Problem Set 2

Applied Stats/Quant Methods 1

Due: October 15, 2023

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.
- This problem set is due before 23:59 on Sunday October 15, 2023. No late assignments will be accepted.

Question 1: 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/manually (even better if you can do "by hand" in R).

Code to calculate the χ^2 test statistic:

```
# I know the formula for the Chi^2 statistic (from Dr. Ziegler's slides)
# And I know that expected value = (rowtot)(columntot)/(grandtot)

# Just saving my f0 values:

f0_ns_u <- bribes[1,1]
f0_ns_l <- bribes[2,1]

f0_b_u <- bribes[1,2]
f0_b_l <- bribes[2,2]

f0_w_u <- bribes[1,3]
f0_w_l <- bribes[2,3]

# Calculating totals:

print(bribes)

colSums(bribes) # Col tots
rowSums(bribes) # Row tots

sum(colSums(bribes)) # Grand tot

# Calculating expected values:

fe_ns_u <- (21*27)/42
fe_ns_l <- (21*15)/42

fe_b_u <- (13*27)/42
fe_b_l <- (13*15)/42

fe_w_u <- (8*27)/42
fe_w_l <- (8*15)/42
```

```r
# Calculating squared difference over expected:

dif_ns_u <- ((f0_ns_u - fe_ns_u)^2)/fe_ns_u
dif_ns_l <- ((f0_ns_l - fe_ns_l)^2)/fe_ns_l

dif_b_u <- ((f0_b_u - fe_b_u)^2)/fe_b_u
dif_b_l <- ((f0_b_l - fe_b_l)^2)/fe_b_l

dif_w_u <- ((f0_w_u - fe_w_u)^2)/fe_w_u
dif_w_l <- ((f0_w_l - fe_w_l)^2)/fe_w_l

# Sum of all:

statistic <- dif_ns_u + dif_ns_l + dif_b_u + dif_b_l + dif_w_u + dif_w_l
   # I get 3.79

# Double checking test (I got the chisq() function from Hannah's code):

chi_test <- chisq.test(bribes,correct = F)
chi_test # Here I get 3.79 as well
```

(b) Now calculate the p-value from the test statistic you just created (in R).[2] What do you conclude if $\alpha = 0.1$?

Code to calculate the p-value:

```r
# I know my df must be (nrow-1)(ncol-1), which is simply 1*2 = 2
# I see in this link I can use the pchisq() function
# https://www.r-bloggers.com/2022/05/calculate-the-p-value-from-chi-
   square-statistic-in-r/

pchisq(statistic, 2, lower.tail = F)

# I get 0.15, which also coincides with what the chisq() function
   generated
```

The resulting p-value is 0.15. Since $0.15 > 0.1$, the conclusion is that I have not found enough evidence to reject the null hypothesis that class and bribing are independent. In other words, there is not enough evidence to support the alternative hypothesis that the two variables are dependent.

[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.

