Lesson 16: Chi-squared test

TB sections 8.3-8.4

Meike Niederhausen and Nicky Wakim

2024-11-20

Learning Objectives

1. Understand the Chi-squared test and the expected cell counts under the null hypothesis distribution.
2. Determine if two categorical variables are associated with one another using the Chi-squared test.

Where are we?

Data
- Collecting data
- Categorical vs. Numeric
- Summary statistics
- Data visualization

Probability
- Probability rules
- Independence, conditional
- Random variables and probability distributions
- Linear combinations
- Binomial, Normal, and Poisson

Sampling Variability, and Statistical Inference
- Sampling distributions
- Central Limit Theorem
- Confidence Intervals
- Hypothesis tests

Inference for continuous data/outcomes
- One sample t-test
- 3+ independent samples
- Simple linear regression / correlation
- 2 sample tests: paired and independent
- Power and sample size
- Non-parametric tests

Inference for categorical data/outcomes
- One proportion test
- Fisher's exact test
- Non-parametric tests
- Chi-squared test
- 2 proportion test
- Power and sample size

R
- Basics
- Reproducibility
- Quarto
- Packages
- Data visualization
- Data wrangling
- R Projects
- …

Last time

- We looked at inference for a **single proportion**
- We looked at inference for a difference in **two independent proportions**

- If there are two groups, we could see if they had different proportions by testing if the difference between the proportions were the same (null) or different (alternative, two-sided, $\neq$)

- What happens when we want to compare two **or more** groups' proportions?
  - Can no longer rely on the difference in proportions
  - Need a new method to make inference (Chi-squared test!)

Learning Objectives

1. Understand the Chi-squared test and the expected cell counts under the null hypothesis distribution.

2. Determine if two categorical variables are associated with one another using the Chi-squared test.

From Lesson 4: Example: hypertension prevalence (1/2)

- US CDC estimated that between 2011 and 2014[1], 29% of the population in America had hypertension

- A health care practitioner seeing a new patient would expect a 29% chance that the patient might have hypertension
  - However, this is **only the case if nothing else is known about the patient**

From Lesson 4: Example: hypertension prevalence (2/2)

- Prevalence of **hypertension varies significantly with age**
  - Among adults aged 18-39, 7.3% have hypertension
  - Adults aged 40-59, 32.2%
  - Adults aged 60 or older, 64.9% have hypertension

- Knowing the age of a patient provides important information about the likelihood of hypertension
  - Age and hypertension status are **not independent - Can we back up this claim??**
- While the probability of hypertension of a randomly chosen adult is 0.29…
  - The **conditional probability** of hypertension in a person known to be 60 or older is 0.649

Question: Is there an association between age group and hypertension?

From Lesson 4: Contingency tables

- We can start looking at the **contingency table** for hypertension for different age groups
  - **Contingency table:** type of data table that displays the frequency distribution of two or more categorical variables

Table: Contingency table showing hypertension status and age group, in thousands.

Age Group	Hypertension	No Hypertension	Total
18-39 yrs	8836	112206	121042
40-59 yrs	42109	88663	130772
60+ yrs	39917	21589	61506
Total	90862	222458	313320

Test of General Association + Hypotheses

- **General research question:** Are two variables (both categorical, nominal) associated with each other?

General wording for hypotheses

Test of "**association**" wording

- H_0: There is no association between the two variables
- H_A: There is an association between the two variables

Test of "**independence**" wording

- H_0: The variables are independent
- H_A: The variables are not independent

Hypotheses test for example

Test of "**association**" wording

- H_0: There is no association between age and hypertension
- H_A: There is an association between age and hypertension

Test of "**independence**" wording

- H_0: The variables age and hypertension are independent
- H_A: The variables age and hypertension are not independent

H_0: Variables are Independent (under the null)

- Recall from Chapter 2, that events A and B are independent if and only if

$$P(A \cap B) = P(A)P(B)$$

- If age and hypertension are independent variables, then *theoretically* this condition needs to hold for *every combination of levels*, i.e.

