Problem Set 2

Applied Stats/Quant Methods 1

Due: October 15, 2021

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub in `.pdf` form.
- This problem set is due before class on Friday October 15, 2021. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (40 points): Political Science

The following table was created using the data from a study run in a major Latin American city.[1] As part of the experimental treatment in the study, one employee of the research team was chosen to make illegal left turns across traffic to draw the attention of the police officers on shift. Two employee drivers were upper class, two were lower class drivers, and the identity of the driver was randomly assigned per encounter. The researchers were interested in whether officers were more or less likely to solicit a bribe from drivers depending on their class (officers use phrases like, "We can solve this the easy way" to draw a bribe). The table below shows the resulting data.

[1] Fried, Lagunes, and Venkataramani (2010). "Corruption and Inequality at the Crossroad: A Multimethod Study of Bribery and Discrimination in Latin America. *Latin American Research Review.* 45 (1): 76-97.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	14	6	7
Lower class	7	7	1

(a) Calculate the χ^2 test statistic by hand (even better if you can do "by hand" in R).

Steps taken as follows...

1) Take 'not stopped' cells from upper and lower class rows to caluclate t stat

2) Observed value for upper class = 14. Observed value for lower class = 7

3) Calculate expected value for upper class first using function: 27/42*21 = fe 13.5

4) Calculate expected value for lower class 'not stopped' cell: 15/42*21 = fe 7.5

5) Use function to calculate t stat: 14 - 13.5/13.5 squared + 7 - 7.5/7.5 squared = 1.37 = 4.44 = 5.81

ANS = 5.81

(b) Now calculate the p-value from the test statistic you just created (in R).[2] What do you conclude if $\alpha = .1$?

PART 2: Calculate p-value from t-stat. Interpret it. Use pt() function

- 2*pt(5.81, df = 2, lower.tail = FALSE)

ANS = 0.03

[2]Remember frequency should be > 5 for all cells, but let's calculate the p-value here anyway.

(c) Calculate the standardized residuals for each cell and put them in the table below.

Calculate standardized residuals for each cell and place in table.

Use formula for each cell to get standardized residuals.

- ANS for Upper Class = O.76, -3.32, 2.82 - ANS for Lower Class = -0.57, 2.46, 2.11

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.76	-3.32	2.82
Lower class	0.57	2.46	2.11

(d) How might the standardized residuals help you interpret the results?

Standardised residuals differentiate expected and oberved values. Expected values come from the null hypothesis.

They are important to interpretation of data by showing how important or unimportant each cell is to the overall result (i.e. how much a cell impacts a statistic in a dataset).

Question 2 (20 points): Economics

Chattopadhyay and Duflo were interested in whether women promote different policies than men.[3] Answering this question with observational data is pretty difficult due to potential confounding problems (e.g. the districts that choose female politicians are likely to systematically differ in other aspects too). Hence, they exploit a randomized policy experiment in India, where since the mid-1990s, $\frac{1}{3}$ of village council heads have been randomly reserved for women. A subset of the data from West Bengal can be found at the following link: `https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv`

Each observation in the data set represents a village and there are two villages associated with one GP (i.e. a level of government is called "GP"). Figure 1 below shows the names and descriptions of the variables in the dataset. The authors hypothesize that female politicians are more likely to support policies female voters want. Researchers found that more women complain about the quality of drinking water than men. You need to estimate the effect of the reservation policy on the number of new or repaired drinking water facilities in the villages.

Figure 1: Names and description of variables from Chattopadhyay and Duflo (2004).

[3]Chattopadhyay and Duflo. (2004). "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica*. 72 (5), 1409-1443.

(a) State a null and alternative (two-tailed) hypothesis. Read the csv file and name it economicdata economicdata ¡- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PF economicdata

PART 1: State Null and Alternative Hypothesis.

Interested only in reserved and water variables as we mean to test the effect of the reservation policy on new water facilities.

Subset data to just reserved and water variables - reserved ¡- economicdata[, 3]

- reserved

- water ¡- economicdata[, 6]

- water

Find avg no of new water services to make data assumption and form null and alt hypothesis - mean(water) = 17.84

- sd(water) = SD is 33.68, quite high.

Null Hypothesis = The number of new water facilities will be less than 17.84 in regions with the reservation policy

Alt Hypothesis = The no of new water facilities will be greater than 17.84 in regions where the reservation policy is introduced.

