

# MS-C1620 Statistical inference

## 1 Descriptive statistics

Jukka Kohonen

Department of Mathematics and Systems Analysis  
School of Science  
Aalto University

Academic year 2020–2021  
Period III–IV

# Course personnel

Lectures on Thursday 8–10 in ZOOM (link on course page).

- Lecturer: Jukka Kohonen

Office hours: by email, [jukka.kohonen@aalto.fi](mailto:jukka.kohonen@aalto.fi)



Weekly exercise sessions.

- Head assistant: Paavo Raittinen  
[paavo.raittinen@aalto.fi](mailto:paavo.raittinen@aalto.fi)



See to the course page for materials and announcements.

<https://mycourses.aalto.fi/course/view.php?id=29630>

## Grading

The course grade (0-5) is determined based on the total of exam points (0-24) and exercise points (0-6).

The total points correspond to the following course grades,

1 = 15 total points

2 = 16 total points

3 = 19 total points

4 = 22 total points

5 = 25 total points

Additionally, grade 1 is also awarded to those who get 12 points from the exam alone.

The exercise points are valid during the year 2021.

## Exercises

The course exercises consist of two kinds of problems.

- **Homework problems:** the first problems of each week's exercise sheet are homework problem which must be completed before that week's exercise session. **Exception: the first exercise.**
- **Class problems:** the remaining problems of each week's exercise sheet are class problems that will be solved together in the exercise sessions.

Exercise points will be awarded from the sessions as follows:

- 1/2 exercise points: both active attendance and homework.
- 1/4 exercise points, only active attendance.

There are a total of 12 exercise sessions, meaning that one can obtain a total of 6 exercise points.

**The exercise points are rounded up to the nearest integer.**

## Exercises

If none of the exercise group times is suitable for you but you would still like to get the exercise points, contact the head assistant. Note that you need an extremely good reason for getting points without attendance!

# Lecture topics

- ① Descriptive statistics
- ② Confidence intervals and hypothesis testing
- ③
- ④
- ⑤
- ⑥
- ⑦
- ⑧
- ⑨
- ⑩
- ⑪
- ⑫

# Contents

- 1 Introduction
- 2 Visual descriptive statistics
- 3 Numerical descriptive statistics

# Contents

- 1 Introduction
- 2 Visual descriptive statistics
- 3 Numerical descriptive statistics

## Statistics as a field

Statistics is the science of collecting, organizing, analyzing and interpreting data.

Statistical models are mathematical and based on probability theory.

The practice of statistics can be considered to have begun in ancient Babylon, Egypt and Rome, as population statistics. Data was collected for the government, for example, about birth rates. The word *statistics* comes from the Latin words *statisticum collegium* (council of state).

# Descriptive vs. inferential statistics

Statistical methods and procedures can be divided roughly into two categories.

- **Descriptive statistics** aims at providing a concise **summary** of the **data**. The summary may be numerical or graphical or of some other form.
  - ▶ Examples of descriptive statistics: numerical tables, average values, deviations and visualizations.
- **Inferential statistics** draws conclusions about a **population** based on a **sample**. Statistical inference is based on mathematical modeling and probabilities.
  - ▶ Examples of inferential statistical methods: confidence intervals, hypothesis testing, linear regression.

Note that there still exists methods that do not really fit into either of the above categories. For example *principal component analysis* can be considered a method of **exploratory data analysis**.

# Population and sample

Most of statistical projects revolve around trying to understand a population based on a sample.

- Population is the collection of all the people, items, or events about which one wants to make inferences (students at Aalto University).
- Sample, is a subset of the population (i.e. the people, items, or events) that one collects and analyzes to make inferences on the population (200 randomly chosen students).
- Observation is an element of the sample (Helena, a student at Aalto University).

In a typical project, descriptive statistics are first used to understand the sample and select suitable methods of inferential statistics, using which we then try to understand the population.

# Variables and data

In statistical research, a sample consists of **data** which is made up of the observed values of selected **variables**. Sometimes the data points (the values of the selected variables) are also called **observations**.

Examples of variables:

- Temperature, height, blood pressure (**numerical** variables, perhaps also *continuous*)
- Clothing size (... S, M, L, ...), (**ordinal** variable)
- Gender, eye colour (**categorical** variables)

Variables of different types usually need to be analyzed with different methods.

(Look up also “levels of measurement” )

## Statistical research projects

Statistical research projects usually consist more or less of the following steps:

1. Define the research topic and the relevant research questions.
2. Define of the population and the variables of interest.
3. Plan of the sample collection such that it is representative of the population.
4. Collect the sample.
5. Organize the sample.
6. Study the sample using descriptive statistics.
7. Model and analyze the sample using inferential statistics.
8. Evaluate the results critically.
9. Communicate the (lack of) findings.

# Different statistical studies

Statistical research projects can be conducted in several different ways, depending on the research questions, population, goals and resources.

- **Observational research** simply observes the sample without changing any existing conditions.
  - ▶ The lung cancer risk of smokers is compared to the lung cancer risk on non-smokers.
  - ▶ The effect of the reputation of an university to the salaries of its graduates is studied.
- **Controlled experiment** examines the effect of one variable to another by controlling existing conditions.
  - ▶ The effect of allergy medicine is compared to the effect of placebo by randomizing patients to two groups.
  - ▶ The effect of the type of soil to the growth rate of plants is studied by randomly planting plants of the same species to different types of soil.

## Different statistical studies

- **Simulations** use mathematical modeling to mimic natural conditions or processes.
  - ▶ The spread of the Ebola virus is predicted by applying computer simulations.
  - ▶ The safety of a new car model is tested using crash test dummies.

The previous types of studies have various sub-cases. For example, a **survey** is an observational study where a representative sample of the population answers some particular questions.

# Contents

- 1 Introduction
- 2 Visual descriptive statistics
- 3 Numerical descriptive statistics

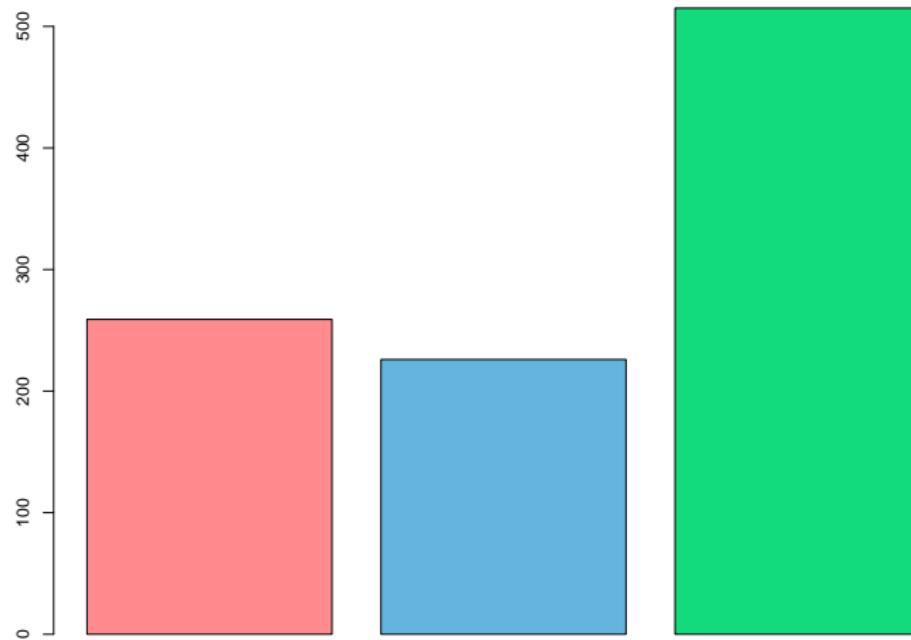
# Visualization

The choice of a suitable visualization methods depends on the number of variables (univariate, bivariate, multivariate) and the types of the variables (continuous, discrete). Some examples:

- Discrete variable:
  - ▶ Bar plot
  - ▶ Pie chart
  - ▶ Dot chart
- Continuous variable:
  - ▶ Box plot
  - ▶ Histogram
- Bivariate (continuous  $\times$  continuous):
  - ▶ Scatter plot
  - ▶ Two dimensional histogram
- Bivariate (continuous  $\times$  categorical):
  - ▶ Multiple boxplots
  - ▶ Colored scatter plots

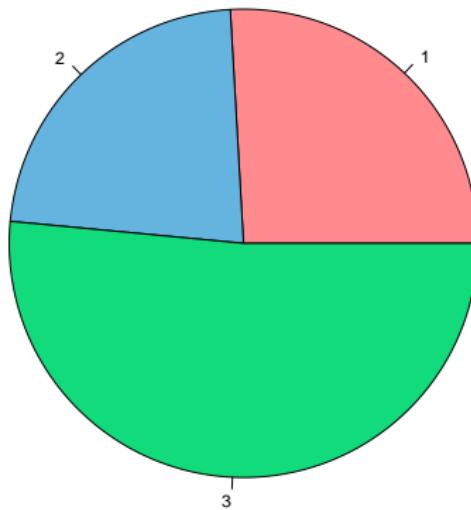
## Bar plot

Single categorical variable.



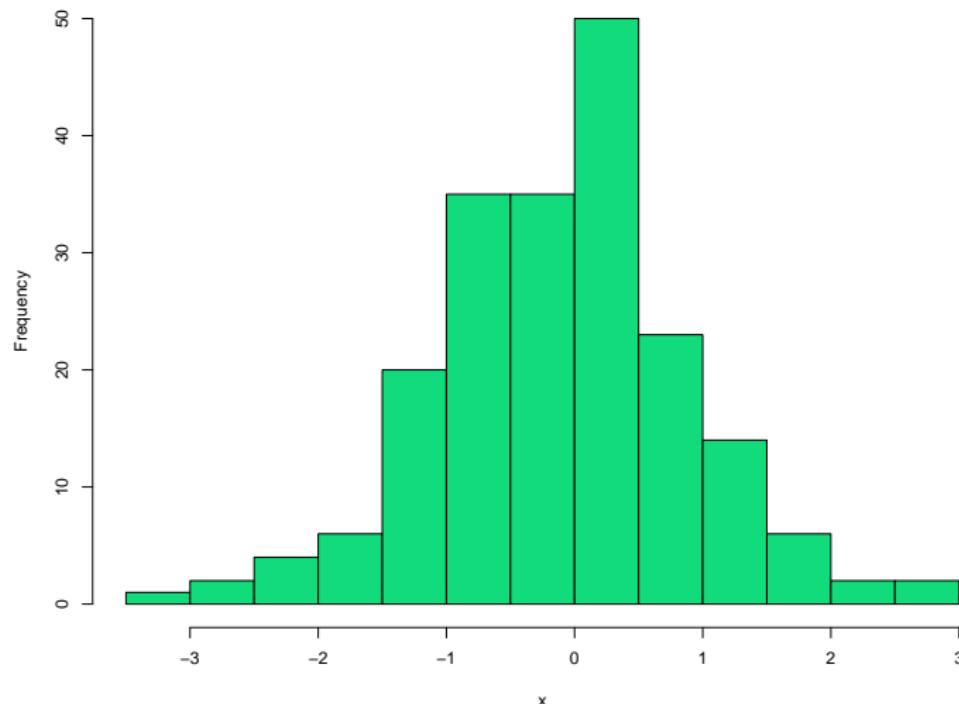
## Pie chart

Single categorical variable.