$$\begin{aligned}
P(18-39 \cap \text{hyp}) &= P(18-39)P(\text{hyp}) \\
P(18-39 \cap \text{no hyp}) &= P(18-39)P(\text{no hyp}) \\
P(40-59 \cap \text{hyp}) &= P(40-59)P(\text{hyp}) \\
P(40-59 \cap \text{no hyp}) &= P(40-59)P(\text{no hyp}) \\
P(60+ \cap \text{hyp}) &= P(60+)P(\text{hyp}) \\
P(60+ \cap \text{no hyp}) &= P(60+)P(\text{no hyp})
\end{aligned}$$

Age Group	Hypertension	No Hypertension	Total
18-39 yrs	8836	112206	121042
40-59 yrs	42109	88663	130772
60+ yrs	39917	21589	61506
Total	90862	222458	313320

$$P(18-39 \cap \text{hyp}) = \frac{121042}{313320} \cdot \frac{90862}{313320}$$

$$\cdots$$

$$P(60+ \cap \text{no hyp}) = \frac{61506}{313320} \cdot \frac{222458}{313320}$$

With these probabilities, for each cell of the table we calculate the **expected** counts for each cell under the H_0 hypothesis that the variables are independent

Expected counts (if variables are independent)

- The expected counts (if H_0 is true & the variables are independent) for each cell are
  - $np = \text{total table size} \cdot \text{probability of cell}$
  - $\text{expected count} = \dfrac{\text{column total} \cdot \text{row total}}{\text{table total}}$

Age Group	Hypertension	No Hypertension	Total
18-39 yrs	8836	112206	121042
40-59 yrs	42109	88663	130772
60+ yrs	39917	21589	61506
Total	90862	222458	313320

Expected count of 40-59 years old and hypertension:

$$\begin{aligned}\text{expected count} &= \frac{\text{column total} \cdot \text{row total}}{\text{table total}} \\ &= \frac{90862 \cdot 130772}{313320} \\ &= 37923.55\end{aligned}$$

- If age group and hypertension are **independent** variables
  - (as assumed by H_0),
- then the **observed counts should be close to the expected counts** for each cell of the table

- Test to see how likely is it that we observe our data given the null hypothesis (no association)

Observed vs. Expected counts

- The **observed** counts are the counts in the 2-way table summarizing the data

Age Group	Hypertension	No Hypertension
18-39 yrs	8836	112206
40-59 yrs	42109	88663
60+ yrs	39917	21589

- The **expected** counts are the counts the we would expect to see in the 2-way table if there was no association between age group and hypertension

Age Group	Hypertension	No Hypertension
18-39 yrs	35101.87	85940.13
40-59 yrs	37923.55	92848.45
60+ yrs	17836.58	43669.42

Expected count for cell i, j :

$$\text{Expected Count}_{\text{row } i, \text{ col } j} = \frac{(\text{row } i \text{ total}) \cdot (\text{column } j \text{ total})}{\text{table total}}$$

Poll Everywhere Question 2

Using R for expected cell counts

- R calculates expected cell counts using the `expected()` function in the `epitools` package
- Make sure dataset is in `matrix` form using `as.matrix()`

```
hyp_data2
```

```
          Hypertension No_Hypertension
18-39 yrs         8836          112206
40-59 yrs        42109           88663
60+ yrs          39917           21589
```

```
library(epitools)
expected(hyp_data2)
```

```
          Hypertension No_Hypertension
18-39 yrs     35101.87        85940.13
40-59 yrs     37923.55        92848.45
60+ yrs       17836.58        43669.42
```

Learning Objectives

1. Understand the Chi-squared test and the expected cell counts under the null hypothesis distribution.

2. Determine if two categorical variables are associated with one another using the Chi-squared test.

Reference: Steps in a Hypothesis Test

1. Check the **assumptions**
2. Set the **level of significance** α
3. Specify the **null** (H_0) and **alternative** (H_A) **hypotheses**
    1. ~~In symbols~~
    2. In words
    3. ~~Alternative: one- or two-sided?~~
4. Calculate the **test statistic**.
5. Calculate the **p-value** based on the observed test statistic and its sampling distribution
6. Write a **conclusion** to the hypothesis test
    1. Do we reject or fail to reject H_0?
    2. Write a conclusion in the context of the problem

