Stat 21 Homework 2

Solutions

Due: Saturday, Feb 5th by midnight

Contents

Part I: Non-R Problems **1**
Problem 1 1
Problem 2 2
Problem 3 2
Problem 4 2

Part II: R coding problems **3**
Problem 5 3
Problem 6 3
Problem 7 4
Problem 8 4
Problem 9 5
Problem 10 5

Use this file as the template for your submission. Do not delete anything from this template unless you are prompted to do so (e.g. where to write your name above, where to write your solutions or code below). Make sure you have installed the following packages in your version of RStudio: `tidyverse`, `knitr` **before** you attempt to knit this document.

Your completed assignment should be submitted as a single **PDF** using the link under Week 3 titled "Submit HW 2 to Gradescope". You must use R markdown to write up your solutions. For any homework problems that involve coding in R, you must provide **both** the code and a written answer interpreting the output within the context of the problem. You are allowed to work with your classmates on this homework assignment but you are expected to write up your own solutions. Every answer must be supported by a written statement unless otherwise specified. *A good rule of thumb is to make sure your answer is understanable to someone who hasn't read the problem question (or code output associated with it).*

Part I: Non-R Problems

Problem 1

Recall the census data problem from HW 1. Data for a certain country shows that 19% of the adult residents are Latinx and 72 people are called for jury duty. Out of these 72, only 9 people are Latinx and we want to know if this apparent under-representation of Latinx jurors calls into question the fairness of the jury selection system. For this problem setting answer the following questions.

(a) What is the population under study and what is the population parameter we are interested in?

(b) What is the sample estimate for the parameter?

(c) State a null and alternative hypothesis test we can conduct to answer the question posed above.

(d) We could answer this question with a confidence interval. Suppose you calculate a lower bound (LB) and upper bound (UB) for the parameter. Referencing the interval, [LB, UB], how would you determine if there is statistical evidence of under-representation of Latinx jurors?

Solution Problem 1:

(a) p = proportion of juorors who are Latinx, population is all potential jury members in this county

(b) $\hat{p} = 9/72$

(c) $H_0 : p = 0.19$ vs $H_A : p < 0.19$ (or could do two sided)

(d) If the value 0.19 is contained in the interval then there is not statistical evidence of under-representation

Problem 2

In a survey of 988 men aged 18˘24, the regression equation for predicting height from weight was:

$$\hat{height} = 62.4 + (0.047)(weight),$$

where height is measured in inches and weight is measured in lbs. Is the following statement correct: "If someone gains 10 pounds, he will get taller by (0.047)(10) = 0.47 inches"? If not, provide a better explanation for the meaning of the slope in the regression equation.

Solution Problem 2: "will get taller" is incorrect. not an exact statement, is an approximation that will hold only "on average"

Problem 3

Student's investigating the packaging of potato chips purchased 6 bags of Lay's Ruffles marked with a net weight of 28.3 grams. They carefully weighed the contents of each bag, recording the following weights (in grams): 29.3, 28.2, 29.1, 28.7, 28.9, 28.5

(a) State the null and alternative hypotheses to test whether or not the net weight of these chips is different from 28.3 g, defining the parameter in terms of the population.

(b) Explain the conditions needed to appropriately conduct this test.

Solution Problem 3:

(a) $H_0 : \mu = 28.3, \quad H_A : \mu \neq 28.3$ where μ is the average of all of the Lay's Ruffles bags of this size

(b) because small n not using CLT; need to assume distribution of weight of bags in population is symmetric and unimodal and that the sample is random

Problem 4

Having done poorly on their Biology final exams in June, six students repeat the course in summer school and take another exam in August. If we consider these students to be representative of all students who might attend this summer school in other years, what procedure can we use to determine if these results provide evidence that the program is worthwhile?

