Stats. Inference Assignment 3

Naeem Chowdhury

5/19/2020

##1. Setup ### options Set up global options

libraries

Load in needed libraries

```
library(tidyverse)
```

```
## -- Attaching packages ----------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts -------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(RColorBrewer)
library(haven)
```

2. File management

Create variables for directories

```
project.dir <- getwd() #naeem
output.dir <- "/Output"
data.dir <- "C:/Users/Naeem Cho/Desktop/School Work/Statistical Inference/Datasets"
setwd(project.dir)
getwd()
```

```
## [1] "C:/Users/Naeem Cho/Desktop/School Work/Statistical Inference/Statistical_Inference/2020-05-19_HW
```

3. Importing Data

```
cpu.data <- read_csv(file.path(data.dir, "Intel_CPUs.csv"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   nb_of_Cores = col_double(),
##   nb_of_Threads = col_double(),
##   Max_nb_of_Memory_Channels = col_double(),
##   Processor_Graphics_ = col_logical(),
```

```
##   Support_4k = col_logical(),
##   OpenGL_Support = col_logical(),
##   Max_nb_of_PCI_Express_Lanes = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

Problem #1

1. Code up your own *my.chisq.test()* function that will perform a χ^2 test. As a single argument, it should just take a contingency table of arbitrary size. As output, it should provide:

- Calculated χ^2 statistic
- p-value

Caclulating the expected cell counts under H_0 hypothesis should constitute a critical part of your function definition. Don't use neither *chisq.test()* nor *prop.test()*, nor any other "fancy cheat" built-in functions inside your function's definition.

```
data <- matrix(c(18, 20, 15, 15, 10, 55, 65, 70, 30), nrow=3)

my.chisq.test <- function(ctable) {
  rows <- nrow(ctable)
  cols <- ncol(ctable)

  rowsums <- rowSums(ctable)
  colsums <-colSums(ctable)

  total <- sum(ctable)
  df <- (rows-1)*(cols-1)

  expected <- matrix(0, nrow = rows, ncol = cols)

  for(i in 1:rows){

    for(j in 1:cols){

      expected[i,j] <- (rowsums[i]*colsums[j])/ total

    }
  }
  chi_sq <- ((ctable - expected)^2)/expected
  chi_sq <- sum(chi_sq)

  p <- pchisq(chi_sq, df, lower.tail = FALSE)


  print(c("X-Squared: ",  chi_sq))
  print(c("Degrees of Freedom: ", df))
  print(c("p-value: ", p))
  # print(c("Matrix of expected counts: ", expected))
}
```

2. Testing on a CPU/GPU Dataset

a. What variables are we interested in?

```
vert_stat <- cpu.data %>% select(Vertical_Segment, Status) %>% drop_na()
table(vert_stat)
```

```
##                 Status
## Vertical_Segment Announced End of Interactive Support End of Life Launched
##         Desktop          3                        357         152      116
##         Embedded         3                          0           4      170
##         Mobile           5                        370           3      382
##         Server           4                         64         275      375
```

I've selected two categorical variables from the Intel_CPUs dataset which each have 4 categories. However, since the 'Announced' category and 'Embedded' category have very few variables, I've decided to drop them in order to simplify the analysis.

```
vert_stat <- cpu.data %>% select(Vertical_Segment, Status) %>% filter(Vertical_Segment != 'Embedded' & S
ctable <- table(vert_stat)
```

b. What are the hypotheses?

H_0 := The launch Status of a CPU is independent of its Vertical_Segment type.

H_a := The launch status of an Intel CPU is dependent on its Vertical Segment type.

c. Print the contingency table. Under H_0 hypothesis, proceed to calculate expected counts for two arbitrary cells of the contingency table.

```
ctable
```

```
##                 Status
## Vertical_Segment End of Interactive Support End of Life Launched
##          Desktop                        357         152      116
##          Mobile                         370           3      382
##          Server                          64         275      375
```

```
# Calculating the expected counts for two arbitrary cells.
rows <- nrow(ctable)
cols <- ncol(ctable)
rowsums <- rowSums(ctable)
colsums <-colSums(ctable)
total <- sum(ctable)

# Produce the pairs of row and column location of arbitrary position
r.indices <- sample(1:rows, 2, replace= T)
c.indices <- sample(1:cols, 2, replace= T)
n <- sum(ctable)

n.r1 <- as.numeric(rowsums[r.indices[1]])
n.c1 <- as.numeric(colsums[c.indices[1]])

n.r2 <- as.numeric(rowsums[r.indices[2]])
n.c2 <- as.numeric(colsums[c.indices[2]])

expected1 <- (n.r1*n.c1)/n
```

```r
expected2 <- (n.r2*n.c2)/n

#Final expected counts for the two points
expected1
```

```
## [1] 285.1982
```

```r
expected2
```

```
## [1] 155.0382
```

d. Proceed to apply your *my.chisq.test()* and interpret the results. As a sanity check, also run *R*'s built-in *chisq.test()* function on that same data, make sure the outputted χ^2 and p-values match with those provided by *my.chisq.test()*.

