Statistics 201 – Fall 2019

Exam 1 – Practice Exam (from Fall 2016)

KEY

Disclaimer:

This practice exam is provided solely for the purpose of familiarizing you with the format and style of the Stat 201 exams. There is no explicit or implicit guarantee that the upcoming exam will ask similar questions. If you use the practice exam as your only tool to help you prepare for the upcoming exam, you most likely will not do well on the exam. You should still do the things you would have done if you did not have access to this practice exam, such as re-read the text, go over your class notes, re-work the online homework problems, and look at the list of exam topics provided and make sure that you understand all the concepts listed within it.

NOTE: Question 15 on this practice exam is from an older exam. As such, the points on this practice exam total more than 100 points.

HASLAM
COLLEGE OF BUSINESS
THE UNIVERSITY OF TENNESSEE, KNOXVILLE

100 Score

1. A study of potential age discrimination considered promotions among middle managers in a large company.

	Age		
	Up to 39	40 and Over	Total
Promoted	38	43	81
Not Promoted	82	88	170
Total	120	131	251

i) (2 points) What percent of employees were promoted?

$$\frac{81}{251} \times 100\% = 32.27\%$$

ii) (2 points) What percent of employees up to age 39 were promoted?

$$\frac{38}{120} \times 100\% = 31.67\%$$

iii) (2 points) What percent of employees were 40 and over?

$$\frac{131}{251} \times 100\% = 52.19\%$$

iv) (2 points) What percent of promoted employees were 40 and over?

$$\frac{43}{81} \times 100\% = 53.09\%$$

2. The demand for bottled water increases during hurricane season in Florida. The number of 1-gallon bottles of water sold for a random sample of n=9 hours in one store during hurricane season is:

64, 80, 74, 85, 82, 63, 67, 65, 75

i) (3 points) What is the average hourly number of 1-gallon bottles of water sold?

$$\bar{y} = \frac{\Sigma y}{n} = \frac{655}{9} = 72.78$$

ii) (3 points) What is the median number of 1-gallon bottles of water sold?

63, 64, 65, 67, $\boxed{74}$, 75, 80, 82, 85 Median = 74

iii) (2 points) Partial JMP output is provided below. From this output, calculate the IQR of these data.

Quantiles

100.0%	maximum	85	
99.5%		85	
97.5%		85	
90.0%		85	
75.0%	quartile	81	$= Q_3$
50.0%			
25.0%	quartile	64.5	$= Q_1$
10.0%		63	
2.5%		63	
0.5%		63	
0.0%	minimum	63	

$$IQR = Q_3 - Q_1$$
$$= 81 - 64.5$$
$$= 16.5$$

$\mu = 4.9 \qquad \sigma = 1.4$

3. Suppose the number of minutes a customer is put on hold at a particular software company's technical support call center is approximately normally distributed, with a mean of 4.9 minutes and a standard deviation of 1.4 minutes. Betty called this call center and was on hold for 7.5 minutes. $y = 7.5$

i) (2 points) Calculate the z-score for Betty's hold time.

$$z = \frac{y-\mu}{\sigma} = \frac{7.5-4.9}{1.4} = 1.86$$

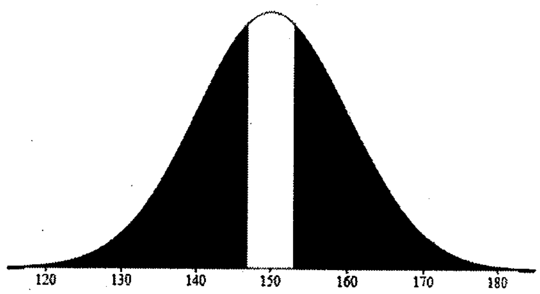
ii) (3 points) Interpret the number you calculated in part (i) above. Don't comment on the magnitude of this number, rather explain what this number means. [Note: If you have no answer for part (i), use 2.34, which is NOT the correct answer to part (i)]

Betty's hold time was 1.86 standard deviations above average.

