Untitled

Rex Cheung

3/20/2019

There is a nice built-in function in R to conduct the Pearson's Chi-Squared Test. In this write up we will demonstrate using this built-in function, as well as performing the test by calculating each step of the test. Before we start, let's try to replicate the gender vs trouble status data from the article.

```
gender = c(replicate(117, 'boys'), replicate(120, 'girls'))
trouble = c(replicate(46, 'trouble'), replicate(71, 'no trouble'), replicate(37, 'trouble'), replicate(8
o.table = table(gender, trouble)
print(o.table)
```

```
##        trouble
## gender  no trouble trouble
##   boys          71      46
##   girls         83      37
```

Using Built-in Function

```
chisq.test(gender, trouble, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  gender and trouble
## X-squared = 1.8733, df = 1, p-value = 0.1711
```

```
#chisq.test(o.table, correct = FALSE) #Inputting the data as a contingency table
```

There are two options of inputting the data into the `chisq.test` function. We can either input the two variables into the function (the `x=` and `y=` arguments), or simply supply the contingency table of the two variables (the variables `tbl` above). If `correct = TRUE`, the test will apply the Yates' correction for continuity.

Using Basic Calculations

Even though the built-in function is simply to use, going through the basic calculations allow us to gain a deeper understanding of the testing procedure. Recall that the test statistics for the Pearson's Chi-Squared Test is

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

where observed is the observed counts (or observed relative frequency), and expected is the expected counts (or expected relative frequency) when the two variables are independent.

```
gender.prob = table(gender)/length(gender)
trouble.prob = table(trouble)/length(trouble)
e.table = matrix(0, nrow = length(gender.prob), ncol = length(trouble.prob))
for(i in 1:length(gender.prob)){ #Create expected count table
  for(j in 1:length(trouble.prob)){
    e.table[i,j] = gender.prob[i] * trouble.prob[j] * length(gender)
```

```
  }
}
colnames(e.table) = c('no trouble', 'trouble')
rownames(e.table) = c('boys', 'girls')
print(o.table)
```

```
##        trouble
## gender  no trouble trouble
##   boys          71      46
##   girls         83      37
```

```
print(e.table)
```

```
##       no trouble  trouble
## boys    76.02532 40.97468
## girls   77.97468 42.02532
```

```
test.stat = sum((o.table - e.table)^2/e.table)
print(test.stat)
```

```
## [1] 1.873294
```

For the Pearson's Chi-Square test, we assume the test statistics has a χ^2 distribution with degrees of freedom (c-1)(r-1). The critical value (assume $\alpha = 0.05$) can be found by using the following command:

```
crit.val = qchisq(p = 0.95, df = 1)
print(crit.val)
```

```
## [1] 3.841459
```

Since the test statistics is less than the critical value, we failed to reject the null hypothesis. The p-value can be calculated by

```
pchisq(q = test.stat, df = 1 , lower.tail = F)
```

```
## [1] 0.1710983
```