STAT 135-01

Final Exam

Spring 2017

Name: _____

You may use a calculator and a two-sided 8.5" by 11" sheet of notes, which you will turn in with your exam. Please show all your work, including all calculations, and explain your answers. Whenever needed, please round numbers (including intermediate calculations) to the nearest 0.001. **Cell phones and any other electronic devices are NOT permitted.** No interaction of any sort is allowed with your classmates.

This Exam consists of two parts:

- **Part I. Multiple Choice Questions (Q1–Q5):** There is only ONE correct response per question, and there are a total of 5 questions in this part.
- **Part II. Word Problems (Q1–Q8):** You must show sufficient work in order to receive full credit, and there are a total of 8 problems in this part.

Part	Points Scored	Out of
I		10
II 1.		10
II 2.		12
II 3.		15
II 4.		18
II 5.		12
II 6.		24
II 7.		12
II 8.		17
Total		130

The following information may be helpful to you:

$P(Z < -1.921) = pnorm(-1.921) = 0.027,$	$qnorm(0.027) = -1.921$
$P(Z < 0.950) = pnorm(0.950) = 0.829,$	$qnorm(0.829) = 0.950$
$P(Z < 1.282) = pnorm(1.282) = 0.900,$	$qnorm(0.900) = 1.282$
$P(Z < 1.645) = pnorm(1.645) = 0.950,$	$qnorm(0.950) = 1,645$
$P(Z < 1.822) = pnorm(1.822) = 0.966,$	$qnorm(0.966) = 1.822$
$P(Z < 1.960) = pnorm(1.960) = 0.975,$	$qnorm(0.975) = 1.960$
$P(Z < 2.759) = pnorm(2.759) = 0.9971,$	$qnorm(0.9791) = 2.759$
$P(Z < 2.901) = pnorm(2.901) = 0.99814,$	$qnorm(0.99814) = 2.901$

$P(t_{120} < -1.921) = pt(-1.921, df = 120) = 0.02855,$	$qt(0.02855, df = 120) = -1.921$
$P(t_{118} < -1.921) = pt(-1.921, df = 118) = 0.02857,$	$qt(0.02857, df = 118) = -1.921$
$P(t_{123} < 1.2885) = pt(1.2885, df = 123) = 0.900,$	$qt(0.900, df = 123) = 1.2885$
$P(t_{120} < 1.2886) = pt(1.2886, df = 120) = 0.900,$	$qt(0.900, df = 120) = 1.2886$
$P(t_{123} < 1.6573) = pt(1.6573, df = 123) = 0.950,$	$qt(0.950, df = 123) = 1.6573$
$P(t_{120} < 1.6577) = pt(1.6577, df = 120) = 0.950,$	$qt(0.950, df = 120) = 1.6577$
$P(t_{123} < 1,9794) = pt(1.9794, df = 123) = 0.975,$	$qt(0.975, df = 123) = 1.9794$
$P(t_{120} < 1.9799) = pt(1.9799, df = 120) = 0.975,$	$qt(0.975, df = 120) = 1.9799$

I Multiple Choice Questions

1. **True or False**: In hypothesis testing, the p-value measures the probability that the data were produced by random chance alone.
   A. True. B. False.

2. Absorption rates into the body are important considerations when manufacturing a generic version of a brand-name drug. A pharmacist read that the absorption rate into the body of a new generic drug (G) is the same as its brand-name counterpart (B). She has a researcher friend of hers run a small experiment to test H_o: $\mu_G - \mu_B = 0$ against H_A: $\mu_G - \mu_B \neq 0$. Which of the following would be a Type I error?

   A. Deciding that the absorption rates are the same, when in fact they are.
   B. Deciding that the absorption rates are different, when in fact they are.
   C. Deciding that the absorption rates are the same, when in fact they are not.
   D. The researcher cannot make a Type I error, since he has run an experiment.
   E. Deciding that the absorption rates are different, when in fact they are not.

3. Which statement is NOT true about confidence intervals?

   A. A confidence interval is an interval of values computed from sample data that is likely to include the true population parameter value.
   B. An approximate formula for a confidence interval is sample estimate $\pm$ margin of error.
   C. A confidence interval between 20% and 40% means that the population proportion definitely lies between 20% and 40%.
   D. A 99% confidence interval procedure has a higher probability of producing intervals that will include the population parameter than a 95% confidence interval procedure.
   E. Confidence intervals are (by definition) statistical inference procedures.

