Data Analysis in Sociology

Lecture 2. Chi-square test of independence

January 2023

A couple of organizational announcements

1. Office hours
2. Data clinic ran by TAs
3. Join our class on canvas (if you have not done that already)
4. Register on DataCamp
5. Description and assessment criteria for Project 1 will be posted today
6. Please, attend practice session with your sub-group (or warn Dmitry in a case of emergency)

Agenda for today

1. Statistical hypothesis
2. P-value
3. Chi-square test

Research questions and hypotheses

- The research process starts with **research questions**, e.g. „Do computer games affect personal relations?“

- A research-based answer to that question assumes the following:
    - A research hypothesis
    - Data collection
    - Data analysis
    - Conclusion as to the research hypothesis
    - Reporting the results to the audience

Research questions and hypotheses

A research hypothesis is a statement about the expected or predicted relation between variables.

Research questions and hypotheses: examples

Research Question	Research Hypothesis
Is students' performance on tests more influenced by their motivation or their learning strategies?	Students who are taught effective learning skills will perform better on tests than students offered incentives to do well.
Do college students and faculty differ in their beliefs about the prevalence of student academic misconduct (plagiarism)?	Faculty members' beliefs about the frequency of student academic misconduct will be lower than students' beliefs.
Is there a link between adolescents' exposure to violence in their family and their academic achievement?	The more adolescents are exposed to violence in their family, the lower their levels of academic achievement.

Research Hypothesis vs. Statistical Hypothesis

- **Research hypothesis** is a statement of what the researcher believes will be the outcome of an experiment or a study

- **Statistical hypothesis** is a formal structure used to scientifically test the research hypothesis

Sometimes those two go in line sometimes they contradict to each other. And that’s okay.

Null Hypothesis vs Alternative Hypothesis

There are two types of statistical hypotheses

- **Null hypothesis**. The null hypothesis, denoted by H_0, is usually the hypothesis that sample observations result purely from chance.

  In all statistical tests that we will cover NH is about "no difference", "equality", "absence of relationship"

- **Alternative hypothesis**. The alternative hypothesis, denoted by H_1 or H_a, is the hypothesis that sample observations are influenced by some non-random cause.

Together, H_0 and H_a cover all possible outcomes.

