L32: 2x2 tables, Epidemiologic terms and the chi-squared goodness of fit

April 22, 2020

Goals for today

- Introducing some terms from Epidemiology
- Goodness of fit: looking at one variable with multiple categories
- Introduce the chi-squared

An Interlude for Epidemiology

What is a risk?

P(event) in some defined time period such as:

- probability of acquiring malaria in a transmission season
- probability of death in the five years following a diagnosis
- probability of developing type 2 diabetes in a lifetime

Two by two table

These numbers represent the probability of ever dying from lung cancer in men over the age of 35

Group	Lung Cancer	No Lung Cancer	Total
Smoker	13	4987	5000
non-smoker	1	4999	5000
Total	14	9986	10000

So what is the risk of Lung cancer?

Marginal probability of Lung Cancer

Group	Lung Cancer	No Lung Cancer	Total
Smoker	13	4987	5000
non-smoker	1	4999	5000
Total	14	9986	10000

The marginal probability of Lung Cancer is:

$$P(LungCancer) = \frac{14}{10,000} = 0.0014$$

This is the lifetime risk of death from lung cancer among men over the age of 35

Comparing groups

We are very often interested in comparing the risk in two groups based on some exposure of interest:

P(Disease | exposure) vs. P(Disease | no exposure)

remember that we could also write the probability of disease in the unexposed group as:

$P(Disease|exposure^C)$ or $P(Disease|\overline{exposure})$

Probability of Lung Cancer conditional on smoking status

Group	Lung Cancer	No Lung Cancer	Total
Smoker	13	4987	5000
non-smoker	1	4999	5000
Total	14	9986	10000

$$P(D|E) = P(LungCancer|Smoker) = \frac{13}{5000} = 0.0026$$

$$P(D|\overline{E}) = P(LungCancer|\overline{smoker}) = \frac{1}{5000} = 0.0002$$

Absolute vs relative

When we make comparisons we usually do this in one of two ways:

- absolute difference
- relative difference

Which one of these comparisons have we covered in hypothesis testing?

Absolute: Example Lung cancer among men in the U.S.

These numbers represent the probability of ever dying from lung cancer in men over the age of 35

Group	Lung Cancer	No Lung Cancer	Total
Smoker	13	4987	5000
non-smoker	1	4999	5000
Total	14	9986	10000

What is the absolute difference in risk?

Risk Difference (RD) = P(LC | Smoker) - P(LC | non-smoker) = ?

Absolute: Example Lung cancer among men in the U.S.

$$RD = P(LungCancer|Smoker) - P(LungCancer|\overline{smoker}) = 0.0026 - 0.0002 = 0.0024$$

Relative difference

Another common way you will see differences presented is in relative terms:

- Risk ratio

$$RR = \frac{P(D|E)}{P(D|\overline{E})}$$

- Odds ratio

$$OR = \frac{\frac{P(D|E)}{1-P(D|E)}}{\frac{P(D|\overline{E})}{1-P(D|\overline{E})}}$$

Absolute: Example Lung cancer among men in the U.S.

Group	Lung Cancer	No Lung Cancer	Total
Smoker	13	4987	5000
non-smoker	1	4999	5000
Total	14	9986	10000

What are the RR and OR for these data?

Relative difference

Another common way you will see differences presented is in relative terms:

- Risk ratio

$$RR = \frac{P(D|E)}{P(D|\overline{E})} = \frac{0.0026}{0.0002} = 13$$

- Odds ratio

$$OR = \frac{\frac{P(D|E)}{1-P(D|E)}}{\frac{P(D|\overline{E})}{1-P(D|\overline{E})}} = \frac{\frac{0.0026}{.9974}}{\frac{0.0002}{.9998}} = \frac{0.00260678}{0.00020004} = 13.03$$

One variable with multiple categories

One categorical variable with more than 2 categories

- With one continuous variable we tested whether the mean was equal to a hypothesized null (Z or one sample T)
- With one categorical variable with two categories (binary, yes/no) we tested that the proportion was equal to a hypothesized null (one sample test of proportions)
- What do we do with a categorical variable when there are more than 2 categories?