Person: Aaron, Brittney, Chloe, Drake, Edward, Frankie

June scores: 54, 49, 68, 66, 62, 62

August scores: 50, 65, 74, 64, 68, 72

Solution Problem 4: paired two sample t-test or CI for difference in paired means

Part II: R coding problems

Problem 5

Consider the regression equation in Problem 2. Suppose the variance of our model error is $\sigma^2 = 2$. Use the R code chunk below and the `pnorm` function to calculate the percentage of all 200-pound men that we expect to be taller than 74 inches? (You can assume the regression model assumptions are met.)

```
pnorm(74, mean=71.8, sd = sqrt(2), lower.tail=FALSE)
```

```
## [1] 0.05989747
```

Solution Problem 5: response (height) is normally distributed with mean $E(\text{height}) = 62.4 + (0.047 \times 200) + E(\epsilon) = 62.4 + (0.047 \times 200) = 71.8$ and $Var(\text{height}) = Var(\epsilon) = 2$.

Problem 6

Do the following data suggest that there is a significant difference in calories between servings of strawberry and vanilla yogurt? Following the code from this week's reading assignment, import the data and then test an appropriate hypothesis. State your conclusion within the context of the problem.

Brand	Strawberry (cal/serving)	Vanilla (cal/serving)
America's Choice	210	200
Breyer's Lowfat	220	220
Columbo	220	180
Dannon Light 'n Fit	120	120
Dannon Lowfat	210	230
Dannon la Creme	140	140
Great Value	180	80
La Yogurt	170	160
Mountain High	200	170
Stonyfield Farm	100	120
Yoplait Custard	190	190
Yoplait Light	100	100

Solution Problem 6: paired two sample test for difference in means; not enough evidence to reject the null of no difference in calories between strawberry and vanilla

```
yogurt <- tibble(strawberry = c(210, 220, 220, 120, 210, 140, 180, 170, 200, 100, 190, 100),
                 vanilla = c(200, 220, 180, 120, 230, 140, 80, 160, 170, 120, 190, 100))
t.test(yogurt$strawberry - yogurt$vanilla, alternative= "two.sided")
```

```
##
##  One Sample t-test
##
## data:  yogurt$strawberry - yogurt$vanilla
## t = 1.332, df = 11, p-value = 0.2098
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -8.155068 33.155068
## sample estimates:
## mean of x
##      12.5
```

Problem 7

A survey of 430 randomly chosen college students found that 21% of the 222 full-time students and 18% of the 208 part-time students had purchased books in person from their campus book store. Is there statistical evidence that full-time students are more likely than part-time students to purchase their books in person (rather than online)? Test an appropriate hypothesis and state your conclusion in the context of the problem.

Solution Problem 7: data do not suggest that full time students are more likely than part time to purchase in person

```r
prop.test(x=c(222*.21, 208*.18), n=c(222, 208), alternative = "greater")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(222 * 0.21, 208 * 0.18) out of c(222, 208)
## X-squared = 0.43857, df = 1, p-value = 0.2539
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.03743938  1.00000000
## sample estimates:
## prop 1 prop 2
##   0.21   0.18
```

Problem 8

Researchers comparing the effectiveness of two pain medications randomly selected a group of patients who had been complaining of a certain kind of joint pain. They randomly divided these people into two groups, then administered the pain killers. Of the 112 people in the group who received medication A, 84 said this pain reliever was effective. Of the 108 people in the other group, 66 reported that pain reliever B was effective.

(a) First, find a 95% CI for the percent of people who may get relief from their joint pain my using medication A. Next, find a 95% CI for the percent of people who may get relief from their joint pain my using medication B. Do these two CIs overlap? What do you think this means about the comparative effectiveness of these medications?

(b) Find and interpret a 95% CI for the difference in the proportions of people who may find these medications effective. Explain what it means if you interval contains or does not contain zero.

