STAT 135

Final Exam

Fall 2016

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–Q6):** There is only ONE correct response per question, and there are a total of 6 questions in this part.
- **Part II. Word Problems (Q1–Q8):** You must show all/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		12
II 1.		12
II 2.		10
II 3.		15
II 4.		24
II 5.		22
II 6.		10
II 7.		25
II 8.		10
Total		140

The following information may be helpful to you:

$$
\begin{array}{ll}
P(Z < -1.691) = pnorm(-1.691) = 0.045, & qnorm(0.045) = -1.691 \\
P(Z < 1.282) = pnorm(1.282) = 0.900, & qnorm(0.900) = 0.161 \\
P(Z < 1.645) = pnorm(1.645) = 0.950, & qnorm(0.950) = 0.313 \\
P(Z < 1.695) = pnorm(1.695) = 0.955, & qnorm(0.955) = 1.085 \\
P(Z < 1.822) = pnorm(1.822) = 0.966, & qnorm(0.966) = 1.822 \\
P(Z < 1.960) = pnorm(1.960) = 0.975, & qnorm(0.975) = 1.960 \\
P(Z < 2.025) = pnorm(2.025) = 0.979, & qnorm(0.979) = 2.025 \\
P(Z < 2.807) = pnorm(2.807) = 0.997, & qnorm(0.997) = 2.807
\end{array}
$$

$$
\begin{array}{ll}
P(t_{87} < 1.291) = pt(1.291, df = 87) = 0.900, & qt(0.900, df = 87) = 1.291 \\
P(t_{94} < 1.291) = pt(1.291, df = 94) = 0.900, & qt(0.900, df = 94) = 1.291 \\
P(t_{129} < 1.288) = pt(1.288, df = 129) = 0.900, & qt(0.900, df = 129) = 1.288 \\
P(t_{87} < 1.663) = pt(1.663, df = 87) = 0.950, & qt(0.950, df = 87) = 1.663 \\
P(t_{94} < 1.661) = pt(1.661, df = 94) = 0.950, & qt(0.950, df = 94) = 1.661 \\
P(t_{129} < 1.657) = pt(1.657, df = 129) = 0.950, & qt(0.950, df = 129) = 1.657 \\
P(t_{87} < 1,988) = pt(1.988, df = 87) = 0.975, & qt(0.975, df = 87) = 1.988 \\
P(t_{94} < 1.986) = pt(1.986, df = 94) = 0.975, & qt(0.975, df = 94) = 1.986 \\
P(t_{129} < 1.979) = pt(1.979, df = 129) = 0.975, & qt(0.975, df = 129) = 1.979
\end{array}
$$

$$
\begin{array}{l}
P(\chi^2_1 < 2.861) = pchisq(2.861, df = 1) = 0.909 \\
P(\chi^2_1 < 2.861) = pchisq(2.861, df = 2) = 0.761 \\
P(\chi^2_2 < 6.832) = pchisq(6.832, df = 2) = 0.967 \\
P(\chi^2_3 < 6.832) = pchisq(6.832, df = 3) = 0.923
\end{array}
$$

I Multiple Choice Questions

1. Suppose that a Normal model describes fuel economy (miles per gallon) for automobiles and that a Saturn has a standardized score (z-score) of +2.2. This means that Saturns

   A. get 2.2 miles per gallon.
   B. achieve fuel economy that is 2.2 standard deviations better than the average car.
   C. have a standard deviation of 2.2 mpg.
   D. get 2.2 mpg more than the average car.
   E. get 2.2 times the gas mileage of the average car.

2. Suppose a local school district decides to randomly test high school students for attention deficit disorder (ADD). There are three high schools in the district, each with grades 9-12. The school board pools all of the students together and randomly samples 250 students. Is this a simple random sample?

   A. No, because we cant guarantee that there are students from each school in the sample.
   B. No, because we cant guarantee that there are students from each grade in the sample.
   C. Yes, because the students were chosen at random.
   D. Yes, because each student is equally likely to be chosen.

3. All but one of these statements contain a blunder. Which could be true?

   A. The correlation between the amount of fertilizer used and the yield of beans is 0.42.
   B. There is a correlation of 0.63 between gender and political party.
   C. The correlation between a football players weight and the position he plays is 0.54.
   D. There is a high correlation (1.09) between height of a corn stalk and its age in weeks.
   E. The correlation between a cars length and its fuel efficiency is 0.71 miles per gallon.

4. True or False: The p-value is the probability that the null hypothesis is true.
   A. True B. False

5. Bat Co., a company that sells batteries, claims that 99.5% of their batteries work. How many batteries would you expect to buy, on average, to find one that does not work?
   A. 994 B. 5 C. 200 D. 995 E. 199

6. In a survey, students are asked how many hours they study in a typical week. A five-number summary of the responses is: 2, 9, 14, 20, 60. Which interval describes the number of hours spent studying in a typical week for about 25% of the students sampled?

