INSH5301 Intro Computational Statistics

Ali Banijamali

02/17/2020

```
# Library for drawing tables:
library(knitr)
```

1. You conduct an exit poll after an election, with the shown below:

	X18_29	X30_44	X45_59	above_60
Democrat	86	72	73	71
Independent	52	51	55	54
Republicn	61	74	70	73

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.

ANS.
Our Hypothesis are:
H_0 : *Age is independent from party*
H_a : *Age and party are dependent*

Now let's calculate f_e for cells, but before that I am using R to calculate the sums of rows and columns:

```
age.vs.party <- data.frame("18_29"=c(86,52,61),
                           "30_44"=c(72,51,74),
                           "45_59"=c(73,55,70),
                           "above_60"=c(71,54,73),
                           row.names=c("Democrat","Independent","Republicn"))

age.vs.party_with.sums <- rbind(age.vs.party,
                                colSums(age.vs.party, na.rm=F, dims=1))
age.vs.party_with.sums <- cbind(age.vs.party_with.sums,
                                rowSums(age.vs.party_with.sums, na.rm=FALSE, dims=1))
row.names(age.vs.party_with.sums) <- c("Democrat","Independent","Republicn", "Total")
colnames(age.vs.party_with.sums) <- c("18_29","30_44","45_59", "above_60", "Total")

# Drawing a table:
kable(age.vs.party_with.sums, format='markdown')
```

	18_29	30_44	45_59	above_60	Total
Democrat	86	72	73	71	302
Independent	52	51	55	54	212
Republicn	61	74	70	73	278
Total	199	197	198	198	792

Now we can move on to calculate the probabilities of individual variables:

$P_{democrat} = \frac{302}{792} = 0.3813131$
$P_{independent} = \frac{212}{792} = 0.2676768$
$P_{republican} = \frac{278}{792} = 0.3510101$

$P_{18-29} = \frac{199}{792} = 0.2512626$
$P_{30-44} = \frac{197}{792} = 0.2487374$
$P_{45-59} = \frac{198}{792} = 0.25$
$P_{60+} = \frac{198}{792} = 0.25$

Now we can calculate the expected probabilities and f_es for cells:
$P_{EXP\ (dem\ and\ 18-29)} = P_{dem} \times P_{18-29} = 0.3813131 * 0.2512626 = 0.09580972$
$f_e = P_{expected} \times total = 75.8813\ f_o = 86$

$P_{EXP\ (dem\ and\ 29-44)} = P_{dem} \times P_{29-44} = 0.3813131 * 0.2487374 = 0.09484683$
$f_e = P_{expected} \times total = 75.11869$
$f_o = 72$

$P_{EXP\ (dem\ and\ 45-59)} = P_{dem} \times P_{45-59} = 0.3813131 * 0.25 = 0.09532828$
$f_e = P_{expected} \times total = 75.5$
$f_o = 73$

$P_{EXP\ (dem\ and\ 60+)} = P_{dem} \times P_{60+} = 0.3813131 * 0.25 = 0.09532828$
$f_e = P_{expected} \times total = 75.5$
$f_o = 71$

$P_{EXP\ (ind\ and\ 18-29)} = P_{ind} \times P_{18-29} = 0.2676768 * 0.2512626 = 0.06725717$
$f_e = P_{expected} \times total = 53.26768$
$f_o = 52$

$P_{EXP\ (ind\ and\ 29-44)} = P_{ind} \times P_{29-44} = 0.2676768 * 0.2487374 = 0.06658123$
$f_e = P_{expected} \times total = 52.73233$
$f_o = 51$

$P_{EXP\ (ind\ and\ 45-59)} = P_{ind} \times P_{45-59} = 0.2676768 * 0.25 = 0.0669192$
$f_e = P_{expected} \times total = 53.00001$
$f_o = 55$

$P_{EXP\ (ind\ and\ 60+)} = P_{ind} \times P_{60+} = 0.2676768 * 0.25 = 0.0669192$
$f_e = P_{expected} \times total = 53.00001$
$f_o = 54$

$P_{EXP\ (rep\ and\ 18-29)} = P_{rep} \times P_{18-29} = 0.3510101 * 0.2512626 = 0.08819571$
$f_e = P_{expected} \times total = 69.851$
$f_o = 61$

