Biostatistics & Epidemiological Data Analysis using R

6

Hypothesis testing & Study planning

Stefan Konigorski

Health Intervention Analytics Group, HPI

November 23, 2022

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Content

Block	Class	Content	Date
R, Data manipulation, Descriptives	1	Overview & Introduction to R and data analysis	2022.10.19
	2	First steps in data analysis using R	2022.10.26
	3	Second steps in data analysis using R	2022.11.02
Epidemiology & Statistics: concepts	4	Epidemiological study designs	2022.11.09
	5	Estimation	2022.11.16
	6	Hypothesis testing and study planning	2022.11.23
	7	Missing data	2022.11.30
Data analysis w/ regression models	8	Linear regression I	2022.12.07
	9	Linear regression II	2022.12.14
	10	Regression models for binary and count data	2023.01.11
	11	Analysis of variance & Linear mixed models I	2023.01.18
	12	Linear mixed models II & Meta analysis	2023.01.25
	13	Survival analysis	2023.02.01
	14	Causal inference & Data analysis challenge	2023.02.08

(see full schedule online)

1. Review & Introduction
2. Hypothesis testing
   - Overview
   - Examples
   - Details
3. Study planning

Learning objectives

- Review content of last weeks class, probability concepts, and homework 5
- Understand the concept of statistical hypothesis testing theory, with selected examples.
- Get an overview about the statistical part in planning a study in form of sample size calculation.

Review of class 5

Estimation

- Point estimation (examples, properties)
- Standard errors (with, without bootstrap)
- Confidence intervals

Review of homework 5

Exercise 1: Probability distributions

- The functions `rnorm`, `rt`, `runif`, `rbinom` in R allow you to generate random numbers from the normal, t-, uniform, and binomial distribution.
- See file `R_5_homework_solutions.Rmd`.
- Remember: **The integral under the curve (= area of the bars in the histogram) is the probability!**
- To get the area under the curve (probability!) in the tails, the functions `qnorm` etc. can be used which computes the quantile (of the normal distribution).
- See file `R_5_homework_solutions.Rmd`.

Review of class 5

To remember

- Point estimation = give a best guess $\hat{\theta}$ for an unknown parameter of interest θ.
- Examples: estimators for expected value, variance, proportion, odds ratio.
- There are different ways how to derive the estimators, e.g. maximum likelihood estimation.
- These estimators can be evaluated with respect to different desirable properties (e.g. on average correct, low variance).
- Standard error = precision of point estimate = standard deviation of point estimate (computable by bootstrap)
- 95% confidence interval = you get a CI with 95% probability that contains the true parameter.

Questions?

Hypothesis testing - Overview

Motivating simulation example

Exercise 1

- Aim: Mimick the process of an empirical study.
- Let's consider the 17,640 children in the KiGGS dataset as the population of interest (children in Germany).
- Study question: Does the BMI of boys and girls differ?

- See `R_6_exercise_1.Rmd`.
- Load the KiGGS dataset.
- Take a random sample of 100 children.
- How can you answer the study question? $\longrightarrow$ Compute mean BMIs.

$\longrightarrow$ How can we make a decision that the BMIs are different/same?

Overview

Goal

The goal of a hypothesis test is to make a decision between a null hypothesis and a (complementary) alternative hypothesis[1].

Reasoning

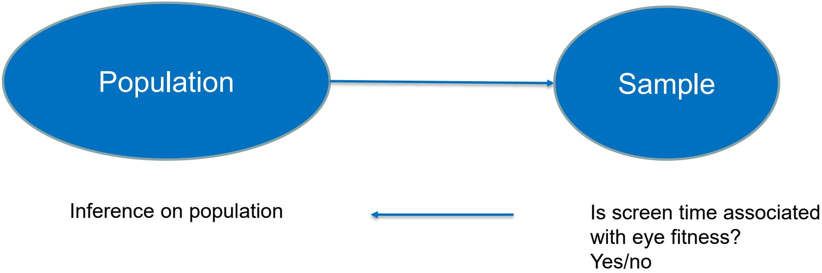
- Evaluate the evidence that the (null) hypothesis is compatible with the empirical observations.
- Depending on this evidence, decide for the null hypothesis or the alternative hypothesis.

