4211 Homework 3

Matthew DeSantis

2023-02-15

1

(a)

```
n = (20+40+16+18+6)
lambdahat = (40+2*16+3*18+4*6)/n
data = c(20,40,16,18,6)
teststat = 0
for (x in 0:4){
  expected = dpois(x, lambdahat)*n
  observed = data[x+1]
  sqdif = ((observed-expected)**2)/expected
  teststat= teststat+sqdif
}
teststat
```

```
## [1] 7.535457
```

(b)

```
(df = 5-1-1)
```

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

(c)

```
1 - pchisq(teststat, df)
```

```
## [1] 0.05665438
```

We do not reject the Poisson model at the alpha=0.05 significance level (but just barely).

2

(a)

```r
data("crimealk")
(test2 = chisq.test(crimealk))
```

```
##
##  Pearson's Chi-squared test
##
## data:  crimealk
## X-squared = 49.731, df = 5, p-value = 1.573e-09
```

Assuming that we are using an alpha=0.05 significance level, we reject the null hypothesis that type of crime and alcoholic status are independent, because the pvalue is (much) lower than 0.05.

(b)

```r
test2$residuals
```

```
##             drinker nondrinker
## arson     0.1271882 -0.1345355
## rape      0.9879282 -1.0449977
## violence  1.2736649 -1.3472405
## theft     1.0801887 -1.1425878
## coining   0.2681762 -0.2836679
## fraud    -4.4291318  4.6849888
```

From the output, it seems clear that fraud has the most information regarding independence, as the residuals in the fraud row are much larger than the other residuals.

(c)

```r
1 - pchisq(test2$residuals[6]**2+test2$residuals[12]**2, 1)
```

```
## [1] 1.139389e-10
```

Performing a test using only the data in the fraud row, we obtain an even smaller pvalue than before, confirming our suspicions from (b).

3

(a)

```r
covid = matrix(c(21338, 21492, 162, 8), ncol = 2)
colnames(covid) = c("no covid", "covid"); rownames(covid) = c("no vaccine", "vaccine")
covid
```

```
##            no covid covid
## no vaccine    21338   162
## vaccine       21492     8
```

(b)

```r
expectedcaserate = 170/43000
expectedcases = expectedcaserate*21500
expectedsafes = 21500 - expectedcases
(teststat3 =
    ((21338-expectedsafes)**2)/expectedsafes +
    ((21492-expectedsafes)**2)/expectedsafes +
    ((162-expectedcases)**2)/expectedcases +
    ((8-expectedcases)**2)/expectedcases
)
```

```
## [1] 140.0596
```

```r
1 - pchisq(teststat3, 1)
```

```
## [1] 0
```

Under the null hypothesis, the test statistic has a Chi-square distribution with 1 degree of freedom.

(c)

```r
phat1 = 8/21500
phat2 = 162/21500
var1 = phat1*(1-phat1)
var2 = phat2*(1-phat2)
n = 21500
(teststat3c = (phat1-phat2)/sqrt((var1/n)+(var2/n)))
```

```
## [1] -11.854
```

```r
pnorm(teststat3c)*2
```

```
## [1] 2.051605e-32
```

The pvalue here is much lower than the pvalue from the Chi-square test, which is expected, as z tests have much greater power than Pearson Chi-square tests.

(d)

```
chisq.test(covid, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  covid
## X-squared = 140.06, df = 1, p-value < 2.2e-16
```

```
prop.test(x = c(8, 162), n = c(21500,21500), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(8, 162) out of c(21500, 21500)
## X-squared = 140.06, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##   -0.008347101 -0.005978481
## sample estimates:
##       prop 1       prop 2
## 0.000372093 0.007534884
```

4

(a)

```r
ca = c(8.27,8.20,8.25,8.14,9.00,8.10,7.20,8.32,7.70)
na = c(8.50,9.48,8.65,8.16,8.83,7.76,8.63)
```

I will use a Mann-Whitney test here, because the sample sizes are very small and n!=m.

(b)

```r
(stat4 = sum(rank(c(na,ca))[1:9])-(9*(10)/2))
```

```
## [1] 46
```

(c)

```r
1 - pwilcox(stat4,9,7,lower.tail = T)
```

```
## [1] 0.05708042
```

At the alpha = 0.05 level of significance, there is not significant evidence to say that Native Americans have higher msce than Caucasians.

5

```r
testcompare = function(ntrials = 1000, n1 = 100, n2 = 200, mean2 = 0, sd2 = 1, exp = FALSE){
zfails = 0
tfails = 0
wfails = 0
for(x in 1:ntrials){
  if (exp){
    x1 = rexp(n1)
    x2 = rexp(n2)
  }
  else {
    x1 = rnorm(n1)
    x2 = rnorm(n2, mean = mean2, sd = sd2)
  }


  zstat = (mean(x1)-mean(x2))/sqrt(var(x1)/n1+var(x2)/n2)
  if (zstat < qnorm(0.025) || zstat > qnorm(0.975)){
    zfails = zfails + 1
  }

classict = t.test(x1,x2,var.equal = TRUE)
if (classict$p.value < 0.05){
  tfails = tfails + 1
}

welcht = t.test(x1,x2)
if (welcht$p.value < 0.05){
  wfails = wfails + 1
}

}
results = matrix(c(zfails/ntrials,tfails/ntrials,wfails/ntrials), ncol = 3)
colnames(results) = c("z test | ", "classical t test | ", "Satterthwaite-Welch t test |")
rownames(results) = "rejection rate"
return(results)

}
```

(a)

```r
testcompare()
```

```
##                z test |  classical t test |  Satterthwaite-Welch t test |
## rejection rate    0.052               0.054                         0.051
```

(b)

```
testcompare(sd2 = sqrt(2))
```

```
##                z test |  classical t test |  Satterthwaite-Welch t test |
## rejection rate    0.041              0.021                         0.04
```

(c)

```
testcompare(mean2 = 0.3)
```

```
##                z test |  classical t test |  Satterthwaite-Welch t test |
## rejection rate    0.708              0.707                        0.708
```

(d)

```
testcompare(mean2 = 0.3, sd2 = sqrt(2))
```

```
##                z test |  classical t test |  Satterthwaite-Welch t test |
## rejection rate    0.562              0.451                        0.558
```

(e)

```
testcompare(n1 = 5, n2 = 10)
```

```
##                z test |  classical t test |  Satterthwaite-Welch t test |
## rejection rate    0.096              0.055                        0.059
```

```
testcompare(n1 = 5, n2 = 10, sd2 = sqrt(2))
```

```
##                z test |  classical t test |  Satterthwaite-Welch t test |
## rejection rate    0.096              0.043                         0.06
```

The z test has a rejection rate above 5% each time.

(f)

```
testcompare(n1 = 5, n2 = 10, exp = TRUE)
```

```
##                z test |  classical t test |  Satterthwaite-Welch t test |
## rejection rate    0.071              0.038                        0.039
```

Either of the t tests are better here, because the sample size is small and the data is not normal.