Step 1: Check assumptions

- **Independence**
  - All individuals are independent from one another
    - In particular, observational units cannot be represented in more than one cell
    - For example, someone cannot be in two differnt age groups
- **Sample size**
  - In order for the distribution of the test statistic to be appropriately modeled by a chi-squared distribution we need
  - 2×2 table
    - expected counts are at least 10 for each cell
  - Larger tables
    - No more than 20% of expected counts are less than 5
    - All expected counts are greater than 1

```
expected(hyp_data2)
```

```
            Hypertension No_Hypertension
18-39 yrs       35101.87        85940.13
40-59 yrs       37923.55        92848.45
60+ yrs         17836.58        43669.42
```

All expected counts > 5

Step 2 and 3: Significance level and Hypotheses

- Set $\alpha = 0.05$

Hypotheses test for example

Test of "**association**" wording

- H_0: There is no association between age and hypertension
- H_A: There is an association between age and hypertension

Test of "**independence**" wording

- H_0: The variables age and hypertension are independent
- H_A: The variables age and hypertension are not independent

Step 4: Calculate the χ^2 test statistic (1/2)

Test statistic for a test of association (independence):

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

- When the variables are independent, the observed and expected counts should be close to each other

Step 4: Calculate the χ^2 test statistic (2/2)

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$
$$= \frac{(8836 - 35101.87)^2}{35101.87} + \frac{(112206 - 85940.13)^2}{85940.13} +$$
$$\ldots + \frac{(21589 - 43669.42)^2}{43669.42}$$
$$= 66831$$

Is this value big? Big enough to reject H_0?

Observed:

Age Group	Hypertension	No Hypertension
18-39 yrs	8836	112206
40-59 yrs	42109	88663
60+ yrs	39917	21589

Expected:

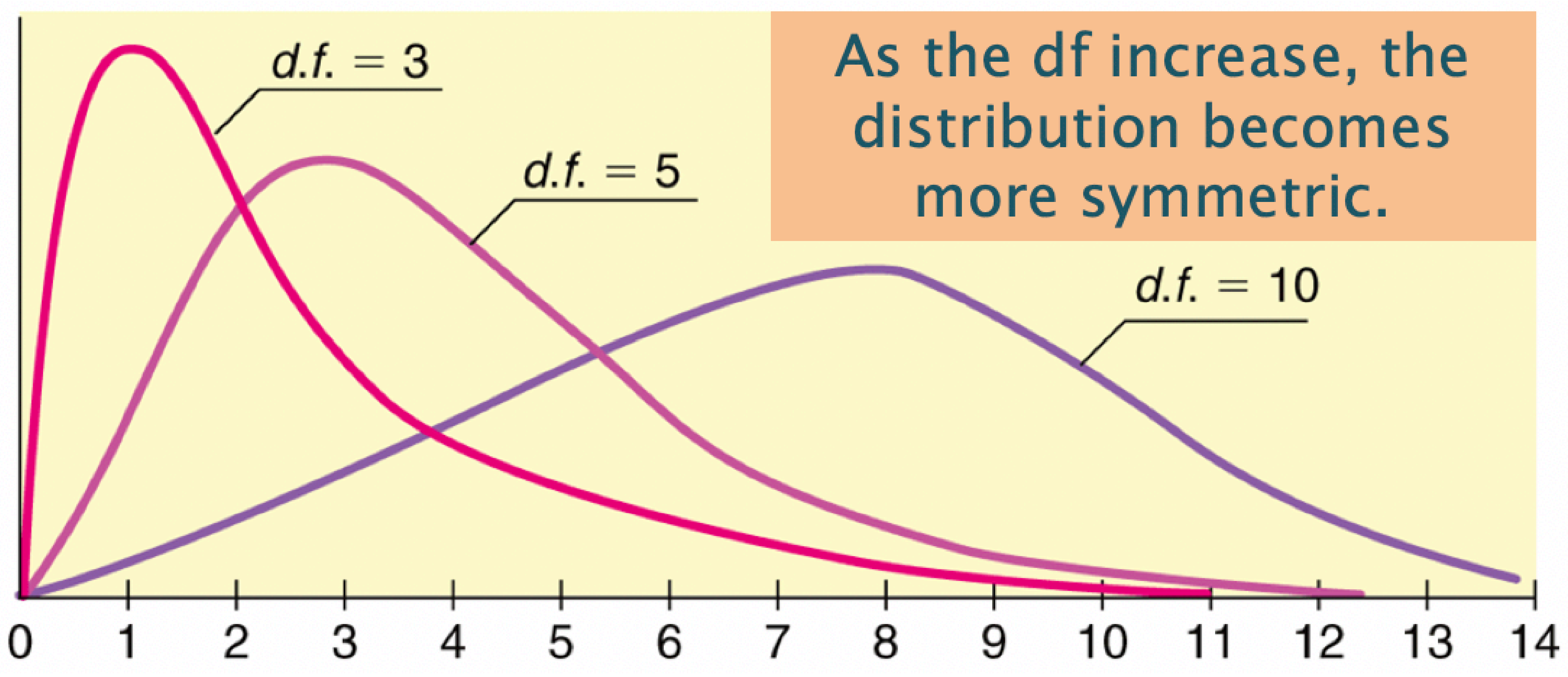
Age Group	Hypertension	No Hypertension
18-39 yrs	35101.87	85940.13
40-59 yrs	37923.55	92848.45
60+ yrs	17836.58	43669.42