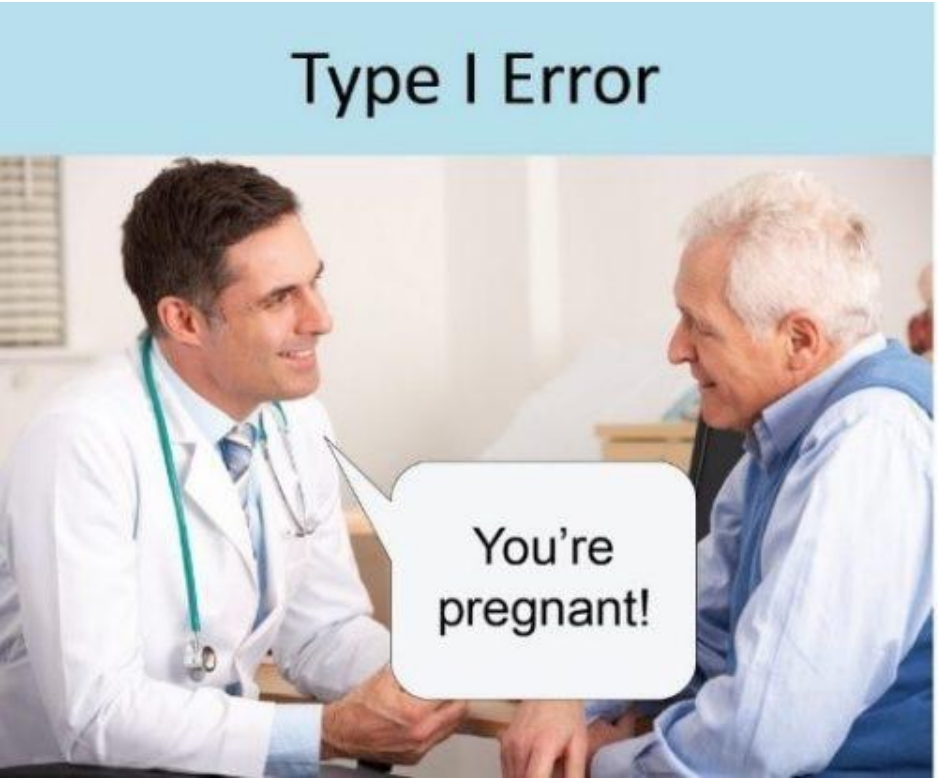
Type I and type II errors

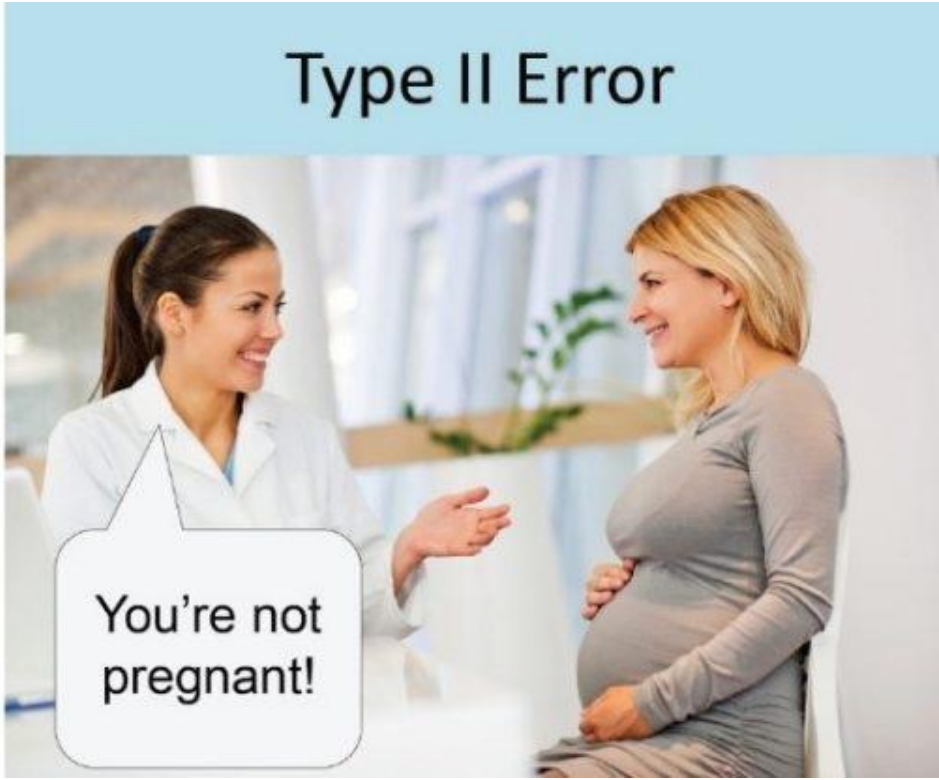
Two types of errors can result from a hypothesis test.

- **Type I error**. A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level**. This probability is also called **alpha**, and is often denoted by α.

- **Type II error**. A Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called **Beta**, and is often denoted by β. The probability of *not* committing a Type II error is called the **Power** of the test.

Type I and type II errors





P-value

In null-hypothesis significance testing (NHST), the ***p*-value** is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.

Low p-value indicates that it is unlikely to get a results like that if H_0 is correct, therefore it can be an evidence to reject H_0.

A traditionally accepted alpha level (p-value threshold) is 0.05 therefore:

P-value < 0.05 ⟶ we can reject H_0 in favor of H_1.

P-value > 0.05 ⟶ we cannot reject H_0.

