Public Policy 529
Fall 2023 Problem Set #8

Francisco Brady

2023-11-20

Due Monday, November 20th

1. In a 2014 survey of Americans, respondents were asked "How important is it for an American to be able to speak English?" They were also asked about their political party identification. The following table shows a raw frequency distribution of these two variables:

Party Identification

Speak English	Democratic	Independent	Republican	Total
Very Important	401	172	312	885
Fairly Important	134	57	66	257
Not Important	51	12	13	76
Total	586	241	391	1218

(a) Calculate f_e, the expected frequency in each cell under the scenario that the variables are independent. Do you have sufficient sample size to perform a χ^2 test?

To find f_e we need to calculate the expected frequencies for each category and apply that to each cell in the table:

Party Identification: Expected Frequencies

Speak English	Democratic	Independent	Republican	Total
Very Important	425.78820	175.11080	284.10100	885
Fairly Important	123.64700	50.85140	82.50164	257
Not Important	36.56486	15.03777	24.39737	76
Total	586.00000	241.00000	391.00000	1218

Based on the expected frequencies, there are at least 5 expected observations per cell, so there is sufficient sample size to conduct a χ^2 test.

(b) Use the χ^2 table from Canvas or lecture slides to look up the critical value of χ^2 that would be necessary to reject the null hypothesis that the variables are independent ($\alpha = .05$). You will first have to find the correct degrees of freedom.

$$df = (\text{\# of rows - 1})(\text{\# of columns - 1}) = (3-1)(3-1) = 4$$

At 4 df, using $\alpha = 0.05$, the critical value of $\chi^2 = 9.49$.

(c) Calculate the χ^2 statistic from the data presented above. Can you reject the null hypothesis? What can you say about the p-value?

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

Party Identification: Expected Frequencies

Speak English	Democratic	Independent	Republican	Total
Very Important	$\frac{(401-425.7882)^2}{425.7882}$	$\frac{(172-175.1108)^2}{175.1108}$	$\frac{(312-284.101)^2}{284.101}$	885
Fairly Important	$\frac{(134-123.647)^2}{123.647}$	$\frac{(57-50.8514)^2}{50.8514}$	$\frac{(66-82.50164)^2}{82.50164}$	257
Not Important	$\frac{(51-36.56486)^2}{36.56486}$	$\frac{(12-15.03777)^2}{15.03777}$	$\frac{(13-24.39737)^2}{24.39737}$	76
Total	586	241	391	1218

$$\begin{aligned}\chi^2 = 1.44309978 + 0.86685976 + 5.69873006 +\\ 0.05526259 + 0.74344624 + 0.61365791 +\\ 2.73970947 + 3.30059042 + 5.32434615 = 20.7857\end{aligned}$$

The observed test χ^2 statistic is 20.7857. Since this exceeds the critical value of 9.49, we can reject the null hypothesis that Party Identification and the response to “How important is it for an American to speak English?” are independent. To find the p-value, we locate our test statistic along the same row we found our critical value. Using the table, the closest value is 18.47, so we can say that our p-value is: $p < .001$.

(d) You are informed by a statistician that the gamma statistic for this relationship is -0.21 with an ASE of 0.05. The Kendall’s tau-b statistic is -0.11 with an ASE of .026. Explain the meaning of these findings.

This suggests a mild negative relationship between the two variables. As party identification moves from Democratic to Republican, a participant is more likely to respond “Very Important” to the question “How important is it for an American to speak English?”.

2. Use the dataset `PEW_HigherEd_subset` for this question. The variable `JobSatis` asks how satisfied a person is with their job (very dissatisfied, somewhat dissatisfied, somewhat satisfied, very satisfied). The variable `OverQual` asks if the person feels overqualified for their job (not overqualified, overqualified).

(a) What are the measurement levels of these variables?

`OverQual` is a binary variable, with 0 indicating “Not overqualified”, and 1 indicating “Overqualified”.

