L33: Extending the Chi-square to two way tables

April 24, 2020

Announcements

Statistics is Everywhere

Air pollution regulations

Opinion

Don't Let a Killer Pollutant Loose

The Trump administration is moving to ease standards on a particularly deadly air contaminant.

By John Balmes

Dr. Balmes is a medical professor and member of the California Air Resources Board.

April 14, 2019

Air pollution regulations

from the NYtimes April 14 2019 op-ed:

Last week, a recently reconstituted panel of science advisers to the E.P.A., the Clean Air Scientific Advisory Committee, sharply questioned the agency's longstanding position that particulate pollution is causally linked with premature death, and it called for a new assessment of the pollutant. In a letter to the E.P.A. administrator, Andrew Wheeler, the committee's chairman, Louis Anthony Cox Jr., said the agency had not provided "a sufficiently comprehensive, systematic assessment of the available science."

Air pollution regulations

The Environmental Protection Agency's own website said at the time: "Numerous scientific studies have linked particle pollution exposure to a variety of problems, including: premature death in people with heart or lung disease, nonfatal heart attacks, irregular heartbeat, aggravated asthma, decreased lung function, increased respiratory symptoms."

Air pollution regulations

All Epidemiologists have been removed from the panel and epidemiologic evidence is being dismissed.

What is the argument against Epidemiology here?

In our framework (PPDAC) what kind of a problem is this?

How would you approach this problem? Would it be possible to do a randomized controlled trial in this case?

Air pollution regulations

Example study from JAMA 2017 Dec 26; 318(24): 2446–2456. Association of Short-Term Exposure to Air Pollution with Mortality in Older Adults, Di et al.

- entire Medicare population from January 1, 2000, to December 31, 2012, residing in 39,182 zip codes
- looked at mortality on days when air pollution was higher vs lower

Air pollution regulations

Table 2

Relative Risk and Absolute Risk Difference of Daily Mortality Associated with Each 10 $\mu g/m^3$ Increase in $PM_{2.5}$ and Each 10 ppb Increase in Ozone

	Relative Risk (Percentage Change)		Absolute Risk Difference in Daily Mortality Rates (No. Per 1 Million Persons at Risk Per Day)[a]	
Air Pollutant	**$PM_{2.5}$**	**Ozone[b]**	**$PM_{2.5}$**	**Ozone[b]**
Main Analysis[c]	1.05% (0.95%, 1.15%)	0.51% (0.41%, 0.61%)	1.42 (1.29, 1.56)	0.66 (0.53, 0.78)

Recap of Chi-squared

Recap

- Last class we learned about the chi-square 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.

- This test is called the chi-square goodness of fit test
  - How good 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 McDonald" test statistic: "E-i, E-i, O!"):

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

Mea culpa

In last lecture I mistakenly described the chi-squared as a parametric test.

While there are some uses of the chi-squared that DO require assumptions about the distribution of the underlying population values, the goodness of fit test actually does not require that there be a specific data-generating process - so the chi-squared goodness of fit can be considered a NON-parametric test.

Today's lecture

- We can also use the chi-square test to investigate the relationship between two categorical variables
- The form of the test statistic is the same!

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 lung cancer among smokers vs. among non-smokers?
- Recall also the definition of explanatory and response variables. In the case of smoking and lung cancer, 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 lung cancer among smokers is equal to the probability distribution among non-smokers
- If you remember our probability independence rules P(A|B)=P(A)... how does this apply here?

Alternative hypothesis:

- H_a : Response and explanatory variables are dependent.
- H_a : The probability distribution for lung cancer among smokers is different from the probability among non-smokers.
  - The alternative hypothesis is not one-sided or two-sided. It is non-specific and allows for any kind of difference from the null. Does this mean we look at 2 sides of the distribution?

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.

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 conditions are met to conduct the test.

Sample size 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 populations, 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.

Example : gastroenteritis outbreak

From Gross et al. Public health reports vol 104, March-April 1989, 164-169

Group	Sandwich	No Sandwich	Row total
Ill	109	4	113
Not Ill	116	34	150
Column total	225	38	263

- 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 : gastroenteritis outbreak

Group	Sandwich	No Sandwich	Row total
Ill	109	4	113
Not Ill	116	34	150
Column total	225	38	263

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

Example : gastroenteritis outbreak

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

Group	Sandwich	No Sandwich	Row total
Ill	?	?	113 (43%)
Not Ill	?	?	150 (57%)
Column total	225 (85.6%)	38 (14.4%)	263

- Recall that if A and B are independent then $P(A\&B) = P(A)P(B)$. That is, if sandwiches and illness are independent, then $P(Sandwich\&Illness) = P(S)P(I) = .855 * .43 = .368 = 36.8\%$
- What is the expected count for the S&I cell under the null hypothesis?
  - 0.368*263 = 96.7

Example : gastroenteritis outbreak

- What is the expected count for the S&I cell under the null hypothesis?
  - 0.368*263 = 96.7

Group	Sandwich	No Sandwich	Row total
Ill	96.7	16.3	113 (43%)
Not Ill	128.3	21.7	150 (57%)
Column total	225 (85.6%)	38 (14.4%)	263