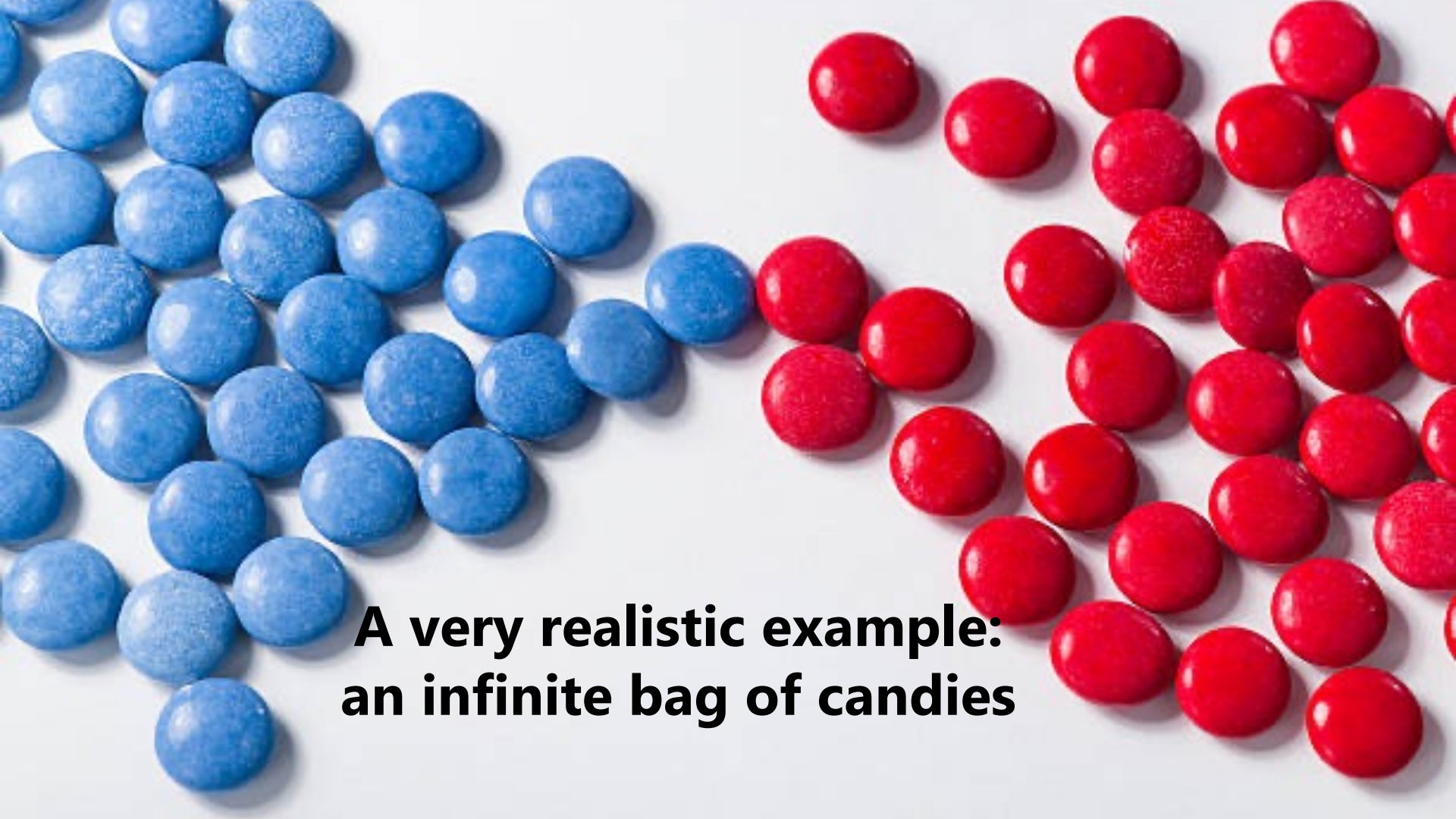
The 0.05 threshold of the p-value is historical

Another step forward from the binary logic of point estimates is the confidence intervals (CIs) — a range of values that covers the true value at a given probability (the confidence level).
Typical confidence levels: 95%, 90%, 99%.
CIs depend on the standard error, SD, and sample size.
If zero is covered by the confidence interval, then H0 is retained.

P-value is the probability to observe the statistic value as extreme or higher, given that H_0 is true.

Using NHST, H_0 can be either retained, or rejected in favour of H_1.

- NHST does not test H_1, so it is formally incorrect to claim that H_1 can be accepted.


A very realistic example:
an infinite bag of candies

A very realistic example: an infinite bag of candies

- We want to know red to blue ratio
- We cannot count them all
- We can estimate the ratio in the bag (population) using a sample
- But first, let’s start with a null hypothesis

A very realistic example: an infinite bag of candies

- H_0: red to blue ratio is not different from 50/50 (you can test for any proportion here)
- H_a : red to blue ratio is different 50/50

A very realistic example: an infinite bag of candies

- Let’s draw a sample of 10
- What is the probability of drawing a sample with such parameter when H0 is true for the population?

A very realistic example: an infinite bag of candies

4 red vs. 6 blue
Probability of such a proportion when H0 is true in the population (p-value) is 0.75

3 red vs. 7 blue
Now, the p-value is 0.34

2 red vs. 8 blue
Now, the p-value is 0.1

1 red vs. 9 blue
Now, the p-value is 0.02 That’s an evidence to reject H0

Any questions so far?