[1] This can be extended to more than 2 hypotheses.

Overview

Goal

The goal of a hypothesis test is to make a decision between a null hypothesis and a (complementary) alternative hypothesis[1].



Population

Sample

Inference on population

Is screen time associated with eye fitness?
Yes/no

[1] This can be extended to more than 2 hypotheses.

Approach

1. Assume: The null hypothesis is correct.
2. Calculate the probability (= p-value), that you obtain such (or more extreme) observations as you have in your sample, given that the null hypothesis is true.
3. If this probability ...
   - is small (e.g. $< 5\% = \alpha$), then the empirical observations are hardly compatible with the assumption.
     $\longrightarrow$ Assumption must be wrong
     $\longrightarrow$ Reject null hypothesis
     $\longrightarrow$ Accept alternative hypothesis
   - is not small (e.g. larger than α), then there is not a strong evidence against the null hypothesis, therefore don't reject the null hypothesis.

Steps of doing a hypothesis test

1. Formulate your study question.
2. Translate this into a testable (null and alternative) hypothesis, i.e. formulate this in terms of parameters.
3. Choose your test.
4. Compute the probability (= p-value) through a test statistic (e.g. 't value', 'F value', χ^2 value) that captures the hypothesis.
5. The value of the test statistic can be computed from the data (e.g. you need sample mean and standard deviation) and compared to a theoretical distribution (e.g. using `qnorm` or automatically in R) to get the p-value.
6. Make a decision (based on p-value or test statistic).

Examples of (1) and (2)?

Hypothesis testing - Examples

2-sample t-test

Introduction

- Reopen exercise 1.
- Study question: Does the BMI differ between boys and girls?
- Formulate this in terms of parameters.
- Which (test) statistic can you imagine that captures what you want to test?
- Which distribution does this test statistic have?
- How can you visualize this distribution and the p-value?

2-sample t-test

Hypothesis

- 2-sided test: $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$
- 1-sided test: $H_0 : \mu_1 < \mu_2$ vs. $H_1 : \mu_1 \geq \mu_2$
- where μ_1 and μ_2 are the expected values of a random variable X in each group, respectively.

Test statistic

... for comparing the expected values of two independent groups:

- $T = \frac{\overline{x_1} - \overline{x_2}}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
- with the pooled standard deviation $s = \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}}$
- where n_1, n_2 and s_1, s_2 are the sample size and standard deviation (of X) in the two groups, respectively.

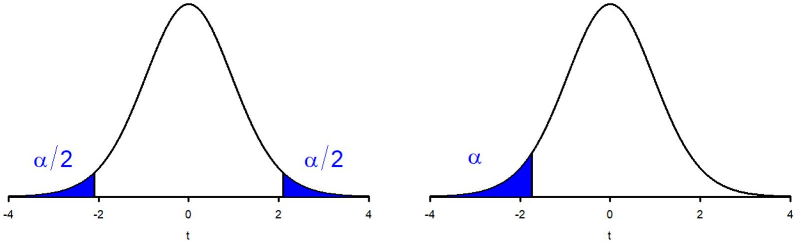
2-sample t-test

Distribution of test statistic

Statistical derivations tell you that T has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

P-value

The p-value for a 2-sided test is then the area under the probability density function of the t-distribution with $n_1 + n_2 - 2$ degrees of freedom, from (the absolute value of) T to infinity, multiplied by 2.



2-sample t-test

Assumptions

- X is normally distributed, in both groups.
- The variances in both groups are the same.
- The observations are independent.

Variations/Alternatives

- 1-sample t-test.
- 1-sample and 2-sample t-test for dependent groups.
- Mann-Whitney U-test based on ranks.

In R

- Use the `t.test` function for all t-tests.
- Use the `wilcox.test` function for the Mann-Whitney U-test.

χ^2 Test of independence

Aim

Test the independence of two categorical variables X, Y.