Poll Everywhere Question 2

Step 5: Calculate the *p*-value

The χ^2 distribution shape depends on its degrees of freedom

- It's skewed right for smaller df,
  - gets more symmetric for larger df
- **df = (# rows-1) x (# columns-1)**



- The **p-value** is always the **area to the right** of the test statistic for a χ^2 test

- We can use the `pchisq` function in R to calculate the probability of being at least as big as the χ^2 test statistic:

```
pv <- pchisq(66831, df = 2,
        lower.tail = FALSE)
pv
```

```
[1] 0
```

Step 4-5: Calculate the test statistic and p-value

- Data need to be in a matrix or table: use `as.matrix()` or `table()`
  - Use matrix if data already in contingency table form
  - Use table if data are two columns with each row for each observation (tidy version)
- Notice that age groups are rownames! Age does not have its own column

```r
hyp_data2
```

```
          Hypertension No_Hypertension
18-39 yrs         8836          112206
40-59 yrs        42109           88663
60+ yrs          39917           21589
```

- Run `chisq.test()` in R

```r
chisq.test(x = hyp_data2)
```

```
	Pearson's Chi-squared test

data:  hyp_data2
X-squared = 66831, df = 2, p-value < 2.2e-16
```

Step 6: Conclusion

Recall the hypotheses to our χ^2 test:

- H_0: There is **no association** between age and hypertension
- H_A: There is **an association** between age and hypertension

- Recall the p-value = 0.0402
- Use α = 0.05
- Do we reject or fail to reject H_0?

Conclusion statement:

- There is sufficient evidence that there is an association between age group and hypertension (p-value < 0.0001`)

Warning!!

If we fail to reject, we DO NOT say variables are independent! We **can** say that we have insufficient evidence that there is an association.

Chi-squared test: Example all together

1. Check expected cell counts threshold

```
expected(hyp_data2)
```

```
           Hypertension No_Hypertension
18-39 yrs      35101.87        85940.13
40-59 yrs      37923.55        92848.45
60+ yrs        17836.58        43669.42
```

All expected cells are greater than 5.

2. $\alpha = 0.05$

3. Hypothesis test:

- H_0 : There is no association between age group and hypertension
- H_1 : There is an association between age group and hypertension

4-5. Calculate the test statistic and p-value for Chi-squared test in R

```
chisq.test(x = hyp_data2)
```

```
	Pearson's Chi-squared test

data:  hyp_data2
X-squared = 66831, df = 2, p-value < 2.2e-16
```

6. Conclusion

We reject the null hypothesis that age group and hypertension are not associated ($p < 2.2 \cdot 10^{-16}$). There is sufficient evidence that age group and hypertension are associated.