iii) (2 points) Another customer (David) called the same technical support call center and had a hold time that had a z-score of -0.55. The percent of customers that wait longer than David is approximately (circle the best answer):

a) 97.5%
b) 71%
c) 50%
d) 29%
e) 2.5%
f) .3%

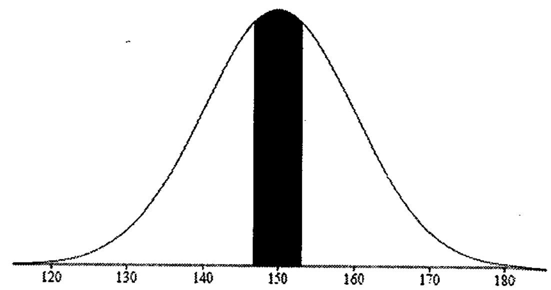
4. (3 points) The reaction time of a professional hockey goalie was measured many times, and the distribution of these reaction times was approximately normally distributed, with an average of 150 milliseconds and a standard deviation of 10 milliseconds. What is the 75th percentile of this hockey goalie's reaction times? Use the correct screen shot below from our Normal Curve Calculator to answer this question (**hint**: only one of the images below can help you answer this question).



○ Area from a value (Use to compute p from Z)
◉ Value from an area (Use to compute Z for confidence intervals)

Specify Parameters:
Area 0.75
Mean 150
SD 10

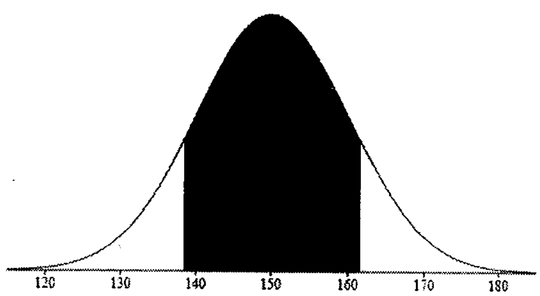
Results:
Recalculate
○ Above
○ Below
○ Between
◉ Outside 146.818 and 153.182



○ Area from a value (Use to compute p from Z)
◉ Value from an area (Use to compute Z for confidence intervals)

Specify Parameters:
Area 0.25
Mean 150
SD 10

Results:
Recalculate
○ Above
○ Below
◉ Between 146.818 and 153.182
○ Outside

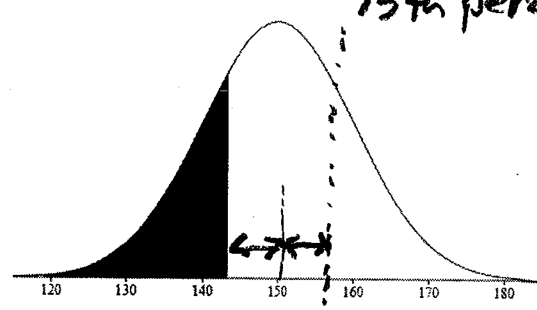


○ Area from a value (Use to compute p from Z)
◉ Value from an area (Use to compute Z for confidence intervals)

Specify Parameters:
Area 0.75
Mean 150
SD 10

Results:
Recalculate
○ Above
○ Below
◉ Between 138.496 and 161.504
○ Outside

75th percentile



○ Area from a value (Use to compute p from Z)
◉ Value from an area (Use to compute Z for confidence intervals)

Specify Parameters:
Area 0.25
Mean 150
SD 10

Results:
Recalculate
○ Above
◉ Below 143.258
○ Between
○ Outside

25th percentile

$150 - 143.258 = 6.742$

$150 + 6.742 = \underline{156.742}$

$\mu = 2.5 \qquad \sigma = .32$

5. A particular brand of cell phone has a battery that will last, under normal conditions, 2.5 years on average, with a standard deviation of 0.32 years. The distribution of these lifetimes is approximately normally distributed.

$y = 2$

i) (2 points) Calculate the z-score for a battery that lasts 2 years under normal conditions.