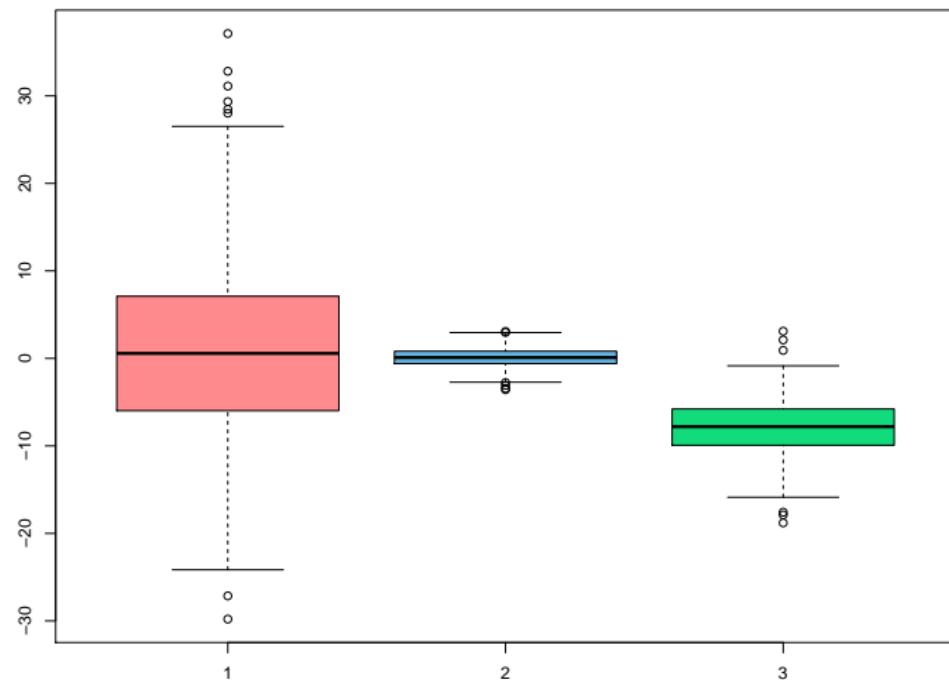
# Histogram

Single continuous variable.



# Box plot

Continuous  $\times$  Categorical



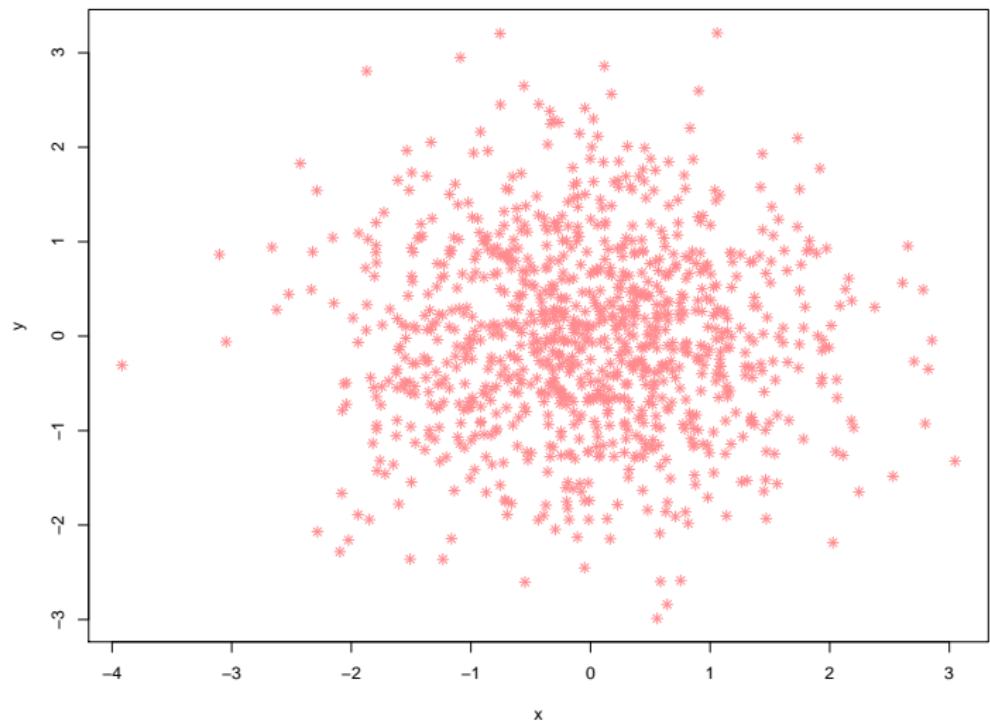
## Interpretation of a box plot.

Let  $Q_1$  and  $Q_3$  denote the 25% and 75% sample quantiles and let  $x_{min}$  and  $x_{max}$  denote the minimum and maximum values of the sample

- The line in the middle of the box is the sample median.
- The lower and upper endpoints of the box are  $Q_1$  and  $Q_3$ , i.e., the box contains 50% of the data.
- The upper “whisker” is located at  $\min\{x_{max}, Q_3 + 1.5(Q_3 - Q_1)\}$ .
- The lower “whisker” is located at  $\max\{x_{min}, Q_1 - 1.5(Q_3 - Q_1)\}$ .
- Outlying points (not inside the whiskers) are marked using circles.

The box plot allows the simultaneous inspection of location, scatter, symmetry and outliers.

# Scatter plot



## Good plotting practices

Remember to make your plots easy to understand.

Numerous examples of the kinds of plots you should not make can be found online. Check for example

[https://en.wikipedia.org/wiki/Misleading\\_graph](https://en.wikipedia.org/wiki/Misleading_graph)

# Contents

- 1 Introduction
- 2 Visual descriptive statistics
- 3 Numerical descriptive statistics

## Measures of location

Measures of location describe where the *center* of the data lies.

They are used to summarize the typical behavior in the data set.

- The average height of the sample.
- Salary level such that 50% of people have salaries above that.
- The most common number of children.

The following slides list various measures of location, with robust (tolerant to outliers) measures marked in **red**.

## Different means

Let  $x_1, x_2, \dots, x_n$  be independent and identically distributed (i.i.d.) observations of a random variable  $x$ .

- The sample **mean**,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

The sample mean estimates the expected value  $\mu = E(x)$  of the random variable  $x$ .

- The  **$\alpha$ -trimmed mean** is the mean of the sample after discarding the proportion  $\alpha$  of both smallest and largest observations,
- **Weighted means** give variable weights for different observations,

$$\sum_{i=1}^n w_i x_i,$$

such that  $\sum_{i=1}^n w_i = 1$ .

# Median

Let  $y_1 < y_2 < \dots < y_n$  be ordered values of the data.

- The sample **median** is the middle value of the ordered values.
- If the number of observations is even, the sample median is the average of the two middle observations.
- The sample median estimates the population median  $m_x$ , the value with the following property

$$P(x < m_x) \leq \frac{1}{2} \quad \text{and} \quad P(x \leq m_x) \geq \frac{1}{2}.$$

# Quantiles

- The sample  **$\beta$ -quantile**,  $0 < \beta < 1$ , is the data point  $y_k$ , where  $k = \lceil \beta n \rceil$  and  $n$  is number of observations.
- The sample  $\beta$ -quantile estimates the population  $\beta$ -quantile  $\beta_x$ , defined as

$$P(x < \beta_x) \leq \beta \quad \text{and} \quad P(x \leq \beta_x) \geq \beta.$$

- 0.25- and 0.75-quantiles are called first and third **quartiles**, and
- the **mid-hinge** is their average

$$\frac{y_{\lceil 0.25n \rceil} + y_{\lceil 0.75n \rceil}}{2}$$

## Mode

The sample **mode** is the (possibly non-unique) value that has the highest frequency in the sample.

Mode estimates a value of a qualitative variable or discrete quantitative variable that has the highest probability.

Mode is rarely useful outside of categorical data.

## Measures of scatter

Measures of scatter describe how far away from its center the data lies.

They are used to summarize the spread of the data set.

- The variability of height in a population
- The magnitude of measurement errors.

The following slides list various measures of scatter, with robust (tolerant to outliers) measures marked in **red**.

# Variance

- The sample **variance**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

The sample variance estimates the population variance

$$\sigma^2 = E[(x - \mu)^2].$$

- The sample **standard deviation**,

$$s = \sqrt{s^2},$$

is often preferred over variance as it is measured in the same units as the data.

## Median absolute deviation

The **median absolute deviation**, MAD, is the median of the sample  $|x_1 - m_x|, |x_2 - m_x|, \dots, |x_n - m_x|$ .

MAD is often multiplied with by the constant 1.4826 to make it a consistent estimator of the standard deviation in a normal model.

## Range

- The **sample range** is the interval  $[x_{min}, x_{max}]$  and its length is

$$x_{max} - x_{min}.$$

- The **interquartile range**, IQR, is the distance between the first and third quartile,

$$y_{\lceil 0.75n \rceil} - y_{\lceil 0.25n \rceil}$$

IQR is often multiplied with by the constant 0.7413 to make it a consistent estimator of the standard deviation in a normal model.

## Measures of skewness and kurtosis

Thus far, roughly:

- First moment = measures of location
- Second moment = measures of spread

If we continue onward, the next two moments give us measures of *skewness* and *kurtosis*

Skewness describes the deviation of the data from symmetry.

Kurtosis describes the heaviness of the tails of the data.

The following slides list various measures of skewness and kurtosis.

## Skewness

The sample **skewness** is

$$\hat{\gamma} = \frac{m_3}{s^3},$$

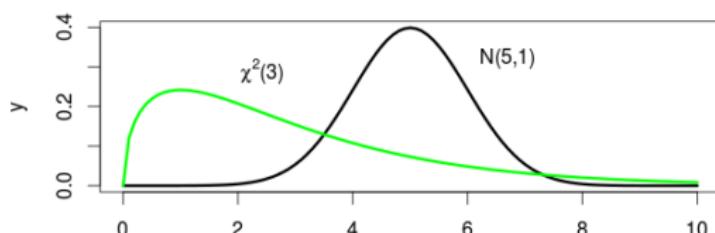
where

$$m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3.$$

Sample skewness coefficient estimates the population skewness,

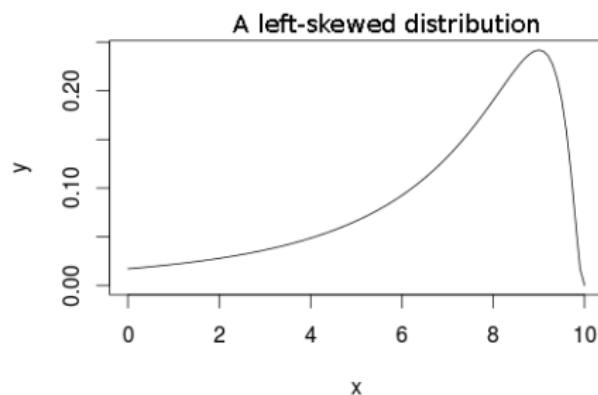
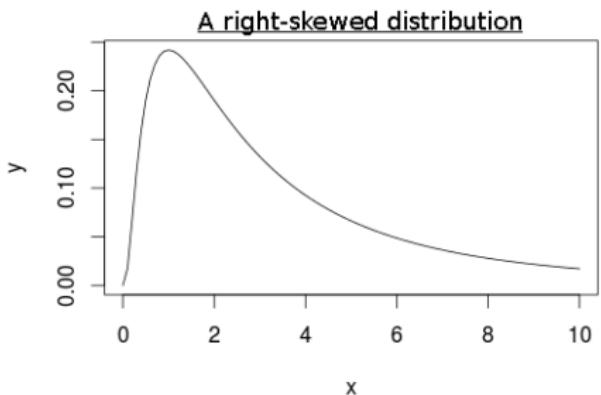
$$\gamma = E \left[ \left( \frac{x - \mu}{\sigma} \right)^3 \right].$$

A symmetric and a skewed distribution



## Interpretation of skewness

- If the skewness coefficient  $\hat{\gamma} > 0$ , then the distribution is *skewed to the right* (positively skewed). I.e. the distribution has a long right tail and the mass of the distribution is concentrated on the left.
- If  $\hat{\gamma} < 0$ , then the distribution is *skewed to the left* (negatively skewed). I.e. the distribution has a long left tail and the mass of the distribution is concentrated on the right.



## Median skewness

The **median skewness**,

$$v_2 = \frac{3(\bar{x} - m_x)}{s}.$$

The underlying reasoning is that for symmetrical distributions the sample mean and the sample median estimate the same population value.

The mean and the median in the median skewness could be replaced with any two measures of location to obtain different measures of skewness.

