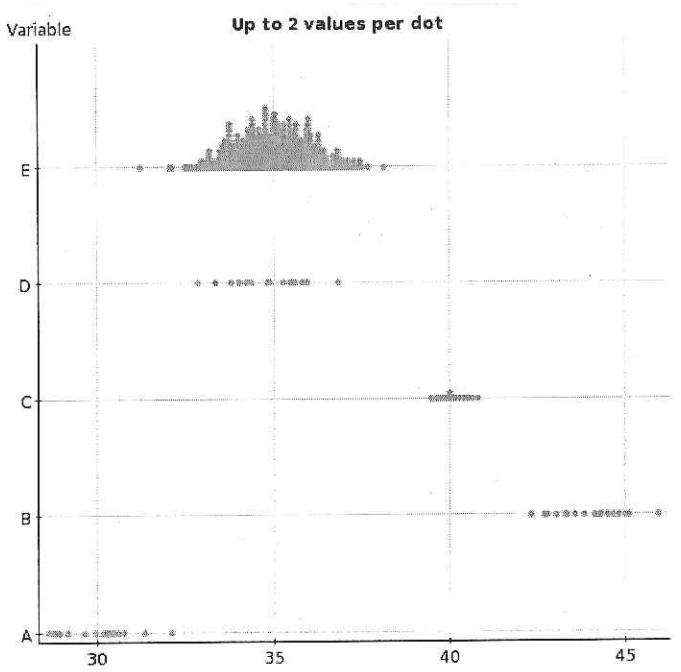


Figure 1: Dotplots of 5 different samples. Samples A-D all have sample size 20, whereas sample E has sample size 1000.

(1) Which sample has the largest sample mean?

[A], [B], [C], [D], [E].

(2) Which sample has the smallest sample standard deviation?

[A], [B], [C], [D], [E].

(3) Suppose for each of the 5 data sets, a 95% confidence intervals for the mean is constructed. Which data set will give the smallest margin of error?

[A], [B], [C], [D], [E].

(4) Dave is sitting in the orthopedic waiting room at the Scott and White clinic in College Station. He observes that all the elderly people there are overweight. He concludes that elderly people tend to be overweight (he may be a bigot!).

Comment on the accuracy of his observation.

(A) It is plausible, but he needs to test this statistically.

(B) He needs to increase the sample size to really see if this is true.

(C) There is a bias in the sample, as he is only looking at elderly people who have visited Scott and White clinic in College Station.

(D) Two of the above.

(E) None of the above.

(5) A biochemistry professor wants to see if there is any difference in the performance of honors and regular sections. The Professor teaches two sections of biochemistry. One section is a regular class and the other is an honors class. He wants to compare the grades in both classes. What test should he do?

(A) An independent two-sample t-test.

(B) A paired t-test.

(C) A one sample t-test.

(D) A test for association.

(E) None of the above.

(6) A couple want to compare the heights of their 21 year old twin son and daughter. The son is 70 inches tall and the daughter is 63.5 inches tall. It is known that the heights of males between 20-30 years are approximately normal with mean 68 inches tall and standard deviation 5 inches, whereas the heights of females between 20-30 years old is approximately normal with mean 65 inches and standard deviation 3 inches. Relative to their gender which sibling is taller?

(A) The daughter is in the 69.1% percentile and the son is in 65.5% percentile, thus relative to their gender the daughter is taller.

(B) The daughter is in the 30.8% percentile and the son is in the 65.5%, thus relative to their gender the son taller.

(C) The daughter is in the 30.8% percentile and the son is in the 34.5% percentile, thus relative to their gender the son is taller.

(D) The daughter is in the 50% percentile and the son is in the 40% percentile, therefore relative to their gender the daughter is taller.

(E) The daughter is in 50% percentile and the son is in the 60% percentile, therefore relative to their gender the son taller.

(7) Which statement(s) about correlation is correct?

(A) There is a perfect correlation between the scores of a mathematics and an english exam. Therefore if a student scored 80 in her mathematics exam she will also score 80 in her mathematics exam.

(B) There is a high correlation between the gender of a person and their occupation.

(C) We found a high correlation of 1.23 between the rating Netflicks gave a film and the rating that Rotten Tomatoes gave a film.

(D) All of the above.

(E) None of the above.

(8-9) Toyota have installed an airbag system at the back of their minivan, which is only activated if the weight on the backseat is over 130 pounds.

