Name: Feedback

SM339 · Applied Statistics Spring 2023 · Uhan

Exam 1 – Part 1 – 2/15/2023

Instructions

- This part is worth 60 points total. The exam (both parts) is worth 100 points total.
- You have until the end of the class period to complete this exam.
- You may use your plebe-issue TI-36X Pro calculator.
- You may refer to notes that you have handwritten, not to exceed one side of an 8.5" × 11" piece of paper.
- You may not use any other materials.
- **No collaboration allowed.** All work must be your own.
- **Show all your work.** To receive full credit, your solutions must be completely correct, sufficiently justified, and easy to follow.
- Keep this booklet intact.
- **Do not discuss the contents of this exam with any midshipmen until it is returned to you.**
- Copy and sign the honor statement below. This exam will not be graded without a signed honor statement.

> The Naval Service I am a part of is bound by honor and integrity. I will not compromise our values by giving or receiving unauthorized help on this exam.

Problem	Weight	Score
1a	0.4	
1b	0.4	
2a	0.4	
2b	0.4	
2c	0.4	
2d	0.4	
2e	0.4	
3a	0.4	
3b	0.4	
3c	0.4	
3d	0.4	
4a	0.4	
4b	0.4	
4c	0.4	
4d	0.4	
Total		/ 60

Name:

Problem 1. You are studying customer behavior on YouView, a new streaming video service. You want to estimate the average length of a video on YouView. You take a random sample of 17 videos and find that the sample mean video length is 12.74 minutes with a sample variance of 24.1.

a. Construct a 95% confidence interval for the mean length of videos on YouView. Provide your answer rounded to three decimal places.

You may find some of the following R output helpful:

Code	Output
`qnorm(1 - 0.05/2, mean = 0, sd = 1)`	1.959964
`qnorm(1 - 0.95/2, mean = 0, sd = 1)`	0.062706
`qt(1 - 0.05/2, df = 16)`	~~0.063697~~ 2.119905
`qt(1 - 0.95/2, df = 16)`	~~2.119905~~ 0.063697

See the example in Lesson 3 Part 2 and Problems 1 and 2 in the Lesson 3 Part 2 Exercises for similar problems.

- Note that $t_{\alpha/2,n-1}$ is given by `qt(1 - alpha / 2, df = n - 1)` in R.
- Note that the sample variance is given, and that the population variance is not given. See Lesson 3 Part 1 for details on the difference between computing CI for population mean when the population variance is known versus not known.
- A confidence interval is just like any other interval: it should be written in the form (lower value, upper value).

b. You write in your report that you are "95% confident" that the interval you found in part a contains the true mean video length. Briefly explain what this means.

See page 3 of Lesson 3 Part 1 for an explanation of what it means to be "95% confident."

Problem 2. The Blackwell Brewery has recently hired you as an analyst. In the past, the brewery produced an average of 725 thousand barrels of beer per day. You've been asked to test whether the average daily production has dropped significantly over the past year.

Using the records for the past year, based on 260 operating days, the sample mean of daily production is 712 thousand barrels per day, and the sample standard deviation is 118 thousand barrels per day.

Let μ be the average number of barrels of beer produced per day, in thousands.

Perform a hypothesis test by answering the following prompts.

For parts a-e: note that this problem is about performing a t-test for population mean, not simple linear regression slope.

a. State the null and alternative hypotheses.

Note that you are testing whether average daily production has dropped significantly over the past year. See Problem 3 in the Lesson 4 Part 2 Exercises for a similar problem.

b. Calculate the test statistic. Provide your answer to three decimal places.

Most of you had the right idea here. See Lesson 4 Part 1 for details on how to compute the test statistic.

c. Suppose the p-value is 0.038. Using a significance level of 0.05, do you reject or fail to reject the null hypothesis? Briefly explain.

Most of you had the right idea here. See Lesson 4 Part 1 for details on how to use the p-value. See Problem 3 in the Lesson 4 Part 2 Exercises as well.

d. Based on your answer to part c, state your conclusion about the average number of barrels of beer produced per day.