4. What statement is true about both p and $\bar{y}$?

   A. They are both parameters.
   B. They are both statistics.
   C. They are both symbols pertaining to means.
   D. $\bar{y}$ is a statistic and p is a parameter.
   E. $\bar{y}$ is a parameter and p is a statistic.

5. Which statement correctly compares the t-models to the normal models?

   (I) The t-models are also mound shaped and symmetric.
   (II) The t-models are more spread out than the normal distribution.
   (III) As degrees of freedom increase, the variance of the t-models becomes larger.

   A. (I) only. B. (II) only. C. (I) and (II). D. (II) and (III). E. All.

II Word Problems

Professor Liao in her first year returning to teach at Amherst College would like to conduct a thorough investigation on the college's student population. A committee of twenty eight has been formed to design and analyze a campus-wise survey, which consists of responses from $n = 124$ randomly selected students and is believed to be representative of the student body of Amherst College.

Please note that the *Independence Assumption* is assumed to be satisfied, so you do NOT need to worry about checking the randomization condition or the 10% condition in the questions below. You will, however, need to check OTHER Assumption(s) and condition(s) whenever necessary.

1. A subgroup of committee (Natalia, Yasmeen, Irish, and Ezra) would like to know what proportion of Amherst students are first generation, and they find that among 124 respondents, 28 identify themselves as first generation college students.

   (a) Construct and *interpret* a 90% confidence interval for the proportion of Amherst College students that are first generation.

   (b) According to the College's website, 14% of Class 2020 are first generation college students. Some committee members argue that this class year seems to have significantly less first generation students, compared to the others. Does your interval above support this statement? Briefly explain.

   (c) What is the significance level associated with your conclusion in (b)?

2. Another subgroup of three members (Emily, Yrenly, and Jekabs) is working on understanding the athlete non-athlete divide on campus, and would like to know if athlete students tend to go to bed earlier than students who don't participate in any sport teams. In particular, they want to test if the proportion of athlete students who sleep before 1AM is greater than that of non-athlete students.

	Athlete	Non-Athlete
Sleep Before 1AM	36	30
Sleep After 1AM	13	33

(a) Test appropriate hypotheses at $\alpha = 0.1$ and state your conclusion.

(b) Explain what your p-value means in the context of the problem.

3. While it's important to sleep early, it's equally important to have enough sleep! Three committee members, Bishesh, Mark, and Michael, devote themselves to investigating students' sleeping habits and would like to know if there is a gender difference in the average number of sleep hours per night. Here are some descriptive statistics they learned from the survey:

```
  Gender  min  Q1  median   Q3  max       mean         sd    n   missing
1 Female    4   6       7  7.5   10   6.771186   1.218825   59       0
2   Male    4   7       7  8.0   10   7.193548   1.198889   62       0
```

   (a) Write a few sentences to compare the number of sleep hours per night between two groups.

   After making sure that all assumptions and conditions are okay, they performed a hypothesis test, followed by a confidence interval for this question of interest and got the following output (partial) from `R`:

```
data:  SleepHours by Gender (Female - Male)
t = -1.9206, df = 118.48, p-value = 0.05719
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -0.78693608 -0.05778781
```

   (b) State appropriate hypotheses for this test.

   (c) Use descriptive statistics above to verify the value of `t = -1.9206` in the output.

(d) Is this difference statistically significant at $\alpha = 0.1$? If no, can we conclude that females have the same average number of sleep hours as males? If yes, carefully interpret the confidence interval above in the context of this problem and then briefly comment on whether you think such a difference is also *practically* significant/meaningful.

4. Three committee members, Josh, Alexander, and Eric, spend their time exploring possible factors that may impact students' academic performances. One of their missions is to check if the average GPA is the same regardless of where a student was born. They run a one-way ANOVA to compare students' GPA among five POB (Place of Birth) regions: International, Midwest, Northeast, South, and West. The corresponding ANOVA table is shown below:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
POB	(1)	1.1333	(2)	(4)	0.01858 *
Residuals	117	10.7363	(3)		

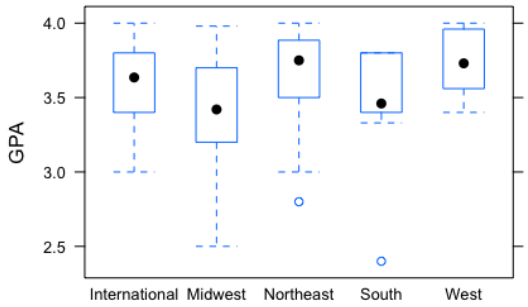
(a) What are the missing values for cells (1) – (4) in the above ANOVA table?