One categorical variable with more than 2 categories

The general pattern we will follow for these types of variables is:

- estimate how many observations we would expect in each category under our null hypothesis
- compare the number of observations in each category to the expected value
- summarize these differences and compare them to a theoretical distribution

Jury Selection example

Suppose that the following number of people were selected for jury duty in the previous year, in a county where jury selection was supposed to be random.

Ethnicity	White	Black	Latinx	Asian	Other	Total
Number selected	1920	347	19	84	130	2500

You read online about concerns that jury was not selected randomly. How can you test this evidence?

- Example derived from this video.

Jury Selection example

Consider the distribution of race/ethnicity in the county overall:

Ethnicity	White	Black	Latinx	Asian	Other	Total
% in the population	42.2%	10.3%	25.1%	17.1%	5.3%	100%

How do we determine the counts that are expected (E) under the assumption that selection was random?:

Ethnicity	White	Black	Latinx	Asian	Other	Total
Expected count						2500

Jury Selection example

Ethnicity	White	Black	Latinx	Asian	Other	Total
% in the population	42.2%	10.3%	25.1%	17.1%	5.3%	100%

- To fill in the table, multiple the total size of the jury by the % of the population of each race/ethnicity:

Expected counts under the assumption that selection is random from the county:

Ethnicity	White	Black	Latinx	Asian	Other	Total
Expected count =	2500×0.422 1055	2500×0.103 257.5	2500×0.251 627.5	2500×0.171 427.5	2500×0.053 132.5	2500 2500

Jury Selection example

How far off does the observed counts of race/ethnicities in the sample differ from what we would expect if the jury had been selected randomly?

Jury Selection example

Here are the counts we observed (O):

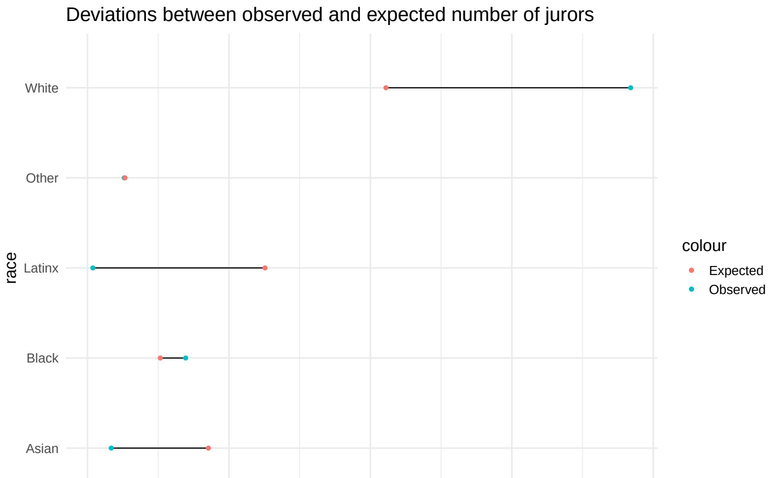
Ethnicity	White	Black	Latinx	Asian	Other	Total
Observed count	1920	347	19	84	130	2500

Which we can compare to our expected (E):

Ethnicity	White	Black	Latinx	Asian	Other	Total
Expected count	1055	257.5	627.5	427.5	132.5	2500

Jury Selection example

This plot shows the deviations between the observed and expected number of jurors. What is the chance of observed deviations of these magnitudes (or larger) under the null hypothesis?



Jury Selection example

- Recall the usual form of the test statistic:

$$\frac{estimate - null}{SE}$$

- We want an estimate that somehow quantifies how different the observed counts (O) are from the expected counts (E) across the 5 race/ethnicities.

The Chi-square test statistic

The chi^2 test statistic quantifies the magnitude of the difference between observed and expected counts under the null hypothesis. It looks like this:

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

- k is the number of cells in the table. Here, k is the number of race/ethnicity groups. That is, $k = 5$
- O_i is the observed count for the i^{th} group (here race/ethnicity)
- E_i is the expected count for the i^{th} group
- χ^2 is a distribution, like t or Normal.