(8) The mean weight of a child below the age of 5 years old is 40 pounds and the standard deviation is 10 pounds. What is the probability that weight of 4 children (all below the age of 5 years old) placed in the backseat of the Toyota will trigger the airbag.

(A) The z-transforms is $z = (32.5 - 40)/5 = -1.5$, thus the probability that 4 children (under 5) will weigh more than 130 pounds is about 6.6%.

(B) The z-transform is $z = (40 - 130)/5 = -9$, therefore the probability that 4 children (under 5) will weigh more than 130 pounds is nearly 100%.

(C) The z-transform is $z = (32.5 - 40)/10 = -4.55$, therefore the probability that 4 children (under 5) will weigh more than 130 pounds is almost 100%.

(D) The z-transform is $z = (32.5 - 40)/5 = -1.5$, therefore the probability that 4 children (under 5) will weigh more than 130 pounds is 93.3%.

(E) The z-transform is $z = (40 - 130)/10 = -9$, therefore the probability that 4 children (under 5) will weigh more than 130 pounds is close to zero.

(9) What was the main assumption in the calculation above?

(A) The distribution of the weights each of child is about the same.

(B) The normal tables are accurate.

(C) The distribution of weights is random.

(D) The weights of the children in the sample are not too small. If this were the case it would be impossible to activate the airbag regardless of the method of calculation.

(E) That the distribution of under 5 year old weights does not deviate far from normal, such that the average weight of 4 children is almost normally distributed.

(10) Your friend is confused about when to use the t-distribution when constructing confidence intervals for the mean. Which is the correct explanation?

(A) The central limit theorem states that the distribution of the sample mean is first close to a t-distribution and then converges to the normal distribution.

(B) In small sample sizes it is used to correct for the lack of normality of the data.

(C) It is used when the standard deviation of the population is unknown and has to be estimated from the sample.

(D) All of the above.

(E) None of the above.

(11) A survey is done in two towns, Hearn and Calvert, to see which College football teams they prefered (UT or A&M).

The population of Hearn is 4500 and the population of Calvert is 1190.

To find out which team each of these towns allegiance lay, a simple random sample of 100 residents was take from Hearn and a simple random sample of 100 residents was taken from Calvert. Each person in the sample was asked which team they supported.

Which sample gave the most reliable estimate for the proportion (which supported A&M).

(A) Since both sample sizes are the same, and Calvert has a much smaller population than Hearn, the sample taken from Calvert is more reliable.

(B) The reliability of the estimates is determined by the standard error. As both sample sizes are the same, the reliability will be roughly the same (though there will be a small influence from the proportion who votes for A&M).

(C) Since both samples are relatively small compared with the population size, neither sample will be that reliable.

(D) When doing a survey it is recommended that the sample size is <u>at least</u> 10% of the population size. As neither sample satisfies this criterion, both the standard errors will be large, however the sample from Calvert will be slightly more reliable.

(E) Two of the above.

(12-13) A survey is taken of the starting salaries of student who have recently graduated. The summary statistics is given below.

Summary statistics:

Column	n	Mean	Std. Dev.	Median	Range
Salary	30	44999.453	7481.5786	45361.15	30784.768

(12) What is the standard error for the sample mean of the salaries (rounded to the nearest dollar)?

(A) $7482/30 = 249$

(B) $7482/\sqrt{30} = 1366$

(C) 7482

(D) $7842/30^2 = 8.7$

(E) There is not enough information answer this question.

(13) Suppose that in the previous question the standard error is 1000.

Construct a 99% confidence interval for the mean (using the standard error 1000).

(A) $[44999 \pm 2.756 \times 1000]$

(B) $[44999 \pm 2.045 \times \frac{1000}{\sqrt{30}}]$

(C) $[45361 \pm 2.756 \times 1000]$

(D) $[44999 \pm 2.750 \times \frac{1000}{30}]$

(B) $[45361 \pm 2.045 \times \frac{1000}{\sqrt{30}}]$

(14) You read the following news article:

'The probability of a couple having fraternal twins (non-identical twins) is 1 in 60. Mary and Joe, a couple from Forth Worth, had a pair of fraternal twins in 2011 and this year gave birth to another pair of fraternal twins. The probability of this extraordinary event happening is about 1 in 3600.'

Comment on the accuracy of the probabilities in the article.

(A) The probability quoted in the article is incorrect, because the probability (under the assumption of independence) of two pairs of twins being born is 1 out of 60.

