Chapter 21: Inference about a categorical variable with > 2 levels

Corinne Riddell

November 6, 2020

### Motivation

* When the data is binary (i.e., a categorical variable with only two levels), we know how to make confidence intervals and conduct hypothesis tests for (one sample) and for (two sample)
* What do we do for categorical variables with > 2 levels?

### 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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ethinicity | White | Black | Latinx | Asian | Other | Total |
| Number selected | 1920 | 347 | 19 | 84 | 130 | 2500 |

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

* Example derived from this [video](https://www.youtube.com/watch?v=Uk36WGxujkc).

### Jury Selection example

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

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

How far off do 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**):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ethinicity | White | Black | Latinx | Asian | Other | Total |
| Observed count | 1920 | 347 | 19 | 84 | 130 | 2500 |

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ethinicity | White | Black | Latinx | Asian | Other | Total |
| Expected count |  |  |  |  |  | 2500 |

### Jury Selection example

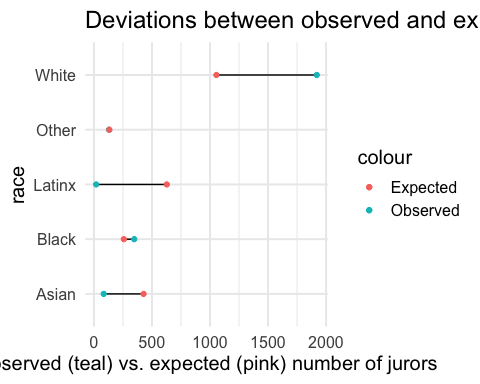
* To fill in the table, multiple the total size of the jury by the % of the population of each race/ethnicity:

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ethnicity | White | Black | Latinx | Asian | Other | Total |
| Expected count | 2500 0.422 | 2500 0.103 | 2500 0.251 | 2500 0.171 | 2500 0.053 | 2500 |
| = | 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:
* We want an estimate that somehow quantifies how different the observed counts () are from the expected counts () across the 5 race/ethnicities.

### The Chi-square test statistic

The chi2 test statistic quantifies the magnitude of the difference between observed and expected counts under the null hypothesis. It looks like this:

* is the number of cells in the table. Here, is the number of race/ethnicity groups. That is,
* is the observed count for the group (here race/ethnicity)
* is the expected count for the group
* is a distribution, like or Normal.

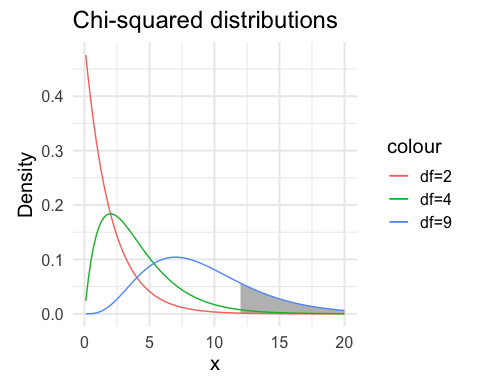
### The Chi-square test statistic

* 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 so that a difference of 100 fewer Latinx jurors  
  counts for more than a difference of 100 fewer white jurors compared to what is expected for each group.

### 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) minus one.

Or, .



* 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 of the .

### 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:

At least one of is different than specified in , for being one of White, Black, Latinx, Asian, or Other.

### Back to the jury example

**Calculate the chi-square statistic using the jury data.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ethinicity | White | Black | Latinx | Asian | Other | Total |
| Observed | 1920 | 347 | 19 | 84 | 130 | 2500 |
| Expected | 1055 | 257.5 | 627.5 | 427.5 | 132.5 | 2500 |

### 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 null

##   
## 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 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 week:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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**

$H\_0: p\_1 = 1/7p\_2=1/7p\_3=1/7p\_4 = 1/7p\_5=1/7p\_6=1/7p\_7=1/7$.

Written another way:

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

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

**Calculate the expected counts under**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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**

* Use the fact that the total number of births equalled 700. Then 700\*(1/7) = 100. We would expect to see around 100 births on each day if the births occurred 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 chi2 test statistic**

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

### 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.

### Conditions to perform a chi-square test

* Fixed number 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 mutually exclusive categories
* The probability of a given outcome is the same for each observation.

### Expected counts requirement

* At least 80% of the cells have 5 or more observations expected ( for 80% of the cells)
* All cells have expected counts > 1 ( for all cells)