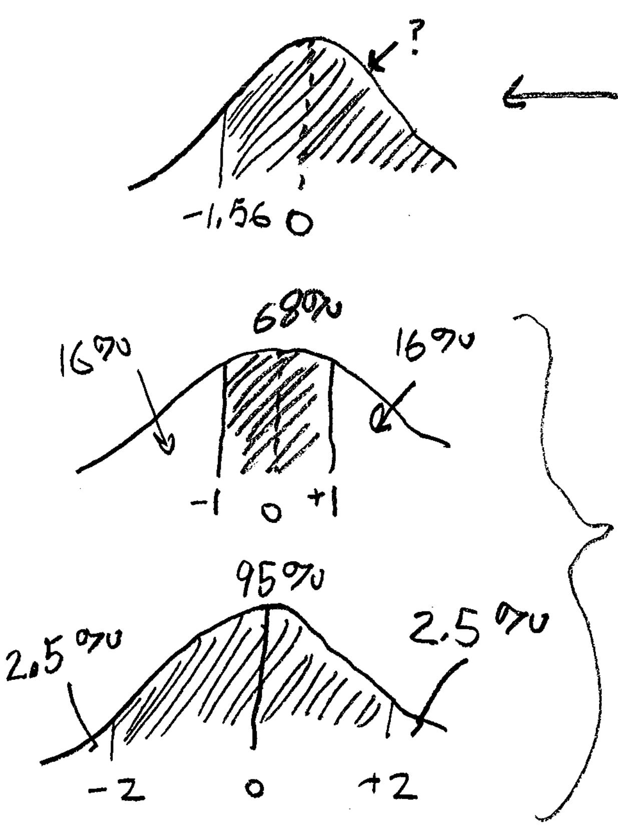
$$z = \frac{y-\mu}{\sigma} = \frac{2-2.5}{.32} = -1.56$$ ← i.e., 1.56 standard deviations below average.

ii) (5 points) What percent of this brand of batteries can be expected to last 2 years or more under normal conditions?

Without the normal curve calculator, you can't give an exact answer here. Base your answer on the 68-95-99.7 rule, and your answer to part [i] above. Give as narrow of an interval as you can that contains the exact answer. Fill in the blanks below. Also, show your work (and/or reasoning) below.

- (Note 1: if you have a calculator that could give you the exact answer, DO NOT use that capability! Use only the 68-95-99.7 rule.)
- (Note 2: if you have NO answer for part [i] above, use -1.2 as your answer there, which is NOT the correct answer to part [i]).

"The exact answer must be between ___84___% and ___97.5___%."



← Shaded area is between:

100% − 16% = 84%

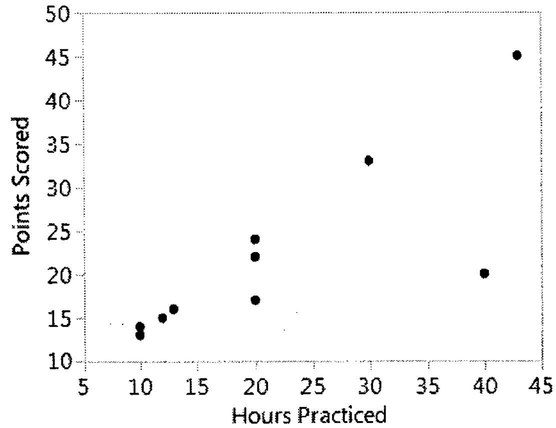
and

100% − 2.5% = 97.5%

6. (4 points) A gas station is trying to increase food sales inside the store. Many customers pay at the pump and leave. The gas station begins collecting data hoping to discover something that will bring customers inside after they pump their gas. A random sample of the data set is shown below. Above each of the eight columns, write a **C** if that variable is categorical, **Q** if that variable is quantitative, or **I** if that variable is an identifier variable.