Chi-square test

[*kai-skwear*]

How to find relationship between different types of variables

	Categorical	Numeric
Categorical		
Numeric		

Pearson’s chi-square test of independence

- A test of proportions
- Helps to tell whether 2 categorical variables are distributed independently
- Works on a contingency table, also known as a cross-tab
- **H0**: the categorical variables are independently distributed.
  H1: the categorical variables are not independently distributed.

Pearson's chi-square test of independence.

RQ: Should Pubs be Open 24/7?

Observed (**Data**)

	Yes	No	Sum (row)
Males	**120**	**90**	210
Females	**80**	**110**	190
Sum (col)	200	200	Total: 400

Pearson's chi-square test of independence.

RQ: Should Pubs be Open 24/7?

Expected frequency: Sum(row) * Sum(column) / Sum(total)

Observed (**Data**) → Expected (**Model**)

	Yes	No	Sum (row)
Males	**120**	**90**	210
Females	**80**	**110**	190
Sum (col)	200	200	Total: 400

	Yes	No	Sum (row)
Males	**105**		**210**
Females			190
Sum (col)	**200**	200	Total: **400**

Pearson's chi-square test of independence.

RQ: Should Pubs be Open 24/7?

Expected frequency: Sum(row) * Sum(column) / Sum(total)

Observed (**Data**) → Expected (**Model**)

	Yes	No	Sum (row)
Males	**120**	**90**	210
Females	**80**	**110**	190
Sum (col)	200	200	Total: 400

	Yes	No	Sum (row)
Males	105	**105**	**210**
Females			190
Sum (col)	200	**200**	Total: **400**

Pearson's chi-square test of independence.

RQ: Should Pubs be Open 24/7?

Expected frequency: Sum(row) * Sum(column) / Sum(total)

Observed (**Data**) → Expected (**Model**)

	Yes	No	Sum (row)
Males	**120**	**90**	210
Females	**80**	**110**	190
Sum (col)	200	200	Total: 400

	Yes	No	Sum (row)
Males	105	105	210
Females	**95**		**190**
Sum (col)	**200**	200	Total: **400**

Pearson's chi-square test of independence.

RQ: Should Pubs be Open 24/7?

Expected frequency: Sum(row) * Sum(column) / Sum(total)

Observed (**Data**) → Expected (**Model**)

	Yes	No	Sum (row)
Males	**120**	**90**	210
Females	**80**	**110**	190
Sum (col)	200	200	Total: 400

	Yes	No	Sum (row)
Males	105	105	210
Females	95	**95**	**190**
Sum (col)	200	**200**	Total: **400**

Pearson's chi-square test of independence.

RQ: Should Pubs be Open 24/7?

Expected frequency: Sum(row) * Sum(column) / Sum(total)

Observed (**Data**) → Expected (**Model**)

	Yes	No	Sum (row)
Males	**120**	**90**	210
Females	**80**	**110**	190
Sum (col)	200	200	Total: 400

	Yes	No	Sum (row)
Males	105	105	210
Females	95	95	190
Sum (col)	200	200	Total: 400

$\chi^2 = (120-105)^2/105 + (90-105)^2/105 + (80-95)^2/95 + (110-95)^2/95 = 8.43$

df = (rows - 1) * (columns - 1) = (2-1)*(2-1) = 1

$\chi^2(1) = 8.43$, $p = 0.004$ (the critical value of χ^2 for df=1 is 3.84, $p<.05$)

- Data = Model + Error
- Observed scores = Expected scores + Residuals
- Chi-square is a *measure of discrepancy between the data and the model:*

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- After the test: *If the test is significant, inspect the standardized residuals and calculate the effect size.*

How to inspect residuals?

The χ^2 statistic is **the sum of standardized residuals.**

Once you have established that the χ^2 is statistically significant and the variables are associated, you may want to know, how exactly?

To understand this, inspect the **standardized residuals**:

stdres = (observed-expected)/ √expected

In the Pub example: = [(120-105)/√105] + [(90-105)/√105] + [(80-95)/√95] + [(110-95)/ √95]

Standardized residuals are z-scores. A residual with a value outside +-1.96 will be significant at p<.05; if it lies outside +-3.29, it is significant at p<.001. For crude estimates, you can use +-2 and +-3 as thresholds.