Solution Problem 8:

```r
prop.test(x=84, n=112, conf.level=0.95)$conf.int
```

```
## [1] 0.6576601 0.8248566
## attr(,"conf.level")
## [1] 0.95
```

```r
prop.test(x=66, n=108, conf.level=0.95)$conf.int
```

```
## [1] 0.5122108 0.7020013
## attr(,"conf.level")
## [1] 0.95
```

```r
prop.test(x=c(84,66), n=c(112,108), conf.level=0.95)$conf.int
```

```
## [1] 0.007794337 0.269983440
## attr(,"conf.level")
## [1] 0.95
```

(a) yes there is overlap, the CI for medication B has a lower lower bound than the CI for medication A but the CI for medication A has a higher upper bound than the upper bound for the CU for medication B. because of this overlap, there is a region (the interval from 0.6576 to 0.702) where there is no discernible difference in the performance of either medication.

(b) the interval for the difference does not contain zero, implying that there is a difference in the comparative effectiveness of the medications (in apparent opposition to answer in part (a)), namely that medication A is more effective. this is the better answer of the two (a or b) because it the standard error of the difference in sample proportions is smaller than adding the standard errors of the individual sample proportions together (as in (a)). (Mathematically, this boils down to: $\sqrt{a} + \sqrt{b} > \sqrt{(a+b)}$)

Problem 9

A nutrition laboratory tests 40 reduced sodium hot dogs, finding that the mean sodium content is 310 mg, with a standard deviation of 36 mg.

(a) Find a 95% CI for the mean sodium content of this brand of hot dog and explain what your internal means.

(b) What assumptions were necessary to find this CI? Are these assumptions reasonable in this situation?

```r
n = 40
xbar = 310
s = 36
LB = xbar - (qt(0.05/2, df=39, lower.tail=FALSE) * (36/ sqrt(n)))
UB = xbar + (qt(0.05/2, df=39, lower.tail=FALSE) * (36/ sqrt(n)))
LB; UB
```

```
## [1] 298.4866

## [1] 321.5134
```

Solution Problem 9:

(a) mean of all hot dogs produced by this company is between [298.4866, 321.5134] mg

(b) sample size is large enough to use CLT or use finite sample method with t-distribution (as I did in calculation above). the only other assumption needed is that the sample of 40 is random

Problem 10

When a professional statistician has information to share with colleagues, they will submit an article to one of several Statistics journals for publication. This can be a lengthy process; typically the article must be circulated for "peer review" and then edited before being accepted for publication. There the article must wait in line with other articles before appearing in print. In the Winter 1998 issue of Change magazine, Eric Bradlow and Howard Wainer reported on this delay for several journals between 1990 and 1994.

For 288 articles published in the journal *The American Statistician*, the mean length of time between initial submission and publication was 21 months with a standard deviation of 12 months. For 209 articles published in the journal *Applied Statistics*, the mean time to publication wa s 31 months with a standard deviation of 12 months.

(a) Use the R code chunk below to calculate and interpret a 90% confidence interval for the difference in mean delay between the two journals by hand. The critical value, t^*, is 1.65.

(b) What are the assumptions needed for part (a)?

(c) State the null and alternative hypotheses for a two-sided hypothesis test that there is a difference in the publication time between the journals. Without doing any calculations in R or by hand, what are the results of this hypothesis test (at an $\alpha = 0.10$ significance level)?

Solution Problem 8:

```
n1 = 288
xbar1 = 21
sd1 = 12
n2 = 209
xbar2 = 31
sd2 = 12
t_star = 1.65
LB = xbar1 - xbar2 - (t_star * sqrt( (sd1^2)/n1 + (sd2^2)/n2  ))
UB = xbar1 - xbar2 + (t_star * sqrt( (sd1^2)/n1 + (sd2^2)/n2  ))
LB; UB
```

```
## [1] -11.79918
```

```
## [1] -8.200823
```

(a) estimated difference of [-11.799, -8.2008] months, so American Statistician publishes on average this much quicker than Applied Stats journal

(b) need to assume the articles submitted to each journal should be independent of one another (independent populations assumption); samples sizes are large enough to justify using the CLT

(c) Because the interval lies entirely below zero, a two sided hypothesis at the $\alpha = 0.1$ confidence level would reject the null in favor of the alternative