

Lesson 18: Chi-squared test

TB sections 8.3-8.4

Meike Niederhausen and Nicky Wakim

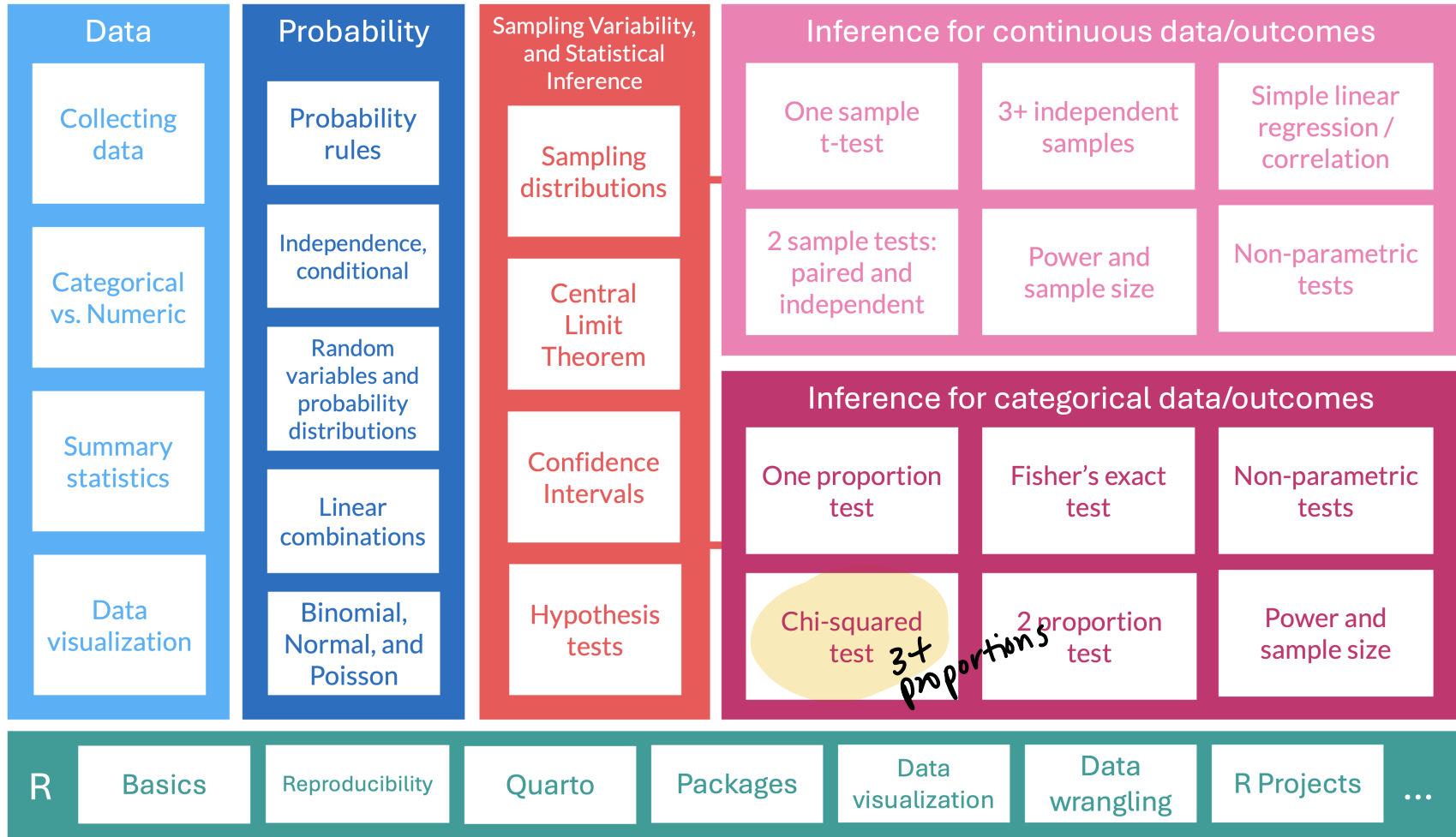
2025-12-01

1 sample	cont out μ t test	binary (cat) outcome p prop test
2 sample	$\mu_1 - \mu_2$ t test	$p_1 - p_2$ prop test
3+ samples	ANOVA	χ^2 test (chi-squared)


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?



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 proportions were the same (null) or different (alternative, two-sided, \neq)
 - What happens when we want to compare **three 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 2014¹, 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

$$\hookrightarrow P(\text{hypertension} \mid \text{60+ yo}) = 0.649$$

"given"

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

$$\hat{P}(\text{hyp} | 18-39 \text{ yrs}) = \frac{8836}{121042}$$

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

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

A & B

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

$$P(18 - 39 \cap \text{hyp}) = P(18 - 39)P(\text{hyp})$$

$$P(18 - 39 \cap \text{no hyp}) = P(18 - 39)P(\text{no hyp})$$

$$P(40 - 59 \cap \text{hyp}) = P(40 - 59)P(\text{hyp})$$

$$P(40 - 59 \cap \text{no hyp}) = P(40 - 59)P(\text{no hyp})$$

$$P(60 + \cap \text{hyp}) = P(60+)P(\text{hyp})$$

$$P(60 + \cap \text{no hyp}) = P(60+)P(\text{no hyp})$$

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

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

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

prop of 18-39
prop of ppl w/ hyp

8836 / 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) under the null

- The expected counts (if H_0 is true & the variables are independent) for each cell are

▪ np = total table size · probability of cell

▪ expected count = $\frac{\text{column total} \cdot \text{row total}}{\text{table total}}$

Expected count of 40-59 years old and hypertension:

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

expected count

expected count for this cell

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

obs count

- 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 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 :

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

used this formula to calc

Poll Everywhere Question 2

09:41 Tue Jan 9

Join by Web PollEv.com/nickywakim275

What does it mean if observed counts are very close to expected counts?