I	C	Q	C	C	C	Q	C
transaction number	type of gas	number of gallons	pay at pump?	inside food sale?	type of payment	Gas purchase in dollars	Day of week
9853	premium	22	y	n	Visa	85.58	Mon
9211	diesel	26	y	n	Am Exp	110.5	Tues
8875	regular	19	y	y	Visa	70.11	Tues
8824	regular	21	y	y	Visa	77.49	Fri
8313	regular	14	y	y	MasterCard	51.66	Wed
7699	premium	22	n	n	cash	85.8	Wed
7645	diesel	45	y	y	Am Exp	191.25	Sat
3145	diesel	38	y	n	Am Exp	161.5	Sat
2588	regular	17	n	y	Visa	62.73	Sun
2499	regular	22	n	n	cash	81.18	Sat
2325	premium	15	y	n	MasterCard	58.35	Fri
2291	diesel	22	y	n	MasterCard	92.4	Mon
2078	regular	14	y	y	Visa	51.66	Thur
1843	regular	35	y	n	Visa	129.15	Thur
2103	regular	25	n	y	cash	92.25	Sat

7. A high school football coach believes that the number of points scored in each game is correlated with how many hours his players spend practicing the week before. He records the hours practiced before 10 games and the points his team scored in those games.



i) (6 points) There are 3 conditions that must be checked before calculating a correlation coefficient (r). List each condition, and briefly comment on whether or not each condition is met in this case.

Condition 1 is: Quantitative variables

Is condition 1 met? (Circle one): Yes No
Briefly explain your answer:

Hours and points are both quantitative variables.

Condition 2 is: Straight enough

Is condition 2 met? (Circle one): Yes No
Briefly explain your answer:

overall, the relationship is linear (and not a "curved" relationship).

Condition 3 is: NO outliers

Is condition 3 met? (Circle one): Yes No
Briefly explain your answer:

There is an outlier in the lower right of the scatterplot.

Hours practiced ≈ 40

Points scored ≈ 20

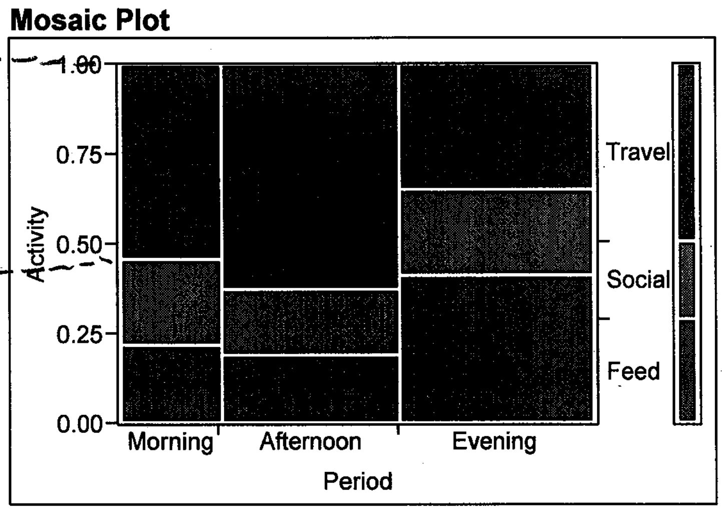
Question 7 (continued)

ii) (2 points) Suppose that the correlation coefficient, r is calculated using all of the data in the scatter plot. In one week, the players practiced 40 hours and scored 20 points. How would the removal of this point change the values of r? Circle the best answer:

a) r would be unchanged
b) r would get closer to -1
c) r would get closer to 0
d) r would get closer to +1
e) The impact on r cannot be determined

8. Are Icelandic dolphins' behavioral activities related to a specific period of the day? Icelandic marine biologists observed the behavior (categorized into 3 behaviors: Travel, Social, Feeding) and time of day (categorized into 3 distinct periods: Morning, Afternoon, and Evening) of 1200 dolphins. They hoped to observe a relationship between these two variables over the course of 2 years. Below is a mosaic plot of the collected data:

About 55%



i) (2 points) Based on the mosaic plot, approximately what percent of these dolphins' mornings were spent traveling?

≈ 55%

ii) (2 points) Based on the mosaic plot, at what time of the day were the Icelandic dolphins observed the least? Briefly explain how you came to this conclusion.

