Problem Set 2

Applied Stats/Quant Methods 1

Due: October 16, 2022

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 16, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

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

```
# part a
# calculate chi-square test statistic
# chi^2 = sum of (diff b/t observed and expected value squared)/(expected
    value)

# step one prepare data in R

PolTre <- matrix(c(14, 7, 6, 7, 7, 1), 2, 3) # assign frequencies to a
    matrix
dimnames(PolTre) = list(
  class = c("Upper", "Lower"),
  treatment = c("Not Stopped", "Bribe Requested", "Warning")
) # name rows and columns
PolTre
addmargins(PolTre)


# step two find proportion of category 'class' representation in total
    sample
# step three calculate and assign expected values

# the expected value of each cell is the column total x sample proportion
    of
# the class i.e. row total over sample total

# if cell1
sum(PolTre[,1])*sum(PolTre[1,])/sum(PolTre)

expv <- matrix(, nrow=2, ncol = 3)
for (r in 1:2){
  expv[r,1] <- sum(PolTre[,1])*sum(PolTre[r,])/sum(PolTre)
  expv[r,2] <- sum(PolTre[,2])*sum(PolTre[r,])/sum(PolTre)
  expv[r,3] <- sum(PolTre[,3])*sum(PolTre[r,])/sum(PolTre)
}
expv # expected values matrix

# step four calculate chi-square values for each cell
# i.e. take expv value from PolTre value for each cell, square, then over
# expv value

# chi square formula in R
```

```
# chsq <- ((x-y)^2)/y where x is observed value and y is expected value

# for cell 1
(((PolTre[1,1])-(expv[1,1]))^2) / (expv[1,1])

exporg <- data.frame(cbind(PolTre, expv))
exporg
cs <- matrix(, nrow=2, ncol = 3)
for (r in 1:2){
  cs[r,1] <- (((exporg[r,1])-(exporg[r,4]))^2) / (exporg[r,4])
  cs[r,2] <- (((exporg[r,2])-(exporg[r,5]))^2) / (exporg[r,5])
  cs[r,3] <- (((exporg[r,3])-(exporg[r,6]))^2) / (exporg[r,6])
}
cs

# step five sum all values of chi-square
sum(cs)
# [1] 3.791168
# on tables, for df=2 ((r-1)(c-1)), p-value is between 0.975 and 0.2
# therefore fail to reject the null hypothesis that
# these two samples have no dependency relationship 'statistical
    independence'

# checking work
chisq.test(PolTre, y=NULL)
# Pearson's Chi-squared test
# X-squared = 3.7912, df = 2, p-value = 0.1502
# same chi square and p range, same conclusion at alpha = 0.1
```

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

```
## part b
# calculate the (right-tail) p-value
pchisq(sum(cs),2, lower.tail = FALSE) # [1] 0.1502306
# fail to reject H0
```

[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.

```
## part c
srs <- matrix(, nrow=2, ncol = 3)
for (r in 1:2){
  srs[r,1] <- (((exporg[r,1])-(exporg[r,4]))^2) / (exporg[r,4])
  srs[r,2] <- (((exporg[r,2])-(exporg[r,5]))^2) / (exporg[r,5])
  srs[r,3] <- (((exporg[r,3])-(exporg[r,6]))^2) / (exporg[r,6])
}
srs

# standardised residuals formula in R
# x - y / sqrt (y(1-rowtotal/total)(1-coltotal/total))

srs <- matrix(, nrow=2, ncol = 3)
rownames(srs) <- c("Upper Class", "Lower Class")
colnames(srs) <- c("Not Stopped", "Bribe Requested",
                   "Given Warning")
for (r in 1:2){
  srs[r,1] <- ((exporg[r,1])-(exporg[r,4])) /
    (sqrt(((exporg[r,4])*
           (1-(sum(PolTre[r,])/
                 sum(PolTre)))*(1-sum(PolTre[,1])/sum(PolTre))
    ))
    )
  srs[r,2] <- ((exporg[r,2])-(exporg[r,5])) /
    (sqrt(((exporg[r,5])*
           (1-(sum(PolTre[r,])/
                 sum(PolTre)))*(1-sum(PolTre[,2])/sum(PolTre))
    ))
    )
  srs[r,3] <- ((exporg[r,3])-(exporg[r,6])) /
    (sqrt(((exporg[r,6])*
           (1-(sum(PolTre[r,])/
                 sum(PolTre)))*(1-sum(PolTre[,3])/sum(PolTre))
    ))
    )
}
srs

# ##           Not Stopped Bribe Requested Given Warning
# Upper Class   0.3220306        -1.641957       1.523026
# Lower Class  -0.3220306         1.641957      -1.523026
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.3220306	-1.641957	1.523026
Lower class	-0.3220306	1.641957	-1.523026

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

Standardised residuals are a relative measure of how different each observed value is from the expected value predicted by the null hypothesis, which predicts there will be no difference between the class groups in treatment by the police from the proportion of class groups in the total sample.

When the null hypothesis is true, standardised residuals have a normal distribution and have a mean of zero, meaning 0 is the balancing point for the data. The equal magnitude in each direction (positive and negative) in the above table indicates an even fluctuation around 0, which would provide some evidence in favour of the null hypothesis wherein there is no association between class and treatment. Though some of the values have what would a appear to be a notable magnitude, none exceed the value of 2, a cutoff which is generally statistically significant at $\alpha = 0.01$. Without too much investigation, the table above gives an indication that we do not have sufficient evidence for an association to reject the null hypothesis, affirming our above finding in part (a) and (b). Thus, we cannot say anything about the difference between the standardised residuals in different cells, as our conclusions would not meet the level of statistical significance anyway.

Question 2 (40 points): Economics

"...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."

(a) State a null and alternative (two-tailed) hypothesis.

H0 : the reservation policy is not associated with the number of new or repaired drinking water facilities i.e. the two variables are statistically independent

H1 : the reservation policy is associated, in an unspecified direction, with the number of new or repaired drinking water facilities in the villages

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

```
fempol <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
summary(fempol)
head(fempol)

# part b
lm1 <- lm(water ~ reserved, data = fempol)
summary(lm1)

# expected value for water projects when reserved policy is in place (=1) is
# 14.738 beta 0
# difference in group specific estimations (beta 1) is 'reserved' 9.252

# t-value: 6.446, p-value: 4.22e-10, reject the null hypothesis
```

We reject the null hypothesis that there is no difference between group means at the 5 per cent significance level, i.e. $\beta_1 \neq 0$

(c) Interpret the coefficient estimate for reservation policy.

The coefficient estimate for reservation policy is represented in the summary() output by the cell 'reserved'/'Estimate', which has a value of 9.252. This value is the difference in group-specific estimations for the output variable 'water'. In other words, there is a sizeable (given the relative values of the water variable) difference in the regression model's predictions for each group, where the groups are separated by villages who maintain a reserved position for female politicians, and villages who don't. Therefore we find evidence suggesting an association between the reservation policy and the number of new or repaired water facilities.