Be careful with your language here. The conclusion of a hypothesis test like the t-test for population mean is never certain. You can say things like, "we see significant evidence that average daily production has dropped" or "we have strong evidence that average daily production has dropped", but avoid things like "we conclude that the average daily production has definitely dropped".

e. Suppose we fail to reject the null hypothesis when in fact it is false. What type of error have we committed?

Almost all of you had the right idea here. See Lesson 4 Part 1 for details on Type I and Type II errors.

Name:

Problem 3. Back to studying YouView. Suppose you are interested in predicting the number of *Likes* (in 1000s) a video on YouView has based on its *Length* (in minutes). You have collected this data for 43 videos. With this data, you fit a simple linear regression model. You obtain the following least squares line:

$$\widehat{Likes} = 17.626 + 1.24Length$$

a. Interpret the slope in the context of the problem.

Be careful with your language here. Remember that the fitted model, or least squares line, represents how the expected value or average value of the response variable is related to the explanatory variable. Your explanation should include the words "average", "expected", or "expectation". See Example 3b in Lesson 6, and STAT2 Exercises 1.19c, 1.45c assigned for homework for similar problems.

b. Predict the number of likes for a video that is 5 minutes long.

Most of you had the right idea here. Be careful with your units. See STAT2 Exercise 1.21a assigned for homework for a similar problem.

c. Calculate the residual for a video that is 5 minutes long and has 20,194 likes.

Most of you had the right idea here. See page 2 of Lesson 6 for the definition of residual. In addition, see STAT2 Exercise 1.21b assigned for homework for a similar problem.

d. Suppose the model results in a sum of squared error of 3,456. Calculate the size of a "typical" error.

Many of you were on the right track here. Note that the estimated standard deviation of the error term $\hat{\sigma}_\varepsilon$ is interpreted as the size of a "typical error", not the estimated variance $\hat{\sigma}_\varepsilon^2$. See Example 1 in Lesson 7 and STAT2 Exercise 1.9 assigned for homework for similar problems.

Problem 4. Back to Blackwell Brewery. You have been asked to evaluate your competition. You have collected data on the *ABV* (alcohol by volume) and *Rating* from the Beer Advocate website for 44 beers. Using R, you fit a simple linear regression model with *Rating* as the response variable and *ABV* as the explanatory variable. The output is below. Note that some parts are missing.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  85.4707     1.6279   52.50   <2e-16 ***
ABV           0.2937     0.2623    1.12    0.269
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.72 on 42 degrees of freedom
Multiple R-squared:  ???????, Adjusted R-squared:  ????????
F-statistic: 1.254 on 1 and 42 DF,  p-value: 0.2692
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
ABV	1	17.34882	17.34882	1.253707	0.2692081
Residuals	42	581.19664	13.83802	NA	NA

a. Write the fitted model.

Be careful with notation! Recall that for simple linear regression, $Y = \beta_0 + \beta_1 X + \varepsilon$ is the population-level model, and $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ is the fitted model or least squares line. See Example 3a in Lesson 6, and STAT2 Exercises 1.19b, 1.27d, 1.29c, 1.33b, 1.45c assigned for homework for similar problems.

b. Suppose you perform a t-test for simple linear regression slope. Using a significance level of 0.05, do you reject or fail to reject the null hypothesis? Briefly explain. Circle the R output you used to make your decision.

Be careful with where to find the p-value for the t-test for simple linear regression slope: it is in the highlighted row above. See Example 1 in Lesson 10 and STAT2 Exercise 2.15 assigned for homework for similar problems.

c. Based on your decision in part b, state your conclusion about the relationship between *Rating* and *ABV*.

Just like with Problem 2d, be careful with your language here. The conclusion of a hypothesis test like the t-test for simple linear regression slope is never certain, and so your language should reflect this. See Example 1 in Lesson 10 for some examples of how to write about the relationship between two variables based on this hypothesis test.

d. What percent of the variability in *Rating* is explained by the model? Provide your answer rounded to three decimal places.

See Example 1 in Lesson 12 and STAT2 Exercise 2.21 assigned for homework for similar problems.