   A. 2 to 14
   B. 2 to 60
   C. 9 to 20
   D. 14 to 60
   E. 20 to 60

II Word Problems

Professor Liao in her first year returning to teach at Amherst College would like to conduct a thorough investigation to re-understand the college's student population better. A committee of twenty seven experts has been asked to design and analyze a campus-wise survey. The sample consists of responses from $n = 129$ 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 its associated conditions in the questions below. You will, however, need to check OTHER Assumptions and Conditions whenever necessary.

1. Three committee members, Sharline, Aleksandar, and Marissa, would like to know if Amherst College has grade inflation, and they found that the mean GPA of all private college in the United States is 3.33. Use the following summary statistics about GPA to answer the questions below.

min	Q1	median	Q3	max	mean	sd	n	missing
2.7	3.405	3.6	3.718	4	3.557	0.2785	90	0

   (a) Given that it's unimodal, what is the shape of the distribution?

   (b) Are there any outliers? Please explain.

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

   (d) Comment on whether it's a good idea to use a one-sample t-test for the mean here. If yes, please state appropriate hypotheses. If not, please list your concern(s).

2. David, Ashdon, and Esteban would like to investigate what impact Age might have on Weight for different Gender. They run multiple regression (or more specifically, ANCOVA in this case) with an indicator variable, `GenderM`, which is coded as 1 for Males, and 0 for Females. Here is the corresponding R output of the first model they run (Model 1):

```
lm(formula = Weight ~ Age * GenderM, data = Q2ds)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   77.572     50.190   1.546   0.1248
Age            3.072      2.563   1.199   0.2329
GenderM      122.788     67.811   1.811   0.0726 .
Age:GenderM   -4.578      3.445  -1.329   0.1864
```

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

Male:

Female:

(b) What is the role of the interaction term (`Age:GenderM`), i.e. what does it allow us to do in modeling?

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):

```
lm(formula = Weight ~ Age + GenderM, data = Q2ds)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 127.0837    33.7307   3.768 0.000254 ***
Age           0.5391     1.7178   0.314 0.754166
GenderM      32.8893     4.6776   7.031 1.21e-10 ***
```

(c) Briefly explain what the estimated coefficient, 32.8893, for the indicator variable `GenderM` in Model 2 tells us.

3. Amalia, Amber, and Ian spent their time investigating if there is a difference in happiness between Humanities majors and Science/Math majors. From their preliminary analysis (shown below), it seems that the proportion of Science/Math majors who feel happy or very happy is greater than that of Humanities majors in the sample. They would like to test if the Science/Math majors are happier than the Humanities majors in the population.

	Humanities	Science/Math
Happy or Vary Happy	19	34
Okay or Sad	16	13

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

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

4. Three committee members, Allison, Alizeh, and Carlos would like to find out how many more hours per week, on average, students would use for studying than exercising. Two inference procedures are performed and the corresponding R outputs (partial) are given below:

   Output I: Two-sample t-interval

```
data:  StudyingHrPerWeek and ExerciseHrPerWeek
90 percent confidence interval:   9.796585 13.591012
sample estimates:   mean of StudyingHrPerWeek    mean of ExerciseHrPerWeek
                              18.147287                     6.453488
```

   Output II: Paired t-interval

```
data:  difference
90 percent confidence interval:   9.807696 13.579901
sample estimates:   mean of difference
                         11.6938
```

   (a) Indicate which inference procedure you would use to answer this question. Explain.

   (b) Assuming all assumptions are satisfied (and you don't need to check them), carefully interpret the interval of the inference procedure you chose above.

   (c) Based on the interval you chose above, would you say that Amherst students on average spend 10 more hours per week on studying than exercising? Explain.

   (d) What is the significance level (α-level) associated with your conclusion in (c)?

