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`).

The χ^2 test result is 3.791168. Below is the R code:

```r
############### Question 1 ###############

####### (a) #######

# Formulate the row of cross table
class_type <- c(rep("Upper class", 27), rep("Lower class", 15))
class_type <- factor(class_type)
# Formulate the columns of cross table
bribe_type <- c(rep("Not stopped", 14), rep("Bribe requested", 6),
                rep("Stopped warning", 7),
                rep("Not stopped", 7), rep("Bribe requested", 7),
                rep("Stopped warning", 1))
bribe_type <- as.factor(bribe_type)
# Check cross table
table(class_type, bribe_type)
# Calculate row and columns counts
row_1 <- 15 ; row_2 <- 27
col_1 <- 13 ; col_2 <- 21; col_3 <- 8
n <- 42
# Calculate expected frequencies
e11 <- (row_1 * col_1)/n; e12 <- (row_1 * col_2)/n; e13 <- (row_1 * col_
    3)/n
e21 <- (row_2 * col_1)/n; e22 <- (row_2 * col_2)/n; e23 <- (row_2 * col_
    3)/n
# Calculate Chi-squared Value
x11 <- (7 - e11)^2/e11; x12 <- (7 - e12)^2/e12; x13 <- (1 - e13)^2/e13
x21 <- (6 - e21)^2/e21; x22 <- (14 - e22)^2/e22; x23 <- (7 - e23)^2/e23
# Calculate chi-square value and print
x_square_value <- x11 + x12 + x13 + x21 + x22 + x23
print(x_square_value)
```

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

Here is the null hypothesis and alternative hypothesis:
H0: These 2 variables in the population are not related and are independent of each others.
H1: These 2 variables in the population are related to each others.
Below is the R code:

```r
####### (b) #######

# Calculate degree of freedom
df <- (3 - 1) * (2 - 1)
# Calculate p value
p_value <- 1 - pchisq(x_square_value, df)
print(p_value)
# Set alpha
alpha <- 0.1
# Check chi-square value with 0.1 alpha
critical_chisq <- qchisq(1 - alpha, df)
print(critical_chisq)
# Compared with p value and alpha
if (p_value > alpha) {
  print("Reject H1, Accept H0")
} else {
  print("Reject H0, Accept H1")
}
# Another way: compared with critical level and chi-square results
if (x_square_value < critical_chisq) {
  print("Reject H1, Accept H0")
} else {
  print("Reject H0, Accept H1")
}
```

p value = $0.1502306 > 0.1 =$ alpha;
and $\chi^2_{0.1} = 3.791168 < 4.60517 = \chi^2$(K=3)

so we reject H1 and accept H0, we can think these 2 variables in the population are not related and are independent of each others.

[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.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.14	-0.82	0.82
Lower class	-0.18	1.09	-1.10

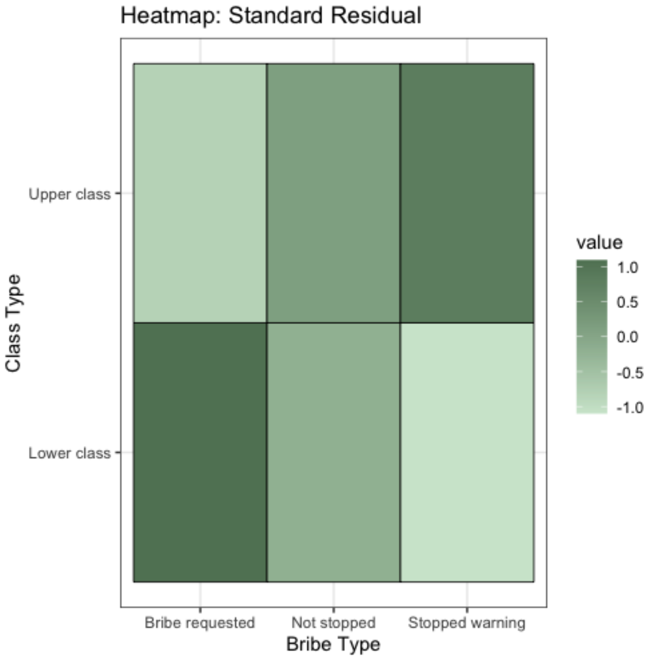
Below is the R code:

```r
####### (c) #######

# According to (a), we have calculated expected frequencies
# Now we can calcualte residual by (frequencies - expected values)
resi11 <- 7 - e11; resi12 <- 7 - e12; resi13 <- 1 - e13;
resi21 <- 6 - e21; resi22 <- 14 - e22; resi23 <- 7 - e23;
# Calculate standard residual by [residual / sqrt(expected value)]
s_resi11 <- resi11/sqrt(e11); s_resi12 <- resi12/sqrt(e12);
s_resi13 <- resi13/sqrt(e13)
s_resi21 <- resi21/sqrt(e21); s_resi22 <- resi22/sqrt(e22);
s_resi23 <- resi23/sqrt(e23)
# Print and check the results
cat(s_resi11, s_resi12, s_resi13,"\n", s_resi21, s_resi22, s_resi23)
```

