INTRODUCTION

APPLIED STATISTICS (STAT200)

ORGANIZATION

SOME TECHNICAL INFORMATION

ORGANIZATION

My name: Jan Bulla

Email: jan.bulla@uib.no

Questions: best asked during / after lecture

Material: slides, assignments, solutions, literature… on Mitt UiB
Remind me if something is missing, but give me 1-2 days

Goals

- Provide a good understanding of basic and more advanced statistical methods in real-life type situations
- Allow you to take decisions: For given sample and open questions, you decide how to analyze data (and what not to do!)
- Carry out all analyses with the statistical software R and interpret the results

ORGANIZATION - SCHEDULE

- First week: starting today (17.1.)
  Last week: after ca. 11 weeks of lectures / labs
- Lecture: Tuesdays 16h-18h
  Thursdays 16h-18h
- Lab: Fridays 14h-16h
  Attention: the data labs take place at three different places
  - Datalab 2 (Høyteknologisenteret)
  - Seminarrom 1 (Realfagbygget)
  - Pi i fjerde (4th floor Realfagbygget) on the 18.1. (tomorrow!)
- No lectures / labs: - during the Winter holidays (25.02. - 01.03.)
  - during the Easter holidays (15.04. - 22.04.)
- Exam: 06.06.2019, see official site for updates!

Deviations from the schedule on MittUiB will be announced via the messaging system

ORGANIZATION – EXAMINATION

- Assignments
  - ~7 in total
  - R-based
  - one hand-in at the middle of the semester (no. 4 or no. 5)
- Exam
  - written
  - no documents allowed
- Grades (2016):

A	1
B	11
C	13
D	6
E	1
F	1

LITERATURE LIST

1. Dalgaard (2008) Introductory Statistics with R
2. Zuur et al. (2009), Mixed Effects Models and Extensions in Ecology with R
3. Binham & Fry (2010), Regression Linear Models in Statistics
4. Shahbaba (2012) Biostatistics with R
5. Fahrmeir et al. (2013) Regression - Models, Methods and Applications
6. Devore & Berk (2012) Modern Mathematical Statistics with Applications

Long, but:

- Only selected chapters are required
- We focus on the applied part
- Parts of one book (6.) serve for more than 50% of the course

MOTIVATING EXAMPLE: R

Why R? Not Excel? Really, no Excel??

NO!

- R is free
- R is easy to use
- R is the most common program used for statistical analyses today
- R forces you to understand what you are doing
- Assignment 1 will get you started, help available tomorrow

Some of the basic topics we will talk about:

- Simple hypothesis testing, power,…
- One-way analysis of variance (ANOVA)
- Two-way ANOVA, ANCOVA,…
- Simple and multiple linear regression
- Non-parametric procedures

PROGRAM – ADVANCED PART

Sometimes, the methods mentioned above are not optimally suited for a particular setting, and more appropriate techniques exists. For example:

- Mixed effects models (for longitudinal data)
- Generalized linear models (for count data,…)
- Bootstrap (for confidence interval)
- Survival analysis (for determining time until the occurrence of an event)
- Hidden Markov models
- …

Some of them will be treated after the basic part!

MOTIVATING EXAMPLE: STATISTICAL EXPERIMENTS

Let's start with a typical real-life motivating example!

SECTION 0

DEFINITIONS AND NOTATIONS, KNOWN FROM ELEMENTARY COURSES

TERMINOLOGY

Let's establish some terms you will need throughout this course. In case some of the terms or concepts are unclear:

- Review old courses, use Wikipedia, Google, YouTube, text book of your choice,…
- Do it immediately, don't wait!
- If the time allows it: ask during lab sessions

TERMINOLOGY

- Population: a group of phenomena (people or things) that have something in common
- Individual: an element or member of the population
- Sample: a smaller group of members of a population selected to represent the population
- Variable: a characteristic of an individual to be measured or observed, e.g. eye color / weight of a person, number of TV's in a home,…

Note: in many textbooks, "population" and "sample" are used in an (almost) interchangeable way. Don't get confused by this – keep in mind that in real life the population cannot be observed in general.

