Lecture 30:The chi-square test for two-way tables

Chapter 22

Corinne Riddell (Instructor: Alan Hubbard)

November 6, 2023

Recap

- Last class we learned about the chi-square test (χ^2 test)
- We used the test to look at the distribution of one categorical variable to test the null hypothesis

$$H_0 : p_1 = \#_1, p_2 = \#_2, ..., p_k = \#_k$$

where $\#_1$, $\#_2$, ..., $\#_k$ were provided in the question or could be derived from percentages provided in the question.

- This test that we learned last class is called the **chi-square goodness of fit test**
- It asks, "how well do the expected counts 'fit' the observed counts?"

Recap of the chi-square goodness of fit test (for one categorical variable)

- The chi-square test statistic (Or, the "Old MacDonald" test statistic: "E-i, E-i, O!"):

$$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$$

Today's lecture

- We can also use the chi-square test to investigate the relationship between two categorical variables
- We will show that the form of the test statistic is the same! That is,

$$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$$

  for the chi-square test for two-way tables as well.

Think back to Chapter 5...

- In Chapter 5, we learned about two-way tables and talked about how to calculate the conditional probability of one variable given another.
- For example, what is the conditional probability of vaping among teens exposed to a JUUL advertisement vs. teens unexposed?
- Recall also the definition of **explanatory** and **response** variables. In the case of seeing JUUL advertisements and vaping, which was explanatory and which was response?

Hypotheses for the chi-square test for two categorical variables

- H_0 : Response and explanatory variables are independent.

Stated another way:

- H_0 : The probability distribution for vaping among teens who saw the advertisement is equal to the probability distribution among teens who did not see the advertisement

Hypotheses for the chi-square test for two categorical variables

Alternative hypothesis:

- H_a : Response and explanatory variables are dependent.
- H_a : The probability distribution for vaping among teens who saw the advertisement is different from the probability among teens who did not see the advertisement.
- The alternative hypothesis is not one-sided or two-sided. It is non-specific and allows for any kind of difference from the null probability distribution.
- That is, if you find a difference, you cannot tell from the chi-square test alone what is driving the difference.

Chi-square test of independence

- Just like last class, we compare observed cell counts (O_i) to expected cell counts (E_i), but this time we have a two-way table showing the distribution of data across two variables.
- This might remind you of Chapter 5, when we first learned about two-way tables.

Chapter 5 example: smoking and lung cancer

Group	Lung Cancer	No Lung Cancer	Row total
Smoker	12	238	250
Non-smoker	7	743	750
Column total	19	981	1000

- The inner four cells are the observed cell counts
- The outer row and column are the table **margins**
- The margins are important for the computations, so be sure to calculate the marginal counts if they aren't computed for you.

Example: smoking and lung cancer

Group	Lung Cancer	No Lung Cancer	Row total
Smoker	12	238	250
Non-smoker	7	743	750
Column total	19	981	1000

- What would these data look like under the null hypothesis of no association between smoking and lung cancer?
- That is, what are the expected counts under the null hypothesis?

Example: smoking and lung cancer

To help us get the expected counts, add the marginal percentages to the table and remove the data from the inner cells

Group	Lung Cancer	No Lung Cancer	Row total
Smoker	?	?	250 (25%)
Non-smoker	?	?	750 (75%)
Column total	19 (1.9%)	981 (98.1%)	1000

- Recall that if A and B are independent then $P(A\&B) = P(A)P(B)$. That is, if smoking and lung cancer are independent, then $P(S\&L) = P(S)P(L) = 0.25 * 0.019 = 0.0047 = 0.47\%$
- What is the expected count for the S&L cell under the null hypothesis?
  - 0.0047*1000 = 4.75

Example: smoking and lung cancer

- What is the expected count for the S&L cell under the null hypothesis?
  - 0.0047*1000 = 4.75

Group	Lung Cancer	No Lung Cancer	Row total
Smoker	4.75	245.25	250 (25%)
Non-smoker	14.25	735.75	750 (75%)
Column total	19 (1.9%)	981 (98.1%)	1000

- What are the expected counts for the other cells under H_0?
  - S' & L: $0.019 \times 0.75 \times 1000$
  - S & L': $0.981 \times 0.25 \times 1000$
  - S' & L': $0.981 \times 0.75 \times 1000$
- Note that once you compute two of the cells you can use subtraction from the marginal counts to get the other two values. Thus, only do as much calculation as you need and then get the rest by subtracting from the margins.