(B) The probability 1 out of 3600 was calculated under the assumption that two pairs of fraternal twins are independent events, which is unlikely to be true given that the two pairs of twins share the same parents.

(C) The event is relatively rare, since the chance of this happenings is 1 in 3600 births.

(D) Since 1/3600 is such a small probability it is statistically highly significant (less than the 5% significance level). Therefore, the probability of this happening by random chance is slim, therefore it is likely the couple had fertility treatment.

(E) None of the above.

(15) What is closest correlation coefficient for this data set.

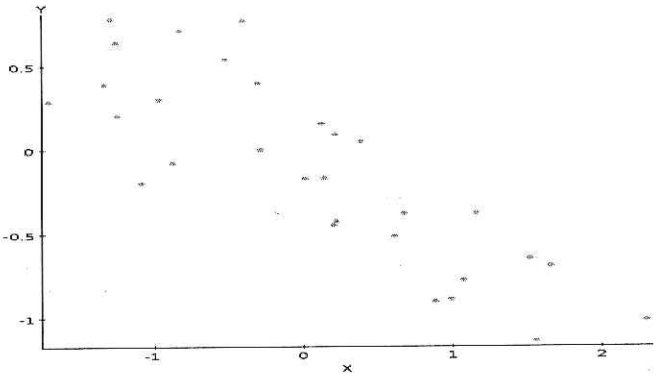
(A) 0.75,

(B) 0.25

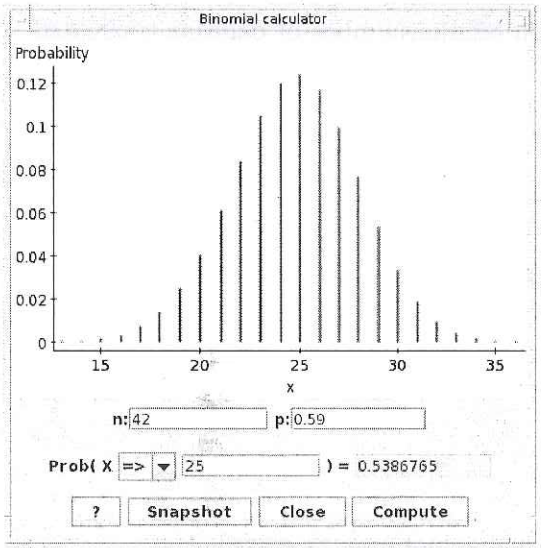
(C) 0

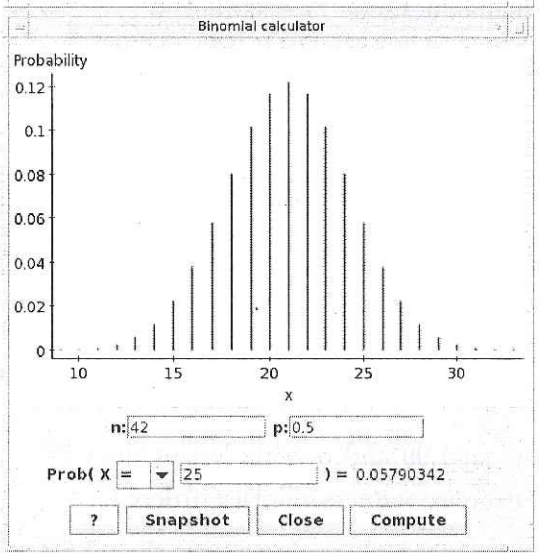
(D) -0.25

(E) -0.75



(16) The Dean of Science is wondering if female students are more likely to major in the biological sciences than male students. In a random sample of 44 biological science students she found that 25 of these students were female. Is there any evidence to support her view (do the test at the 5% level).





(A) The hypothesis of interest is $H_0 : p = 0.59$ against $H_A : p > 0.59$, since the p-value is less than 5% there is evidence to suggest that there tends to be more females majoring in the biological sciences.

(B) The hypothesis of interest is $H_0 : p = 0.5$ against $H_A : p > 0.5$, since the proportion of females doing biological sciences is $p = 25/42 \approx 0.59$, there is statistical evidence to suggest more females do the biological sciences than males.

(C) The hypothesis of interest is $H_0 : p = 0.5$ against $H_A : p < 0.5$, the p-value $= (1 - 0.057) = 94.3\%$, this is so large, that at sensible significance level there is no evidence to reject the null.