Data

	Tall	**Short**	**Total**
Hypertensive	n_{11}	n_{12}	$n_{11} + n_{12}$
Not Hypertensive	n_{21}	n_{22}	$n_{21} + n_{22}$
Total	$n_{11} + n_{21}$	$n_{12} + n_{22}$	Total n

χ^2 Test of independence

Data

	Tall	Short	Total
Hypertensive	n_{11}	n_{12}	$n_{1.}$
Not Hypertensive	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Hypothesis

- $H_0 : N_{ij} = N_{i.} \cdot N_{.j}/N$ for all i, j
- $H_1 : N_{ij} \neq N_{i.} \cdot N_{.j}/N$ for at least one i, j
- where $N_{..}$ are the (unknown) frequencies underlying the table.
- Note: This is here equivalent to testing that the OR $= 1$.

χ^2 Test of independence

Data

	Tall	Short	Total
Hypertensive	n_{11}	n_{12}	$n_{1.}$
Not Hypertensive	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Test statistic

- $\chi^2 = \sum_i \frac{(Observed_i - Expected_i)^2}{Expected_i}$ for all cells i
- $\chi^2 = \frac{(n_{11} - n_{.1} \cdot n_{1.}/n)^2}{n_{.1} \cdot n_{1.}/n} + \frac{(n_{21} - n_{.1} \cdot n_{2.}/n)^2}{n_{.1} \cdot n_{2.}/n} + \cdots + \frac{(n_{22} - n_{.2} \cdot n_{2.}/n)^2}{n_{.2} \cdot n_{2.}/n}$
- has a χ^2 distribution with $(k-1) \cdot (l-1)$ degrees of freedom, where k, l are the number of categories of the two variables.

χ^2 Test of independence

In R

- Use the `chisq.test` function for χ^2-tests.

Overview of tests

Association between two categorical variables

- χ^2-test or Fisher's exact test for independent samples.
- McNemar's test for dependent samples.

Association between binary and ordinal/metric variable

- Respective t-test for dependent/independent samples and normally-distributed outcome variable.
- Mann-Whitney U-test or Wilcoxon test for ordinal/not-normally-distributed outcome variable and dependent/independent samples.

Association between ordinal or metric variable

- Respective correlation coefficient.

Decisions and errors

Goal

When testing a null hypothesis H_0 versus an alternative hypothesis H_1, there can be 4 different scenarios:

		Truth	
		H_0	H_1
Decision	H_0	correct	wrong
	H_1	wrong	correct

Decisions and errors

Goal

When testing a null hypothesis H_0 versus an alternative hypothesis H_1, there can be 4 different scenarios:

		Truth	
		H_0	H_1
Decision	H_0	$1-\alpha$	β
	H_1	α	$1-\beta$

- α: type I error, significance level
- β: type II error
- $(1-\beta)$: power

Summary - how do you do a hypothesis test?

- Formulate your study question, set α.
- Translate this into a testable (null and alternative) hypothesis, i.e. formulate this in terms of parameters.
- Determine which test is appropriate (variable scales).
- Test the assumptions of the respective test.
- Perform the test i.e. compute the test statistic and p-value.
- Make a decision.
- Depending on goal of study: look at effect size, do multiple testing correction, perform other tests.
- Remember: statistical significance $\neq$ clinical significance.

Exercise 2

- In the KiGGS dataset, select one metric and one binary variable (or create one) and perform a 2-sample t-test.
- In the KiGGS dataset, select two categorical variables (or create them) and perform a χ^2 test.

Hypothesis testing - Details

Construction of hypothesis tests

How are the underlying test statistics derived?

- You can propose any test statistic that captures your study question.
- But the question is: which distribution does it have, and is the according test "good"?

Which test do you use?

- Parametric, if you know the distribution of the test statistic from theory (Wald test, Score test, likelihood ratio test, etc.)
- Nonparametric: using rank or permutation tests

Comparison of nonparametric & parametric tests

Nonparametric tests generally:

- don't make any assumption on the distribution of the variables.
- can be used for different measuring scales.
- don't need theoretical derivations of the distribution of the test statistic.
- are more robust against outliers (parametric tests, too, sometimes).
- are computationally more intensive.
- have lower power compared to parametric tests, if their assumptions are satisfied.