A trick for calculating the expected counts

- On the previous slides, we first calculated the marginal probabilities and multiplied them together and with the sample size to calculate the expected counts.
- We started with this calculation so you could see the intuition for why it worked.
- But there is a quicker way!:

$$E_i = \frac{\text{row total} \times \text{col total}}{\text{overall total}}$$

Worked calculations for the inner four cells:

- S&L = $(19 \times 250)/1000 = 4.75$
- S&L' = $(981 \times 250)/1000 = 245.25$
- S'&L = $(19 \times 750)/1000 = 14.25$
- S'&L' = $(981 \times 750)/1000 = 735.75$
- **Use this trick for faster calculation**

Compare E_i and O_i

Group	Lung Cancer	No Lung Cancer
Smoker	E=4.75 vs. O=12	E=245.25 vs. O=238
Non-smoker	E=14.25 vs. O=7	E=735.75 vs. O=743

- Think about the direction of the deviations. When is the observed higher than the expected? When is it the other way around? Does this jibe with the association you're expecting?

Calculate the chi-square test statistic

$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$

$\chi^2 = \frac{(4.75-12)^2}{4.75} + \frac{(14.25-7)^2}{14.25} + \frac{(245.25-238)^2}{245.25} + \frac{(735.75-743)^2}{735.75}$

$\chi^2 = 11.06579 + 3.688596 + 0.2143221 + 0.07144071 = 15.04015$

Calculate the degrees of freedom

- Like last class, we need a degrees of freedom for the test statistic.
- When we only had one variable the degrees of freedom equalled $k - 1$
- Here we have two variables. The degrees of freedom equals $(r - 1)(c - 1)$, where r is the number of inner row cells and c is the number of inner column cells (here $r = 2$ and $c = 2$)
- For these data, df = (2-1)(2-1) = 1

Calculate the p-value for the chi-square test

```
pchisq(q = 15.04015, df = 1, lower.tail = F) #df = (2-1)(2-1) = 1
```

```
## [1] 0.0001052481
```

- **Remember for the chi-squared test we always do an upper tail test!**

Interpret the p-value: Assuming no association between smoking and lung cancer, there is a 0.01% chance of the chi-square value we calculated or a larger one. This probability is small enough that there is evidence in favor of the alternative hypothesis of dependence between smoking and lung cancer.

Chi-square test of independence in R

To compute the chi-square test in R, we need to first put this two-way table into a data frame:

```
library(tibble)
two_way <- tribble(~ smoking, ~ non_smoking,
                   12,        238, #row for lung cancer
                   7,         743) #row for no lung cancer
```

Chi-square test of independence in R

Then, we use `chisq.test()`. We set `correct=F` to get the same value as what we calculated by hand:

```
chisq.test(two_way, correct = F) #not using Yates' correction for continuity
```

```
## Warning in chisq.test(two_way, correct = F): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  two_way
## X-squared = 15.04, df = 1, p-value = 0.0001052
```

Chi-square test of independence in R

Compare to the result where `correct = T` (the default):

```
chisq.test(two_way, correct = T) #using Yates' correction for continuity
```

```
## Warning in chisq.test(two_way, correct = T): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  two_way
## X-squared = 13.037, df = 1, p-value = 0.0003054
```

- A common practice is to incorporate the Yate's continuity correction when n < 100 or any $O_i < 10$.
- Reference

Expected counts conditions for the chi-square test of independence

- $E_i \geq 5$ for at least 80% of the cells
- All $E_i > 1$
- If table is 2X2, all four cells need $E_i \geq 5$

Statistical assumptions for the chi-square test of independence

Must have either data arising from:

- Independent SRSs from ≥ 2 population, with each individual classified according to one category (i.e., each individual can only belong to one cell in the table so the categories need to be mutually exclusive)
- A single SRS, with each individual classified according to each of two categorical variables.

Relationship between the chi-square test and the two-sample z test

- Recall the null hypothesis for the two sample z test:

$$H_0 : p_1 = p_2$$

For this test, the variable of interest is binary and can be summarized in a two-way table:

Group	Success	Failure
Group 1	#	#
Group 2	#	#

Relationship between the chi-square test and the two-sample z test

- If you were to calculate the z-test statistic using these data and call it z, then z^2 would be equivalent to the χ^2 test statistic for these same data.
- Thus, the p-value for the **two-sided** z-test and the chi-squared test are the same.
- When you have data that looks like this, you may want to use a z-test because you can test a **one-sided** alternative, which you cannot test using the chi-squared test.

Another example: brain trauma among individuals who did and did not play contact sports

- Chronic traumatic encephalopathy (CTE) is a progressive neurodegenerative disorder caused by repetitive brain trauma.
- A case-control study examined the brains of deceased men with a diagnosed non-genetic neurodegenerative disorder.
- The study compared individuals with and without a history of participation in amateur contact sports. Of the 66 men who had participated in contact sport, 21 had CTE. None of the 132 control men who had not participated in a contact sport had CTE.