(D) The hypothesis of interest is $H_0 : p = 0.59$ against $H_A : p \neq 0.59$, since the p-value is greater than 53% there is no evidence to reject the null at the 5% level.

(E) The hypothesis of interest is $H_0 : p = 0.5$ against $H_A : p > 0.5$, since the p-value is greater than 5.79% there is no evidence to reject the null at the 5% level.

(17) Which of the following about the meaning of the p-value is true?

(A) The p-value measures the power in the test.

(B) The p-value is negative if the alternative is pointing to the left.

(C) The p-value is the same as the significance level of the test.

(D) The p-value measures the probability of observing the data under the null hypothesis.

(E) The larger the p-value the stronger the evidence *against* the null hypothesis.

(18-19) In past few years fast food places in New York are required by law to give the number of calories for each item on the menu.

Dieticians are sceptical that this will aid customers in selecting the less calorific options. Instead they suggest that placing the amount of exercise required to burn off the number of calories consumed may be a better method of guiding customers.

To test this theory they monitor the average number of calories consumed per customer at a fast food place. In the first week only the the number of calories are given in the menu and the following week the amount of exercise required is also placed on the menu.

The average number of calories consumed when only the number of calories were given on the menu was 1053 (over 30 customers).

The average number of calories consumed when the amount of exercise is placed on the menu is 962 (over 35 customers).

Hypothesis test results:
μ_1 : mean of population 1
μ_2 : mean of population 2
$\mu_1 - \mu_2$: mean difference
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 \neq 0$
(without pooled variances)

Difference	Sample Mean	Std. Err.	DF
$\mu_1 - \mu_2$	91	45.682495	54.42036

The following critical values may be useful.

prob.	0.15	0.10	0.05	0.025	0.01	0.005
t^*	1.046	1.297	1.673	2.005	2.400	2.669

(18) State the hypothesis of interest and give the p-value and conclusion for the test. Let μ_E denote the mean number of calories consumed when the amount of exercise is given and μ_C denote the mean number of calories consumed when only the number of calories is given on the menu. Which statement gives the correct hypothesis, t-value, p-value and result of test at the 5% level?

(A) $H_0 : \mu_E - \mu_C = 0$ against $H_A : \mu_E - \mu_C < 0$ the t-value is -1.99 and p-value is less than 5%, there is evidence to suggest that putting the amount of exercise reduces calories ordered.

(B) $H_0 : \mu_E - \mu_C < 0$ against $H_A : \mu_E - \mu_C = 0$ the t-value is -1.99 and p-value is less than 2.5%, there is no evidence to suggest that putting the amount of exercise reduces calories ordered.

(C) $H_0 : \mu_E - \mu_C = 0$ against $H_A : \mu_E - \mu_C > 0$ the t-value is 1.99 and p-value is less than 10%, there is evidence to suggest that putting the amount of exercise reduces calories ordered.

(D) $H_0 : \mu_E - \mu_C = 0$ against $H_A : \mu_E - \mu_C \neq 0$ the t-value is 1.99 and the p-value is less than 10%, there is evidence to suggest that putting the amount of exercise reduces calories ordered.

(E) $H_0 : \mu_E - \mu_C = 0$ against $H_A : \mu_E - \mu_C < 0$ the t-value is -1.99 and p-value is less than 5% , there is no evidence to suggest that putting the amount of exercise reduces calories ordered.

(19) To implement the law that all fast food places must include the amount of exercise in the menu is a costly. Therefore it only makes sense if the implementation leads to at least a mean reduction of a 100 calories. Is there evidence of this?

(A) As the 95% confidence interval for the mean reduction is [-0.588,182], and since 100 lies in this interval, it is likely that the mean reduction is calories is over a 100.

