PS02

Rebecca Katherine Armstrong, 15320332

2024-10-14

```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)

#     Question 1: Political Science

  ##############

# Creating Table

  ##############


Class<-c("Upper Class", "Lower Class")

Not_Stopped<-c(14, 7)

Bribe_Requested<-c(6,7)

Stopped<-c(7,1)


table<-data.frame(Class,Not_Stopped,Bribe_Requested,Stopped)

print(table)


print("(a) Calculating the x^2 test statistic manually")

          #############
  # Finding totals for rows and columns
          ############


totalUClass<-sum(table[1,-1])

totalLClass<-sum(table[2,-1])

totalnStopped<-sum(table$Not_Stopped)

totalBribe<-sum(table$Bribe_Requested)

totalStopped<-sum(table$Stopped)
```

```
print("The total presentations from upper class across all results was:")

print(totalUClass)

print("The total presentations from lower class across all results was:")

print(totalLClass)

print("The summary of the values in the columns 'Not Stopped', 'Bribe Requested'
and 'Stopped or Given Warning'
respectively were:")

print(totalnStopped)

print(totalBribe)

print(totalStopped)


      #########


total1<-totalLClass+totalUClass

total2<-totalBribe+totalnStopped+totalStopped

print(total1)

print(total2)

print("Having found all row and
column totals, next the expected
frequency for each cell must be
found")


nStopped1<-c((totalUClass*totalnStopped/total1))

nStopped2<-c((totalLClass*totalnStopped/total1))

bribe1<-c((totalUClass*totalBribe/total1))

bribe2<-c((totalLClass*totalBribe/total1))

stop1<-c((totalUClass*totalStopped/total1))

stop2<-c((totalLClass*totalStopped/total1))

Expected_Frequency<-c("Upper Class", "Lower Class")

Not_StoppedF<-c(nStopped1,nStopped2)

BribeF<-c(bribe1,bribe2)
```

```
StoppedF<-c(stop1,stop2)

EFtable<-data.frame(Expected_Frequency,Not_StoppedF
,BribeF,StoppedF)

print(EFtable)


print("Expected_Frequency Not_StoppedF   BribeF StoppedF
1        Upper Class         13.5 8.357143 5.142857
2        Lower Class          7.5 4.642857 2.857143")


print("Now to find x^2 statistic
for each cell, take the expected
frequency of that cell from the
actual value, square it and divide
by the expected frequency of the
cell:")

ns1<-((Not_Stopped[1]-nStopped1)^2)/nStopped1

ns2<-((Not_Stopped[2]-nStopped2)^2)/nStopped2

brb1<-((Bribe_Requested[1]-bribe1)^2)/bribe1

brb2<-((Bribe_Requested[2]-bribe2)^2)/bribe2

stp1<-((Stopped[1]-stop1)^2)/stop1

stp2<-((Stopped[2]-stop2)^2)/stop2



Chi_Table<-c("Upper Class", "Lower Class")

Not_StoppedC<-c(ns1,ns2)

BribeC<-c(brb1,brb2)

StoppedC<-c(stp1,stp2)

chi.table<-data.frame(Chi_Table,Not_StoppedC,BribeC,StoppedC)

print(chi.table)


print("Summing all cell values from chi-table to find X^ statistic:")

XSquared<-sum(chi.table[,-1])

print(XSquared)
```

```
print("The X^2 statistic is: 3.791168")
```

```
###########
```

```
print("(b) Calculating the p-value:")

df<-(2-1)*(3-1)

print(df)

pValue<-pchisq(XSquared,df=2,lower.tail = FALSE)

print(pValue)
```

```
print("The p-value is: 0.1502306")
```

```
print("The p-value for this dataset
is greater than 0.1.  Therefore,
at the 10% significance level we
fail to reject the null hypothesis
- that is, from the evidence
provided by this sample, you
cannot determine if police
officers are more or less likely
to solicit a bribe from drivers
depending on their class")
```

```
###########
```

```
print("(c) Calculating the
standardised residuals for
each cell:")
```

```
StandRes<-c("Upper Class", "Lower Class")
```

```
srnotstop1<-((Not_Stopped[1]-nStopped1)/(nStopped1^
(1/2)))

srnotstop2<-((Not_Stopped[2]-nStopped2)/(nStopped2^
(1/2)))

srbribe1<-((Bribe_Requested[1]-bribe1)/(bribe1^(1/2)))

srbribe2<-((Bribe_Requested[2]-bribe2)/(bribe2^(1/2)))

srstop1<-((Stopped[1]-stop1)/(stop1^(1/2)))

srstop2<-((Stopped[2]-stop2)/(stop2^(1/2)))

Not_StoppedSR<-c(srnotstop1,srnotstop2)

BribeSR<-c(srbribe1,srbribe2)

StoppedSR<-c(srstop1,srstop2)

SRtable<-data.frame(StandRes,Not_StoppedSR,BribeSR,StoppedSR)

print(SRtable)


print("The Standard Residuals for
each cell are as follows:
Upper Class, Not Stopped = 0.1360828;
Upper Class, Bribe = -0.8153742;
Upper Class, Stopped or Given Warning = 0.818923;
Lower Class, Not Stopped = -0.1825742;
Lower Class, Bribe = 1.0939393;
Lower Class, Stopped or Given Warning = -1.098701")

##########

print("(d) How do the standardised
residuals help to interpret results?")
```

```
print("Looking at the standardised
residuals from this dataset, the
first considereation of note is that
the values all reside between the
values of -2 and +2 non-inclusive,
indicating that there are no
significant deviations from the
expected values.  They are mixed
positive and negative values,
showing that the observed frequencies
while not moving remarkably beyond
what was expected, the observed
frequencies were both higher and
lower than what was statistically
anticipated.  The observations
categorised as being 'not stopped'
lay closest to what the expected
observations would have been, with
the observations of bribes being
requested and individuals stopped or
given warnings being equivalently
close to 1 unit above or below the
expected frequency (Lower class
observations of bribes or being
stopped/given warnings holding the
strongest deviation from expected
at over 1.09 positive and negative
difference from the expected counts
respectively.  The equal split of
positive and negative values across
the standardised residuals with no
discernable relation to class or
outcome create difficultly in
ascertaining if the outcomes or
classes were more or less represented
than expected.  Overall, the
standardised residuals of the
combinations in this dataset,
were not conspciously large enough
or consistently positive or negative
enough to infer that the observed
frequency was notably different from
the expected frequency.")
```