## Kurtosis

The sample **kurtosis coefficient** is

$$\hat{\kappa} = \frac{m_4}{s^4} - 3,$$

where

$$m_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4.$$

The sample kurtosis coefficient estimates the population kurtosis

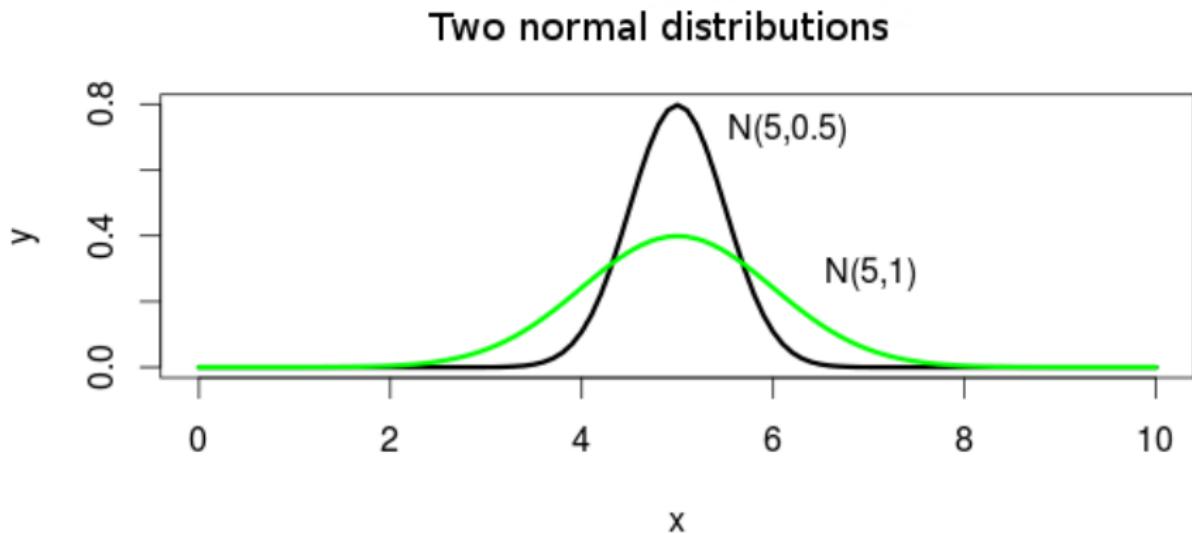
$$\kappa = E \left[ \left( \frac{x - \mu}{\sigma} \right)^4 \right] - 3.$$

## Interpretation of kurtosis

- A random variable with normal distribution has kurtosis value 0.
- If the kurtosis value is  $\kappa > 0$ , then the distribution has heavier tails than the normal distribution.
- If  $\kappa < 0$ , then the distribution has lighter tails than the normal distribution.

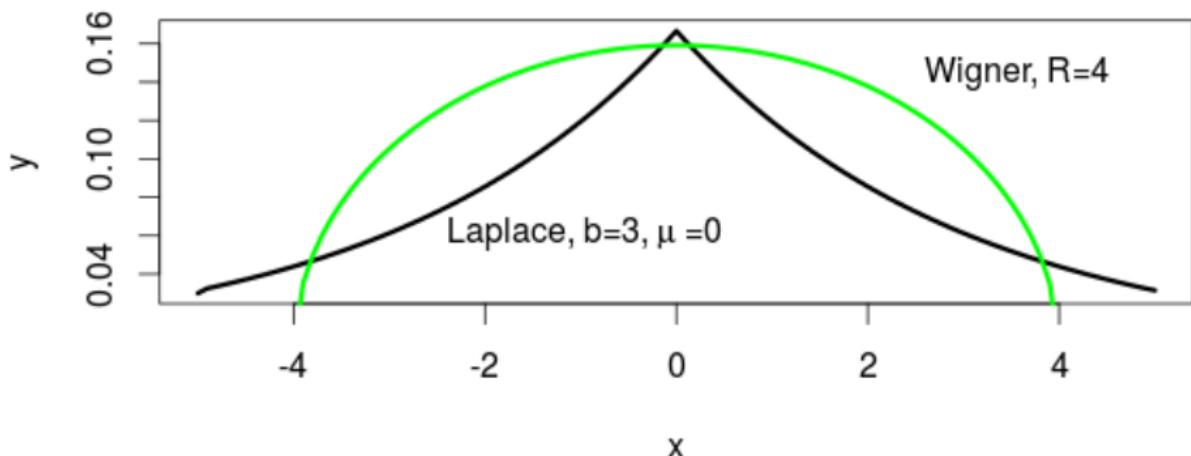
## Zero kurtosis

Two normal distributions with different parameters. Kurtosis is the same for both.



## Small and large kurtosis

Wide and sharp distributions



## Descriptive statistics for multivariate data

Numerous different descriptive statistics exist also for multivariate data.

The measures are commonly only defined for bivariate data and then computed for all possible pairs of variables.

The most common bivariate descriptive statistic is the *correlation*.

## Correlation

Let  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  be i.i.d. observations of a bivariate random variable  $(x, y)$ .

- The **sample covariance**

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

estimates the population covariance  $\sigma_{xy} = E[(x - E[x])(y - E[y])]$ .

- The **sample correlation**

$$\hat{\rho}(x, y) = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

estimates the population correlation  $\rho(x, y) = \sigma_{xy}/(\sigma_x \sigma_y)$ .

Correlation measures the linear dependence between two random variables.  
The coefficient is always in the interval  $[-1, 1]$

# MS-C1620 Statistical inference

## 2 Confidence intervals and hypothesis testing

Jukka Kohonen

Department of Mathematics and Systems Analysis  
School of Science  
Aalto University

Academic year 2020–2021  
Period III–IV

# Contents

- 1 Confidence intervals
- 2 Hypothesis testing
- 3  $t$ -tests
- 4 Variance tests

## Point estimates

With parametric models, we often want to estimate the value of some **parameter** using the sample  $x_1, \dots, x_n$ .

- We estimate the expected value  $\mu$  of a normal distribution  $\mathcal{N}(\mu, \sigma^2)$  by the maximum likelihood estimate  $\bar{x}$ .
- We estimate the population skewness coefficient  $\gamma$  by the corresponding sample estimate  $\hat{\gamma}$ .

Such estimates are called **point estimates** of the parameter.

(A *parameter* is a quantity characterizing the population / generating distribution, similar to *statistic* which is property of a sample.)

A point estimate on its own rarely gives us enough information.

To gain some idea of the **precision** of a point estimate, they are usually accompanied with some measures of their accuracy.

## Confidence interval

A **confidence interval** gives an estimated range of values which is likely to contain the value of an unknown population parameter.

The **confidence level** of a confidence interval determines the probability that the confidence interval produced (interpreted as a random interval) will contain the true parameter value.

E.g. if 95% confidence intervals for an unknown parameter are computed from 100 independent samples, approximately 95 of the these will contain the true parameter value — but we do not know which!

Note that any particular realized confidence interval either contains the true value or not; the 95% frequency concerns the probability *in the sampling process*

## The bootstrap

The standard formulas for confidence intervals either make heavy parametric assumptions or work only for parameters estimable by means (CLT).

The standard non-parametric procedure for estimating confidence intervals is known as the **bootstrap**.

Bootstrap creates pseudo-samples by drawing  $n$  observations *from the data, with replacement, repeating* the procedure for a large number of times.

If  $n$  is large enough, the pseudo-sampling approximates true sampling from the population.

## Bootstrap confidence intervals

Let  $x_1, x_2, \dots, x_n$  be independent and identically distributed (i.i.d.) sample from a distribution (parametric model)  $F_x$ .

Let  $\theta$  be a parameter of the distribution  $F_x$  and assume that  $\hat{\theta}$  is a point estimate of  $\theta$ .

An approximate confidence interval for  $\theta$  can now be obtained by bootstrap resampling as follows:

## Bootstrap confidence intervals

- ① Select  $n$  data points randomly with replacement from the original sample  $\{x_1, x_2, \dots, x_n\}$ . Each data point can be selected once, multiple times, or not at all. (Note that the sample size of the new sample is the same as the sample size of the original sample.)
- ② Use this new sample to calculate a new estimate for the parameter  $\theta$ .
- ③ Repeat the previous steps  $B$  times.
- ④ After the replications, order the  $B$  estimates from the smallest to the largest.
- ⑤ A  $100(1 - \alpha)\%$  confidence interval is now obtained by choosing the  $\lfloor B \times (\alpha/2) \rfloor$  ordered estimate as the lower endpoint and the  $\lfloor B \times (1 - \alpha/2) \rfloor$  ordered estimate as the upper endpoint.

## Exact confidence intervals

Often, when the type of the distribution is known, also exact confidence intervals can be calculated.

Bootstrap, however, while an approximation, makes no assumption on the distribution of the data.

## Exact confidence intervals, normal distribution

Let  $x_1, x_2, \dots, x_n$  be an i.i.d. sample from the normal distribution  $\mathcal{N}(\mu, \sigma^2)$  where both  $\mu \in \mathbb{R}$  and  $\sigma^2 > 0$  are unknown.

A level  $100(1 - \alpha)\%$  **confidence interval for  $\mu$**  is obtained as,

$$\left( \bar{x} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \right),$$

where  $t_{n-1, \alpha/2}$  is the  $(1 - \alpha/2)$ -quantile of the Student's  $t$ -distribution with  $n - 1$  degrees of freedom.

For large values of  $n$ , the Student's  $t$ -distribution with  $n - 1$  degrees of freedom approaches the standard normal distribution and its corresponding quantile can be substituted in place of  $t_{n-1, \alpha/2}$ .

## Exact confidence intervals, normal distribution

A level  $100(1 - \alpha)\%$  **confidence interval for  $\sigma^2$**  is obtained as,

$$\left( \frac{(n-1)s^2}{\chi_{n-1,\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{n-1,1-\alpha/2}^2} \right),$$

where  $\chi_{n-1,\alpha/2}^2$  is the  $(1 - \alpha/2)$ -quantile of the  $\chi_{n-1}^2$ -distribution and  $\chi_{n-1,1-\alpha/2}^2$  is the  $(\alpha/2)$ -quantile of the  $\chi_{n-1}^2$ -distribution.

# Contents

1 Confidence intervals

2 Hypothesis testing

3  $t$ -tests

4 Variance tests

# Hypothesis testing

Statistical tests are applied extensively in various fields of science.

## Examples of statistical testing situations

- Testing whether psychic can predict the winner of a sports match,
- Testing whether a treatment works better than the old one,

# Hypothesis testing

Statistical hypothesis testing is based on

- ① Selecting a statistical model/assumptions and
- ② Setting a null hypothesis and often also an alternative hypothesis,
- ③ Choosing a suitable test statistic, the value of which is calculated from a sample of observations.

The result of statistical hypothesis testing is a  $p$ -value, based upon which the conclusions are drawn.

## Statistical model

- Statistical model/assumptions casts the problem in a mathematical context and defines the rules of probability governing it.
- Statistical models are usually of the form:

*"Let  $x_1, \dots, x_n$  be an i.i.d. sample from the distribution  $F$  with the unknown parameter  $\theta$ ".*
- The validity of the model can, and should, be tested separately.

### Examples of statistical models/assumptions

- A psychic guesses the winner of each of the  $n$  sports matches correctly with the probability  $p$ , independent of his previous guesses.
- The treatment group responses  $x_1, \dots, x_n$  are an i.i.d. sample from  $\mathcal{N}(\mu_1, \sigma^2)$  and the control group responses  $y_1, \dots, y_m$  are an i.i.d. sample from  $\mathcal{N}(\mu_2, \sigma^2)$ .

## Null hypothesis

- The statement of interest about a model parameter is called the **null hypothesis**  $H_0$ .
- $H_0$  is assumed to be true unless there is strong evidence that indicates otherwise, in which case it is rejected.
- In simple statistical tests the null hypothesis can often be stated as an *equality*,  $H_0 : \theta = \theta_0$ , where  $\theta$  is the parameter being tested and  $\theta_0$  is a fixed value of the parameter.
- The null hypothesis is often conceptually of the form “*is the same*” or “*there is no difference*”.

### Examples of null hypotheses

- $H_0 : p = 0.5$ .
- $H_0 : \mu_1 = \mu_2$ .

## Alternative hypothesis

- The null hypothesis is usually accompanied by an alternative hypothesis  $H_1$ , which is often the logical opposite of  $H_0$  (though not always).
- If  $H_0$  is rejected, then  $H_1$  is accepted.
- The alternative hypothesis is often conceptually of the form “*is not the same*” or “*there is a difference*”.

### Examples of alternative hypotheses

- $H_1 : p \neq 0.5$ .
- $H_1 : \mu_1 > \mu_2$ .

Most tests in these lecture slides are for simplicity formulated using *two-tailed alternative hypotheses*,

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0.$$

## Test statistic

- **Test statistic** measures deviation of the observed sample from the null hypothesis  $H_0$ .
- A test statistic is a random variable and its value depends on the observations.
- The distribution of the test statistic under the null hypothesis  $H_0$  must be known for assessing the compatibility of the observations and the null hypothesis  $H_0$ .

