Final Practice Problems

The following is a compilation of problems from past ST314 Final exams. These are examples of what could be asked on the exam but will not be a direct reflection of the exact problems or structure of the current exam.

Use for the following 5 questions. For each scenario indicate the matching test by filling in the corresponding letter in the ________. Each test can only be used once.

A. Two-sample t test	B. Single-factor ANOVA	C. Paired t test	D. One Sample t test	E. Simple linear Regression.

1. Matching test letter: __D__
   Jenny would like to know whether she has budgeted enough money per month for groceries while she is in college. She takes a representative random sample of 50 college students and asks them how much they spend a month on groceries per month. Jenny has budgeted $200. Which test can she use to test whether she has budgeted enough for groceries per month?

2. Matching test letter: __E__
   Research has found that there is a relationship between neck circumference and body fat percentage. A nationwide fitness franchise would like to be able to predict body fat percentage from neck circumference. Which type of analysis is appropriate for this scenario?

3. Matching test letter: __B__
   The lumen output is measure on 6 light bulbs each for 4 different brands of comparable bulbs. A contractor is interested in finding whether there is significant difference in the average lumen output for the different bulb brands. What type of procedure would you recommend?

4. Matching test letter: __C__
   Hexavalent chromium has been identified as an inhalation carcinogen in several different locations. Investigators would like to compare the concentration of Hexavalent chromium in nanograms/m3 for inside a home and outside a home to see if there is a mean difference other than zero. Which type of analysis is appropriate for this scenario?

5. Matching test letter: __A__
   A student would like test whether the average rent per room is higher for dwellings that are closer to campus versus further away. They take a random sample of 40 dwellings within 1 mile of their University campus, and another random sample of 40 dwellings more than 1 mile from campus. For each sample they record the cost of rent per bedroom. Which type of analysis is appropriate for this scenario?

6. If the p-value for a test is small, it means it is very unlikely the null hypothesis is false.

o True

o False

7. The probability density function for the continuous random variable for X is $f(x) = \frac{x}{6}$ when x is between 2 and 4. What is the cumulative density function?

a. $F(x) = \frac{x^2}{12}$ for x from 2 to 4

$F(x) = \int_2^x \frac{t}{6} dt = \frac{t^2}{12} \Big|_2^x = \frac{x^2}{12} - \frac{4}{12}$

b. $F(x) = \frac{x^2}{12} - \frac{4}{12}$ for x from 2 to 4

c. $F(x) = \frac{x^2}{12} - 2$ for x from 2 to 4

d. $F(x) = 1$ for x from 2 to 4

Use the following to answer the next two questions. "Blast-o-bike", a bicycle manufacturer, claims their bike will increase speed and hence reduce route times for cyclists. To test their claim, a simple random sample of 5 competitive cyclists were asked to record their times on the same route while first riding a standard racing bicycle and then riding the "Blast-o-bike". The population of cyclists is Normally distributed. The times are recorded below.

Cyclist	***1***	***2***	***3***	***4***	***5***
Standard bicycle route time in minutes	29.1	33.2	31.3	19.0	21.7
"Blast-o-bike" route time in minutes	28.3	31.8	27.6	15.4	19.9

Cyclists want to know. Does, on average, the "Blast-o-bike" make a difference on cyclist route times?

8. What type of test is appropriate to analyze this data?
   - o two-sample t test
   - o paired t test
   - o one sample t test
   - o pooled t test

9. Given the 95% confidence interval of the true population parameter is 0.6 minutes to 3.9 minutes. What statement is true regarding a two-sided hypothesis test for this example?
   - o The null hypothesis is rejected at a 0.01 significance level.
   - o The null hypothesis value falls within the 95% confidence interval
   - o There is at least moderately suggestive evidence in favor of the alternative.
   - o None of the above statements are true.

10. Consider comparing two dependent populations. Determine whether the following statements are true or false.
    a. Statement 1: We should perform a two sample t test.
       o True
       o False
    b. Statement 2: We should compute the differences of each paired observation.

- True
- False

c. Statement 3: The parameter of interest is μ_d.
- True
- False

d. Statement 4: The point estimate is difference than an independent population comparison.
- True
- False

e. Statement 5: The standard error is different than an independent population comparison.
- True
- False

11. When performing the two sample t test, we assume
    a. the population standard deviations are the same
    b. the population means are equal
    c. the two populations are dependent
    d. all the options are correct

Use the following for the next 3 problems. Is there a difference in the amount of airborne bacteria between carpeted and uncarpeted rooms? In an experiment, 7 rooms were carpeted and 7 were left uncarpeted. The rooms are similar in size and function. After a suitable period of time, the concentration of bacteria in the air was measured (in units of bacteria per cubic foot) in all of these rooms. Assume the populations are normally distributed. The following is a summary of the data:

Carpeted rooms: $\bar{x} = 184, s = 22$

Uncarpeted rooms: $\bar{x} = 175, s = 16.9$

12. A 95% confidence interval for the difference in mean bacterial concentration in the air of carpeted rooms versus uncarpeted rooms is (df = 11)
    a. -7.47 to 31.47
    b. -18.89 to 42.89
    c. -16.66 to 34.66
    d. -14.0$\cancel{2}_8$ to 32.0$\cancel{1}_8$

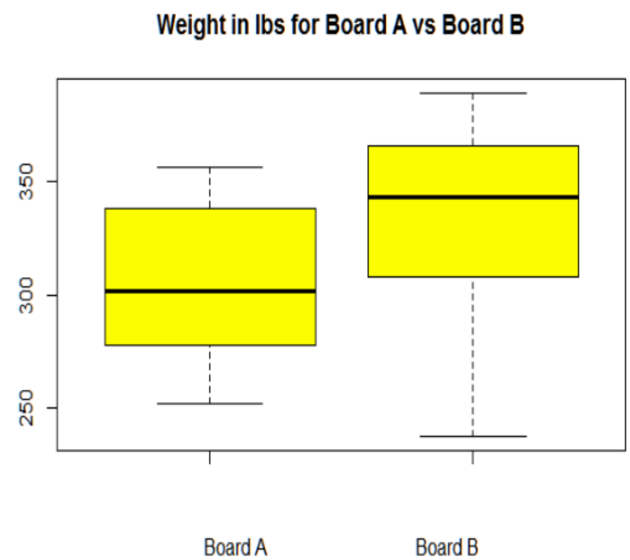
$(184-175) \pm 2.201\sqrt{\frac{22^2}{7}+\frac{16.9^2}{7}} = (-14.08, 32.08)$

13. The researcher wants to investigate whether carpets makes a difference (either increases or decreases) in the mean bacterial concentration in air. The numerical value of the two-sample t statistic for this test is
    a. 0.414
    b. 3.818
    c. 1.312
    d. 0.858

$\frac{184-175}{\sqrt{\frac{22^2}{7}+\frac{16.9^2}{7}}} = 0.858$

14. The p-value for t he test described above is approximately 0.42. What can be concluded from he hypothesis test?
    a. There is no evidence of a difference between the average amount of bacteria in the carpeted and uncarpeted rooms.
    b. Reject the null hypothesis at a significance level of 0.05
    c. The null hypothesis is true
    d. The alternative hypothesis is true.

Use the following to answer the next 3 questions. A lumber manufacturing company would like to compare the average weight capacity of two types of boards they currently manufacture. The sampled boards were obtained using a random mechanism.



Population	Sample mean	Sample Standard deviation	Sample Size
Board A	305.4 *lbs*	36.5	16
Board B	331.7 *lbs*	44.3	26

data: boardA and boardB
t = -2.0875, df = 36, p-value = 0.0439
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-51.8601685 -0.7621273
sample estimates:
mean of x mean of y
305.3603 331.6715

15. Describe the side-by-side box plot. Include a comparison of the two group, the center, shape, and spread of each.

Board B looks to have an overall higher weight, on average, in comparison to board A. The variability of B also looks to be greater for board B. Board A weights look ot be somewhat symmetric, yet board B weights have a longer negative tail.

16. State the null and alternative hypotheses to test whether there is a difference between the two boards.

$H_0: \mu_A - \mu_B = 0$

$H_A: \mu_A - \mu_B \neq 0$

17. Write a 4-part conclusion based on the software output for the hypothesis test.

There is moderately suggestive evidence the weights of board A and board b differ on average.
The null hypothesis is rejected based on a significance level of 0.05 (t = -2.0875, df = 36, p-value = 0.0439). The 95% confidence interval estimates the weights of board A will be 0.76 lbs to 51.9 lbs less than board B, on average.
The best guess for the difference in average weight between the boards is 26 lbs, with board A weighing in less than board B.

18. Suppose you have two single factor ANOVA experiments with the same degrees of freedom. The resulting *F* statistics are:

Experiment 1 *F* = 5.68
Experiment 2 *F* = 20.15

Which statement is true in regards to comparing Experiment 1 and Experiment 2?

a. Procedure 1 has a smaller test statistic and therefore will result in stronger evidence in favor of the alternative hypothesis.
b. Procedure 2 has a larger test statistic and therefore will result in strong evidence in favor of the alternative hypothesis.
c. We should reject the null for both tests.
d. Impossible to know with this information.

Use the following for the next two questions. In an experiment to study automobile engine operating efficiency for five different brands of gasoline, mpg was measured over a controlled distance and speed for eight cars in each group.