Morning. The width of the morning bars is the smallest of the 3 time periods.

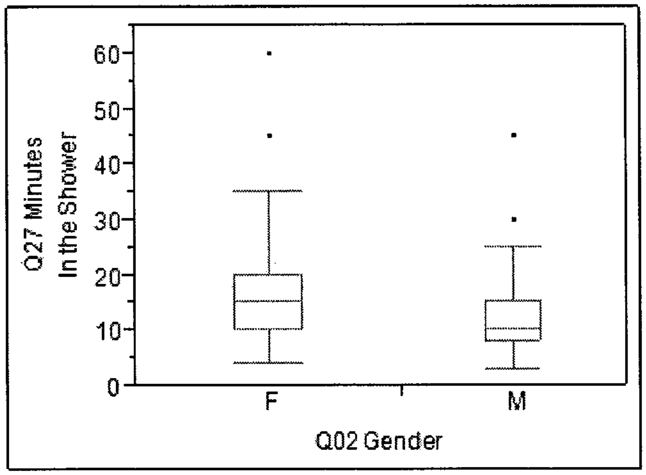
iii) (1 point) Is there a relationship between the variables Activity and Period? CIRCLE your answer.

Relationship   No Relationship

iv) (3 points) Make reference to the mosaic plot to justify your choice above.

* The heights of the bars within each time period are different for different time periods.
* They mostly traveled in the afternoons (≈60% of the time) but traveled much less in the evenings (≈ 35% of the time).

9. Who takes longer showers? The following boxplots show the shower times in minutes for Stat 201 students, as reported in a previous survey.



i) (2 points) The distribution of the male shower times is (Circle the BEST answer)

a) Right-Skewed
b) Left-Skewed
c) Symmetric
d) Unimodal
e) Bimodal

ii) (2 points) What is the approximate range of shower times for these females?

Min ≈ 3
Max ≈ 60 Range $\approx 60 - 3 = 57$

iii) (2 points) Which group has a higher median shower time? CIRCLE the best answer.

males females cannot be determined the median times are the same

iv) (2 points) 75% of males spend less than <u>15</u> minutes in the shower.

v) (2 points) What is the IQR (approximately) for the female shower times?

$Q_1 \approx 10$
$Q_3 \approx 20$ $IQR = Q_3 - Q_1 = 20 - 10 = 10$

10. (2 points) Adding an outlier in a scatterplot will do which of the following to the correlation coefficient? (Circle the best answer)

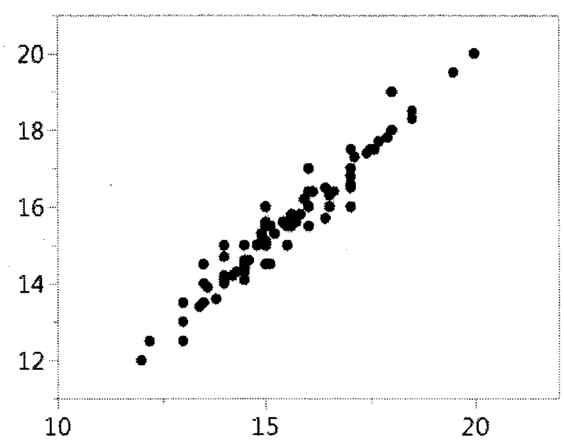
a) Increase the correlation coefficient
b) Decrease the correlation coefficient
c) Change the correlation coefficient from positive to negative
d) Change the correlation coefficient from negative to positive
e) All options above are possible

11. (2 points) When the correlation coefficient is close to +1, it indicates (Circle the best answer)

a) changes in one variable cause changes in the other, but we don't know which one causes the other to change.
b) changes in the variable on the x-axis cause changes in the variable on the y-axis.
c) changes in the variable on the y-axis cause changes in the variable on the x-axis.
d) nothing regarding the possible cause and effect relationship between the two variables.