### Examples of test statistics

- The proportion of correct guesses out of the total  $n$ .
- $\bar{x} - \bar{y}$

## *p*-value

- The *p*-value of a statistical test is the probability of observing at least as deviating value towards  $H_1$  as the observed value of the test statistic under the null hypothesis  $H_0$ .
- Note that what is considered as “deviating” depends on the form of the hypotheses.
- If the *p*-value is *too small* (the observation is too strange to have happened under  $H_0$ ), we reject  $H_0$  in favor of  $H_1$ .
- Note that we can never accept  $H_0$  based on the test, only “continue to believe in it”.

## Significance level and critical values

- Significance level  $\alpha$  is used to make a cut-off between small and large  $p$ -values.
  - ▶ If  $p < \alpha$  we reject  $H_0$ .
  - ▶ If  $p \geq \alpha$  we do not reject  $H_0$ .
- Commonly used significance levels are  $\alpha = 0.05, 0.1, 0.01, 0.001$ . and it **should be set before the study**.
- The set of values of the test statistic for which the null hypothesis is rejected (i.e. the values that yield a  $p$ -value smaller than  $\alpha$ ) is called the **critical region**.
- The threshold values delimiting the regions of non-rejection and rejection for the test statistic are called the **critical values**.

## Errors in statistical hypothesis testing

There are two kinds of errors related to the rejection of the null hypothesis  $H_0$ .

- **Type 1 error:** True null hypothesis is rejected.
- **Type 2 error:** False null hypothesis is not rejected.

The **type 1 error rate** is the probability of rejecting a true  $H_0$ . It is at most  $\alpha$ .

The **type 2 error rate** is the probability of not rejecting a false  $H_0$ . Type 2 error rate is more difficult to control as it is usually a function of the possible distributions under  $H_1$ .

**Power of a test** is equal to  $1 - \text{type 2 error rate}$ . The larger the power, the better the test detects false null hypotheses.

## Steps of statistical hypothesis testing

- ① Select the statistical model and state the hypotheses.
- ② Select a test statistic.
- ③ Pick a sample (for which the model holds).
- ④ Calculate the value of the test statistic from the data.
- ⑤ Calculate the  $p$ -value corresponding to the observed value of the test statistic.
- ⑥ Draw conclusions and reject/do not reject the null hypothesis.

# Contents

1 Confidence intervals

2 Hypothesis testing

3  $t$ -tests

4 Variance tests

## One-sample $t$ -test

One-sample  $t$ -test compares the expected value of a distribution to a given constant.

### One-sample $t$ -test, assumptions

Let  $x_1, x_2, \dots, x_n$  be an i.i.d. sample from  $\mathcal{N}(\mu, \sigma^2)$ .

### One-sample $t$ -test, hypotheses

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0.$$

# One-sample $t$ -test

## One-sample $t$ -test, test statistic

- The  $t$ -test statistic,

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

- follows Student's  $t$ -distribution with  $n - 1$  degrees of freedom under  $H_0$ .
- The expected value of  $t$  under the null hypothesis  $H_0$  is 0 and if the value of  $t$  has **large absolute value**, evidence against the null hypothesis  $H_0$  is found.

If the sample size is large, then the one-sample  $t$ -test is not very sensitive to moderate deviations from normality.

## Two-sample $t$ -test

Two-sample  $t$ -test compares the expected values of two distributions.

### Two-sample $t$ -test, assumptions

Let  $x_1, x_2, \dots, x_n$  be an i.i.d. sample from  $\mathcal{N}(\mu_x, \sigma_x^2)$  and let  $y_1, y_2, \dots, y_m$  be an i.i.d. sample from  $\mathcal{N}(\mu_y, \sigma_y^2)$ . Furthermore, let the two samples be independent.

### Two-sample $t$ -test, hypotheses

$$H_0 : \mu_x = \mu_y \quad H_1 : \mu_x \neq \mu_y.$$

## Two-sample $t$ -test

### Two-sample $t$ -test, test statistic

- The  $t$ -test statistic,

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n + s_y^2/m}},$$

follows approximately the Student's  $t$ -distribution with

$$\frac{(s_x^2/n + s_y^2/m)^2}{((s_x^2/n)^2/(n-1)) + ((s_y^2/m)^2/(m-1))}.$$

degrees of freedom under  $H_0$ .

- The expected value of  $t$  under  $H_0$  is 0 and if the value of the test statistic has **large absolute value**, evidence against the null hypothesis  $H_0$  is found.

If the sample size is large, then the two-sample  $t$ -test is not very sensitive to moderate deviations from normality.

## Paired *t*-test

- The two-sample *t*-tests assumes that the samples are independent.  
What if this is not the case?
  - ▶ A comparison of two measurement devices where both devices are used to measure the same subject under same circumstances.
  - ▶ A drug study where the subjects' responses are measured both before and after the treatment.
  - ▶ A comparison of health-related life style choices of matched pairs, such as spouses.

## Paired *t*-test

### Paired *t*-test, assumptions

Observations consist of an i.i.d. sample of pairs  $(x_{i1}, x_{i2}), i = 1, 2, \dots, n$  (the values **within** a pair need not be independent). The differences  $d_i = x_{i1} - x_{i2}$  have the normal distribution  $\mathcal{N}(\mu_d, \sigma_d^2)$ .

### Paired *t*-test, hypotheses

$$H_0 : \mu_d = 0 \quad H_1 : \mu_d \neq 0.$$

The recipe is now simple: Apply the *one-sample t-test* **to the differences**  $d_i$ , to test whether their expected value is zero (whether there is no systematic difference between the values in a pair).  
(Stop for a moment to think why this works.)

# Contents

- 1 Confidence intervals
- 2 Hypothesis testing
- 3  $t$ -tests
- 4 Variance tests

# Variance test

The variance test compares the variance of a distribution to a given constant.

## Variance test, assumptions

Let  $x_1, x_2, \dots, x_n$  be an i.i.d. sample from  $\mathcal{N}(\mu, \sigma^2)$ .

## Variance test, hypotheses

The null hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \quad H_1 : \sigma^2 \neq \sigma_0^2.$$

# Variance test

## Variance test, test statistic

- The  $\chi^2$ -test statistic,

$$\chi^2 = \frac{(n - 1)s^2}{\sigma_0^2},$$

follows  $\chi^2$ -distribution with  $n - 1$  degrees of freedom under  $H_0$ .

- The expected value of the test statistic under  $H_0$  is  $n - 1$  and both **large** and **small** values of the test statistic suggest that the null hypothesis  $H_0$  is false.

The variance test is sensitive to deviations from normality and does not work, even for large samples, if the underlying distribution is skewed.

## Variance comparison test

The variance comparison test compares the variances of two distributions.

### Variance comparison test, assumptions

Let  $x_1, x_2, \dots, x_n$  be an i.i.d. sample from  $\mathcal{N}(\mu_x, \sigma_x^2)$  and let  $y_1, y_2, \dots, y_m$  be an i.i.d. sample from  $\mathcal{N}(\mu_y, \sigma_y^2)$ . Furthermore, let the two samples be independent.

### Variance comparison test, hypotheses

$$H_0 : \sigma_x^2 = \sigma_y^2 \quad H_1 : \sigma_x^2 \neq \sigma_y^2.$$

## Variance comparison test

### Variance comparison test, test statistic

- The  $F$ -test statistic,

$$F = \frac{s_x^2}{s_y^2},$$

follows the  $F$ -distribution with  $n - 1$  and  $m - 1$  degrees of freedom under  $H_0$ .

- The expected value of the test statistic under  $H_0$  is  $\approx 1$  and both **large** and **small** values of the test statistic suggest that the null hypothesis  $H_0$  is false.

Also the variance comparison test is sensitive to deviations from normality and does not work, even for large samples, if the underlying distribution is skewed.

# MS-C1620 Statistical inference

## 3 Non-parametric tests

Jukka Kohonen

Department of Mathematics and Systems Analysis  
School of Science  
Aalto University

Academic year 2020–2021  
Period III–IV

## Non-parametric statistics

Statistical models can be divided roughly into two classes, parametric and non-parametric.

A model is **parametric** if it is fully defined by a set of parameters  $\theta_1, \dots, \theta_K$  (we assume a certain “family” of distributions, e.g. “some normal” or “some exponential”).

A model is **non-parametric** (roughly) if it does not assume a specific family of distributions.

*Stricter assumptions  $\leftrightarrow$  more powerful results  $\leftrightarrow$  more narrow area of application.*

# How to understand a statistical test

or: How it works, why it works, when it works

Rule 1: Suppose the hypothesis is true. Try to understand how the test statistic is then distributed. What kinds of values do you expect?

Rule 2: Suppose the hypothesis is false. Try to understand how the test statistic is then distributed.

# Contents

1 Sign tests

2 Rank tests

# One-sample sign test

One-sample sign test is applied in similar testing problems as the one-sample  $t$ -test. However, it makes much milder distributional assumptions.

## One-sample sign test, assumptions

Let  $x_1, x_2, \dots, x_n$  be an i.i.d. sample from a continuous distribution with the median  $m$ .

## One-sample sign test, hypotheses

$$H_0 : m = m_0 \quad H_1 : m \neq m_0.$$

# One-sample sign test

## One-sample sign test, test statistic

- The test statistic  $S$  equals the number of cases with  $x_i > m_0$  (alternatively, the number of the cases with  $x_i < m_0$ ) and follows binomial distribution with the parameters  $n$  and  $1/2$  under  $H_0$ .
- Under  $H_0$ , the expected value of the test statistic is  $\frac{1}{2}n$  (its variance is  $\frac{1}{4}n$ ) and both **large** and **small** values of the test statistic suggest that the null hypothesis  $H_0$  is false.

The distribution of the test statistic  $S$  is tabulated and statistical software gives exact  $p$ -values of the test.

## One-sample sign test, asymptotic distribution

- If the sample size is large, then the standardized test statistic,

$$Z = \frac{S - n/2}{\sqrt{n/4}},$$

follows approximately the standard normal distribution under  $H_0$ .

- For large  $n$ , the  $p$ -value of the test can thus be retrieved from a normal table.
- The approximation is usually good enough if  $n > 20$ . For smaller samples, it is better to rely on the exact distribution of the test statistic  $S$  (which is the binomial distribution).

## One sample sign test and discrete distributions

We assumed above that the observations come from a continuous distribution. However, the sign test can be used for discrete variables as well.

In that case it is possible that for some of the differences,  $x_i - m_0 = 0$ , are neither positive nor negative.

If the number of zeros is small compared to the sample size, these observations can be deleted and the sample size can be modified accordingly.

If the number of zeros is large, then the zeroes should be dealt with such that they are against rejecting the null hypothesis (conservative approach).

- For example: Consider the case of two-tailed alternative hypothesis, 3 negative differences, 15 positive differences and 6 zero differences. Now the test should be conducted as there were 9 negative differences and 15 positive ones.

## Paired sign test

Paired sign test is a non-parametric version of the paired  $t$ -test.

### Paired sign test, assumptions

Observations consist of an i.i.d. sample of pairs  $(x_{i1}, x_{i2}), i = 1, 2, \dots, n$  (the values **within** a pair need not be independent). The distribution of the differences  $d_i = x_{i2} - x_{i1}$  is continuous (denote the median of this distribution by  $m_d$ ).

### Paired sign test, hypotheses

$$H_0 : m_d = 0 \quad H_1 : m_d \neq 0.$$

The test is conducted by applying the one-sample sign test to the differences  $d_i$ .

# Contents

1 Sign tests

2 Rank tests

## One-sample signed rank test/Wilcoxon test

More sophisticated versions of the sign tests are given by **signed rank tests**, which consider not only the signs of the observations but also their relative order (and thus use more information).

### One-sample signed rank test, assumptions

Let  $x_1, x_2, \dots, x_n$  be an i.i.d. sample from continuous, **symmetric** distribution with the median  $m$ .

### One-sample signed rank test, hypotheses

$$H_0 : m = m_0 \quad H_1 : m \neq m_0.$$

## One-sample signed rank test

- Calculate the absolute values of the differences  $|d_i| = |x_i - m_0|$ ,  $i = 1, 2, \dots, n$ , and order the absolute values from the smallest to the largest.
- Define signed ranks  $r_i$  such that  $r_i$  is the rank of the absolute value  $|d_i| = |x_i - m_0|$  multiplied with the sign of the difference  $(x_i - m_0)$ .