In principle there are two types of variables:

- Qualitative
  - are represented by attributes such as gender or marital status
  - are either ordinal and nominal
- Quantitative
  - take numerical measurements such that addition or averages makes sense
  - are either continuous or discrete

VARIABLES

Qualitative (categorical)

- Nominal: there's no ranking/ordering (gender, eye color)
- Ordinal: they can be arranged in some order (degree of satisfaction with cell phone provider,…)

Quantitative

- Discrete: the set of possible values is finite or countable (number of children in a family)
- Continuous: they take an infinite number of possible values, they are defined over an interval of values (height, weight)

MORE DEFINITIONS

- Parameters: a key role in statistics is played by parameters, i.e. quantities that represent a particular characteristic of a population.
  The same transfers to statistical models, which are usually also specified by parameters
- Statistics (estimator): a statistics (or estimator) is a function of sample values which does not involve any parameters
- Hypothesis: a hypothesis usually constitutes an assumption about population parameters
- Statistical inference : by statistical inference we understand the process of learning about what we do not observe (parameters) using what we observe (data)

STATISTICAL INFERENCE

The main purposes of statistical inference are:

1. Estimation of parameters on the basis of sample observations through a statistic
   ⇒ estimation theory
2. Evaluation of a hypothesis about population parameters, by using their estimates
   ⇒ hypothesis testing

Definition: Given a random variable (r.v.) X, a sample of length / size n denoted by $X_1, \dots, X_n$ represents *n* (usually) independent individuals / experiments for which the same quantity (X) is measured.

Example: Let X represent the height of an individual, and assume *n* individuals are measured. Then X_i will be the height of the i^{th} individual

Definition: Let X be a random variable (r.v.) and $X_1, \dots, X_n$ a sample.
An estimation / estimator, often denoted by f / $\hat{\theta}$ /…, is a function of the sample, i.e. $f(X_1, \dots, X_n)$ / $\hat{\theta}(X_1, \dots, X_n)$.

Examples

- Mean estimation: $\bar{X} = \frac{X_1 + \cdots X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$
- Variance estimation: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- Covariance estimation: for X, Y r.v. with observations $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$, the covariance is $cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$ and their correlation is $r(X, Y) = \frac{cov(X,Y)}{S_X S_Y}$.

HYPOTHESIS TESTING – EXAMPLE

We explain the principle of hypothesis testing by means of a simple one-sample experiment. Suppose:

- We randomly give a sample of participants from our population a pill
- Then, their IQ is evaluated. We know that the IQ of people not taking the pill is 100.

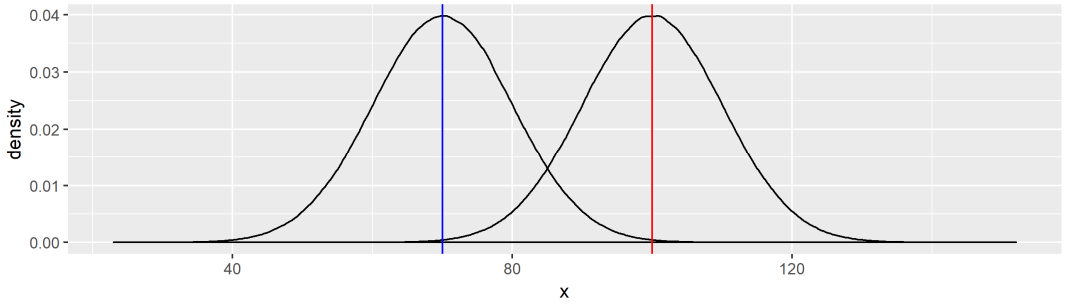
Here, the amount of pill taken (0 / 1) is an independent or explanatory variable, also called predictor. The IQ is the dependent or response variable.

We can investigate the following experimental hypotheses:

1. The pill works, by increasing / decreasing the IQ score
2. The pill does not work, because IQ scores do not change

HYPOTHESIS TESTING – EXAMPLE

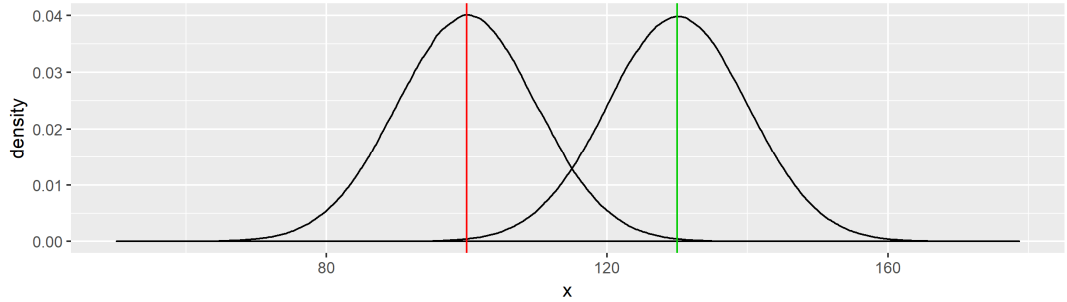
- In practice one is generally interested to show an effect. Hence, assume that we expect to see an effect of the pill, our experimental hypothesis
- The experimental hypothesis is directly connected statistical analysis as it leads to the formulation of the alternative hypothesis: the experiment is working as predicted, i.e. $H_1: \mu \neq 100$