`JobSatis` is an ordinal variable going from 0 to 3, with 0 indicating “Very dissatisfied” and 3 indicating “Very satisfied”.

```r
# note: code for 2.b:
jft <- table(PEW_HigherEd_subset$JobSatis, PEW_HigherEd_subset$OverQual)
# add margins but don't add margins when running chi square
rownames(jft) <- c('Very dissatisfied', 'Somewhat dissatisfied',
                   'Somewhat satisfied', 'Very satisfied')
colnames(jft) <- c('Not overqualified', 'Overqualified')
jft
```

(b) Make a table showing the joint frequency distribution of these two variables, with `JobSatis` forming the rows. Have your software report a χ^2 test for this table. Interpret the reported statistics. See lecture slides for code.

```
##
##                           Not overqualified Overqualified
##   Very dissatisfied                      23            46
##   Somewhat dissatisfied                  46            54
##   Somewhat satisfied                    260           173
##   Very satisfied                        432           219
```

```r
# code for chi sq test:
chisq.test(jft, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  jft
## X-squared = 39.551, df = 3, p-value = 1.327e-08
```

The output reports that the test statistic is 39.551 for these two variables. At the default $\alpha = 0.05$, this would support rejecting the null hypothesis that the two questions are independent.

(c) What is the critical value of the χ^2 statistic in the test that was just performed? The output indicates that the degrees of freedom for this test is 3, which we can verify with $(4-1)(2-1) = 3$. At the default level of $\alpha = .05$, the critical value would be: 7.81.

3. Use the dataset `anes2016subset` for this question. The variable `ScientistsTherm` contains the respondent's feeling thermometer score for scientists. The variable `PartyID` measures whether the respondent identifies as a Democrat, Republican, Independent, or Other. The variable `ReligImpt` indicates whether respondents say religion is an important part of their life (yes, no).

(a) Use the tabulate command with the summarize option to obtain the mean values of the variable `ScientistsTherm` across the categories of `PartyID`:
Stata: `tabulate PartyID, summarize(ScientistsTherm)`
R: `aggregate(ScientistsTherm ~ PartyID, data = anes2016, FUN = mean)`
Interpret the findings from this table.

```r
# note: this code is for question 3a:
aggregate(ScientistsTherm ~ PartyID, data = anes2016, FUN = mean)
```

```
##   PartyID ScientistsTherm
## 1       1        80.93805
## 2       2        72.72109
## 3       3        75.95913
## 4       5        74.23016
```

The table reports average feelings toward scientists by party identification. 1 = Democrat, 2 = Republican, 3 = Independent, 5 = Other Party. The highest mean feeling for Scientists is among Democrat-identified respondents.

(b) Use ANOVA to test whether these means are different from each other. Interpret the output to perform the significance tests. The correct command is:
Stata: `anova ScientistsTherm PartyID`
R: `summary(aov(ScientistsTherm ~ PartyID, data = anes2016))`

```
# note: this code is for question 3b:
summary(aov(ScientistsTherm ~ as.factor(PartyID), data = anes2016))
```

```
##                       Df  Sum Sq Mean Sq F value Pr(>F)
## as.factor(PartyID)     3   40043   13348   36.33 <2e-16 ***
## Residuals           3593 1320247     367
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 674 observations deleted due to missingness
```

The output reports that the F statistic is 36.33. The output also reports degrees of freedom for the between group (Party ID), which is 3, and the Residuals, which is the degrees of freedom for the within group means: 3593. Using an $\alpha = 0.05$, this would give a critical value of between 2.68 and 2.60. Since our F-statistic from the output is 36.33, we can reject the null hypothesis that the means are the same across groups. The output also reports a p-value, which in this case is: 0.0000000000000002, this is lower than $\alpha = 0.05$, so we can reject the null hypothesis.

(c) Now add `ReligImpt` to the ANOVA analysis, using a twoway ANOVA to test whether the category means of each variable are different from each other when controlling for the effect of the other variable.
Stata: `anova ScientistsTherm PartyID ReligImpt`
R: `summary(aov(ScientistsTherm ~ PartyID + ReligImpt, data = anes2016))`

```
# note: this code is for question 3c:
summary(aov(ScientistsTherm ~ as.factor(PartyID) + ReligImpt, data = anes2016))
```

```
##                       Df  Sum Sq Mean Sq F value Pr(>F)
## as.factor(PartyID)     3   39596   13199   36.85 <2e-16 ***
## ReligImpt              1   28520   28520   79.64 <2e-16 ***
## Residuals           3578 1281355     358
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 688 observations deleted due to missingness
```

In this output, the F-statistic increases to 36.85. The F-statistic for the `ReligImpt` variable is 79.64, and the p-values for both variables are `<2e-16`, which is much lower than $\alpha = 0.05$. Given this, it seems that controlling for Religious Importance does not change the interpretation of the ANOVA test on these means, and we can reject the null hypothesis that the mean thermometer value across Party Identifications is the same, even controlling for Religious Importance.