Lecture 29: Inference about a categorical variable with > 2 levels

Chapter 21

Instructor: Tomer Altman

November 7, 2025

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 $\hat{p}$ (one sample) and for $\hat{p}_1 = \hat{p}_2$ (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.

Ethnicity	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.

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 far off do the **observed counts** of race/ethnicity 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

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

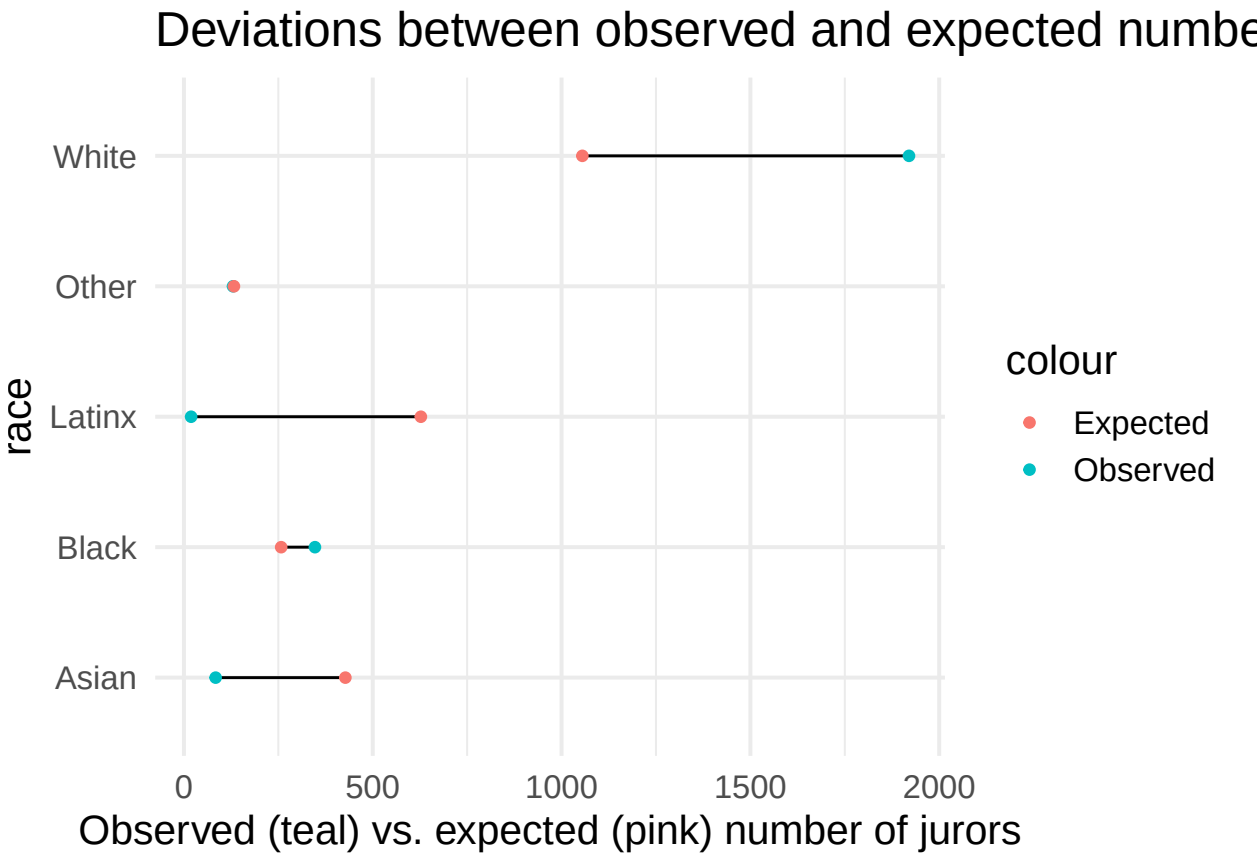
- To fill in the table, multiply the total size of the jury by the % of the population of each race/ethnicity:

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:

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

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

Check your understanding!