⇐ Without pill

With pill

OR

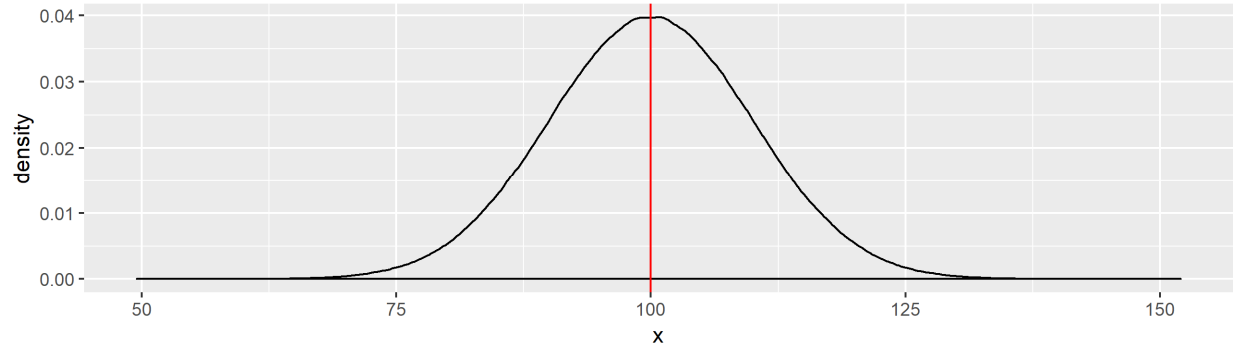


⇐ Without pill

With pill

HYPOTHESIS TESTING – EXAMPLE

On the other hand, the null hypothesis - that we denote by H_0 - corresponds to the (less interesting / irrelevant / … experimental) hypothesis that the independent variable (pill) has no effect, i.e. H_0: $\mu = 100$.



⇐ With pill AND without pill

Remark: together, H_0 and H_1 include all possible values of μ

Let's additionally assume that the following conditions are fulfilled:

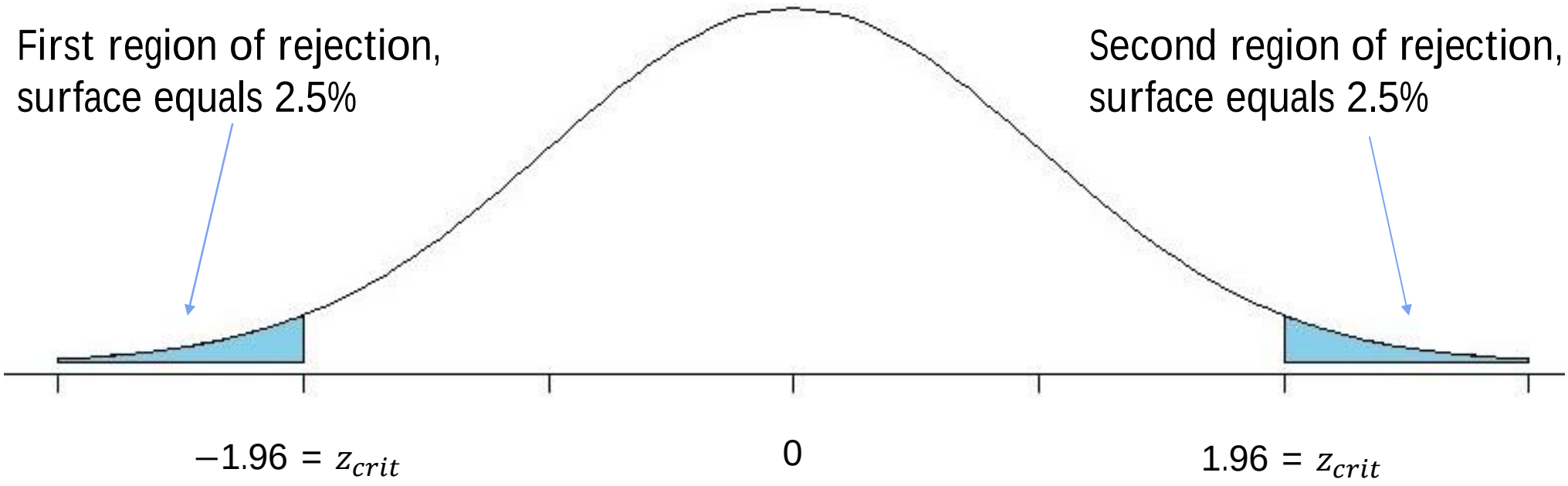
1. Our sample has been selected randomly
2. The dependent / response variable (IQ) is (approximately) normally distributed
3. We know the true standard deviation (σ_X) of the population – e.g., we know from previous researches that the IQ of the population is 100 with s.d. 15.