19. How many observations are in this experiment?

a. 8
b. 5
c. 39
d. 40

SSG
20. If [redacted] = 90 and MSE = 12 what is the value for the overall F statistic for a single factor ANOVA analysis?

a. 7.500
b. 1.875
c. 1.500
d. 65.620

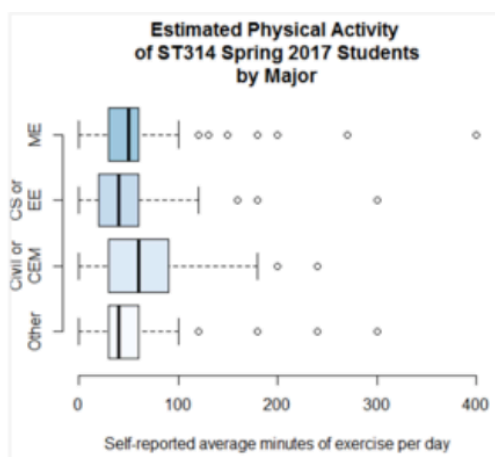
$$F = \frac{MSG}{MSE} = \frac{\frac{SSG}{I-1}}{MSE} = \frac{\frac{90}{4}}{12} = 1.875$$

21. In a single-factor ANOVA, the __________ is a measure of the average between samples variation and is denoted by _________.

a. Mean squared error, MSE
b. Mean squared treatment, [redacted] MSG
c. Sum of squared error, SSE
d. Sum of squared treatment, [redacted] SSG

Use the following for the next 5 questions. The following data is from the ST314 Student Information Survey from the Spring of 2017. Use the output to answer the following questions.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Major_new	3	30339	10113	4.258	0.00561
Residuals	401	952309	2375		



22. Based on the side-by-side box plots, which statement is a true description of the data?

a. There are no observable outliers.
b. There are more Civil/CEM engineering majors than other majors.
c. For each major, the activity variable is positively skewed.
d. The median time spent exercising is the same for each major.

23. From the single-factor ANOVA table from R, which value represents the average between group variation?

a. 10113
b. 2375
c. 4.258
d. 0.00561

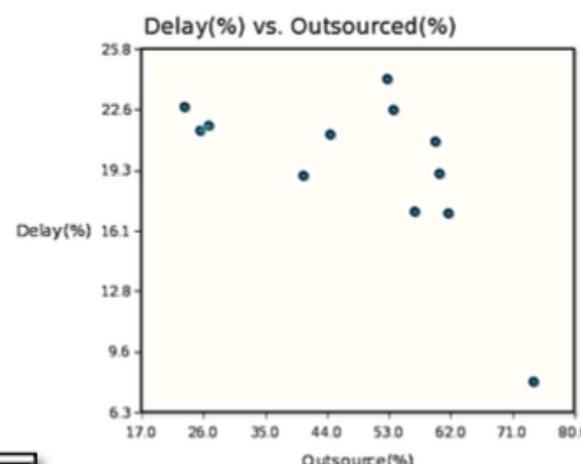
24. What is the null hypothesis that is assumed to be true when performing this ANOVA F test?

a. $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$
b. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
c. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$
d. $H_0 : \bar{x}_1 = \bar{x}_2 = \bar{x}_3 = \bar{x}_4$

25. Why should we be concerned about making inference to a larger population based on this data?

a. The data could be biased based on students self-reporting daily activity.
b. The sample is not random.
c. The sampled distributions are skewed; it is probable the population distributions are non-normal.
d. All of the options are correct.

Use the following for the next 7 questions. Airlines have increasingly outsourced the maintenance of their planes to other companies. Flight delays are often due to maintenance problems. Critics are concerned that the maintenance may be done less carefully, such that outsourcing creates safety hazards and delays.The following is a simple linear regression analysis on data from 2005 and 2006 on the percent of outsourcing for 12 airlines and their respective percent of delayed flights. Does the data support the concerns of the critics?



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.48	3.296	8.338	<0.0001
Outsource	-0.1636	0.06488	-2.522	0.03030

27. Assuming the conditions are met, what is the least squares regression equation for estimated percent of flights delayed given the percent of maintenance outsourced by the airline.

   a. $y = 27.48 - 0.1636x$
   b. $x = 0.1636 - 27.48y$
   c. $\hat{y} = 0.1636 - 27.48x$
   d. $\hat{y} = 27.48 - 0.1636x$

28. Use the regression equation. Given 50% of the maintenance has been outsourced by an airline, the estimated average percent of flights delayed is approximately:

   a. 19.3%
   b. 20.5%
   c. 1374%
   d. -22.6%

$\hat{y} = 27.48 - 0.1636(50) = 19.3$

29. From the regression analysis, which statement is true about the relationship between the percent of outsourced maintenance and percent of flight delays?

   a. The critics were right. Outsourcing causes delays.
   b. The LSRL estimates that, on average, for every 1% increase in outsources maintenance, flight delays increase by 27.48%.
   c. There is a very strong correlation between percent of outsources maintenance from airlines and percent of flight delays.
   d. The LSRL estimates that for every 1% increase in outsources maintenance the average percent of flight delays decrease by 0.1636.

