Module 3 - Statistical Inference Part II

1. Inference for Proportions
2. Table Analysis

Valeria Vitelli
Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
valeria.vitelli@medisin.uio.no

MF9130E – Introductory Course in Statistics
27.04.2023

Overview for Module 3, part II

Schedule for today: Lectures in flipped classroom style (FOR*)

- 12:45 – 13:30 Introductory lecture
- 13:30 – 14:30 Self study session
- 14:30 – 14:45 Break
- 14:45 – 15:30 Group work session
- 15:30 – 15:45 Closing session (Wrap-up / Q&A)

Tomorrow morning: Lab in flipped classroom style (SEM*)

- Point estimates and CIs for proportions with **R**
- Table Analysis with **R**

Introductory Lecture for Module 3, part II

Key concepts

1. **Analysis of Proportions**
   - Proportions and the Binomial distribution
   - Inference for one population
2. **Comparing two proportions**
   - Effect estimates (risk difference, relative risk, odds ratio)
   - 2×2 contingency tables
3. **Table Analysis**
   - Pearson's Chi-squared test
   - Exact tests (Fisher's exact test)
   - Larger Tables

Key concept 1. Analysis of Proportions

Yesterday:

- Analysis of **continuous data**: data measured on a continuous scale
- Used t-tests to test for differences between groups

Today:

- **Binomial data**
- Testing for differences in proportions between groups
- New measures of the effect: Relative Risk and Odds Ratio

Key concept 1. Proportions and the Binomial distribution

Risk

- Binary variable with two possible outcomes:
  D (disease) and **H (healthy)**
- Study the **probability** or **risk**, π, that D occurs in the population

Sample proportion

- **sample proportion *p*** is the proportion of individuals in the sample in category D:

$$p = \frac{d}{n},$$

  where $d =$ number of subjects who experience D, and
  $n =$ sample size
- p is an **estimate** of the probability or **risk for D in the population**

Key concept 1. Proportions and the Binomial distribution

Recap from Module 2: the binomial distribution!

- **Assumptions:** independent experiments, two outcomes (success / not), probability of success same in all experiments
- Therefore, the **sampling distribution of a proportion** is the binomial distribution

The normal approximation to the binomial distribution

- When n is large, the binomial distribution can be approximated by a **normal distribution** with the same mean and standard error as the binomial distribution
(**Rule of thumb**: $n \times \pi \geq 10$ and $n \times (1 - \pi) \geq 10$)
- This is useful for:
  - calculating **confidence intervals**
  - carrying out **hypothesis tests**

Key concept 1. Inference for one population

Confidence interval for a proportion

Given the normal approximation to the binomial distribution, the **CI for the population proportion**, π, is

$$\mathrm{CI} = \left(p - z' \times \sqrt{\frac{p(1-p)}{n}}, p + z' \times \sqrt{\frac{p(1-p)}{n}} \right),$$

where z' is the appropriate percentage point of the standard normal distribution (1.96 if 95% CI)

Testing a hypothesis about one proportion

To test the **null hypothesis that the population proportion equals a particular value**, π_0:

$$\mathrm{H}_0 : \pi = \pi_0, \mathrm{H}_a : \pi \neq \pi_0,$$

we perform a z-test using the approximating normal distribution

Key concept 2. Comparing two proportions

Exposed versus unexposed

We want to compare **two exposure (or treatment) groups** with respect to the occurrence of a binary outcome

- **group 1:** individuals *exposed* to a risk factor
  group 0: *unexposed* individuals
- Clinical trials:
  group 1: *treatment* group
  group 0: *control (or placebo) group*

Different Measures

for comparing the outcome between the two groups

- **Risk difference** (not that much used in practice)
- Risk ratio, or **relative risk**
- **Odds ratio**

Each measure has an associated **confidence interval**

Key concept 2. Contingency Tables

2×2 contingency table

- Individuals are classified according to their exposure and outcome categories
- Cross tabulation is used to display the data in a **2×2 contingency table**

