HW_06_Gupta_S

Sumit Gupta

October 18, 2017

Problem 1. a. Based on the exit poll results, is age independent of Party ID or not? Conduct a chi-squared test by hand, showing each step in readably-formatted latex.

We will conduct a hypothesis test based on Chi-square testing to determine the age group's dependency on Party ID.

$H_0 = Age\ is\ independent\ of\ PartyID\ H_a = Age\ is\ not\ independent\ of\ PartyID$

Assuming null hypothesis, we can find expected value for each elemet of the sample table. Let's calculate the number of participants in our sample. $n = (86+52+61)+(72+51+74)+(73+55+70)+(71+54+73) = 792$

Expected values: $fe = \frac{(rowtotal)\times(columntotal)}{overalltotal}$

So, manually calculating using above formula we get the expected values as:

Age Group 18-29 30-44 45-59 60+ Democrat 75.79 75.1 75.24 75.24 Independent 53.06 52.27 52.83 52.83 Republican 69.69 68.9 69.46 69.46

Now we calculate our test statistic (chi-squared) as:

$\chi^2 = \sum \frac{(f_o-f_e)^2}{f_e}$

Thus we get, $\chi^2 = \frac{(86-75.88)^2}{75.88} + \frac{(72-75.1)^2}{75.1} + \frac{(73-75.5)^2}{75.5} + \frac{(71-75.5)^2}{75.5} + \frac{(52-53.26)^2}{53.26} + \frac{(51-52.73)^2}{52.73} + \frac{(55-53)^2}{53} + \frac{(54-53)^2}{53} + \frac{(61-69.85)^2}{69.85} + \frac{(74-69.15)^2}{69.15} + \frac{(70-69.5)^2}{69.5} + \frac{(73-69.5)^2}{69.5}$

$\chi^2 = 1.35+0.13+0.083+0.27+0.029+0.057+0.075+0.02+1.12+0.34+0.036+0.18 = 3.69$ So, $\chi^2 = 3.69$

Degrees of freedom: $df = (r-1)(c-1) = 2\times 3 = 6$

```
pchisq(3.69,6,lower.tail=FALSE)
```

```
## [1] 0.7185431
```

```
qchisq(0.95,6,lower.tail=FALSE)
```

```
## [1] 1.635383
```

Thus, from the computations our test statistic is smaller (0.71<1.63) and hence we fail to reject the null hypothesis.

b. Verify your results using R to conduct the test.

lets compute using R and verify:

```
AgeGroup_Party<-data.frame(AgeGroup_18to29=c(86,52,61),AgeGroup_30to44=c(72,51,74),AgeGroup_45to59=c(73
chisq.test(AgeGroup_Party)
```

```
##
##  Pearson's Chi-squared test
##
## data:  AgeGroup_Party
## X-squared = 3.6529, df = 6, p-value = 0.7235
```

Thus, we can see that the p-value (0.7235>0.05) and hence it is verified that Null hypothesis cannot be rejected.

Problem 2. a. Now test for independence using ANOVA (an F test). Your three groups are Democrats, Independents, and Republicans. The average age for a Democrat is 43.3, for an Independent it's 44.6, and for a Republican it's 45.1. The standard deviations of each are D: 9.1, I: 9.2, R: 9.2. The overall mean age is 44.2. Do the F test by hand, again showing each step.

Here, N total samples in G different groups of one same parameter x: $(\bar{x_1}, s_1, n_1), (\bar{x_1}, s_1, n_1), ..., (\bar{x_g}, s_g, n_g)$

$H_0 = \mu_1 = \mu_2 = ... = \mu_g = \bar{x}(average\,of\,all\,samples)$ $H_a = at\,least\,one\,group\,has\,different\,response\,to\,x$

$F - statistic = \frac{average\,variance\,between\,groups}{average\,variance\,within\,groups}$

N=total number of all observations G=the number of groups

$average\,variance\,between\,groups = \frac{n_1(\bar{x_1}-\bar{x})^2+...+n_G(\bar{x_G}-\bar{x})^2}{G-1}$ $average\,variance\,within\,groups = \frac{(n_1-1)\times(s_1^2)+...+(n_1-1)\times(s_1^2)}{N-G}$

Degrees of freedom: $df1 = G - 1$ $df2 = N - G$

Given data: Democrates(mean=43.3,s=9.1,n=302) Independent(mean=44.6,s=9.2,n=212) Republicans(mean=45.1,s=9.2,n=278) Overal Mean=44.2 N=792, G=3

$average\,variance\,between\,groups = \frac{302\times(43.3-44.2)^2+212\times(44.6-44.2)^2+278\times(45.1-44.2)^2}{3-1} = 251.86$

$average\,variance\,within\,groups = \frac{(302-1)\times(9.1^2)+(212-1)\times(9.2^2)+(278-1)\times(9.2^2)}{792-3} = 83.94$

$F - statistic = \frac{251.86}{83.94} = 3.0004$ $df1 = 3 - 1 = 2$ $df2 = 792 - 3 = 789$

```
pf(3.0004,2,789)
```

```
## [1] 0.9496646
```

```
qf(0.05,2,789)
```

```
## [1] 0.05129663
```

Thus, we can see that the test statistic is very marginally smaller. Also the p value (0.0512>0.05) is ver marginally greater than alpha. Hence, it is difficult to reject Null hypothesis indicating the means are equal for the three groups.

b.Check your results in R using simulated data. Generate a simulated dataset by creating three vectors: Democrats, Republicans, and Independents. Each vector should be a list of ages, each with a length equal to the number of Democrats, Independents, and Republicans in the table above, and the appropriate mean and sd based on 2.a (use rnorm to generate the vectors). Combine all three into a single dataframe with two variables: age, and a factor that specifies D, I, or R. Then conduct an F test using R's aov function on that data and compare the results to 2a. Do your results match 2a? If not, why not?

```
Democrats <- rnorm(302,43.3,9.1)
Independents <- rnorm(212,44.6,9.2)
Republicans <- rnorm(278,45.1,9.2)

#Combining into a single dataframe
DC <- cbind("D",Democrats)
ID <- cbind("I",Independents)
RC <- cbind("R",Republicans)
Age_Party <- rbind(DC,ID,RC)
Age_Party <- cbind(as.factor(Age_Party[,1]),as.numeric(Age_Party[,2]))
```

```
df <- data.frame(Age_Party)
aov.ex1 <- aov(df[,2]~df[,1], data = df )
summary(aov.ex1)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
```

```
## df[, 1]       1    349   348.7   4.257 0.0394 *
## Residuals   790  64712    81.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This gives us our p-value < 0.05 and hence we reject Null Hypothesis that the groups are equal. However, when we compare to the results of 2.a it does not match as the p-value in former is slightly greater than significance level. Also, the significant codes above indicate that the p-value is not that significant which makes us to wonder the accuracy of rejecting/accepting Null hypothesis in both cases (2.a and 2.b)