### One-sample signed rank test, test statistic

- The test statistic,

$$W = \sum_{r_i > 0} r_i$$

is the sum of the positive ranks (alternatively, the sum of the negative ranks) and follows a certain distribution under  $H_0$ .

- Under  $H_0$ , the expected value of the test statistic is  $\frac{n(n+1)}{4}$  (its variance is  $\frac{n(n+1)(2n+1)}{24}$ ) and both **large** and **small** values of the test statistic suggest that the null hypothesis  $H_0$  is false.

## Asymptotic one-sample Wilcoxon signed rank test

- The distribution of the test statistic  $W$  is tabulated and statistical software gives exact  $p$ -values of the test.
- If the sample size is large the standardized test statistic,

$$Z = \frac{W - E(W)}{\sqrt{\text{var}(W)}},$$

where  $E(W) = \frac{n(n+1)}{4}$  and  $\text{var}(W) = \frac{n(n+1)(2n+1)}{24}$  follows approximately the standard normal distribution under  $H_0$ .

- For large  $n$ , the  $p$ -value of the test can thus be retrieved from a normal table.
- The approximation is usually good enough if  $n > 20$ . For smaller samples, it is better to rely on the exact distribution of the test statistic  $W$ .

## One-sample signed rank test

We assumed above that the observations come from a continuous distribution but the Wilcoxon signed rank test can be applied for discrete observations as well.

However, it is then possible that some points share the same rank.

In that case, all tied points are assigned rank equal to the median of the corresponding ranks.

- If two observations have the same rank corresponding to ranks 7 and 8, then both points are assigned to have rank 7.5.
- If three sample points have the same rank corresponding to ranks 3, 4, and 5, then each is assigned to have rank 4.

## Paired signed rank test

- Also the one-sample signed rank test can be used for paired data, in a way analogous to paired  $t$ -test and paired sign test.

### Paired signed rank test, assumptions

Observations consist of an i.i.d. sample of pairs  $(x_{i1}, x_{i2}), i = 1, 2, \dots, n$  (the values **within** a pair need not be independent). The difference  $d_i = x_{i1} - x_{i2}$  has a symmetric distribution, the median of which is denoted by  $m_d$ .

### Paired signed rank test, hypotheses

$$H_0 : m_d = 0 \quad H_1 : m_d \neq 0.$$

The test is conducted by applying the one-sample signed rank test to the differences  $d_i$ .

## When to use sign tests and signed rank tests

Both families of tests are non-parametric counterparts of the  $t$ -test and worthy alternatives when normality cannot be assumed.

Both types of tests are appropriate in similar problems:

- one sample — comparison of the location to a constant,
- paired samples — comparison of the locations.

The assumptions of the two types of tests differ:

- sign test — continuous distribution,
- signed rank test — continuous symmetric distribution.

If normality can be assumed, use  $t$ -tests. If symmetry (but no normality) can be assumed, use signed rank tests. Otherwise, use sign tests.

## Two-sample rank test/Wilcoxon rank-sum test

The two-sample rank sum test (a.k.a. Mann-Whitney U-test) is as the two sample  $t$ -test, but requires milder assumptions.

### Two-sample rank test, assumptions

- Let  $x_1, x_2, \dots, x_n$  and  $y_1, y_2, \dots, y_m$  be mutually independent i.i.d. samples from continuous distributions.
- Assume that the distributions of the samples are equal up to a location shift and denote their medians by  $m_x$  and  $m_y$ , respectively.

### Two-sample rank test, hypotheses

$$H_0 : m_x = m_y \quad H_1 : m_x \neq m_y$$

## Two-sample rank test

- The two-sample rank test is based on analyzing the order of all the observations.
- Assume without loss of generality that  $n \leq m$  and combine  $x_1, \dots, x_n$  and  $y_1, \dots, y_m$  into one sample  $z_1, \dots, z_{n+m}$ . Order then the observations  $z_i$  from smallest to largest and let  $r_i$  be the rank of  $z_i$  in the combined sample  $z_1, z_2, \dots, z_{n+m}$ .

### Two-sample rank test, test statistic

- The test statistic

$$W = \sum_{i=1}^n r_i$$

is the sum of the ranks of the smaller sample and it follows a certain distribution under  $H_0$ .

- Under  $H_0$ , the expected value of the test statistic is  $n(n + m + 1)/2$  (its variance is  $nm(n + m + 1)/12$ ) and both **large** and **small** values of the test statistic suggest that the null hypothesis  $H_0$  is false.

## Two-sample rank test, asymptotic distribution

- The distribution of the test statistic  $W$  is tabulated and statistical software gives the exact  $p$ -values.
- If the sample size is large the standardized test statistic,

$$Z = \frac{W - E(W)}{\sqrt{\text{var}(W)}},$$

where  $E(W) = n(n + m + 1)/2$  and  $\text{var}(W) = nm(n + m + 1)/12$ , follows approximately the standard normal distribution, under  $H_0$ .

- For large  $n$ , the  $p$ -value of the test can thus be retrieved from a normal table.
- The approximation is usually good enough if  $n, m > 10$ . For smaller samples, the exact distribution of the test statistic  $W$  should be relied upon.

## Two-sample rank test

- The two-sample rank test can be used also when the observations are discrete, however then it is possible that some of the sample points have the same rank. The ties are handled as in one-sample signed rank test.
- The two sample rank test is a non-parametric counterpart of the two-sample  $t$ -test.
- If normality can be assumed, use the two-sample  $t$ -test. Otherwise, use the two-sample rank test.

## Final note on rank tests

Note that ranks can be used even when a variable cannot be measured numerically but the observations can be ranked (for example, one could rank bands, or qualities of apartments, without measuring them numerically).

# MS-C1620 Statistical inference

## 4 Inference for binary data

Jukka Kohonen

Department of Mathematics and Systems Analysis  
School of Science  
Aalto University

Academic year 2020–2021  
Period III–IV

# Contents

1 Binary data

2 Single binary sample

3 Two binary samples

4 Lecture quiz

## Binary observations

In many applications the observations are binary.

- Something is true/false.
- Something happened/did not happen.
- Someone belongs/does not belong to a group.

In such a case the observations are most conveniently coded as 0/1.

Recall that if we have a iid sample of binary observations, their distribution is necessarily the *Bernoulli distribution*.

## Bernoulli distribution

The random variable  $x$  is said to obey the Bernoulli distribution with the probability of success  $p$  if,

$$\mathbb{P}(x = 1) = p \quad \text{and} \quad \mathbb{P}(x = 0) = 1 - p.$$

The expected value and variance of  $x$  are,

$$\begin{aligned}\mathbb{E}(x) &= p \\ \text{Var}(x) &= p(1 - p).\end{aligned}$$

That is, the Bernoulli distribution has only a single parameter to estimate.

The sum of  $n$  i.i.d. Bernoulli random variables with the success probability  $p$  has the binomial distribution with the parameters  $n$  and  $p$ .

# Contents

- 1 Binary data
- 2 Single binary sample
- 3 Two binary samples
- 4 Lecture quiz

## Approximate confidence interval

Central limit theorem can be used to obtain a confidence interval for the success probability  $p$  of a Bernoulli distribution.

Let  $x_1, x_2, \dots, x_n$  be an i.i.d. sample from the Bernoulli distribution with the success probability/expected value  $p$ .

For large  $n$ , a level  $100(1 - \alpha)\%$  confidence interval for the success probability  $p$  is obtained as

$$\left( \hat{p} - z_{\alpha/2} \frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}}, \hat{p} + z_{\alpha/2} \frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}} \right),$$

where  $\hat{p}$  is the observed proportion of successes and  $z_{\alpha/2}$  is the  $(1 - \alpha/2)$ -quantile of the standard normal distribution.

## One-sample proportion test

To test whether the success probability of a Bernoulli distribution equals some pre-specified value, we employ **one-sample proportion test**.

### One-sample proportion test, assumptions

Let  $x_1, x_2, \dots, x_n$  be an i.i.d. sample from a Bernoulli distribution with the success probability  $p$ .

### One-sample proportion test, hypotheses

$$H_0 : p = p_0 \quad H_1 : p \neq p_0.$$

## One-sample proportion test

### One-sample proportion test, test statistic

- The test statistic,

$$C = \sum_{i=1}^n x_i,$$

follows the binomial distribution with parameters  $n$  and  $p_0$  under  $H_0$ .

- Under  $H_0$ , the test statistic has  $E[C] = np_0$  and  $\text{Var}(C) = np_0(1 - p_0)$  and both large and both **large** and **small** values of the test statistic suggest that the null hypothesis  $H_0$  is false.

The distribution of the test statistic  $C$  is tabulated and statistical software calculate exact  $p$ -values of the test.

## Asymptotic one-sample proportion test

If the sample size is large, then under the null hypothesis  $H_0$  the standardized test statistic,

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}},$$

where  $\hat{p}$  is the unbiased estimator  $\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$  of the parameter  $p$ , follows approximately the standard normal distribution.

The approximation is usually accurate enough if  $n\hat{p} > 10$  and  $n(1 - \hat{p}) > 10$ . For smaller sample sizes one should rely on the exact distribution of the test statistic  $C$ .

# Contents

- 1 Binary data
- 2 Single binary sample
- 3 Two binary samples
- 4 Lecture quiz

## Two-sample proportion test

The one-sample proportion test can be seen as the equivalent of  $t$ -test when the normal distribution is replaced by the Bernoulli distribution.

As with  $t$ -test, a two-sample version readily follows and in **two-sample proportion test** parameters of two independent Bernoulli-distributed samples are compared.

### Two-sample proportion test, assumptions

Let  $x_1, x_2, \dots, x_n$  be an i.i.d. sample from a Bernoulli distribution with the success probability  $p_x$  and let  $y_1, y_2, \dots, y_m$  be an i.i.d. sample from a Bernoulli distribution with the success probability  $p_y$ . Furthermore, let the two samples be independent.

### Two-sample proportion test, hypotheses

$$H_0 : p_x = p_y \quad H_1 : p_x \neq p_y.$$

## Two-sample proportion test

### One-sample proportion test, test statistic

- Calculate the sample proportions

$$\hat{p}_x = \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{p}_y = \frac{1}{m} \sum_{i=1}^m y_i, \quad \hat{p} = \frac{n\hat{p}_x + m\hat{p}_y}{n+m}.$$

- The test statistic,

$$Z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}},$$

follows for large  $n$  under  $H_0$  the standard normal distribution.

- Both **large** and **small** values of the test statistic suggest that the null hypothesis  $H_0$  is false.

The normal approximation is usually good enough if  $n\hat{p}_x > 5$ ,  
 $n(1 - \hat{p}_x) > 5$ ,  $m\hat{p}_y > 5$  and  $m(1 - \hat{p}_y) > 5$ .

## Frequency tables

Assuming a “paired binary sample”, the previous test is no longer valid.

id	X	Y
1	0	1
2	0	0
3	0	1
4	1	1
:	:	:

This kind of data is most conveniently represented in a **frequency table**.

	Y = 0	Y = 1
X = 0	173	40
X = 1	65	53

Inference for frequency tables is discussed next time.

# Contents

1 Binary data

2 Single binary sample

3 Two binary samples

4 Lecture quiz

## Lecture quiz

A lecture quiz to determine what you have learned thus far!

Answer the following questions on your own or in small groups.

# Lecture quiz

## Question 1

Consider the following random sample: 5, -4, -2, 2. Calculate the following sample quantities:

- ① Sample mean
- ② Sample standard deviation
- ③ Sample median
- ④ Sample median absolute deviation
- ⑤ Sample range
- ⑥ Signs of the sample points
- ⑦ Ranks of the sample points
- ⑧ Signed ranks of the sample points with respect to distance to 0.