######################### "' # Question 2: Economics

#########################

print("Uploading dataset for question:")

bengalData<-read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")

print("Null Hypothesis: There is no relationship between the presence of female leaders and the quality of drinking water in a village.")

print("Alternate Hypothesis: The level of female leaders in a village has an impact on the quality of drinking water in that village")

#################

print("Testing the hypothesis with a bivarate regression model:")

waterQual<-lm(water~GP+village+female+irrigation+reserved, data = bengalData)

summary(waterQual)

print("lm(formula = water ~ GP + village + female + irrigation + reserved, data = bengalData) Residuals: Min =-89.140; 1Q = -12.628; Median = -5.277; 3Q = 5.035; Max = 285.749.") print("For the coefficients, the Estimate, Std. Error, t value, and Pr(>|t|) for each variable were as follows:")

print("(Intercept): 4.30135, 6.29921, 0.683, 0.495")

print("GP: -0.02493, 0.03684, -0.676, 0.499")

print("village: 5.00291, 3.40251, 1.470, 0.142")

print("female: -0.08429, 7.96740, -0.011, 0.992")

print("irrigation: 1.46776, 0.17995, 8.157, 8.21e-15")

print("reserved: 9.82875, 8.21005, 1.197, 0.232")

##########

print("Residual standard error: 30.52 on 316 degrees of freedom, Multiple R-squared: 0.1915,
Adjusted R-squared: 0.1787 F-statistic: 14.97 on 5 and 316 DF,
p-value: 3.406e-13")

print("From the model summary, we can see that only the presence of any new or repaired irrigation system has any statisticaly important bearing on the quality of drinking water in a village.
When all other variables are held constant, the increase of each additional unit of a repaired or new irrigation system has a positive correlation of increasing the quality of drinking water by 1.46776 units.
As the p-value for this variable is 8.21e-15, this is statistically meaningful at the 5% level (p<0.05). No other variable has any statistically significant impact on the dependent variable (the quality of drinking water). Considering first the lack of statistical significance of the presence of female leaders in a village, and the lack of statistical significance of the reservation of female places in government at the level of GP, we fail to reject the null hypothesis that the presence of female leaders has any impact on the quality of drinking water in a village. Moreover, the Multiple R-Squared value of 0.1915 indicates that only 19.15% of variablility in the quality of drinking water is explained by this model, thus the model is clearly missing one or more variables which will contribute to the explanation of variability in drinking water quality by village.")

#######################

print("(d) Discuss the coefficient estimate for the reservation policy:")

print("The reservation policy (denoted in the model by 'reserved') infers that there should be a substantial positive relationship between the GP being reserved for women leaders and the quality of drinking water in the area (that is, with all other variables held constant, for each additional unit GP being reserved for female leaders, the coefficient of reserved indicates that there should be a positive increase in the units of quality of drinking water by 9.82875).
However, the p-value of 0.232 highlights that this relationship holds no statistically significant correlation at the 10% level of significance, as p>0.1, so no meaningful relationship between the two variables can be inferred from the data provided in this model.") "'