This setting will lead to the well-known *z*-test

We proceed as follows:

1. Choose α, the criterion or level, that defines samples as being too unlikely to represent the underlying population. Usually, the criterion takes the value 5% (but can also be 1%, 0.1% ...)
2. Locate the region of rejection, which may be in both tails (two-tailed test) or in one tail (one-tailed test)
3. Determine the critical value z_{crit} for our given α. In our case, for $\alpha = 0.05$, $z_{crit} = \pm 1.96$ (Gaussian standard distribution)
4. Calculate the *z*-score z_{obs} by the data, with the formula:
   $z_{obs} = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$, where $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$ with *n* being the number of observations

Graphical presentation



As last step, we compare z_{obs} and z_{crit}. Then:

- Reject H_0 (and accept H_1) if z_{obs} falls beyond the critical value (i.e. if $|z_{obs}| > z_{crit}$).
  In our example, we reject the (statistical) null hypothesis that the pill does not work and thus confirm the experimental hypothesis
- Do not reject H_0 if z_{obs} does not fall beyond z_{crit}.
  This means that our result did not provide any convincing evidence of a relationship between dependent and independent variable (in our example, IQ and taking the pill or not).

Note: not rejecting H_0 does NOT mean that we (should) accept H_0!

TYPE I AND TYPE II ERRORS

Recall that we may commit two types of errors by taking a decision (based on the sample).

- Type I error: Rejecting H_0 when H_0 is true.
  In our example, we conclude that the predictor (amount of pills taken) has an effect, although this is not true. This type of error can be controlled by the criterion α.
- Type II error: Retaining H_0 when H_0 is false.
  In our example, we do not conclude the presence of an effect of the pill, although there is one present. In other words, we fail to identify that the predictor (pills taken or not) has an effect.

Note: The error of type II cannot be controlled directly - an unlucky sample draw can always happen.
However, the number of type II errors decreases with increasing same size.

TYPE I AND TYPE II ERRORS

In any experiment, the results of your inferential procedure (test) will place you in one of the columns of the following table:

		Decision about H_0	
		Reject H_0	Do not reject H_0
Truth about H_0	H_0 is true	Type I error ($p = \alpha$)	Correct, avoid type I error ($p = 1 - \alpha$)
	H_0 is false	Correct, avoid type II error ($p = 1 - \beta$)	Type II error ($p = \beta$)

POWER

In our research (and in most practical situations), the goal is to reject H_0 (for a given α) when it is false.
In our example, this means to conclude that the pill works when the truth is that the pill works.

This ability is called power, the probability of rejecting H_0 when it is false. Thus, power is the probability of not committing a type II error, and equals $1 - \beta$.

- Usually, one seeks to maximize power, such that one is confident of not committing type II errors when retaining H_0
- Power increases with increasing sample size. A "good" sample size can often be determined by simulation studies

P-VALUE

Recall that the first step of hypothesis testing requires:

1. Choosing a criterion (or level) α
2. Locating the region of rejection, being either in both tails or in only one tail
3. Determining the critical value(s) (for the *z*-test it is z_{crit}) for our given α
4. Checking if the value of the test statistic (for the *z*-test it is z_{obs}) is smaller or greater than the critical value(s)

However, in many cases this procedure is eased considerably by statistical software, which commonly provide a so-called p-value (or P-value) when a test is carried out.

P-VALUE – DEFINITIONS

Roughly speaking, a low p-value shows that the sample considered is unlikely under H_0.
More precisely, the p-value of a test can be defined as follows:

1. The p-value is the probability of obtaining a test statistic at least as extreme as the one actually observed, assuming that H_0 is true
2. For a statistical test, the p-value corresponds to the lowest possible level of the test (the α) for which H_0 can be rejected

The lower the p-value, the less likely the result is if the null hypothesis is true, and consequently the "more significant" the result is, in the sense of statistical significance.

P-VALUE

One often accepts the alternative hypothesis (i.e. rejects the null hypothesis) if the p-value is less than 5% or 1%, corresponding to a 5% or 1% chance, respectively, of rejecting the null hypothesis when it is true (type I error).

Most statistical software directly indicate p-values. Often, particular symbols are shown if a p-value lies below a certain threshold:

P-value	Symbol	Term for description
< 10%	·	tendency
< 5%	*	Significant result
< 1%	* *	Highly significant result
< 0.1%	* * *	Very highly significant result

P-VALUES – BE CAREFUL!

Remarks

- P-values are important and useful tools, but lots of recent literature shows that they must be interpreted with care: mistakes can happen during their interpretation, and some authors argue against the use of p-values at all.
- For example, it is important to remember that – however small the p-value – there is always a finite chance that the (significant) result – obtained by accepting H_1 – is simply a rare accident
- See, e.g., https://www.youtube.com/watch?v=OcJImS16jR4 by Geoff Cumming, author of *Introduction to the New Statistics* – maybe basis for reading assignment during this course