# Lecture quiz

## Question 2

Give concrete examples when you would/would not use the following measures of location:

- ① Sample mean
- ② Sample median
- ③ Mode

## Question 3

Give concrete examples when you would/would not use the following measures of scatter:

- ① Standard deviation
- ② Median absolute deviation
- ③ Sample range

## Lecture quiz

### Question 4

What does it mean in practice if:

- The confidence interval of a parameter is narrow
- The significance level of a test is set low
- The  $p$ -value of a test is high
- Type I error occurs in a statistical test
- Type II error occurs in a statistical test

# Lecture quiz

## Question 5

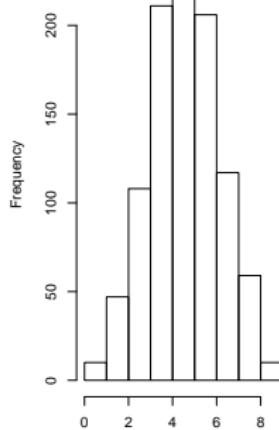
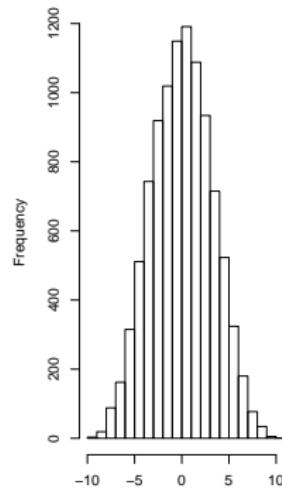
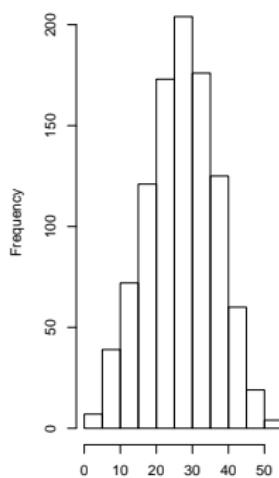
How would you visualize the following samples:

- The heights of the male and female students attending a course.
- The exam points (0-24) on a large course.
- The proportions of faulty products produced by 5 different production lines.
- Stock prices of 3 companies.
- The monthly salaries and postal codes of adults living in Helsinki area.

# Lecture quiz

## Question 6

The following plots show the distributions of the test statistics of  $t$ -test, sign test and signed rank test for the null hypothesis of zero location when the data is a sample of size  $n = 10$  from the standard normal distribution. Which plot corresponds to which test?



# MS-C1620 Statistical inference

## 5 Distribution tests

Jukka Kohonen

Department of Mathematics and Systems Analysis  
School of Science  
Aalto University

Academic year 2020–2021  
Period III–IV

# Contents

1 Testing distributional assumptions

2 Normality tests

3 Chi-squared tests

- In statistics, assumptions on the underlying distribution are done all the time.
- Many statistical methods become ineffective or even give false results if their assumptions do not hold.
- This is why it is very important to test the distributional assumptions separately.
- Assumptions on normally distributed observations are made particularly often, especially with classical statistical methods.

# Contents

1 Testing distributional assumptions

2 Normality tests

3 Chi-squared tests

# Normality testing

## Normality testing, assumptions

Assume that  $x_1, x_2, \dots, x_n$  are i.i.d. observed values of a random variable  $x$ .

## Normality testing, hypotheses

$H_0$  : Random variable  $x$  is normally distributed.

$H_1$  : Random variable  $x$  is not normally distributed.

## Bowman-Shenton normality test

### Bowman-Shenton normality test, test statistic

- The Bowman-Shenton (Jarque-Bera) normality test is a function of skewness and kurtosis,

$$BS = n \left( \frac{\hat{\gamma}^2}{6} + \frac{\hat{\kappa}^2}{24} \right),$$

where  $\hat{\gamma}$  is the sample skewness coefficient and  $\hat{\kappa}$  is the sample kurtosis coefficient discussed in lecture 1.

- The test tests whether the skewness and kurtosis of the data-generating distribution match with the normal distribution.
- If the observed skewness or kurtosis values differ significantly from the skewness and/or kurtosis values of the normal distribution (0 and 0), the test statistic gets large values.

## Bowman-Shenton normality test

### Bowman-Shenton normality test, test statistic

- If  $n$  is large, then under  $H_0$  the test statistic  $BS$  follows approximately  $\chi^2_2$  distribution.
- The expected value of the test statistic under  $H_0$  is approximately 2 and **large values** of the test statistic suggests that the null hypothesis  $H_0$  is false.

**Note that the Bowman-Shenton test is suitable only for large sample sizes.**

## Rank plot / Quantile-quantile (Q-Q) plot

- Let  $y_1 \leq y_2 \leq \cdots \leq y_n$  be the data points  $x_1, x_2, \dots, x_n$  ordered from the smallest one to the largest one.
- Let  $q_i$  be the  $i/(n+1)$  quantile from the standard normal distribution  $\mathcal{N}(0, 1)$  and plot the pairs  $(q_i, y_i), i = 1, 2, \dots, n$ .
- If the observations  $x_i$  do come from a normal distribution, then the points  $(q_i, y_i)$  should approximately lie on a line.
- If the points do not lie on a line, there is evidence of non-normality.
- The plot can be used in detecting skewness of a distribution and in finding outliers.

## Shapiro-Wilk normality test

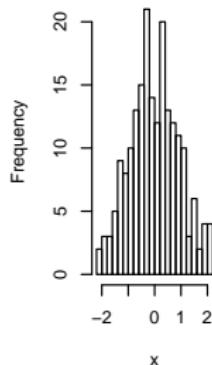
### Shapiro-Wilk normality test, test statistic

- The Shapiro-Wilk normality test statistic is the squared value of the Pearson sample correlation coefficient calculated from the rank plot points  $(q_i, y_i), i = 1, 2, \dots, n$ .
- The null distribution of the test statistic is complicated and the test is usually performed with statistical software.
- **Small** values of the test statistic suggest that the assumption of normality does not hold. **Large** values of the test statistic are in line with the null hypothesis.

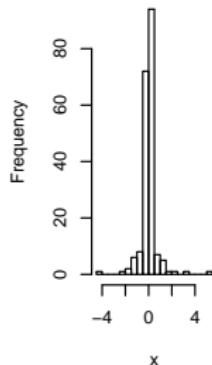
The Shapiro-Wilk normality test requires a large sample size.

## Q-Q plot quiz

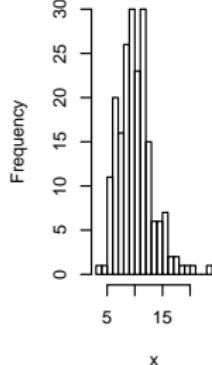
Which of the following histograms corresponds to each of the Q-Q plots on the next slide? (the answers are given on the next slide after that)



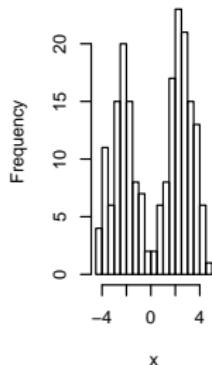
A



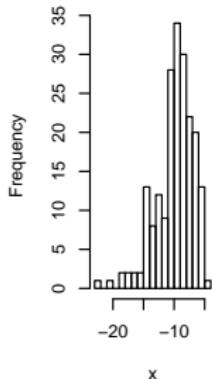
B



C



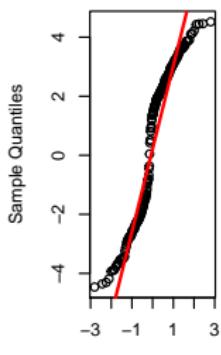
D



E

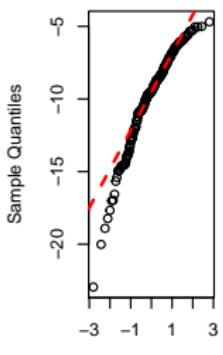
# Q-Q plot quiz

Normal Q-Q Plot



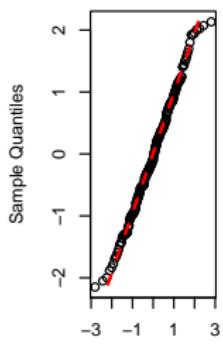
1

Normal Q-Q Plot



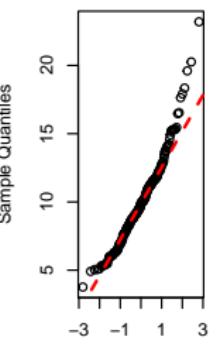
2

Normal Q-Q Plot



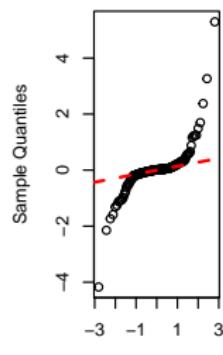
3

Normal Q-Q Plot



4

Normal Q-Q Plot



5

## Q-Q plot quiz, answers

- The correct pairs are (histogram, qqplot):  
 $(A, 3); (B, 5); (C, 4); (D, 1); (E, 2).$

# Contents

- 1 Testing distributional assumptions
- 2 Normality tests
- 3 Chi-squared tests

## Multinomial distribution

Consider a random experiment which has  $k$  mutually exclusive outcomes and which is run independently  $n$  times.

Let the vector  $\mathbf{y} = (y_1, \dots, y_k)$  contain the observed frequencies of the  $k$  outcomes.

The distribution of  $\mathbf{y}$  is known as the **multinomial distribution**, the generalization of the binomial distribution into more than two outcomes.

## Multinomial distribution

The random vector  $\mathbf{y} = (y_1, \dots, y_k)$  follows the **multinomial distribution** with **parameters**  $n, \mathbf{p} = (p_1, \dots, p_k)$ , if its probability mass function is,

$$p(\mathbf{y}) = \frac{n!}{y_1! y_2! \cdots y_k!} p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k},$$

where

$$\sum_{j=1}^k y_j = n \quad \text{and} \quad \sum_{j=1}^k p_j = 1.$$

A useful result in the following is that for a random vector  $\mathbf{y}$  following the multinomial distribution with parameters  $n, \mathbf{p}$ , the normalized sum,

$$\sum_{j=1}^k \frac{(y_j - np_j)^2}{np_j},$$

where  $np_j$  are the expected frequencies of the outcomes follows for large  $n$  approximately the  $\chi^2_{k-1}$ -distribution

## $\chi^2$ goodness-of-fit test

The  $\chi^2$  goodness-of-fit test uses the multinomial distribution to tests whether the distribution of a random variable  $x$  is some particular, arbitrary distribution.

### Goodness-of-fit tests, assumptions

Assume that  $x_1, x_2, \dots, x_n$  are i.i.d. observed values of a random variable  $x$ .

### Goodness-of-fit tests, hypotheses

$H_0$  : Random variable  $x$  follows the distribution  $F_x$  (with or without unknown parameters).

$H_1$  : Random variable  $x$  does not follow the distribution  $F_x$ .

## $\chi^2$ goodness-of-fit test

- Categorize the  $n$  observations into  $k$  categories.
- Calculate the frequencies  $O_1, \dots, O_k$ , where  $O_j$  is the observed frequency of the  $j$ th category (note that  $\sum_{j=1}^k O_j = n$ ).
- Let  $p_j$  be the probability that, under the null hypothesis, the random variable  $x$  belongs to the  $j$ th category.
- Calculate the expected frequencies  $E_j = np_j$  of the  $k$  categories (note that  $\sum_{j=1}^k p_j = 1$  and  $\sum_{j=1}^k E_j = n$ ).
- 

Now, under the null hypothesis, the random vector  $(O_1, \dots, O_k)$  follows the multinomial distribution with the parameters  $n, \mathbf{p} = (p_1, \dots, p_k)$  and the expected category frequencies  $(E_1, \dots, E_k)$

## $\chi^2$ goodness-of-fit test

### $\chi^2$ goodness-of-fit test, test statistic

- The test statistic,

$$\chi_g^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i},$$

follows, for large  $n$ , under  $H_0$  approximately the  $\chi_{k-1-e}^2$ -distribution, where  $e$  is the number of estimated parameters (see the salary example below).

- The expected value of the test statistic under  $H_0$  is approximately  $k - 1 - e$  and **large** values of the test statistic suggest that the null hypothesis  $H_0$  does not hold.
- Note that a very small value of the test statistic could be an indicator of *overfitting*.

## $\chi^2$ goodness-of-fit test, example with unknown parameters

- Consider testing whether the monthly salary of the Finns follows a normal distribution.
- Select randomly  $n$  Finns and document their salaries.
- The null hypothesis is that the observations come from a normal distribution with an unknown expected value and an unknown variance.