Back to the pub example

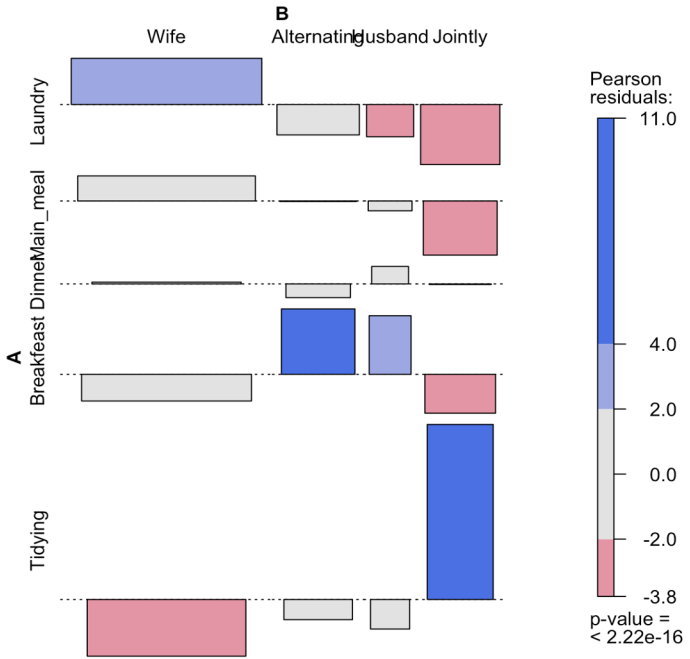
Standardized residuals

	Yes	No
Males	3.004	-3.004
Females	-3.004	3.004

In the cells with standardized residuals above 2 there are significantly more observations than it would be expected if the variables were independent

In the cells with standardized residuals below -2 there are significantly fewer observations than it would be expected if the variables were independent

Inspecto Residulus!



See this for an annotated script: http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r

How to report the results?

Let’s take a more playful example

Can cats be trained to dance?

	Training Reward		
	Food	Love	*Total*
Danced	28	48	*76*
Didn’t dance	10	114	*124*
Total	*38*	*162*	*200*

This example comes from Andy Field’s famous textbook “Discovering Statistics…”, which I recommend.

How to report the results?

The cat example: *'There was a **significant association** between the type of training and whether or not cats would dance **$\chi^2(1) = 25.36$, $p < .001$**. Based on the **odds ratio**, the odds of cats dancing were **6.65** times higher if they were trained with food than if trained with affection.'*

Field, p.688

Broken windows article

2 variables:

- order/disorder
- was the norm violated? (yes/no)

Again there was a clear cross-norm inhibition effect. Of the participants in the order condition (where bicycles were not locked to the fence), 27% stepped through the gap in the fence, compared with 82% of the participants in the disorder condition (where the bicycles were attached to the fence). The difference is significant [$\chi^2(1, 93) = 27.791$, $P < 0.001$].

Yet
Another
Example
To read
At home:

Why do
people
gamble
at casino?

Real research: Casino gamblers (2006)

C.-K. Lee et al. / Tourism Management 27 (2006) 856–866

Table 4
Clusters differences by casino gamblers' characteristics

Characteristics	Cluster I ($n = 75$)	Cluster II ($n = 118$)	Cluster III ($n = 98$)	Cluster IV ($n = 108$)	
	Challenge and winning seekers	*Only winning seekers*	*Light gambling seekers*	*Multi-purpose seekers*	
Primary purpose					
Gambling	74	104	58	60	$\chi^2 = 66.553$, df = 3
No gambling	1	14	40	48	$p < 0.001$

„The results of the study reveal that for the ‘‘challenge/winning seekers’’ and ‘‘only winning seekers,’’ gambling was primary purpose of their visit to casino. Meanwhile, gambling was not the major purpose of the casino visit for light gambling seekers and multi-purpose seekers“

```
#to replicate the results,
X <- matrix(c(74, 1, 104, 14, 58, 40, 60, 48), nrow = 2)
chisq.test(X)
```

Alternatives to chi-square for categorical variables

- Likelihood ratio test (G-test): $L\chi^2 = 2 \Sigma\ observed * ln\ (observed/expected)$, $df=(r-1)*(c-1)$ — it is preferred over chi-square in small samples.
- The classic chi-square is somewhat relaxed about Type I error in 2*2 tables so R applies *Yates's continuity correction* for 2*2 tables by default (and you are safe)
- If there are not enough observations, use *Fisher's exact test* (see the video: https://www.youtube.com/watch?v=I9KsLCc-eiQ) or *Barnard's exact test*.