Tests

Aim: test $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$ for a parameter θ in a model.

Wald Test

The Wald test statistic $\frac{\hat{\theta}}{\widehat{SE}(\hat{\theta})}$ has a standard normal distribution (if $\hat{\theta}$ is asymptotically normally distributed).

Likelihood ratio test

- The likelihood ratio test (LRT) is based on the likelihood function.
- In more detail: the LRT statistic is based on the ratio of the unrestricted likelihood function, divided by the likelihood function restricted to the null hypothesis.

Which test is best?

best = highest power

Neyman-Pearson lemma

To answer this question, the Neyman-Pearson lemma can be used in certain situations, which guarantees that a test has the highest power (for a specified α).

In general

- Compare the power function of the tests.
- Compare the power of tests empirically.

Which test should you use for your analysis?

Look at recommendations in textbooks, or look at papers that have compared different tests empirically

Multiple testing

Problem description

- If you perform (multiple) hypothesis tests each to the level α, the error level α does not hold anymore over all tests.
- That means, the probability that the null hypothesis is falsely rejected in one or more tests (= family-wise error rate, FWER) is larger than α.
- There are different approaches how to adjust the tests in order to keep the FWER intact so that it is at most α.

Multiple testing

Methods to adjust for multiple testing

- Bonferroni correction: for k tests, multiply each p-value by k.
- Benjamini-Hochberg correction: Control the FDR (False Discovery Rate) instead of the FWER, which results in a more liberal correction.
- ...

Study planning

Relevant aspects of study planning

- Study question
- Study design
- Study population and sampling
- Measurement of variables
- Statistical analysis plan (SAP)
- Sample size
- Ethics approval, data protection, ...

⟶ Focus here on calculation of sample size.

Sample size computation - overview

Background

An important part of every study plan (i.e. before conducting the study!) is the question: How large should the sample size be?

How do you choose the sample size?

Determine the sample size by answering the question: How large should the sample size be so that you will find an association (between X and Y) in the statistical test of your main hypothesis with probability ..%, given that there actually is an effect?

Power of a statistical test

Power = P(conclude that there is an association | there really is an association)

Sample size computation - overview

Sample size

- The necessary sample size for finding a true association with power ..% is, among others, dependent on the chosen statistical test, the significance level α, and the size of the true effect.
- It is also dependent on the study design (e.g. of the respective sample size of the treatment arms), on how the groups are sampled, and if a 2-sided (equivalence/difference) or 1-sided (superiority) hypothesis should be tested etc.

Effect size

- ... describes how large an observed effect is.
- ... has to be specified to compute the optimal sample size.
- Example: Mean difference between groups.

Optimal sample size

Optimal sample size

Minimal sample size which is necessary so that an effect, which is at least as large as the specified effect size, will be identified with probability $(1 - \beta)$ for the significance level α.[2]

How to do this in practice?

1. Choose the study question, study design, statistical test.
2. Specify the possible effect size(s) (expected, min, max) through literature search or own thinking.
3. Compute the necessary sample size to achieve a certain power (which might depend on your study goal), e.g. 80% - by hand or using software.

[2] or: Minimal sample size to estimate an effect with at least the specified precision.

Compute the sample size manually

For 2-sided tests with significance level α and power $1-\beta$:

Examples

- Test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$ of $X \sim N(\mu, \sigma^2)$:

$$n \approx 2 \cdot \frac{\sigma^2}{(\mu_1 - \mu_2)^2} \cdot (z_{1-\alpha/2} + z_{1-\beta})^2$$

- Test $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ of $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ with same sample size:

$$n \approx 2 \cdot \frac{\sigma_1^2 + \sigma_2^2}{(\mu_1 - \mu_2)^2} \cdot (z_{1-\alpha/2} + z_{1-\beta})^2$$