The variables are highly associated. 23%

The test statistic will likely be large. 0%

The variables are likely independent. 77%

There is a violation of the assumptions for the test. 0%

Powered by Poll Everywhere

more evidence that observed data also come from null (aka ind)

↳ calc under null that 2 variables are independent

not independent

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()`

```
1 hyp_data2
```

observed data

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

```
1 library(epitools)
```

```
2 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 and each group is independent
 - In particular, observational units cannot be represented in more than one cell
 - For example, someone cannot be in two different age groups

- Sample size

- In order for the distribution of the test statistic to be appropriately modeled by a chi-squared distribution we need

(continuous distribution)

- ▪ 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

rows by cols
3 × 2 table

```
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

6 cells : 1.2 cell (20%)

1 cell can have < 5 expected counts

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):

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

diff small, then χ^2 value is small

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

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

$$\begin{aligned}\chi^2 &= \sum \frac{(O - E)^2}{E} \\ &= \frac{(8836 - 35101.87)^2}{35101.87} + \frac{(112206 - 85940.13)^2}{85940.13} + \\ &\quad \dots + \frac{(21589 - 43669.42)^2}{43669.42} \\ &= 66831\end{aligned}$$

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

09:41 Tue Jan 9

100%



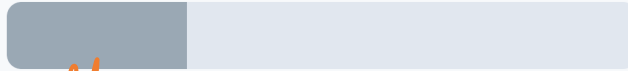
Join by Web PollEv.com/nickywakim275



What does a large Chi-squared test statistic indicate?

sample seems to be under null

Observed counts closely match expected counts.



13%

under the null

Strong evidence that the variables are independent



27%

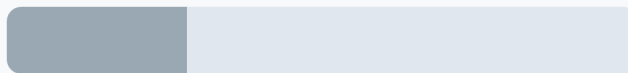
A strong association between the variables. ✓



47%

variables are NOT independent

A small sample size.



13%

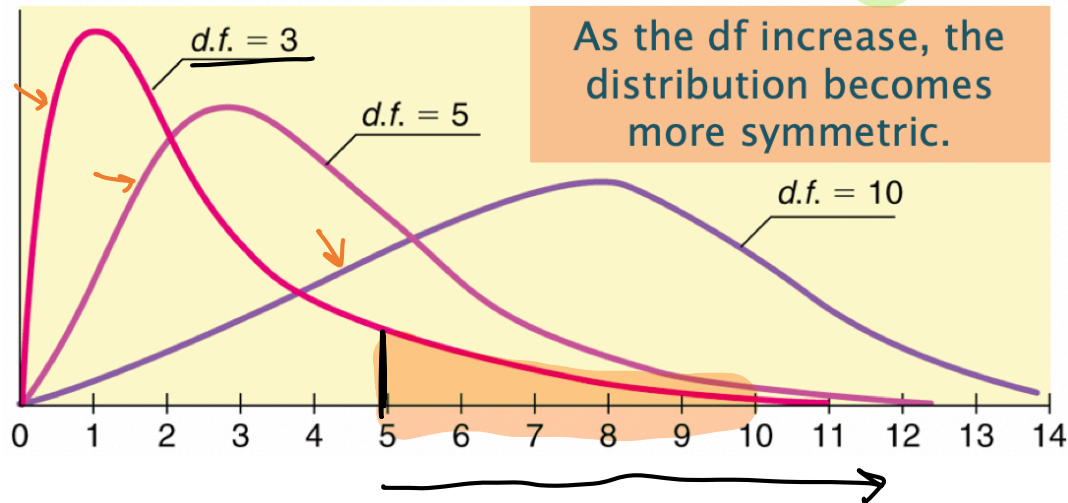
large χ^2 means observed are far from expected

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
- $df = (\# \text{ rows} - 1) \times (\# \text{ columns} - 1)$

$$\begin{aligned} df &= (3-1) \times (2-1) \\ &= (2) \times (1) \\ &= 2 \end{aligned}$$



- 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:

```
1 pv <- pchisq(66831, df = 2,  
2           lower.tail = FALSE)  
3 pv  
[1] 0
```

↳ area to right

↳ showing 0 b/c
so close to 0
& rounding

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

- Run `chisq.test()` in R

```
1 hyp_data2 ↴
```

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

```
1 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 $(0 < 0.001)$

- Use $\alpha = 0.05$

- Do we reject or fail to reject H_0 ?

$$p\text{-val} < 0.001 < 0.05 = \alpha$$

Reject the null

Conclusion statement:

- There is sufficient evidence that there is an association between age group and hypertension ($p\text{-value} < 0.0001$)

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.

do not "accept" the null

that the variables are not independent

Ref.

Chi-squared test: Example all together

1. Check expected cell counts threshold

```
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 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

```
1 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.