Chi-square test assumptions

1. The data in the cells should be frequencies, or counts of cases rather than percentages or some other transformation of the data.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058

Chi-square test assumptions

2. The levels (or categories) of the variables are mutually exclusive. That is, a particular subject fits into one and only one level of each of the variables.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058

Chi-square test assumptions

3. Each subject may contribute data to one and only one cell in the χ^2. If, for example, the same subjects are tested over time such that the comparisons are of the same subjects at Time 1, Time 2, Time 3, etc., then χ^2 may not be used.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058

Chi-square test assumptions

4. The study groups must be independent. This means that a different test must be used if the two groups are related. For example, a different test must be used if the researcher's data consists of paired samples, such as in studies in which a parent is paired with his or her child.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058

Chi-square test assumptions

5. There are 2 variables, and both are measured as categories, usually at the nominal level. However, data may be ordinal data. Interval or ratio data that have been collapsed into ordinal categories may also be used.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058

Chi-square test assumptions

6. The value of the cell *expecteds* should be 5 or more in at least 80% of the cells, and no cell should have an expected of less than one.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058

Andy Field's (2016) tips:

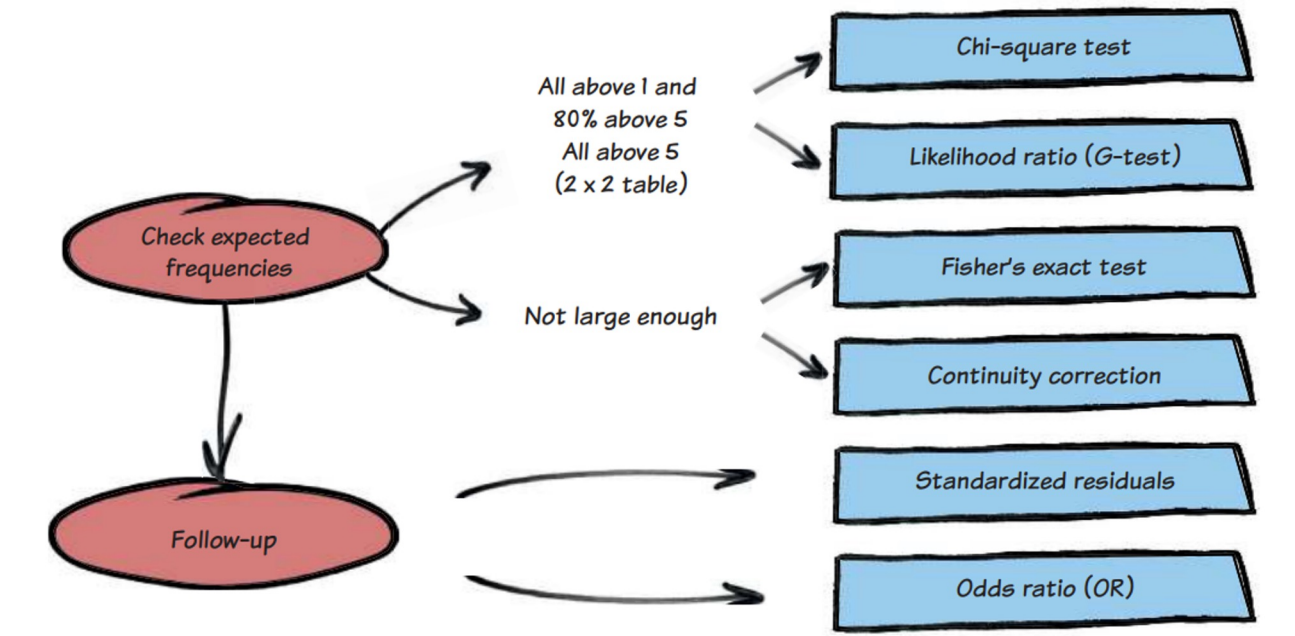


Figure 13.3 The general process for fitting a model looking for a relationship between two categorical variables

Source: "The Adventure in Statistics" by Andy Field (available at our library and in book folder on Yandex disk)

How to find relationship between different types of variables?

	Categorical	Numeric
Categorical	**Chi-square test**	
Numeric		

Resources to use:

Your annotated script for the chi-square: http://rpubs.com/ovolchenko/chisq2

Sample of the chi-square test analysis: http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r

An article reviewing major oints on chi-square test
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058

Online chi-square calculator: http://vassarstats.net/odds2x2.html

Questions?