## $\chi^2$ goodness-of-fit test, example with unknown parameters

- ① Estimate the unknown parameters ( $\mu$  and  $\sigma^2$ ) from the sample.
- ② Discretize the continuous salary variable into  $k$  categories.
- ③ Calculate the observed category frequencies  $O_1, \dots, O_k$ .
- ④ Calculate the category probabilities for the estimated normal distribution, for example,  
 $\dots, \mathbb{P}(1900 < X \leq 2000), \mathbb{P}(2000 < X \leq 2100), \dots$
- ⑤ Calculate the expected category frequencies  $E_1, \dots, E_k$ .
- ⑥ Calculate the test statistic. Under the null hypothesis the test statistic approximately follows  $\chi^2_{k-1-e} = \chi^2_{k-3}$ -distribution, where  $k$  is the number of categories and we estimated  $e = 2$  parameters ( $\mu$  and  $\sigma^2$ ).
- ⑦ Calculate the  $p$ -value and based on that either reject or do not reject the null hypothesis.

## $\chi^2$ homogeneity test

The  $\chi^2$  homogeneity test is used to assess whether multiple samples come from the same distribution.

### $\chi^2$ homogeneity test, assumptions

We observe a total of  $r$  samples such that the samples are independent and the observations within a single sample are i.i.d. Assume that the sample  $i \in \{1, \dots, r\}$  has  $n_i$  observations.

### $\chi^2$ homogeneity test, hypotheses

$H_0$  : The samples come from the same distribution  $F_x$ .

$H_1$  : The samples do not come from the same distribution.

## $\chi^2$ homogeneity test, observed frequencies

- Categorize all observations into  $k$  categories.
- Calculate the frequencies  $O_{ij}$ ,  $i \in \{1, 2, \dots, r\}$ ,  $j \in \{1, 2, \dots, k\}$ , where  $O_{ij}$  is the observed frequency of the observations of the sample  $i$  in category  $j$ .

	1	2	$\dots$	$k$	sum
1	$O_{11}$	$O_{12}$	$\dots$	$O_{1k}$	$n_1$
2	$O_{21}$	$O_{22}$	$\dots$	$O_{2k}$	$n_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$r$	$O_{r1}$	$O_{r2}$	$\dots$	$O_{rk}$	$n_r$
sum	$K_1$	$K_2$	$\dots$	$K_k$	$n$

## $\chi^2$ homogeneity test, expected frequencies

- Let  $p_j = K_j/n$  be an estimate of the proportion of the  $j$ th category under  $H_0$  (under the null hypothesis the probability of the category  $j$  is the same for each sample  $i$ ).
- Calculate the expected frequencies under the null,  $E_{ij} = n_i p_j$ .

	1	2	$\dots$	$k$	sum
1	$E_{11}$	$E_{12}$	$\dots$	$E_{1k}$	$n_1$
2	$E_{21}$	$E_{22}$	$\dots$	$E_{2k}$	$n_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$r$	$E_{r1}$	$E_{r2}$	$\dots$	$E_{rk}$	$n_r$
sum	$K_1$	$K_2$	$\dots$	$K_k$	$n$

## $\chi^2$ homogeneity test

### $\chi^2$ homogeneity test, test statistic

- The test statistic,

$$\chi_h^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

follows, for large  $n$ , under  $H_0$  approximately the  $\chi^2_{(r-1)(k-1)}$  distribution.

- Under  $H_0$  the expected value of the test statistic is approximately  $(r - 1)(k - 1)$  and **large** values of the test statistic suggest that the null hypothesis  $H_0$  is false.

## $\chi^2$ test of independence

$\chi^2$  test of independence is used to study if two random variables (factors) are stochastically independent.

### $\chi^2$ -test of independence, assumptions

We observe an i.i.d. random sample of size  $n$  and the observations are divided into  $r$  classes with respect to a factor  $A$  and into  $k$  classes with respect to a factor  $B$ .

### $\chi^2$ -test of independence, hypotheses

$H_0$  : The variables  $A$  and  $B$  are independent.

$H_1$  : The variables  $A$  and  $B$  are not independent.

## $\chi^2$ test of independence, observed frequencies

- Let  $R_i$  be the frequency of the observations in class  $i$  of the factor  $A$  and let  $K_j$  be the frequency of the observations in class  $j$  of the factor  $B$ .
- Let  $O_{ij}$  be the observed frequency of the observations that are in class  $i$  of the factor  $A$  and in class  $j$  of the factor  $B$ .

	1	2	$\dots$	$k$	sum
1	$O_{11}$	$O_{12}$	$\dots$	$O_{1k}$	$R_1$
2	$O_{21}$	$O_{22}$	$\dots$	$O_{2k}$	$R_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$r$	$O_{r1}$	$O_{r2}$	$\dots$	$O_{rk}$	$R_r$
sum	$K_1$	$K_2$	$\dots$	$K_k$	$n$

## $\chi^2$ test of independence, expected frequencies

- Let  $q_i = R_i/n$  and  $p_j = K_j/n$ . Under the null hypothesis of independence the probability to fall in to the cell  $(i,j)$  is approximately  $q_i p_j$ .
- Calculate the expected frequencies under the null,

$$E_{ij} = nq_i p_j = R_i p_j = K_j q_i.$$

	1	2	...	$k$	sum
1	$E_{11}$	$E_{12}$	...	$E_{1k}$	$R_1$
2	$E_{21}$	$E_{22}$	...	$E_{2k}$	$R_2$
:	:	:	..	:	:
$r$	$E_{r1}$	$E_{r2}$	...	$E_{rk}$	$R_r$
sum	$K_1$	$K_2$	...	$K_k$	$n$

## $\chi^2$ test of independence

### $\chi^2$ -test of independence, test statistic

- The test statistic,

$$\chi_I^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

follows, for large  $n$ , under  $H_0$  approximately the  $\chi^2(r - 1)(k - 1)$ -distribution.

- The expected value of the test statistic under  $H_0$  is approximately  $(r - 1)(k - 1)$  and **large** values of the test statistic suggest that the null hypothesis is false

## The homogeneity test and the test of independence

- Note that the  $\chi^2$  test of independence and  $\chi^2$  homogeneity test have their test statistics and the degrees of freedom calculated identically.
- However, the tests apply to different situations:
  - ▶ If the group sizes of one of the factors are pre-determined, one can not speak of the independence of the factors (since one of them has its frequencies fixed), and the correct interpretation is via the  $\chi^2$  homogeneity test.
  - ▶ If only the overall sample size  $n$  is fixed and the observations are allowed to freely fall into the categories with respect to both factors, the correct interpretation is via the  $\chi^2$ -test of independence.

# MS-C1620 Statistical inference

## 6 Correlation and dependence

Jukka Kohonen

Department of Mathematics and Systems Analysis  
School of Science  
Aalto University

Academic year 2020–2021  
Period III–IV

# Contents

- 1 Linear dependence
  - Pearson correlation coefficient
  - Confidence intervals for Pearson correlation
  - Hypothesis tests for Pearson correlation
  
- 2 Monotonic dependence
  - Confidence intervals for Spearman correlation
  - Hypothesis tests for Spearman correlation

## Independence and dependence

- Two random variables/experiments are **independent** if the result of one does not in any way help us predict the result of the other.
- More formally, the random variables  $x$  and  $y$  are independent if for all (suitable) sets  $A, B$  we have,

$$\mathbb{P}(x \in A \mid y \in B) = \mathbb{P}(x \in A).$$

- If the above does not hold, the random variables  $x, y$  are called **dependent**.
- Saying that two random variables are dependent does not, however, give any indication on the **type of dependence** or **how dependent** they are.

# Dependence in statistics

In statistics, the dependence of random variables is usually of major interest.

- The dependence between unemployment rate and (growth of) GDP in Finland, election promises, etc.
- The dependence between alcohol consumption and alcohol price, income level, availability of alcohol, warning labels, etc.
- The dependence between incidence of lung cancer and smoking (duration, amount of cigarettes) etc.

## Linear dependence

- The simplest form of dependence is linear dependence.
- If the random variables  $x$  and  $y$  satisfy,

$$y = ax + b,$$

for some constants  $a, b \in \mathbb{R}$ ,  $a \neq 0$ , then the variable  $y$  is a linear transformation of the variable  $x$  and the random variables  $x$  and  $y$  are said to be (completely) linearly dependent.

- Linear dependence between two variables can be measured, for example, using the (Pearson) correlation coefficient.

## Pearson correlation coefficient

- Let  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  be i.i.d. observations of a bivariate random variable  $(x, y)$ .
- Then the sample covariance,

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

estimates the population covariance  $E[(x - E[x])(y - E[y])] = \sigma_{xy}$ ,

- and the sample Pearson correlation coefficient,

$$\hat{\rho} = \hat{\rho}(x, y) = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

estimates the Pearson correlation coefficient

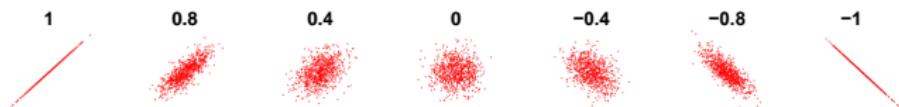
$$\rho = \rho(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

## Interpreting the Pearson correlation

- Pearson correlation coefficient numerically measures the **linear dependence** between two random variables. The coefficient is always in the interval  $[-1, 1]$  and attains the values  $\pm 1$  if and only if  $y = ax + b$  for some  $a, b \in \mathbb{R}$ ,  $a \neq 0$ .
- If the variables  $x$  and  $y$  are independent, then the Pearson correlation coefficient  $\rho(x, y) = 0$ .
- However, the contrary does not hold. That is,  $\rho(x, y) = 0$  does not imply the independence of  $x$  and  $y$  (take, for example,  $x$  standard normal and  $y = x^2$ ).
- *Thus there are also other forms of dependence than linear dependence.*

## Example 1

Examples of data exhibiting linear dependence of various degrees:



## Example 2

Examples of data exhibiting complete linear dependency (correlation coefficients equal to  $\pm 1$ ) (in the middle one  $\rho(x, y)$  cannot be computed due to division by zero):



## Example 3

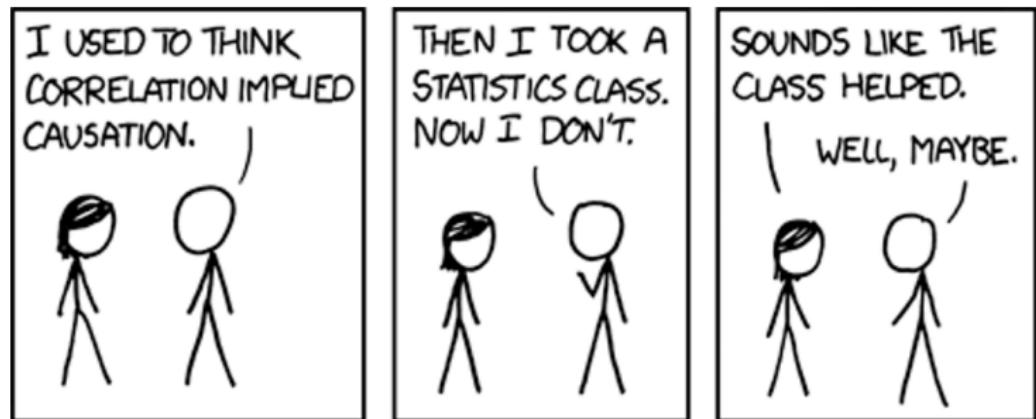
Examples of weird shaped data all having approximately zero linear dependency (but still having clear dependency of some other form):



Check also <http://guessthecorrelation.com/> and  
<https://www.autodeskresearch.com/publications/samestats>

## Correlation vs. causation

Note that showing that two things are dependent reveals nothing about their causal relation (which one caused the other).



Check out: <http://www.tylervigen.com/spurious-correlations>.

## Three different warnings!

- ① Sample correlation only *estimates* the “true” correlation. So even if true distribution has *zero* correlation, a small sample usually has nonzero. → Try confidence intervals or tests.  
E.g. I just rolled two dice, 30 times each. The correlation coefficient was  $-0.12$ .
- ② Correlation could be small or zero, but still there could be *nonlinear* dependence. → See end of this lecture about monotonic dependence.
- ③ Correlation between  $X$  and  $Y$  could be big, (so big that it is not just random noise in sample), but it does not mean that  $X$  *causes*  $Y$ .  
It could be the other way round. Or it could be a third variable that causes both to be big at the same time. Finding causality is very important, but requires stronger tools than we have on this course. At least a high correlation can be taken as a *hint* of possible causality.