The Chi-square test statistic

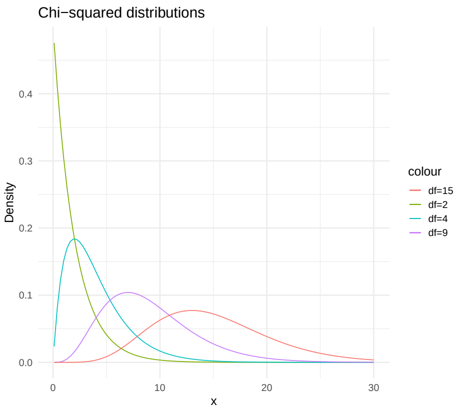
$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- The numerator measures the squared deviations between the observed (O) and expected (E) values. Bigger deviations will make the test statistic larger (which means that its corresponding p-value will be smaller)
- The denominator makes this magnitude *relative* to what we expect. This adjusts for the different magnitude of expected counts. For example, with our example, we would *expect* the number of white jurors to be close to 1055, but we would expect the number of Latinx jurors to be close to 628. Therefore, we divide by these expectations such that a difference of 100 fewer Latinx jurors than expected counts for more than a difference of 100 few white jurors.

The Chi-Square distribution

The Chi-square distribution

The chi-square distribution is a new distribution to us. Like the t-distribution, the chi-square distribution only has one parameter: a degrees of freedom. The degrees of freedom is equal to the number of groups (here, race/ethnicities) - 1. Or, $df = k - 1$.



The shape of the Chi-square

- As the `df` is increased, the distribution's central tendency moves to the right.
- This means that there will be more probability out in the right tail when the degrees of freedom is higher.
- The chi-square distribution is also positive. We only ever compute upper tail probabilities for the chi-square test because there is only one form to the H_a.

Back to the jury example

State the null and alternative hyptoheses.

- The null hypothesis is that the proportions of each race/ethnicity in the jury pool is the same as the proportion of each group in the county. That is:

$H_0 : p_{white} = 42.2\%,\ p_{black} = 10.3\%,\ p_{latinx} = 25.1\%,\ p_{asian} = 17.1\%,\ p_{other} = 5.3\%$

H_a : At least one of p_k is different than specified in H_0, for k being one of white, black, latinx, asian, or other.

Back to the jury example

Calculate the chi-square statistic using the jury data.

Ethnicity	White	Black	Latinx	Asian	Other	Total
O	1920	347	19	84	130	2500
E	1055	257.5	627.5	427.5	132.5	2500

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

$\chi^2 = \frac{(1920-1055)}{1055}^2 + \frac{(347-257.5)}{257.5}^2 + \frac{(19-627.5)}{627.5}^2 + \frac{(84-427.5)}{427.5}^2 + \frac{(130-132.5)}{132.5}^2$

$\chi^2 = 709.218 + 31.10777 + 590.0753 + 276.0053 + 0.04716981$

$\chi^2 = 1606.454$

Back to the jury example

Calculate the p-value (what is the approprate degrees of freedom?).

```
pchisq(q = 1606.454, df = 4,lower.tail = F)
```

```
## [1] 0
```

The probability of seeing this pool of people chosen for jury duty under the null hypothesis of random sampling from the county is so small that R rounded the p-value to 0!

Chi-square test in R

Run the chi-square test using the `chisq.test` command in R.

```
chisq.test(x = c(1920, 347, 19, 84, 130), # x is vector of observed counts
           p = c(.422, .103, .251, .171, .053)) # p is probability under the
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(1920, 347, 19, 84, 130)
## X-squared = 1606.5, df = 4, p-value < 2.2e-16
```

Interpretation

- Which race/ethnicities appear to deviate the most from what was expected under the null hypothesis?
  - Compare the proportion observed vs. proportion expected
  - Compare the count observed vs. the count expected
  - Compare the 5 contributions to the chi-square test from each race/ethnicity. We see that whites, Latinx, and Asians contribute the most to the χ^2 statistic. This agrees with what we saw in the data visualization in terms of the size of the gaps between observed and expected counts.

Example 2: Births by day of the week (Ex. 21.7)

A random sample of 700 births from local records shows the distribution across the days of the weeks:

Day	M	T	W	Th	F	Sa	Su
Births	110	124	104	94	112	72	84

Is there evidence that the proportion of births occurring on any given day of the week is not random?