12. (2 points) What is the most likely correlation coefficient for the scatterplot below?

a) 1.00
b) 0.98
c) 0.50
d) 0.00
e) -0.50
f) -0.98
g) -1.00



13. The following regression equation was developed after a large number of nursing students agreed to share their latest medical records with us:

$$\text{Height (in.)} = 54.55 + 0.085*\text{Weight (lbs.)}$$

i) (2 points) Mary was one of the nursing students that were part of this study. Suppose Mary weighs 110 pounds. Use the regression equation above to predict Mary's height (in inches).

$$\widehat{\text{Height}} = 54.55 + .085(110)$$
$$= 54.55 + 9.35 = 63.9 \text{ inches}$$

ii) (2 points) If Mary is actually 65 inches tall, what would the residual for this observation be? [Note: if you have no answer for part (i), use 60.4 inches, which is NOT the correct answer for part (i).]

$$e = y - \hat{y} = 65 - 63.9 = 1.10$$

iii) (3 points) Interpret the slope of this regression equation in the context of this problem.

For every additional pound a nursing student was, they were .085 inches taller, on average.

iv) (4 points) John, another nursing student that was part of this study, is 68 inches tall. We used the above regression equation to find John's predicted height, and then calculated the residual for this observation, and found that the residual was -1.47. How much does John weigh (in pounds)?

$y = 68$
$e = -1.47$
$x = ?$

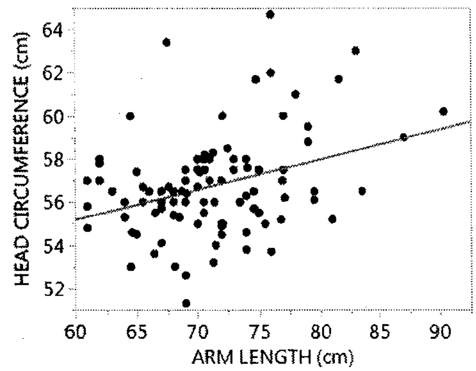
$e = y - \hat{y}$ and $\hat{y} = 54.55 + .085(x)$

$$\downarrow$$

$$-1.47 = 68 - [54.55 + .085(x)]$$
$$-1.47 = 68 - 54.55 - .085(x)$$
$$-1.47 = 13.45 - .085(x)$$
$$-14.92 = -.085(x)$$
$$x = 175.53 \text{ pounds}$$

14. In Stat 201, Fall 2016, a group of students measured their head circumference and their right arm length. Below is some output from JMP for a regression analysis, using X = right arm length (cm) and Y = head circumference (cm):

Bivariate Fit of HEAD CIRCUMFERENCE (cm) By ARM LENGTH (cm)



Linear Fit

Linear Fit

HEAD CIRCUMFERENCE (cm) = 46.784076 + 0.1398624*ARM LENGTH (cm)

Summary of Fit

RSquare	0.115974
RSquare Adj	0.106766
Root Mean Square Error	2.195403
Mean of Response	56.75612
Observations (or Sum Wgts)	98

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	46.784076	2.818699	16.60	<.0001*
ARM LENGTH (cm)	0.1398624	0.039411	3.55	0.0006*

i) (1 point) Based on the output above, is the linear relationship between these two variables "statistically significant"? Circle one: YES NO

ii) (2 points) Circle the ONE value in the output that lead to your conclusion in part (i).

iii) (3 points) Regardless of your answers above, briefly describe what it means for a linear relationship to be "statistically significant". Limit your answer to two or less sentences.

The amount of linear association observed in the data is unlikely due to pure chance.

True/False Questions
Circle the best answer
(2 points each)

T F The normal model can be used to describe all distributions.

T F An observation with a z-score of 0.5 would be considered an outlier.

T F If all of the data in a skewed right data set are converted into z-scores, the distribution of these z-scores will be normally distributed.

T F Bar charts are used to graphically display both quantitative and categorical variables.

T F When doing a Pivot Table in Excel, the default value displayed is the SUM of the selected variable.

Formula Sheet

Regression:

$$\hat{y} = b_0 + b_1 x$$

$$e = y - \hat{y}$$