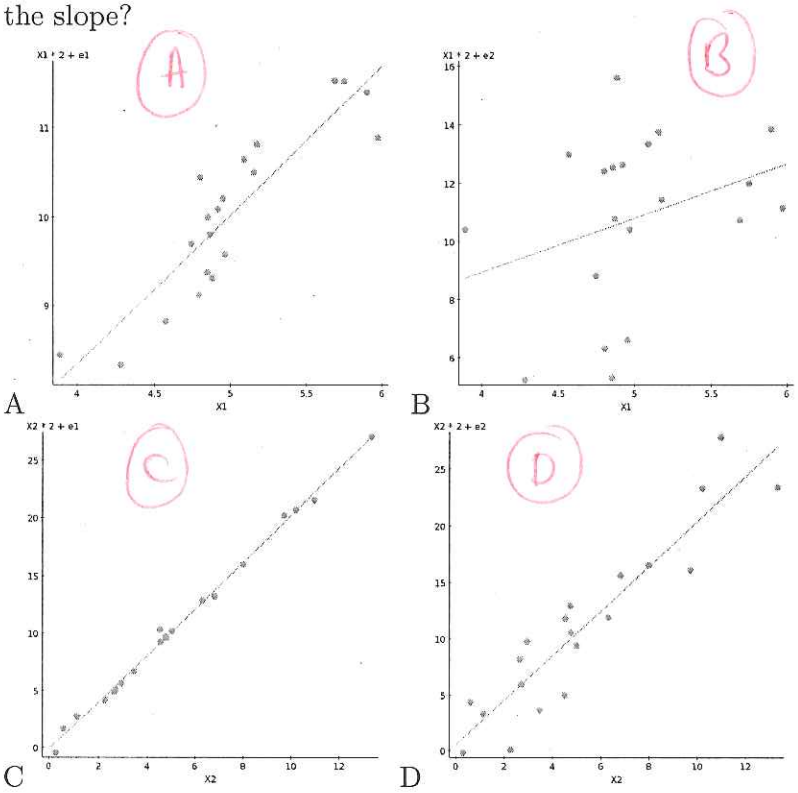
(B) As the sample mean is 91 calories, which is less than 100 it is clear that there is no evidence of this.

(C) As the 95% confidence interval for the mean reduction is [-136,227], and 100 and above is the majority of this interval, there is evidence to suggest that putting the amount of exercise on the menu reduces the number of calories by at least 100.

(D) As the 95% confidence interval for the mean reduction is [-136,227], and zero is in this interval, the mean difference is not statistically significant.

(E) This question cannot be answered without doing a formal test.

(20) Which plots give the smallest standard error and which plot will give the largest standard errors for the slope?



(A) Smallest s.e.= D, Largest s.e= A.

(B) Smallest s.e.= B, Largest s.e.= C.

(C) Smallest s.e.= C, Largest s.e.= B.

(D) Smallest s.e.= C, Largest s.e.= D.

(E) It's not possible to say from the plots alone.

(21-22) A survey was done to see whether the age of a student was a factor in determining whether a student was full time or part time.

Contingency table results:
Rows: Age
Columns: None

Cell format
Count (Total percent)

	Full-Time	Part-Time	Total
15-19	3550 (20.41%)	360 (2.07%)	3910 (22.48%)
20-24	5700 (32.78%)	1210 (6.958%)	6910 (39.73%)
25-34	1820 (10.47%)	1860 (10.7%)	3680 (21.16%)
35 and over	901 (5.181%)	1990 (11.44%)	2891 (16.62%)
Total	11971 (68.83%)	5420 (31.17%)	17391 (100.00%)

(21) Based on this sample, what is the probability a student is in full time education and the probability that a student over 35 is in full time education.

(A) The probability a student is in full time education is 68.83%, the probability that a student over 35 is in full time education is 31.2%.

(B) The probability a student is in full time education is 68.83%, the probability that a student over 35 is in full time education is 5.2%.

(C) The probability a student is in full time education is 68.83%, the probability that a student over 35 is in full time education is also 68.83%.

(D) The probability a student is in full time education is 68.83%, the probability that a student over 35 is in full time education is 16.6%.

(E) The probability a student is in full time education is $68.83 \times 16.62\%$, the probability that a student over 35 is in full time education is 31.2%.

(22) Suppose you want to test for independence between the full/part time status of a student and their age. The chi-squared value for the above test is 4035. What is the result of the test at the 5% level and the conclusions.

(A) We are testing the hypothesis H_0 : There is no association between age and student status against H_A : There is an association between age and student status. As we are rejecting the null, there appears to be a positive correlation between student status and age.

(B) We are testing the hypothesis H_0 : There is no association between age and student status against H_A : There is an association between age and student status. As we cannot reject the null there is no evidence to suggest that there is a positive correlation between student status and age.