Example 2: Births by day of the week (Ex. 21.7)

State the null and alternative hypotheses

$H_0 : p_1 = 1/7,\ p_2 = 1/7,\ p_3 = 1/7,\ p_4 = 1/7,\ p_5 = 1/7,\ p_6 = 1/7,\ p_7 = 1/7,\ .$
Written another way:

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = 1/7$$

H_a : At least one of these p_k differ from $1/7$. Or: not all p_k equal $1/7$.

Example 2: Births by day of the week (Ex. 21.7)

Calculate the expected counts under H_0

Day	M	T	W	Th	F	Sa	Su
Expected births	?	?	?	?	?	?	?

Example 2: Births by day of the week (Ex. 21.7)

Calculate the expected counts under H_0

- Use the fact that the total number of births equaled 700. Then $700*(1/7) = 100$. We would expect to see around 100 births on each day if the births occurring randomly over the course of the week.

Day	M	T	W	Th	F	Sa	Su
Expected births	100	100	100	100	100	100	100

Example 2: Births by day of the week (Ex. 21.7)

Calculate the chi^2 test statistic

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

$\chi^2 =$
$\frac{(110-100)^2}{100} + \frac{(124-100)^2}{100} + \frac{(104-100)^2}{100} + \frac{(94-100)^2}{100} + \frac{(112-100)^2}{100} + \frac{(72-100)^2}{100} + \frac{(84-100)^2}{100}$

$\chi^2 = 1 + 5.76 + 0.16 + 0.36 + 1.44 + 7.84 + 2.56$

$\chi^2 = 19.12$

- Based on the individual contributions of each day to the chi-square statistic, which days were most different from the expected value under H_0?

Example 2: Births by day of the week (Ex. 21.7)

Calculate the p-value

```
pchisq(q = 19.12, df = 6, lower.tail = F)
```

```
## [1] 0.003965699
```

Interpret the p-value

Based on a p-value of 0.39%, there is very strong evidence against the null hypothesis in favor of an alternative hypothesis where the proportion of births across the seven days of the week are not evenly distributed.

Example 3: cheating at dice?

Suppose there is a game in which the objective is to roll sixes as possible using 3 die. Over 100 rolls, one of the players seems to be winning quite often, we see the following

Number of 6s	0	1	2	3
Observed rolls	47	35	15	3

We suspect they are using a loaded die or cheating in some way.

Are they cheating? Or just lucky (within the bounds of chance)?

Example derived from this site

Example 3: cheating at dice?

What would we expect?
The rolls of dice should follow a binomial distribution (# of successes in # trials)

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

What is P here? What is K?

Example 3: cheating at dice?

Remember dbinom?

'dbinom(#successes,size,probability of success)'

This function calculates the probability of observing x successes when $X \sim Binom(n, p)$

```
Expect_0<-dbinom(0,size=3,prob=0.166666667)
Expect_1<-dbinom(1,size=3,prob=0.166666667)
Expect_2<-dbinom(2,size=3,prob=0.166666667)
Expect_3<-dbinom(3,size=3,prob=0.166666667)
Expected<-c(Expect_0,Expect_1,Expect_2,Expect_3)
Expected
```

```
## [1] 0.57870370 0.34722222 0.06944444 0.00462963
```

Example 3: cheating at dice?

Number of 6s	0	1	2	3
Observed rolls	47	35	15	3
Expected rolls	57.9	34.7	6.9	0.46

Exampe 3: cheating at dice?

```
chisq.test(x = c(47,35,15,3), # x is vector of observed counts
           p = Expected) # p is probability under the null
```

```
## Warning in chisq.test(x = c(47, 35, 15, 3), p = Expected): Chi-squared
## approximation may be incorrect
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(47, 35, 15, 3)
## X-squared = 25.292, df = 3, p-value = 1.342e-05
```

Conditions to perform a chi-square test

- Fixed n of observations
- All observations are independent of one another. What does this mean in the first example? In the second example?
- Each observation falls into just one of the k mutually exclusive categories
- The probability of a given outcome is the same for each observation.

Counts requirement

- At least 80% of the cells have 5 or more observations ($O_i \geq 5$ for $\geq 80\%$ of the cells)
- All k cells have expected counts > 1 ($E_i > 1$)

Parting humor, courtesy of the Comedian Erik Tanouye