(d) How might the standardized residuals help you interpret the results?
If we assume that we pass the hypothesis test, and we can say lower class people are more possible to request a bribe, and higher class people are more possbile to stopped or given warning.

Figure 1: Heatmap: Standard Residual



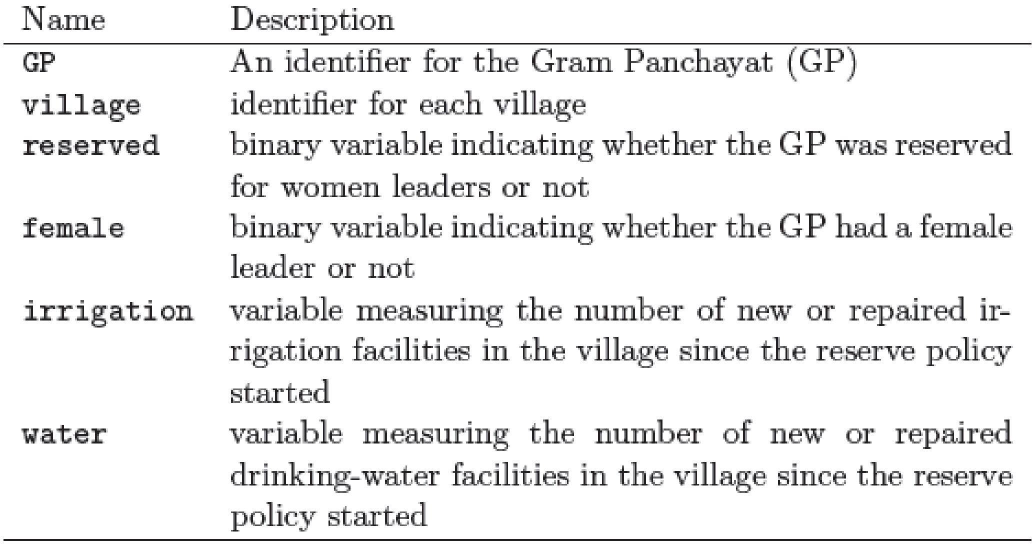
```r
####### (d) #######
# Draw a heat map to visualize the differences between standard residual
# Make a matrix about standard residual
s_resi_mat <- matrix(
  c(s_resi11, s_resi12, s_resi13, s_resi21, s_resi22, s_resi23),
  nrow = 2, ncol = 3, byrow = TRUE)
# Give colnames and rownames in matrix
colnames(s_resi_mat) <- c("Bribe requested", "Not stopped", "Stopped warning")
rownames(s_resi_mat) <- c("Lower class", "Upper class")
# Transfer matrix into long data table to draw ggplot picture
s_resi_long <- melt(s_resi_mat)
print(s_resi_long)
# Use ggplot to draw heatmap
s_resi_heatmap <- ggplot(s_resi_long, aes(x=Var2, y=Var1)) +
  geom_tile(aes(fill=value), color="black", size=0.3) +
  scale_fill_gradient(low="#D0E7D2", high="#618264") +
  labs(title="Heatmap: Standard Residual",
       x="Bribe Type", y="Class Type") +
  theme_bw()
print(s_resi_heatmap)
```

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 2 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 2: 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.

H_0: There was no difference in the number of newly built drinking water facilities between GP with and without female leadership.

H_1: There was difference in the number of newly built drinking water facilities between GP with and without female leadership.

And in math language, we can also make the statement like this:

H_0: $\theta_0 = \theta_1$
H_1: $\theta_0 \neq \theta_1$

θ_0 means the number of newly built drinking water facilities in GP which has a female leader.

θ_1 means the number of newly built drinking water facilities in GP which has not have a female leader.

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

Put "water" as dependent variable and binary variable "female" as independent variable.
Because binary variable "female" is already a dummy variable, so we don't need to operate anything on it.

Use R to get the data frame and create regression model:

(c) Interpret the coefficient estimate for reservation policy.