HW06_Sampathirao_A

Anvita Sampathirao

6/13/2019

R Markdown

#1a.

$$H_0 : P(Y\&X) = P(Y) * P(X)$$

$$H_1 : P(Y\&X) \neq P(Y) * P(X)$$

#1b.

$$\chi^2 \text{ test}$$

As, this test involves testing 2 variables that take categorical values like Unemployed/Employed and Non-College Grads/College Grads. Chi square test is a suitable statistical test when testing the independence of two means that summarize categorical variables. #1c. To calculate p-value, we need X^2 value.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where O is observed frequency E is the expected frequency

$$E = \frac{(RowTotal)(ColumnTotal)}{OverallTotal}$$

Also, Degree of freedom is:

$$df = (R-1)(C-1)$$

```
O<-matrix(c(11179,2720,187920,100305),2,2)
O
```

```
##       [,1]   [,2]
## [1,] 11179 187920
## [2,]  2720 100305
```

```
X_0<-sum(O[1,])
X_0
```

```
## [1] 199099
```

```
X_1<-sum(O[2,])
X_1
```

```
## [1] 103025
```

```
Y_0<-sum(O[,1])
Y_0
```

```
## [1] 13899
```

```r
Y_1<-sum(O[,2])
Y_1
```

```
## [1] 288225
```

```r
Total<-sum(O)
Total
```

```
## [1] 302124
```

```r
E_1<-X_0*Y_0/Total
E_2<-X_1*Y_0/Total
E_3<-X_0*Y_1/Total
E_4<-X_1*Y_1/Total
E<-matrix(c(E_1,E_2,E_3,E_4),2,2)
E
```

```
##          [,1]      [,2]
## [1,] 9159.408 189939.59
## [2,] 4739.592  98285.41
```

```r
chisquare<-sum((O-E)^2/E)
chisquare
```

```
## [1] 1368.851
```

```r
df<-(2-1)*(2-1)
df #degree of freedom
```

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

```r
1-pchisq(chisquare,df) #P-value of chi square test
```

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

Thus, we reject the null hypothesis as:

$$P - value \leq 0.05$$

#1d.

```r
chisq.test(O)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  O
## X-squared = 1368.2, df = 1, p-value < 2.2e-16
```

#2a.

$$H_0 : \mu_D = \mu_I = \mu_R$$

$$H_A : \text{At least one mean is different}$$

#2b.

F-test

Because we are testing the difference for more than 2 means, we wont be able to rewrite the null (difference between means) as a function of a single variable. Hence, F test that takes the ratio of variances between the groups and within the groups helps in giving us a single test statistic when analyzing the variance between more than 2 groups.

#2c.

$$BetweenVariance = \sum_{i=1}^{3} \frac{n_i * (\bar{y_i} - \bar{y})^2}{g-1}$$

$$BetweenVariance = \frac{302 * (43.0 - 44.1)^2 + 212 * (43.6 - 44.1)^2 + 278 * (45.8 - 44.1)^2}{3-1} = 610.92$$

$$WithinVariance = \sum_{i=1}^{3} \frac{(n_i - 1) * s_i^2}{N-g}$$

$$WithinVariance = \frac{(302-1) * 9.1^2 + (212-1) * 9.3^2 + (278-1) * 8.8^2}{792-3} = 81.90885$$

$$F = \frac{BetweenVariance}{WithinVariance} = \frac{610.92}{81.90885} = 7.458535$$

```
N<-792
g<-3
alpha<-0.10
df1<-g-1 #degree of freedom for numerator
df2<-N-g #degree of freedom for denominator
BetweenVariance<-(302*(43.0-44.1)^2+212*(43.6-44.1)^2+278*(45.8-44.1)^2)/df1
BetweenVariance
```

```
## [1] 610.92
```

```
WithinVariance<-((302-1)*9.1^2+(212-1)*9.3^2+(278-1)*8.8^2)/df2
WithinVariance
```

```
## [1] 81.90885
```

```
F_Calc<-BetweenVariance/WithinVariance
F_Calc #F-Calculated
```

```
## [1] 7.458535
```

```
f_Thr<-qf(1-alpha,df1,df2)
f_Thr #F-Threshold
```

```
## [1] 2.309318
```

Since,

$$F_{\text{Calculated}} > F_{\text{Threshold}}$$

We reject the null hypothesis.

```
1-pf(7.458535,df1,df2)
```

```
## [1] 0.0006180726
```

Also,

$$P - value \leq 0.05$$

Therefore,

we reject the null hypothesis at $\alpha = 0.10$

#2d.

```
set.seed(1)
Dem<-cbind("Democrat",rnorm(302,43.0,9.1))
Ind<-cbind("Independent",rnorm(212,43.6,9.3))
Rep<-cbind("Republican",rnorm(278,45.8,8.8))
ExitPoll<-data.frame(rbind(Dem,Ind,Rep),stringsAsFactors=FALSE)
colnames(ExitPoll)<-c("Party","Mean_AgeParty_Pair")
ExitPoll$Party<-as.factor(ExitPoll$Party)
ExitPoll$Mean_AgeParty_Pair<-as.numeric(ExitPoll$Mean_AgeParty_Pair)
head(ExitPoll)
```

```
##      Party Mean_AgeParty_Pair
## 1 Democrat           37.29927
## 2 Democrat           44.67115
## 3 Democrat           35.39578
## 4 Democrat           57.51706
## 5 Democrat           45.99852
## 6 Democrat           35.53374
```

```
aov.ex= aov(ExitPoll[,2]~ExitPoll[,1],data=ExitPoll)
summary(aov.ex)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## ExitPoll[, 1]   2    661   330.3   3.831 0.0221 *
## Residuals     789  68021    86.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$P - value \leq 0.05$$

Hence, we reject our null hypothesis of respective means of Democrat, Independent and Republican being equal.