5. Many committee members are very interested in exploring the so-called Athlete-Nonathlete divide at Amherst. Among them, Maggie, Mia, and Emma would like to know if the average number of nights per week that students attend parties varies among 3 different athlete groups: Varsity Athlete, Club Athlete, and Non-Athlete. They performed ANOVA, and partial output from R is shown below:

```
                  Df   Sum Sq   Mean Sq   F value    Pr(>F)
VarsityAthlete   (1)    10.75     (2)       (4)     0.00233 **
Residuals        126   106.44     (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) What assumptions are necessary for ANOVA?

(d) Assuming that the assumptions are all met, what is the corresponding P-value? State your conclusion at $\alpha = 0.1$.

(e) Since there are some concerns about certain required assumptions for ANOVA, a few committee members suggest to run a non-parametric test for further verification of the above conclusion. What would be a nonparametric alternative to one-way ANOVA?

(f) Based on the descriptive statistics below, the largest difference appears to exist between Varsity Athlete and Non-Athlete. But how big is this difference? Assuming all required assumptions are okay (and you don't need to check any of them) and the degrees of freedom of the model is $df = 94$, construct a 90% confidence interval for the difference in the average number of party nights per week between Varsity Athlete and Non-Athlete. Interpret your interval carefully.

	VarsityAthlete	min	Q1	median	Q3	max	mean	sd	n	missing
1	Non Athlete	0	0	1.0	1	4	1.004902	1.0105570	51	0
2	Club Athlete	0	1	1.5	2	3	1.469697	0.8285850	33	0
3	Varsity Athlete	0	1	2.0	2	3	1.655556	0.8714031	45	0

6. While also working on the Athlete-Nonathlete divide, another six committee members (Kelly, Sabrina, Uzoma, Philippe, Jack, and Jayde) are more interested in the association between students' athletic status and how healthy they consider their diet to be. The two-way table below summaries the results from the survey. **Observed count** (Expected count) is the setup.

Athletic Status Diet	GOOD or EXCELLENT	FAIR or POOR	Total
Non-Athlete	**21** ( )	**30** ( )	51
Club Athlete	**20** (18.16)	**13** (14.84)	33
Varsity Athlete	**30** (24.77)	**15** (20.23)	45
Total	71	58	129

(a) Do you think the Athletic Status is independent of how healthy students consider their diet to be? Please use ONLY the **observed** counts in the above table to answer this question. (DO NOT do any inference yet.)

(b) They decide to test an appropriate hypothesis for the association between these two variables. Name an appropriate inference procedure to perform (be specific) and state appropriate hypotheses.

Name of Analysis -

H_0:

H_A:

(c) Two expected counts for inference are missing in the table. Compute and fill in those missing expected counts in the parentheses.

(d) Check the assumptions and conditions.

(e) The test statistic works out to be 6.832 and the P-value is 0.033. Report the degrees of freedom of the test, and state your conclusion in context at $\alpha = 0.1$.

(f) The two cells with missing expected counts turn out to be the ones which contribute the most to the above test statistics. Compare their observed and expected counts carefully, and report their χ^2-components. What additional message do those cells try to tell us?

7. Jasmine, William, and Sam notice an interesting relationship between the amount of desserts Amherst students eat and their GPA. They run a simple linear regression on the data and get the following output from R:

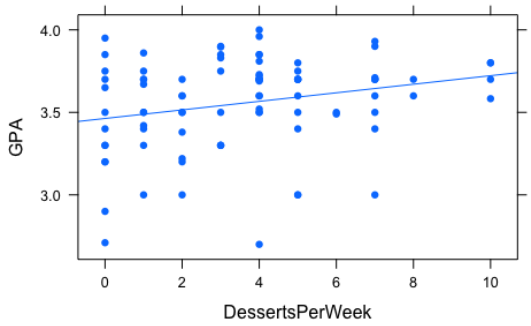
```
Call:  lm(formula = GPA ~ DessertsPerWeek, data = survey)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.46298    0.04891  70.810   <2e-16 ***
DessertsPerWeek  0.02593    0.01093   2.372   0.0199 *

Residual standard error: 0.2729 on 87 degrees of freedom
Multiple R-squared:  0.06073,Adjusted R-squared:  0.04993
F-statistic: 5.625 on 1 and 87 DF,  p-value: 0.01992
```

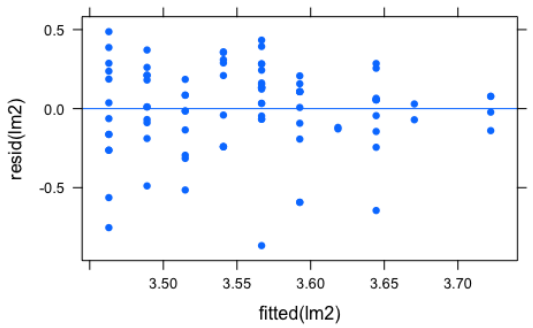
Three plots attached below are the scatterplot of the two variables of interest, the residuals vs. fitted plot, and the histogram of the residuals.

(a) What is the correlation?

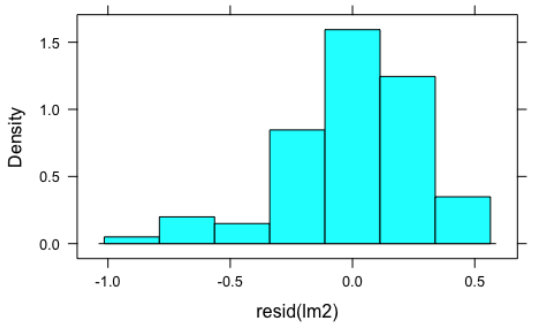
What is the equation of the regression line?



(b) For a student with GPA = 3.5 who usually eat desserts 5 times per week, calculate his/her residual.



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



(d) Is there an association between these two quantitative variables? Write appropriate hypotheses, check and explain if the assumptions for regression satisfied, provide the corresponding test statistic and P-value, 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.

(f) Do these results suggest that students can improve their GPA by consuming more desserts? Explain.

8. Three committee members, Anri, Daniella, and Elias, explore various social divides at the college. One of their research interest is to find out the population proportion of Amherst students who are currently in a relationship. Suppose that this proportion is 20%.

   (a) What is the probability that one out of four randomly selected Amherst students is in a relationship?

   (b) In a random sample of 129 students, what is the probability that more than 35 students are in a relationship?