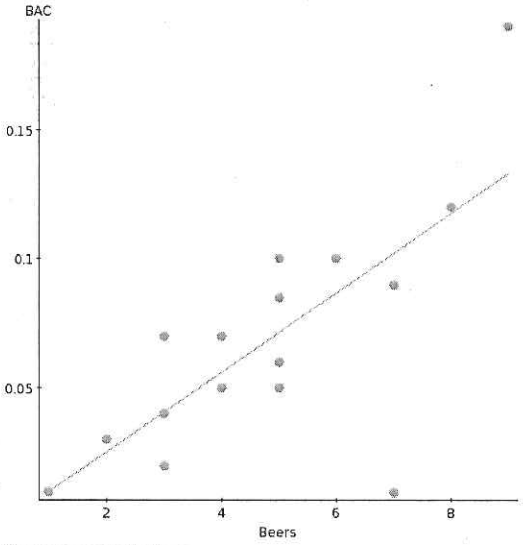
(C) Since $4035 > 7.81$, the p-value is a lot less than 5% and there is evidence to suggest that age has an influence on the status of a student. Indeed it appears that the probability of being enrolled as a part-time decreases with age.

(D) Since $4035 > 7.81$, there is no evidence at the 5% level that there is an association between age and full/part time status.

(E) Since $4035 > 7.81$, the p-value is a lot less than 5% and there is evidence to suggest that age has an influence on the status of a student. Indeed it appears that the probability of being enrolled as a part-time increases with age.

(23-25) We want to investigate the relationship between number of beers drunk and the blood alcohol content (half an hour after taking the beer). A sample of 16 students is taken and the results are summarised below. (BAC = blood alcohol content).

(23) The correlation between beer and BAC is 0.726, and some additional summary statistics and plot is given below.



Summary statistics:

Column	n	Mean	Std. Dev.	Std. Err.	Variance
Beers	16	4.8125	2.1975365	0.5493841	4.829167
BAC	16	0.06840625	0.046507783	0.011626946	0.0021629739

Using the above information, give the equation for the line of best fit.

(A) BAC= $-0.010 \times$beers - 0.004.

(B) BAC= $0.015 \times$beers - 0.004.

(C) beers= $0.015 \times$BAC - 0.004.

(D) beers= $-0.010 \times$BAC - 0.004.

(E) BAC= $0.010 \times$beers - 0.004.

(24) A new sample is taken. A summary of the simple regression is given below.

Simple linear regression results:
Dependent Variable: BAC
Independent Variable: Beers
BAC = -0.028630717 + 0.018183779 Beers
Sample size: 16
R (correlation coefficient) = 0.8355
R-sq = 0.6980703
Estimate of error standard deviation: 0.027202273

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF
Intercept	-0.028630717	0.01978307	≠ 0	14
Slope	0.018183779	0.0031961228	< 0	14

Is there evidence to suggest that the more beers drunk the higher the BAC (at the 5% level)?

(A) Here we are testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 < 0$. $t = -1.05$, so the p-value will greater than 50% and there is no evidence to suggest any dependence between number of beers and BAC. However, does not mean there is no dependence, it just means for this data set there is no statistical significance.

(B) Here we are testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 > 0$. $t = -1.05$, so the p-value will smaller than 20% and there is evidence to suggest the more drinks the lower higher the BAC.

(C) Here we are testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 > 0$. $t = 5.66$, so the p-value will be a smaller than 0.05% and there is evidence to suggest the more drinks the higher the BAC.

(D) Here we are testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 < 0$. $t = -5.66$, so the p-value will be close to 100% and there is evidence to suggest the more drinks the higher the BAC.

(E) Here we are testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 > 0$. $t = 5.66$, so the p-value will be a close to 100% and there is no evidence to suggest the more drinks the higher the BAC.

(25) The BAC is measured in mg/dL (mg per deciliter). Suppose instead that BAC is measured in grams per deciliter what will happen to the correlation and slope?

Hint: You may need to use that 1 gram = 1000mg.

(A) The correlation coefficient stays the same at 0.835 but the slope changes to $0.0181/1000 = 0.0000181$.

(B) The correlation coefficient changes to $0.835/1000 = 0.000835$ and the slope changes to $0.0181/1000 = 0.0000181$.

(C) The correlation coefficient changes to $0.835/1000 = 0.000835$ but the slope stays the same at 0.0181.

(B) The correlation coefficient changes to $0.835 \times 1000 = 835$ and the slope changes to $0.0181 \times 1000 = 18.1$.

(E) Neither the slope nor correlation coefficient change.