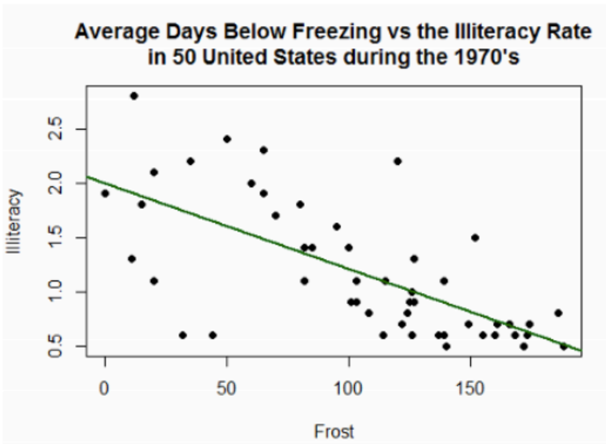
$-0.1636 \pm \overset{df=10}{2.228}(0.06488) = (-0.308, -0.019)$

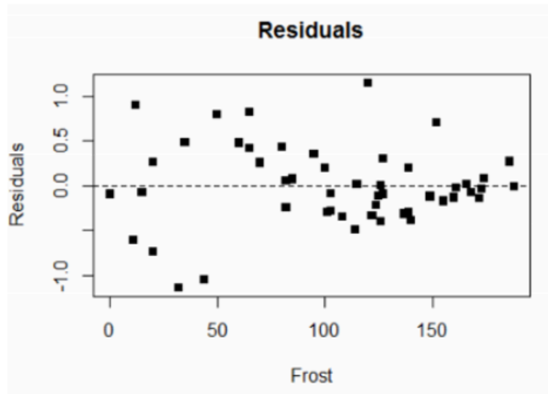
30. Calculate the 95% confidence interval for the slope.
31. Calculate the residual for the continental airlines that outsourced 44.5% and had 21.23% of their flights delayed. $\hat{y} = 27.48 - 0.1636(44.5) = 20.1998$ residual $= y - \hat{y} = 21.23 - 20.1998 = 1.03$
32. Which of the following statements is false about the correlation coefficient, *r*?

   a. The correlation coefficient is a unitless number and must always lie between 0 and 1, inclusive.
   b. The correlation coefficient can only describe the relationship between two quantitative variables.
   c. If r =1, then there is a perfect positive association between x and y.
   d. The correlation coefficient is a unitless number and must always lie between –1 and 1, inclusive.

The US Census Bureau of the 1970's collected data from the 50 United States. Two of the variables were the percent of Illiteracy (inability to read or write) and average number of days below freezing (Frost) in the states most populous city. Consider the data to be randomly obtained.

Use the R output to answer the following questions.






```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.993074   0.145963  13.655  < 2e-16 ***
Frost       -0.007879   0.001253  -6.286 9.16e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4561 on 48 degrees of freedom
Multiple R-squared:  0.4515,    Adjusted R-squared:  0.4401
F-statistic: 39.51 on 1 and 48 DF,  p-value: 9.156e-08
```

33. Describe from the scatterplot the relationship between the two variables, include strength, direction, form, outliers, and context.

There is a moderately strong, negative, linear relationship between average number of days below freezing and illiteracy for the 50 states from the 1970's.

34. State the null and alternative hypothesis for testing whether the explanatory variable is a significant predictor of the response. $H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

35. What is the estimated least squares regression equation? $\hat{y} = 1.993 - 0.008x$

36. Is there evidence the number of freezing days is a significant predictor of illiteracy rate? Given a significance level of 0.05, make a conclusion based on the test results and a statement in terms of the alternative hypothesis. Provide the test statistic, degrees of freedom and p-value associated with the test.

There is convincing evidence the number of frost days is a significant predictor of illiteracy rate. The null hypothesis is rejected at the 0.05 significance level.

37. Calculate and interpret the 95% confidence interval for the slope, β_1.

$-0.0079 \pm 2.011\,(0.0013) = (-0.011, -0.005)$

38. Using the LSRL, calculate the residual for Oregon, which has an observed number of Frost Days of 44 and an Illiteracy rate of 0.6.

$\hat{y} = 1.993 - 0.008\,(44) = 1.641$

residual $= y - \hat{y} = 0.6 - 1.641 = -1.041$

Each additional day of frost on average will decrease the average percent illiteracy rate by 0.011 to 0.005%, with a best guess at a 0.0079% decrease.