Code to calculate the standardized residuals:

```
# To calculate the standardized residual, I also follow Dr. Ziegler's
    formula from his Week 3 slides
# I also check against the results of using my chisq() function and it
    works

std_res_ns_u <- (f0_ns_u - fe_ns_u)/sqrt(fe_ns_u*(1-(27/42))*(1-(21/42)))
std_res_ns_l <- (f0_ns_l - fe_ns_l)/sqrt(fe_ns_l*(1-(15/42))*(1-(21/42)))

std_res_b_u <- (f0_b_u - fe_b_u)/sqrt(fe_b_u*(1-(27/42))*(1-(13/42)))
std_res_b_l <- (f0_b_l - fe_b_l)/sqrt(fe_b_l*(1-(15/42))*(1-(13/42)))

std_res_w_u <- (f0_w_u - fe_w_u)/sqrt(fe_w_u*(1-(27/42))*(1-(8/42)))
std_res_w_l <- (f0_w_l - fe_w_l)/sqrt(fe_w_l*(1-(15/42))*(1-(8/42)))
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.32	-1.64	1.52
Lower class	-0.32	1.64	-1.52

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

The standardized residuals give us a standardized measure of the difference between what we would expect to observe if H0 (that the two variables are independent) were true, compared to what we actually observe. The fact that they are standardized means that they are in the same units as our variables. In this case, we could think of our units as the "interactions" between drivers and police officers. We could make the following interpretations:

Looking at our column of interactions between drivers and police where the police did not stop the driver making the illegal left turn, we can see that the standardized residual is 0.3 for upper class drivers. This means that the observed number of cases where the police did not stop upper class drivers was higher than it would be expected under the null that class and probability of bribery are independent. Specifically, we would expect to see 0.3 less "interactions" or cases where this was the outcome, when it comes to upper class drivers. Likewise, our standardized residual for lower class drivers is -0.03. This means that the number of interactions that we observe where the outcome is not being stopped and the type of driver is lower class is less than we would expect under the null. In other words, for our first column, upper class drivers are stopped less often than we would expect if the variables are independent, or more often than we would expect if they are not, while the exact opposite is true for lower class drivers.

Looking at our other two columns, the standardized residuals are of greater magnitude, which means the observed is even further away from the expected under independence, as compared to our first column. From the second column, we can interpret that upper class drivers are requested a bribe in less interactions than we would expect under independence. In other words, if the variables were independent, upper class drivers would have to be requested more bribes. The opposite is true for lower class drivers: lower class drivers are requested bribes more often than we would expect under independence. For the third column, we obtain that, for upper class drivers, we observe a lot more warnings (after being stopped) than we would expect under independence, and for lower class drivers we observe a lot less warnings than we would expect under independence.

Coming from Mexico, I would think that a) class and the probability of bribery are NOT independent, and b) the police is more likely to target people from lower socioeconomic status, perhaps because they are more vulnerable to extortion (probably among many reasons). The results from all columns seem to be in line with this intuition, where lower class drivers are less less likely to be stopped than they would be under independence, and also less likely to be requested a bribe than they would be under independence.

Question 2: 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).

Name	Description
`GP`	An identifier for the Gram Panchayat (GP)
`village`	identifier for each village
`reserved`	binary variable indicating whether the GP was reserved for women leaders or not
`female`	binary variable indicating whether the GP had a female leader or not
`irrigation`	variable measuring the number of new or repaired irrigation facilities in the village since the reserve policy started
`water`	variable measuring the number of new or repaired drinking-water facilities in the village since the reserve policy started

[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.

Our model would be of the type: $y = a + bx$, where:

- x: policy (reserved / not reserved; I assume that 1 means reserved for women, 0 means not)
- y: no. of new or repaired water facilities

And so the relevant hypotheses for our intercept and slope would be (we can hypothesize about effects in particular because the presence of the policy in villages is random):

- H0: $a = 0$ (The no. of new/repaired water facilities when the reservation policy is not in place is zero).
- Ha: $a \neq 0$ (The no. of new/repaired water facilities when the reservation policy is not in place is NOT zero.)
- H0: $b = 0$ (The reservation policy has no effect on the no. of new/repaired water facilities).
- Ha: $b \neq 0$ (The policy does have an effect on the no. of new/repaired water facilities.)

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

I tests both sets of hypotheses by hand, estimating the intercept and the slope, and then confirm.

Code to run the regression:

```
# Trying to run the regression "by hand", following Dr. Ziegler's formula of the estimators
# for b0 and b1, and his code (both from Week 4 slides)

b1 <- sum((data$water - mean(data$water)) * (data$reserved - mean(data$reserved)))/
  sum((data$reserved - mean(data$reserved))^2)

b0 <- mean(data$water) - b1*mean(data$reserved)

# Confirming with lm() (I knew the lm function well from before the course)

bivar_reg <- lm(water ~ reserved, data = data)
bivar_reg # I do get the same results !
```

(c) Interpret the coefficient estimate for reservation policy.

The coefficient estimate for reservation (b) is 9.25. It is positive, and it is quite sizable. This estimate indicates that, on average, places where the reservation policy is in place build/repair 9.25 more water facilities, as compared to villages where the policy is not in place.

Therefore, the reservation policy does have an effect over the no. of new/repaired water facilities, and having government positions reserved for women does lead, in this case, to the implementation of more policies that women support.

Part of the reason why the coefficient is so sizable is that our input variable is a dummy. This is also why we interpret the coefficient in the above way, instead of the 'a one unit change in x causes an increase of b units in y' approach, although they are equivalent, and our 'one unit increase in x' is simply passing from no reservation policy to a reservation policy.

Additionally, our intercept estimate is 14.73. What this means is that, in places where the reservation policy is not in place, 14.73 water facilities are built/repaired, on average. The intercept also helps us make sense of the size of our slope: since our slope is more than half of our intercept, this means that places where the policy is in place build/repair a little over 50% more water facilities, so we could say that the effect of the policy over the no. of new/repaired facilities is quite large.