$P_{EXP\ (rep\ and\ 29-44)} = P_{rep} \times P_{29-44} = 0.3510101 * 0.2487374 = 0.08730934$
$f_e = P_{expected} \times total = 69.149$
$f_o = 74$

$P_{EXP\ (rep\ and\ 45-59)} = P_{rep} \times P_{45-59} = 0.3510101 * 0.25 = 0.08775252$
$f_e = P_{expected} \times total = 69.5$
$f_o = 70$

$P_{EXP\ (rep\ and\ 60+)} = P_{rep} \times P_{60+} = 0.3510101 * 0.25 = 0.08775252$
$f_e = P_{expected} \times total = 69.5$
$f_o = 73$

Now we can calculte the chi squared according to the following formula:
$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} \approx 3.7$

Now, the degree of freedom is: $df = (r-1) \times (c-1) = 2 \times 3 = 6$

Based on this d.o.f and 95% threshold, we have:

```
qchisq(.95, df=6)
```

```
## [1] 12.59159
```

3.7 is way below 12.59159, therefore we can't reject the null hypothesis and that means that these variables are indeed independent.

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

ANS.

```
kable(age.vs.party, format='markdown')
```

	X18_29	X30_44	X45_59	above_60
Democrat	86	72	73	71
Independent	52	51	55	54
Republicn	61	74	70	73

```
chisq.test(age.vs.party)
```

```
Pearson's Chi-squared test
```

data: age.vs.party X-squared = 3.6529, df = 6, p-value = 0.7235

The results verifies our calculations and we can't reject the null hypothesis.

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.

ANS.
Let's start by defining the null and alternative hypothesis:
$H_o : \mu_1 = \mu_2 = ... = \mu_g$
$H_a : At\ least\ one\ group\ is\ different$

Now let's calculate the F-statistic:
F-statistic = $\frac{Average\ variance\ between\ groups}{Average\ variance\ within\ groups}$
Wehere:
Between Variance = $\frac{n_1(\bar{y_1}-\bar{y})^2+...+n_G(\bar{y_G}-\bar{y})^2}{G-1}$
and,
Within Variance = $\frac{(n_1-1)s_1^2+...+(n_G-1)s_G^2}{N-G}$
Overall mean $\bar{y} = 44.2$

BV = $\frac{302(43.3-44.2)^2+212(44-44.2)^2+278(45.1-44.2)^2}{3-1} = 239.14$

WV = $\frac{(302-1)*9.1^2+(212-1)*9.2^2+(278-1)*9.2^2}{792-3} = 83.95$

F-statistic = $\frac{239.14}{83.95} = 2.85$

Degrees of freedom are:

$df_1 = 3 - 1 = 2$
$df_2 = 792 - 3 = 789$

And finally finding the F threshold:

```
qf(0.95,2,789)
```

```
## [1] 3.007136
```

Our value for F-statistic (2.85) is smaller than 3.007136 and therefore not in the rejection region. Therefore again we can not reject the null hypothesis

2.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?

ANS.

```
set.seed(1)
# 3 vectors with ages and respective mean and sd:
dem <- data.frame(Party=as.factor("D"), Age=rnorm(302, 43.3, 9.1))
ind <- data.frame(Party=as.factor("I"), Age=rnorm(212, 44.6, 9.2))
rep <- data.frame(Party=as.factor("R"), Age=rnorm(278, 45.1, 9.2))

party.age <- rbind(dem, ind, rep)

anova <- aov(party.age[,2]~party.age[,1], data=party.age)
summary(anova)
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## party.age[, 1]    2    155   77.55   0.878  0.416
## Residuals       789  69730   88.38
```

As can be seen the p-value (0.416) is a lot larger than 0.05, therefore again the null hypothesis can't be rejected.
Between Variance is 77.55 and Within Variance is 88.38 and therefore the F-statistic is 77.55/88.38=0.878.
Our $df_1 = 2$ and $df_2 = 789$, are the same as part a. Although, the results could turn out in a way that we indeed could reject the null hypothesis. Since we are creating random datasets, we can expect to get random results.