The Chi-square test statistic

The χ^2 (chi square) 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

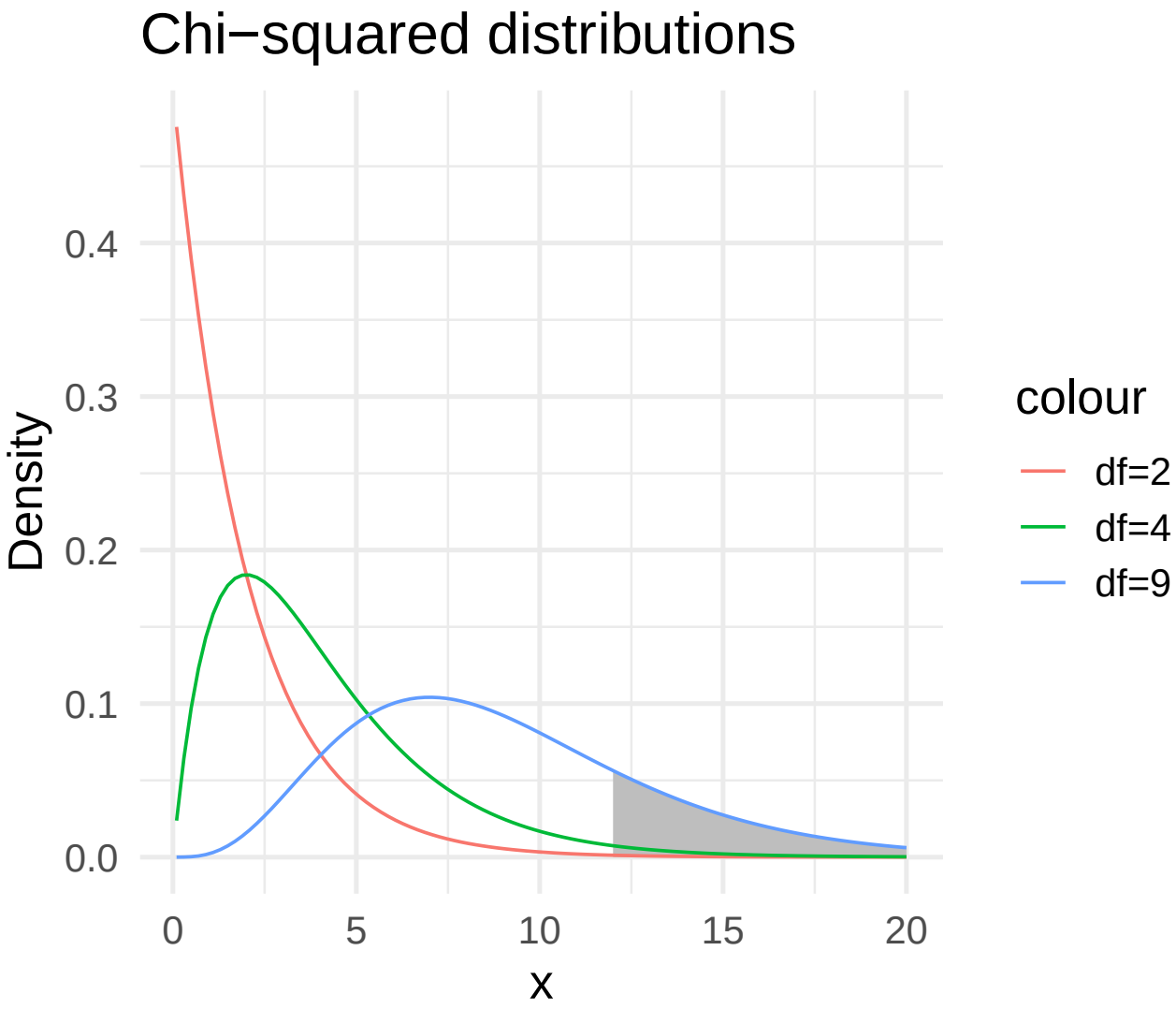
$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- The numerator measures the squared deviations between the observed (O) and expected (E) values
  - Bigger deviations will make the test statistic larger
  - Thus, the corresponding p-value will be smaller
  - We square the difference so that we only care about the **magnitude** of change, not the **direction** of change
- The denominator makes this magnitude *relative* to what we expect
  - It also makes the units of the χ^2 statistic to be the same as the observations
  - This adjusts for the different magnitude of expected counts
  - We would *expect* the number of white jurors to be close to 1,055, and Latinx jurors to be close to 628
  - Therefore, 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

$$\frac{100}{628} > \frac{100}{1,055}$$

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: the degrees of freedom
- The degrees of freedom is equal to the number of groups (here, race/ethnicity) minus one: $\text{df} = k - 1$
- 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 amount is higher
- The chi-square distribution is strictly non-negative. We only ever compute upper tail probabilities for the chi-square test because there is only one form of the H_A.



Back to the jury example

State the null and alternative hypotheses:

- 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:

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

H_A : At least one of p_k is different than specified in H_0, for k being one of the groups: 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
Observed	1920	347	19	84	130	2500
Expected	1055	257.5	627.5	427.5	132.5	2500

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

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

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

$$\chi^2 = 1606.454$$

Back to the jury example

Calculate the p-value:

- Remember: the appropriate degrees of freedom here is $\text{df} = k - 1 = 4$

```
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!
- But this could be an artifact of the χ^2 test, as it has a flaw where the p-value $p \to 0$ as $n \to \infty$

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/ethnicity groups 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 five contributions to the chi-square test from each race/ethnicity group. 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 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

$H_0 : p_1 = \frac{1}{7}, p_2 = \frac{1}{7}, p_3 = \frac{1}{7}, p_4 = \frac{1}{7}, p_5 = \frac{1}{7}, p_6 = \frac{1}{7}, p_7 = \frac{1}{7}$

Written another way:

$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = \frac{1}{7}$

H_A : At least one of these p_k differ from $\frac{1}{7}$. Or: not all p_k equal $\frac{1}{7}$

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 $700 \times \frac{1}{7} = 100$
- We would expect to see around 100 births on each day if the births occurred randomly over the course of the week
- That is, if the null were H_0

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-square 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.

Conditions to perform a chi-square test

- Fixed number 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

Expected counts requirement

- At least 80% of the cells have 5 or more observations expected ($E_i \geq 5$ for $\geq 80\%$ of the cells)
- All k cells have expected counts > 1 ($E_i > 1$ for all cells)
  - Given the same sample size, n, if one did the repeated experiment of drawing an SRS of n a "large number of times", then the average of the counts in a cell is always > 1

Notes about Goodness of Fit tests

- The null is probably almost never true in these circumstances (that is, the underlying data matches the null distribution perfectly)
- Whenever the null is not perfectly true, the p-value will go to 0 as $n \to \infty$
- Thus, it is always important to look at the differences that resulted in the p-value to see if there are important differences rather than make a conclusion just based on p-value

Check your understanding!