Conduct all steps of the chi-squared test based on these data.

1. Make the two-way table.
2. Calculate the expected values.
3. Calculate the test statistic.
4. Calculate the degrees of freedom and p-value.
5. Interpret the p-value and assess the evidence.

Also: assess whether the sample size conditions are met to conduct the test.

Another example: HPV Status and age group

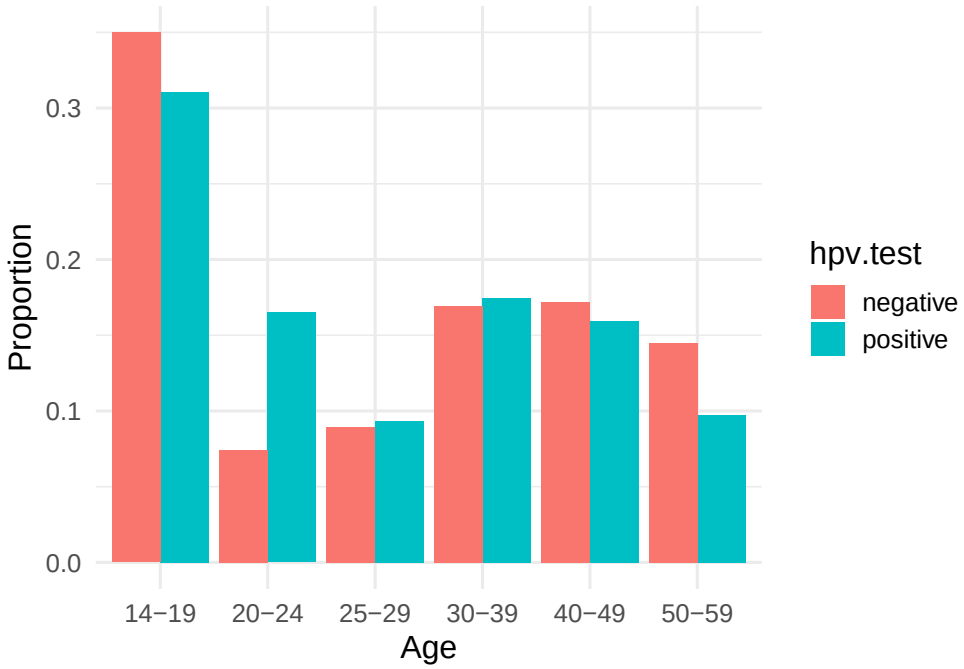
Suppose you had these data of HPV status vs. age group.

Age Group	HPV +	HPV -	Row total
14-19	160	492	652 (33.9%)
20-24	85	104	189 (9.8%)
25-29	48	126	174 (9.1%)
30-39	90	238	328 (17.1%)
40-49	82	242	324 (16.9%)
50-59	50	204	254 (13.2%)
Col total	515 (26.8%)	1406 (73.2%)	1921

- Which variable is explanatory and which is response?
- Can you formulate a null and alternative hypothesis using these data?

Welcome back to the dodged histogram

- Recall that we used dodged histograms to compare the conditional distribution of one variable across the levels of another variable.
- These plots are useful to make before we conduct the hypothesis test. Is there visual evidence of a difference between the conditional distribution of age by HPV status?



Example: HPV Status and age group

- Conduct all stages of the chi-square hypothesis test for independence (state the null and alternative hypotheses, calculate the test statistic, calculate the degrees of freedom and the p-value, interpret the p-value and assess whether there is evidence against the null in favor of the alternative.)

Let's do it in R

```
# make contigency table from data above, in this case I do
# rows age, and columns hpv (positive, negative)
# get counts by age for the negatives and positives
negs <- (hpv.data%>%filter(hpv.test=="negative"))$number.of.women
postvs <- (hpv.data%>%filter(hpv.test=="positive"))$number.of.women
# Get ages to label
ages <- (hpv.data%>%filter(hpv.test=="negative"))$age.group
# Make a table where rows are ages, columns HPV status and entries counts
tble <- cbind(postvs,negs)
# Give rownames to table
rownames(tble) <- ages
tble
```

```
##       postvs negs
## 14-19    160  492
## 20-24     85  104
## 25-29     48  126
## 30-39     90  238
## 40-49     82  242
## 50-59     50  204
```

```r
# Before chi-square test (can use default which is correct=T)
chisq.test(tble)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tble
## X-squared = 40.554, df = 5, p-value = 1.155e-07
```