- What are the expected counts for the other cells under H_0?
  - S' & I: $0.144 \times 0.43 \times 263$
  - S & I': $0.856 \times 0.57 \times 263$
  - S' & I': $0.144 \times 0.57 \times 263$
- 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!:

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

A trick for calculating the expected counts

Worked calculations:

- S&I = 225*113/263 = 96.7
- S&I' = 225*150/263 = 128.3
- S'&I = 38*113/263 = 16.3
- S'&I' = 38*150/263 = 21.7
- Use this trick for faster calculation

Compare E_i and O_i

Group	Sandwich	No Sandwich
Ill	E=96.7 vs. O=109	E=16.3 vs. O=4
Not Ill	E=128.3 vs. O=116	E=21.7 vs. O=34

- 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

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

$$\chi^2 = \frac{(96.7-109)^2}{96.7} + \frac{(16.3-4)^2}{16.3} + \frac{(128.3-116)^2}{128.3} + \frac{(21.7-34)^2}{21.7}$$

$$\chi^2 = 1.5645 + 9.2816 + 1.1792 + 6.972 = 18.9973$$

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 equaled $k - 1$
- 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 = 18.9972, df = 1, lower.tail = F) #df = (2-1)(2-1) = 1
```

```
## [1] 1.309104e-05
```

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

Interpret the p-value: Assuming no association between sandwiches and illness, there is less than 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 that there is a relationship between sandwiches and illness.

Chi-squared test of independence in R

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)
```

```
## Warning: package 'tibble' was built under R version 3.5.3
```

```
two_way <- tribble(~ sandwich, ~ nosandwich,
                   109,         4, #row for Illness
                   116,         34) #row for no Illness
```

Chi-square test of independence in R

Then, we use `chisq.test()`. We set `correct=F` to get a value closer to what we calculated by hand - there will be some differences here because of rounding:

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

```
##
##  Pearson's Chi-squared test
##
## data:  two_way
## X-squared = 19.074, df = 1, p-value = 1.257e-05
```

Yates' continuity correction

Continuity correction

- The χ^2 is a continuous distribution and we are using discrete observations to estimate a χ^2 value.
- When there are many degrees of freedom and/or a large number of observations, this is a reasonable approximation
- In a 2x2 table (df=1) with a small sample size this may be less reasonable.
- The correction looks like this

$$\chi^2 = \sum_{i=1}^{k} \frac{(|E_i - O_i| - 0.5)^2}{E_i}$$

What do you thing this will do to the χ^2 value?

Chi-square test of independence in R

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

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

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  two_way
## X-squared = 17.558, df = 1, p-value = 2.786e-05
```

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

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

- The topic of next week's lab.

Extending the 2 X 2 to a more generic R X C

Extending the 2 X 2 to a more generic R X C

- we have looked at 2 x 2 as an example of how you would compare two categorical variables
- 2 X 2 tables are common as many variables that we look at are classified as binary
- however the chi-squared test works the same way for variables with more than 2 categories

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

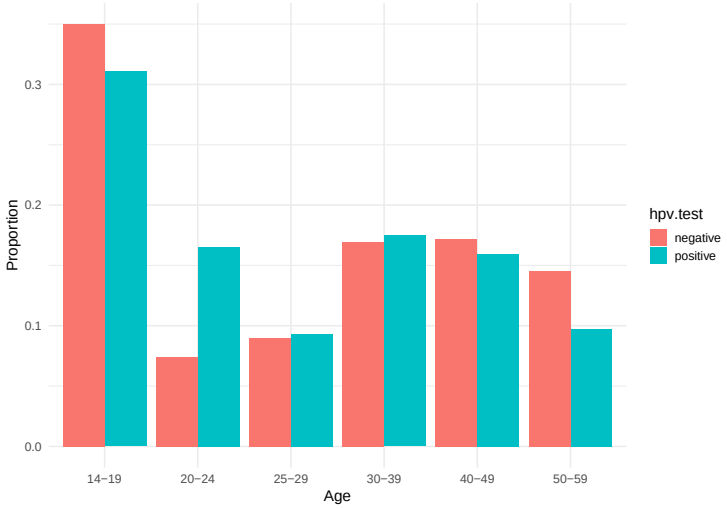
- Recall that we used dodged histograms to compare the conditional distribution of one variable across the levels of another variable.
- These plots are useful to make before we conduct the hypothesis test.

Remember the syntax: geom_bar(aes(fill = outcome), stat = "identity", position = "dodge")

The "identity" option tells R that the values are already calculated

Welcome back to the dodged histogram

Is there visual evidence of a difference between the conditional distribution of HPV status by age group?



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.)

Example: HPV Status and age group

```
pchisq(40.55353, df=5, lower.tail=F)
```

```
## [1] 1.154754e-07
```

Example: HPV Status and age group

```
library(tibble)
n_way <- tribble(~ HPV, ~ noHPV,
                 160,492,
                 85,104,
                 48,126,
                 90,238,
                 82,242,
                 50,204)
chisq.test(n_way, correct=F)
```

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

Parting humor, courtesy of the xkcd