(b) State appropriate hypotheses for this ANOVA test.

(c) Assuming that the assumptions are all met, what is the corresponding p-value for this test? State your conclusion at $\alpha = 0.1$.

(d) Now list the required assumptions for ANOVA and use the plot and numerical summaries provided below to check them. Do you have any concerns? Explain.

	POB	min	Q1	median	Q3	max	mean	sd	n
1	International	3.0	3.40	3.635	3.800	4.00	3.573929	0.2884356	28
2	Midwest	2.5	3.20	3.420	3.700	3.98	3.448750	0.3722880	16
3	Northeast	2.8	3.50	3.750	3.885	4.00	3.669216	0.2887826	51
4	South	2.4	3.40	3.460	3.800	3.80	3.443333	0.4346838	9
5	West	3.4	3.57	3.730	3.925	4.00	3.728333	0.2071728	18

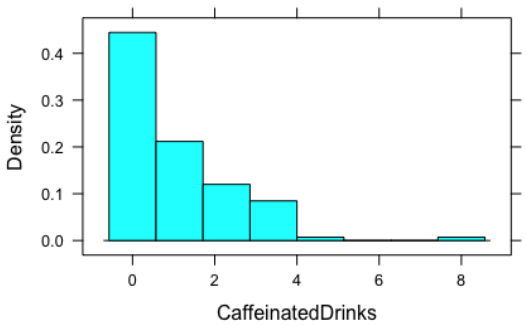


(e) If there is any concern about those assumptions for ANOVA, it would be a good idea to run a non-parametric test instead. What would be an appropriate nonparametric alternative to one-way ANOVA? Name the procedure.

5. Focusing on students' health issues, a subgroup of three (Rana, Kashmeera, and Janelle) investigate the number of cups of caffeinated drinks students at Amherst College consume per day. The following are some descriptive statistics they learned from the survey.

min	Q1	median	Q3	max	mean	sd	n	missing
0	0	---	1.625	8	0.9798387	1.277	124	0

(a) What is the shape of the distribution? Is the median greater or less than the mean?



(b) Are there any outliers? Please explain.

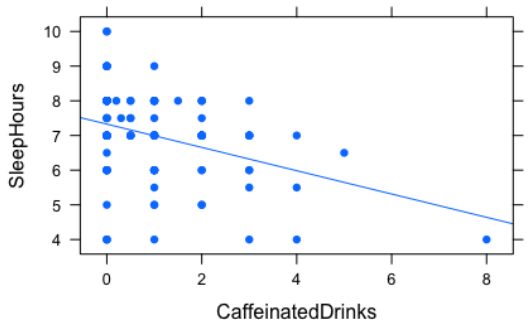
(c) When describing the center and spread of this variable, which set of summary statistics would you prefer? Why?

(d) Comment on whether it's a good idea to use a one-sample t-interval for the mean here. If yes, please report the 90% confidence interval. If no, please list your concern(s).

6. Also interested in students' daily consumption of caffeinated drinks, Araceli, Amelia, and Sarah further check if this variable is related to the number of sleep hours per night.

   They first look at the scatterplot of these two variables (as shown on the right), and run a simple linear regression (SLR) on the data. The corresponding fitted line is

   $\widehat{SleepHours} = 7.328 - 0.336\, CaffeinatedDrinks.$

   After careful considerations, they decided to re-fit the model WITHOUT the point with `CaffeinatedDrinks = 8`.

   Here is the new output (partial) from `R`:

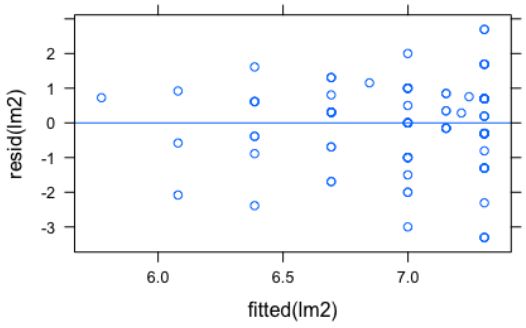
```
Call:  lm(formula = SleepHours ~ CaffeinatedDrinks)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.30557    0.13522   54.03  < 2e-16 ***
CaffeinatedDrinks -0.30643    0.09341   -3.28  0.00136 **

Residual standard error: 1.144 on 120 degrees of freedom
Multiple R-squared:  0.08229,Adjusted R-squared:  0.07464
```