```r
my.chisq.test(ctable)
```

```
## [1] "X-Squared: "      "624.560481019293"
## [1] "Degrees of Freedom: " "4"
## [1] "p-value: "              "7.48768805220064e-134"
```

```r
chisq.test(ctable)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ctable
## X-squared = 624.56, df = 4, p-value < 2.2e-16
```

My test works everywhere except for computing the p-value, where it seems to fail. However, when I test my function on another dummy table , it seems to work just fine. What could be the issue?

```r
my.chisq.test(data)
```

```
## [1] "X-Squared: "      "63.3040389697903"
## [1] "Degrees of Freedom: " "4"
## [1] "p-value: "            "5.85615608615707e-13"
```

```r
chisq.test(data)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 63.304, df = 4, p-value = 5.856e-13
```

e. In case you end up claiming that variables are not independent, proceed to make a few comments on strength of the relationship (as was done for Income & Happiness example in class).

Since we have a low p-value, we can confirm that we have statistically significant evidence of assocation. We cannot say anything about the practical strength of that association without using either the difference of proportions or ratio of proportions.

```r
# Return to later.
```

Problem #2 (I will not complete this part, as I never wrote Problem #1 for HW 11. However, I will use the two-sample proportion z-test that's built into R.)

Subjects were randomly assigned to regularly take aspirin or placebo, and followed up with over a 5-year period on whether they suffered a cancer death.

```
cancer.death <- matrix(c(347, 11188, 327, 13708), ncol = 2, byrow = TRUE)
colnames(cancer.death) <- c("Yes", "No")
rownames(cancer.death) <- c("Placebo", "Aspirin")
cancer.death <- as.table(cancer.death)

cancer.death
```

```
##           Yes    No
## Placebo   347 11188
## Aspirin   327 13708
```

```
prop.yes <- c(347/(347+11188), 327/(327+13708))
prop.yes
```

```
## [1] 0.03008236 0.02329890
```

```
cbind(cancer.death, prop.yes)
```

```
##         Yes    No   prop.yes
## Placebo 347 11188 0.03008236
## Aspirin 327 13708 0.02329890
```

Main question: Is there a difference in cancer death rates between aspirin & placebo groups?

a. Use your two-sample proportion z-test function from Problem #1 HW 11 (previous semester) to conduct appropriate hypothesis test to address the main question. What are the parameters of interest? What are the hypotheses? What's the conclusion?

Are the proportion of deaths the same for placebo and aspirin in the population?

H_0 := the proportion of deaths when taking a placebo are the same as for when taking aspirin

H_a := the proportion of deaths for placebo and aspirin differ.

```
prop.test(x = c(347, 327), n = c(341+11188, 327+13708), alternative = "two.sided", correct = TRUE)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(347, 327) out of c(341 + 11188, 327 + 13708)
## X-squared = 11.135, df = 1, p-value = 0.000847
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.00272572 0.01087252
## sample estimates:
##     prop 1     prop 2
## 0.03009801 0.02329890
```

Since the p-value is 0.000847 we can reject the null hypothesis. Thus there is a statistically significant difference in the proportion of patients who died while taking a placebo and the proportion who died while taking aspirin.

b. Use χ^2-test to address the main question. Formulate the hypotheses. What's the conclusion?

H_0 := the number of deaths of cancer patients is independent of treatment by placebo or aspirin

H_a := cancer death is dependent on treatment by placebo or aspirin

```
chisq.test(cancer.death)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cancer.death
## X-squared = 11.089, df = 1, p-value = 0.0008683
```

We conclude by rejecting the null hypothesis. There is sufficient evidence that the death of cancer patients in the population is dependent on treatment by placebo or aspirin.

c. Are your p-values in parts (a) and (b) equal? there should be direct correspondence between χ^2 test of independence for 2×2 tables, and two-sided proportion test, in big part due to the fact that

The p-values are exactly the same.

$$z^2 \equiv X^2$$

and that

$$Z^2 \equiv \chi^2_1, Z \sim N(0,1).$$

Problem 3

From **Agresti** book, do exercises:

11.84 Degrees of freedom explained

For testing independence in a contingency table of size $r \times c$, the degrees of freedom (df) for the chi-squared distribution equal $df = (r-1) \times (c-1)$. They have the following interpretation: Given the row and column marginal totals in an $r \times c$ contingency table, the cell counts in a rectangular block of size $(r-1) \times (c-1)$ determine all the other cell counts. Consider the following table, which cross-classifies political views by whether the subject would ever vote for a female president, based on the 2010 GSS. For this 3×2 table, suppose we know the counts in the upper left-hand $(3-1) \times (2-1) = 2 \times 1$ block of the table, as shown.

```
femprez <- matrix(c(56, NA, 58, 490, NA, 509, NA, NA, 61, 604, 24, 628), ncol = 3, byrow = TRUE)
colnames(femprez) <- c("Yes", "No", "Total")
rownames(femprez) <- c("Extremely Liberal", "Moderate", "Extremely Conservative", "Total")

femprez <- as.table(femprez)

femprez
```

```
##                       Yes  No Total
## Extremely Liberal      56        58
## Moderate              490       509
## Extremely Conservative            61
## Total                 604  24   628
```

a. Given the cell counts and the row and column totals, fill in the counts that must appear in the blank cells.