- Compare two binomial proportions p_1 and p_2 with same sample size:

$$n \approx \frac{\left(z_{1-\alpha/2} \cdot \sqrt{2\left(\frac{p_1+p_2}{2}\right)\left(1-\left(\frac{p_1+p_2}{2}\right)\right)} + z_{1-\beta} \cdot \sqrt{p_1(1-p_1) + p_2(1-p_2)}\right)^2}{(p_1 - p_2)^2}$$

Compute the sample size using R

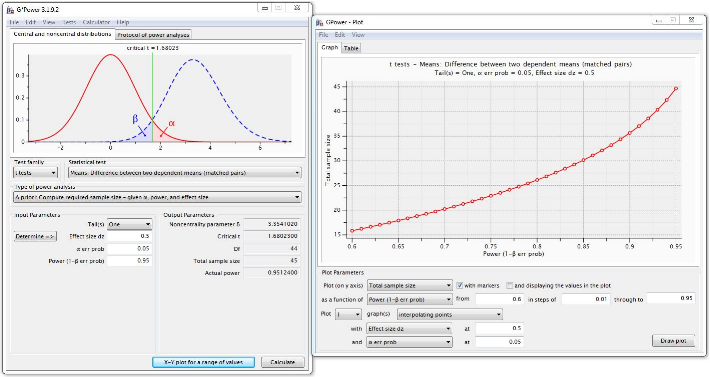
The R Package `pwr`[3] contains functions to compute the power/sample size for the following tests:

- 2 proportions (equal n): `pwr.2p.test`
- 2 proportions (unequal n): `pwr.2p2n.test`
- Balanced 1-way ANOVA: `pwr.anova.test`
- χ^2 test: `pwr.chisq.test`
- Linear model: `pwr.f2.test`
- Proportion (1 group): `pwr.p.test`
- Correlation: `pwr.r.test`
- t-test (one group, 2 groups, dependent/independent): `pwr.t.test`
- t-test (2 groups, unequal n): `pwr.t2n.test`

[3] `https://cran.r-project.org/web/packages/pwr`

Compute the sample size using G*Power

G*Power[4] contains a graphical interface to compute the power and sample size of different tests:



Test family: t tests

Statistical test: Means: Difference between two dependent means (matched pairs)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

- Tail(s): One
- Effect size dz: 0.5
- α err prob: 0.05
- Power (1-β err prob): 0.95

Output Parameters

- Noncentrality parameter δ: 3.3541020
- Critical t: 1.6802300
- Df: 44
- Total sample size: 45
- Actual power: 0.9512400

t tests – Means: Difference between two dependent means (matched pairs)
Tail(s) = One, α err prob = 0.05, Effect size dz = 0.5

Plot Parameters

- Plot (on y axis): Total sample size; ☑ with markers; ☐ and displaying the values in the plot
- as a function of: Power (1-β err prob) from 0.6 in steps of 0.01 through to 0.95
- Plot 1 graph(s) interpolating points
- with Effect size dz at 0.5
- and α err prob at 0.05

[4] http://www.gpower.hhu.de/

Exercise 3

- Aim: Compute the sample size using R or G*Power.
- Study question: investigate if a new drug has the side effect that it increases blood pressure.
- Study design: investigate one group that all get the drug, comparison of blood pressure before/after.
- Statistical test: Analyze using a t-test for two dependent groups.
- Question: Which sample size is necessary, to find a clinically relevant true effect with 80% power at $\alpha = 0.05$?

Exercise 4

- Aim: Compute the sample size using R or G*Power.
- Study question: investigate if biking to the HPI is associated with concentration in class or not.
- Study design: ?
- Statistical test: ?
- Question: Which sample size is necessary, to find a true effect with 80% power at $\alpha = 0.05$?

Questions?

References

Statistical fundamentals

- Knight K (1999). Mathematical statistics. CRC Press
- Rosner B (2010). Fundamentals of biostatistics. Brooks/Cole, Cengage Learning
- Wasserman L (2010). All of statistics. A concise course in statistical inference. Springer.

Benjamini-Hochberg Correction

- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 57, 289-300.

Homework

Homework

See file `R_6_homework.Rmd`.