(b) Run a bivariate regression to test this hypothesis in `R` (include your code!).

Code is as follows to run regression and find coefficient estimate

- reservedwater ¡- lm(water reserved) - reservedwater = Intercept = 14.74, slope = 9.25 - summary(reservedwater)

(c) Interpret the coefficient estimate for reservation policy.
Find Coefficient Estimate. ANS = 14.74 (see above)
Less than 17.84. Cannot reject the null hyothesis

Question 3 (40 points): Biology

There is a physiological cost of reproduction for fruit flies, such that it reduces the lifespan of female fruit flies. Is there a similar cost to male fruit flies? This dataset contains observations from five groups of 25 male fruit flies. The experiment tests if increased reproduction reduces longevity for male fruit flies. The five groups are: males forced to live alone, males assigned to live with one or eight newly pregnant females (non-receptive females), and males assigned to live with one or eight virgin females (interested females). The name of the data set is `fruitfly.csv`.[4]

`No`	serial number (1-25) within each group of 25
`type`	Type of experimental assignment 1 = no females 2 = 1 newly pregnant female 3 = 8 newly pregnant females 4 = 1 virgin female 5 = 8 virgin females
`lifespan`	lifespan (days)
`thorax`	length of thorax (mm)
`sleep`	percentage of each day spent sleeping

1. Import the data set and obtain summary statistiscs and examine the distribution of the overall lifespan of the fruitflies.

   PART ONE: Obtain summary Stats on Lifespan

   Import fruitfly.csv

   - data ¡- read.csv("http://stat2.org/datasets/FruitFlies.csv")

   - data

   Conduct Summary Statistics on longevity incl. mean, sd, etc.

   - longevity ¡- data[, 4] - mean(longevity) - sd(longevity)

   Mean longevity = 57.44 days, sd = 17.56, sample size = 125

[4]Partridge and Farquhar (1981)."Sexual Activity and the Lifespan of Male Fruitflies". *Nature*. 294, 580-581.

2. Plot `lifespan` vs `thorax`. Does it look like there is a linear relationship? Provide the plot. What is the correlation coefficient between these two variables?

3. PART 2: Plot Lifespan and Thorax and Comment Upon Relation. Find Correlation Coefficient

   - lifespanthorax ¡- data[, 5, 4]

   - lifespanthorax

   - plot(lifespanthorax)

   - thorax ¡- data[, 5]

   Find correlation coefficient. ANS = 0.64 cor.test(longevity, thorax)

   Corrleation coefficient 0.64 close enough to 1 to have some evidence of positive relationship between thorax and lifespan.

   Regress `lifespan` on `thorax`. Interpret the slope of the fitted model.

4. PART 3: Use lm() function to regress longevity on thorax. Find slope.

   - longthorregression ¡- lm(longevity thorax)

   - summary(longthorregression)

Intercept = -61.05. Slope for line = 144.33. Slope greater than 0. Therefore, there is a positive relationship between x and y variables.

Test for a significant linear relationship between `lifespan` and `thorax`. Provide and interpret your results of your test.

PART 4: Run and interpet significance test on lifesoan and thorax

Use summary() function of longthorregression to find p.value

- summary(longthorregression)

P value less than 0.5. Relationship significant between lifespan and thorax.

5. Provide the 90% confidence interval for the slope of the fitted model.

   Find 90 Confidence Interval

   Use summary() function to gather info to get 90 confidence interval...

   BY HAND Slope = 144.33. Standard error of slope = 15.77. Find t stat to calculate by hand

   Use confint() function instead. Faster and more efficient.

   - confint(longthorregression, level = 0.90)

   ANS = 118.20 and 170.47

   - Use the formula of confidence interval.
   - Use the function `confint()` in `R` .

6. Use the `predict()` function in `R` to (1) predict an individual fruitfly's lifespan when `thorax`=0.8 and (2) the average `lifespan` of fruitflies when `thorax`=0.8 by the fitted model. This requires that you compute prediction and confidence intervals. What are the expected values of lifespan? What are the prediction and confidence intervals around the expected values?

   PART 6: Use predict function. 58th variable in thorax = 0.8mm. Subset thorax to 58th variable and use predict function

   - newthorax ¡- thorax[58]

   - class(longthorregression)

   - predict(longthorregression, newdata = newthorax)

7. For a sequence of `thorax` values, draw a plot with their fitted values for `lifespan`, as well as the prediction intervals and confidence intervals.