```
femprez.fill <- matrix(c(56,(58-56), 58, 490, (509-490), 509, (61 -(24 - ((58-56)+(509-490)) )), (24 -
colnames(femprez.fill) <- c("Yes", "No", "Total")
rownames(femprez.fill) <- c("Extremely Liberal", "Moderate", "Extremely Conservative", "Total")

femprez.fill <- as.table(femprez.fill)

femprez.fill
```

```
##                         Yes  No Total
## Extremely Liberal        56   2    58
## Moderate                490  19   509
## Extremely Conservative   58   3    61
## Total                   604  24   628
```

b. Now, suppose instead of the preceding table, you are shown the following table, this time only revealing a 2×1 block in the lower-right part. Find the counts in the remaining cells.

```
femprez2 <- matrix(c(NA, NA, 58, NA, 19, 509, NA, 3, 61, 604, 24, 628), ncol = 3, byrow = TRUE)
colnames(femprez2) <- c("Yes", "No", "Total")
rownames(femprez2) <- c("Extremely Liberal", "Moderate", "Extremely Conservative", "Total")

femprez2 <- as.table(femprez2)

femprez2
```

```
##                         Yes  No Total
## Extremely Liberal                  58
## Moderate                     19   509
## Extremely Conservative        3    61
## Total                   604  24   628
```

Solution:

```
femprez2.fill <- matrix(c(604-((509-19)+(61-3)), 58-(604-((509-19)+(61-3))), 58, (509-19), 19, 509, (61
colnames(femprez2.fill) <- c("Yes", "No", "Total")
rownames(femprez2.fill) <- c("Extremely Liberal", "Moderate", "Extremely Conservative", "Total")

femprez2.fill <- as.table(femprez2.fill)

femprez2.fill
```

```
##                         Yes  No Total
## Extremely Liberal        56   2    58
## Moderate                490  19   509
## Extremely Conservative   58   3    61
## Total                   604  24   628
```

11.9 Happiness and gender (and calculate all the expected cell counts, for practice)

For the 2×3 table on gender and happiness in Exercise 11.4 (shown below), software tells us that $\chi^2 = 1.04$ and the P-value $= 0.59$.

```
happy.gender <-matrix(c(154, 592, 336, 123, 502, 257), ncol = 3, byrow =TRUE)
colnames(happy.gender) <- c("Not", "Pretty", "Very")
rownames(happy.gender) <- c("Female", "Male")
happy.gender <- as.table(happy.gender)

happy.gender
```

```
##        Not Pretty Very
## Female 154    592  336
## Male   123    502  257
```

a. State the null and alternative hypothesis, in context, to which these results apply.

- H_0 := Happiness is independent of gender.
- H_a := Happiness is dependent on gender.

b. Interpret the p-value.

There is insufficient evidence to reject the null hypothesis.

11.16 Primary food choice of alligators

For alligators caught in two Florida lakes, the following table shows their primary food choice. The four food categories refer to fish, invertebrates (such as snails, insects, or crayfish), birds and reptiles (such as egrets or turtles), and others, including mammals or plants. Is there evidence that primary food choice differs between the two lakes?

```
lake.food <- matrix(c(30, 4, 8, 13, 55, 13, 18, 12, 10, 53), ncol = 5, byrow = TRUE)
colnames(lake.food) <- c("Fish", "Invertebrates", "Birds & Reptiles", "Others", "n")
rownames(lake.food) <- c("Hancock", "Trafford")

lake.food <- as.table(lake.food)

lake.food
```

```
##          Fish Invertebrates Birds & Reptiles Others  n
## Hancock    30             4                8     13 55
## Trafford   13            18               12     10 53
```

a. Find the conditional sample distributions of primary food choice in lakes Hancock and Trafford.

```
prop.lake <- prop.table(lake.food)

prop.lake
```

```
##                Fish Invertebrates Birds & Reptiles     Others          n
## Hancock  0.13888889    0.01851852       0.03703704 0.06018519 0.25462963
## Trafford 0.06018519    0.08333333       0.05555556 0.04629630 0.24537037
```

b. Set up the hypotheses of interest.

$H_0 := $ primary food choice does not differ between lakes Hancock and Trafford.

H_a := food choice of the animals is dependent on the lake.

c. The X^2 value for this table equals 16.79. Based on the df for the corresponding chi-squared distribution, can this be considered large? Why?

```
df = (2-1)*(4-1)
df
```

```
## [1] 3
```

```
pchisq(16.79, df = 3, lower.tail = FALSE)
```

```
## [1] 0.0007806146
```

For a chi-squared distribution with $df = 3$, 16.79 is considered a large value. This is because such a low df corresponds with a highly right-skewed distribution. As df gets large, the chi-squared distribution will approach the normal distribution.

The p-value for the chi-squared is less than 0.001. Write the conclusion of the test in this context.

There is sufficient evidence to reject the null hypothesis. We conclude that the primary food choice differs between the two lakes.