	Outcome		
Exposure	**Event: D (Disease)**	**No event: H (Healthy)**	**Total**
Group 1 (exposed)	d_1 $(d_1/n_1 \times 100\%)$	h_1 $(h_1/n_1 \times 100\%)$	n_1 (100%)
Group 0 (unexposed)	d_0 $(d_0/n_0 \times 100\%)$	h_0 $(h_0/n_0 \times 100\%)$	n_0 (100%)
Total	d	h	n

- Showing the proportion (or percentage) in each outcome category, within each of the exposure groups can be useful

Key concept 3. Table Analysis

So far:

- Analysis of **proportions** - one and two populations, CIs
- 2×2 contingency tables

Now:

- How to **test for a significant association?**
  - **Chi-square tests** are the most common
  - In case of little data: **exact tests**
- What if you have **more than two** exposure categories? Or outcomes?

Key concept 3. Table Analysis

Pearson's Chi-squared test (for a 2×2 table)

- **Null hypothesis:** no association between exposure and outcome
- The **test statistic** is

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \quad \text{d.f.} = 1,$$

  where O_i and E_i denote the observed and expected values in the ith cell
- The value of the test statistic χ^2 is **extreme** when values in the table are very **unlikely** under the null hypothesis
- For a 2×2 table the test statistic is chi-squared distributed with 1 degree of freedom under the null hypothesis (equivalent to the z-**test** for the difference between two proportions).

Key concept 3. Table Analysis

Test validity

- The **chi-squared test** is valid when:
  - The overall total is more than 40, regardless of the expected values, or
  - The overall total is between 20 and 40 provided all the expected values are at least 5
- The use of the **exact test** is recommended when:
  - The overall total of the table is less than 20, or
  - The overall total is between 20 and 40 and the smallest of the four expected numbers is less than 5

Important

- The chi-squared test produces **only a p-value**
- A **measure of the effect** (RD, RR or OR; with relative CI) is required when publishing, for helping results interpretation

Summary

Key terms and concepts

- Recap from Module 2 (the **Binomial distribution**)
- **Analysis of Proportions:** CI and z-test for one proportion
- **Comparing two proportions:**
  2×2 contingency tables, effect estimates:
  - Risk difference, and associated CI
  - Relative Risk (RR), and associated CI
  - Odds Ratio (OR), and associated CI
- **Table Analysis:**
  - Pearson's Chi-squared test, test validity
  - Fisher's exact test, when to use it

Self study session – Tasks

1. **Deepen your understanding** of each **key concept** from the previous slides by reading the corresponding longer slides:
   - `Module3-PartII-key_concept_1.pdf`
   - `Module3-PartII-key_concept_2.pdf`
   - `Module3-PartII-key_concept_3.pdf`
2. **Verify your learning outcome:**
   - **Review the Summary** (slide 13, "Key terms and concepts") in this presentation, and make sure you understand all terms
   - **IF** you feel you are still not familiar with any terms and concepts from the summary slides, use the provided **Learning Material** for this course module (next slide) to read up more
3. **Prepare for the group work session** by keeping in mind the "Guiding questions for the group work session" (slide 16) when reviewing the material

Self study session

Learning Material

- **Analysis of Proportions**: Aalen chapter 6.1-6.2, Kirkwood and Sterne (K&S) chapter 15
- **Comparing two proportions**: Aalen chapter 6.3, K&S chapter 16
- **Table Analysis:** Aalen chapter 6.5, K&S chapter 17

Group work session

Task

In your group (which should include 4-6 participants), jointly **revise the following guiding questions and provide an answer**

Guiding questions

1. What is the assumption that is the basis for CI and z-test for a proportion? Which are the two situations in which it can fail?
2. How do you interpret the role of the exposure when the associated relative risk (or OR) is larger than 1?
   And how when smaller than 1?
3. When a Chi-squared test with $\alpha = 5\%$ shows evidence to reject the null hypothesis, what does this imply on the 95%-CI for the risk difference? And on the one for the relative risk?