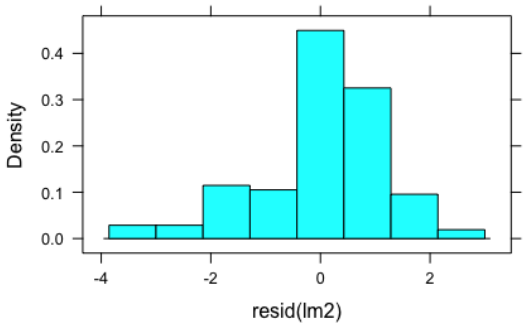
   Two plots attached attached below are the residuals-fitted plot and the residuals histogram.

   (a) What is the equation of the new fitted line?

   What is the correlation?

   (b) Is the point with `CaffeinatedDrinks = 8`, the one being removed above, high leverage, high influence, or neither? Explain.

(c) Explain (in context) what the y-intercept of the new fitted line means.

(d) Is there a relationship between these two quantitative variables? State appropriate hypotheses, check the required assumptions for SLR (proceed with caution if you have any concerns), perform the test and then state your conclusion about the association.

(e) Interpret the value of the sample slope in the context of the problem. Then, create a 90% confidence interval for the true slope and explain in context what your interval means.

7. Sarah D., Angelika, and Caroline are interested in the relationship between students' showering habit (the number of times they shower per week) and their exercise habit (the number of hours they exercise per week). They also wonder if such a relationship would differ by gender, so they run a multiple regression with an indicator variable, `GenderMale`, which is coded as 1 for Males, and 0 for Females. Here is the `R` output of this first model they tried (Model 1):

```
Call:  lm(formula = Shower ~ Exercise * Gender)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.73672    0.44207  10.715  < 2e-16 ***
Exercise             0.16328    0.06571   2.485  0.01438 *
GenderMale           1.74153    0.59677   2.918  0.00422 **
Exercise:GenderMale  0.02584    0.08082   0.320  0.74977
```

   (a) What is the fitted regression equation for each gender?

   Male:

   Female:

   (b) What role does the interaction term (`Exercise:GenderMale`) play in modeling here?

Since the interaction term is not significant at $\alpha = 0.1$, they re-fit the model without the interaction term and get the following output (Model 2):

```
Call:  lm(formula = Shower ~ Exercise + Gender)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.64711    0.34056  13.645  < 2e-16 ***
Exercise     0.18036    0.03811   4.733 6.21e-06 ***
GenderMale   1.88629    0.38726   4.871 3.49e-06 ***
```

   (c) Briefly explain the meaning of the estimated coefficient, 1.88629, for the indicator variable `GenderMale` in Model 2.

(c) Also interested in students' showering habit, another three committee members, Nouraiz, Mikayla, and Jade in fact discovered another regression model for the variable `Shower` using `Exercise` and `Stressed` (students' self-reported stress levels on a scale of 0 to 5, with 5 being the most stressful) as predictors, both of which are statistically significant. The corresponding fitted model is $\widehat{Shower} = 6.702 + 0.19\,Exercise - 0.406\,Stressed$. Does the *negative* coefficient for the predictor `Stressed` suggest that students should spend more time showering if they want to reduce their stress level? Explain.

8. Three committee members, Jackie, Nick, and Vic, are concerned about students' mental health and would like to see if there is an association between their gender (Female/Male) and how stressful they usually feel (Mildly/Moderately/Extremely). The two-way table below summaries the results from the survey. **Observed count** (Expected count) is the setup.

	Mildly Stressful	Moderately Stressful	Extremely Stressful	Total
Female	**5** (9.26)	**27** (31.21)	**27** ( )	59
Male	**14** (9.74)	**37** (32.79)	**11** ( )	62
Total	19	64	38	121

(a) What inference procedure should be used to test for the association between these two variables? State appropriate hypotheses.

Inference Procedure -

H_0:

H_A:

(b) Two expected counts are missing in the table above. Compute and fill in both missing expected counts in the parentheses.

(c) Check the assumptions and conditions.

(d) The test statistic turns out to be 12.496 and the p-value is 0.001935. What are the degrees of freedom for the model being used? State your conclusion at $\alpha = 0.1$.

(e) The two cells with missing expected counts turn out to be the ones which contribute the most to the above test statistic. Compare their observed and expected counts carefully, and report the values of their contributions to the test statistic. What additional message can we obtain from these two cells?