## Bivariate normal distribution

The **bivariate normal distribution** is an extension of the normal distribution into two dimensions.

Let  $(x, y)$  have the bivariate normal distribution. Then its marginal distributions are normal distributions with the expected values  $\mu_x, \mu_y \in \mathbb{R}$  and the variances  $\sigma_x^2, \sigma_y^2 > 0$ .

In addition to these, the bivariate normal distribution has the parameter  $\rho \in [-1, 1]$ , the Pearson correlation between the marginals.

The probability density function of a bivariate normal distribution is,

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_x\sigma_y}.$$

$$\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)}{\sigma_x}\frac{(y-\mu_y)}{\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right)\right).$$

## Parametric confidence interval for Pearson correlation

Let  $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$  be an iid. sample from the bivariate normal distribution and denote,

$$\tanh(t) = \frac{e^{2x} - 1}{e^{2x} + 1}, \quad \operatorname{arctanh}(t) = \frac{1}{2} \log\left(\frac{1+t}{1-t}\right)$$

Then  $\operatorname{arctanh}(\hat{\rho})$  is (for large  $n$ ) approximately normally distributed and this can be used to derive an **approximate  $100(1 - \alpha)$  level confidence interval for  $\rho$ :**

$$\left[ \tanh\left(\operatorname{arctanh}(\hat{\rho}) - z_{\alpha/2} \frac{1}{\sqrt{n-3}}\right), \tanh\left(\operatorname{arctanh}(\hat{\rho}) + z_{\alpha/2} \frac{1}{\sqrt{n-3}}\right) \right],$$

where  $z_{\alpha/2}$  is the  $(1 - \alpha/2)$ -quantile of the standard normal distribution.

Note that the above confidence interval makes the **assumption of bivariate normal distribution**.

Note also that the confidence intervals for the Pearson correlation provided by statistical software are almost always based on normality assumption.

## Non-parametric confidence interval for Pearson correlation

If the iid. bivariate sample  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  does not come from the bivariate normal distribution, a non-parametric alternative to the previous parametric confidence interval is given by the **bootstrap**.

## Non-parametric confidence interval for Pearson correlation

- ① Pick new sample of  $n$  pairs from the observed pairs  $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$  with replacement, such that new pairs are selected one-by-one and the selected pair is each time “returned back” to the original sample (the same pair can be selected multiple times).
- ② Estimate the Pearson correlation coefficient for the new sample formed in the previous step.
- ③ Repeat the previous steps  $B$  times.
- ④ After the replications, order the  $B$  estimates from the smallest to the largest.
- ⑤ A  $100(1 - \alpha)\%$  confidence interval is now obtained by choosing the  $\lfloor B \times (\alpha/2) \rfloor$  ordered estimate as the lower endpoint and the  $\lfloor B \times (1 - \alpha/2) \rfloor$  ordered estimate as the upper endpoint.

## One-sample test for Pearson correlation

The one-sample test for the Pearson correlation coefficient compares the Pearson correlation to a given constant under the assumption of bivariate normality.

### One-sample test for Pearson correlation, assumptions

Let  $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$  be an iid. random sample from a bivariate normal distribution.

### One-sample test for Pearson correlation, hypotheses

$$H_0 : \rho = \rho_0 \quad H_1 : \rho \neq \rho_0.$$

## One-sample test for Pearson correlation

### One-sample test for Pearson correlation, test statistic

- The test statistic,

$$z = \frac{\operatorname{arctanh}(\hat{\rho}) - \operatorname{arctanh}(\rho_0)}{\sqrt{\frac{1}{n-3}}},$$

follows under  $H_0$  (for large  $n$ ) approximately the standard normal distribution.

- The expected value of  $z$  under  $H_0$  is 0 and if the value of the test statistic has **large absolute value**, evidence against the null hypothesis  $H_0$  is found.

## Two-sample test for Pearson correlation

The two-sample test for Pearson correlation compares the Pearson correlation coefficients of two independent samples under the assumption of bivariate normality.

### Two-sample test for Pearson correlation, assumptions

Let  $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$  be an iid. random sample from a bivariate normal distribution with the Pearson correlation  $\rho_1$  and let  $(z_1, w_1), (z_2, w_2) \dots, (z_m, w_m)$  be an iid. random sample from a bivariate normal distribution with the Pearson correlation  $\rho_2$ . Furthermore, let the two samples be independent.

### Two-sample test for Pearson correlation, hypotheses

$$H_0 : \rho_1 = \rho_2 \quad H_1 : \rho_1 \neq \rho_2.$$

## Two-sample test for Pearson correlation

### Two-sample test for Pearson correlation, test statistic

- The test statistic,

$$z = \frac{\operatorname{arctanh}(\hat{\rho}_1) - \operatorname{arctanh}(\hat{\rho}_2)}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}},$$

follows under  $H_0$  (for large  $n$  and  $m$ ) approximately the standard normal distribution.

- The expected value of  $z$  under  $H_0$  is 0 and if the value of the test statistic has **large absolute value**, evidence against the null hypothesis  $H_0$  is found.

## Parametric significance test for Pearson correlation

Often it is of interest to assess whether the Pearson correlation differs statistically significantly from zero (no correlation). Under the assumption of bivariate normality, an approximate test for this can be carried out using the one-sample test for Pearson correlation.

However, an exact test can also be performed.

### Parametric significance test for Pearson correlation, assumptions

Let  $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$ , be an i.i.d. random sample from a bivariate normal distribution.

### Parametric significance test for Pearson correlation, hypotheses

$$H_0 : \rho = 0 \quad H_1 : \rho \neq 0.$$

## Parametric significance test for Pearson correlation

### Parametric significance test for Pearson correlation, test statistic

- The test statistic,

$$t = \sqrt{n-2} \cdot \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}},$$

follows under  $H_0$  the  $t_{n-2}$ -distribution.

- The expected value of  $t$  under  $H_0$  is 0 and if the value of the test statistic has **large absolute value**, evidence against the null hypothesis  $H_0$  is found.

## Non-parametric significance test for Pearson correlation

A non-parametric alternative to the parametric significance test is given by a **permutation test**.

### Permutation test for Pearson correlation, assumptions

Let  $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$ , be an i.i.d. random sample from a bivariate distribution.

### Permutation test for Pearson correlation, hypotheses

$$H_0 : \rho = 0 \quad H_1 : \rho \neq 0.$$

## Permutation test for Pearson correlation

Let  $\hat{\rho}$  be the Pearson correlation of the original sample. The probability of obtaining a value equally or more deviating than  $\hat{\rho}$  under the null hypothesis can be estimated using a permutation test as follows.

- ① Form  $n$  new pairs  $(x_1, y_1^*), (x_2, y_2^*) \dots, (x_n, y_n^*)$  from the original observed pairs  $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$ , such that  $y_1, y_2, \dots, y_n$  are permuted randomly and each original  $y_j$  is used only once in the new sample.
- ② Estimate Pearson correlation using the new sample  $(x_1, y_1^*), (x_2, y_2^*) \dots, (x_n, y_n^*)$ .
- ③ Repeat the steps 1 and 2 several times.
- ④ Estimate the probability of obtaining a value equally or more deviating than  $\hat{\rho}$  under the null hypothesis using the generated distribution of estimates. That is, calculate the percentage of the generated estimates in that have **absolute value** greater than  $|\hat{\rho}|$ .

## Permutation test for Pearson correlation

- The permutation test is based on the idea that the permuted samples  $(x_1, y_1^*), (x_2, y_2^*) \dots, (x_n, y_n^*)$  do not exhibit correlation (as the pairs are chosen randomly) and if  $\hat{\rho}$  differs a lot from the typical correlation coefficient of a permuted sample, we can conclude that the original sample exhibits significant correlation.
- More accurate estimate can be achieved using a permutation test without simulation. In an **exact permutation test**, all possible  $n!$  sample combinations are used, and the probability of obtaining the value  $\hat{\rho}$  or more extreme value under the null hypothesis is estimated exactly using all  $n!$  correlation coefficients.

# Contents

- 1 Linear dependence
  - Pearson correlation coefficient
  - Confidence intervals for Pearson correlation
  - Hypothesis tests for Pearson correlation
  
- 2 Monotonic dependence
  - Confidence intervals for Spearman correlation
  - Hypothesis tests for Spearman correlation

## Monotonic dependence

- A more flexible form of dependency is given by **monotonic dependence**.
- If the random variables  $x$  and  $y$  satisfy

$$y = g(x),$$

where  $g$  is a monotonic (increasing or decreasing) function, then  $y$  is a monotonic transformation of  $x$  and the random variables  $x$  and  $y$  are said to be (completely) monotonically dependent.

- The monotonic dependence between two random variables can be measured using, for example, **Spearman's rank correlation coefficient**.

## Spearman's rank correlation coefficient

- Let  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  be i.i.d. observations of a bivariate random variable  $(x, y)$ .
- Let  $R(x_i)$  denote the rank of the observation  $x_i$  in the sample  $x_1, x_2, \dots, x_n$  and let  $R(y_i)$  denote the rank of the observation  $y_i$  in the sample  $y_1, y_2, \dots, y_n$ .
- Then **Spearman's rank correlation coefficient**  $\rho_S(x, y)$  is the Pearson's correlation coefficient calculated from the ranks.

## Spearman's rank correlation coefficient

- Spearman's rank correlation coefficient measures the monotonic dependence between two random variables. The coefficient is always in the interval  $[-1, 1]$  and (in case of no repeating data values) attains the absolute value 1 if and only if  $y = g(x)$  for some monotonic function  $g$ .
- If the variables  $x$  and  $y$  are independent, then the Spearman correlation  $\rho_S(x, y) = 0$  (using the same counterexample as with Pearson correlation, we see that the contrary does not again hold).
- See, [link 1](#) and [link 2](#) for values of the Spearman correlation for some particular data sets.

## Non-unique ranks

- It is possible that some of the sample points have the same rank.
- In that case, all those points are assigned to have the median of the corresponding ranks.
- For example, if two observations have the same rank, corresponding to ranks 7 and 8, then both are assigned to have rank 7.5. If three observations have the same rank, corresponding to ranks 3, 4, and 5, then each is assigned to have rank 4.

## Confidence intervals for Spearman correlation

Confidence intervals for Spearman's rank correlation coefficient can be estimated using the bootstrap.

## Significance tests for Spearman correlation

Significance test for Spearman correlation can be conducted non-parametrically either via a **permutation test** or through the following.

### Significance test for Spearman correlation, assumptions

Let  $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$ , be an i.i.d. random sample from a bivariate distribution.

### Significance test for Spearman correlation, hypotheses

$$H_0 : \rho_S = 0 \quad H_1 : \rho_S \neq 0.$$

## Significance test for Spearman correlation

### Significance test for Spearman correlation, test statistic

- The test statistic,

$$t = \sqrt{n-2} \cdot \frac{\hat{\rho}_S}{\sqrt{1 - \hat{\rho}_S^2}},$$

follows under  $H_0$  (for large  $n$ ) approximately the  $t_{n-2}$ -distribution.

- The expected value of  $t$  under  $H_0$  is approximately 0 and if the value of the test statistic has **large absolute value**, evidence against the null hypothesis  $H_0$  is found.

# MS-C1620 Statistical inference

## 7 Linear regression I

Jukka Kohonen

Department of Mathematics and Systems Analysis  
School of Science  
Aalto University

Academic year 2020–2021  
Period III–IV

# Contents

- 1 Linear regression model
- 2 Parameter estimation
- 3 Assessing model fit
- 4 Inference for model parameters

## Regression analysis

The aim in regression analysis is to study how the value of a **response** ("dependent variable") changes when the values of one or more **explanatory variables** ("independent variables", covariates) are varied.

The name regression comes from the concept of *regression toward the mean* stating that, given independent replications, extremal values tend to be followed by average sized ones.

## Regression analysis, examples

- Does the number of violent crimes depend on alcohol consumption and if it does, how strong is this dependence?
- Does statistics exam score depend on hours slept on the night prior to the exam and if it does, how strong is this dependence?
- Does salary depend on education level and if it does, how strong is this dependence?
- Does a parent's smoking have an effect on the height of a child and if it does, how strong is this dependence?
- Do crime rates depend on income inequality level and if yes, how strong is this dependence?

## Regression analysis, objectives

Possible aims in regression analysis are for example:

- Description of the dependence between the explanatory and dependent variables. What is the type of the relationship? How strong is the dependence?
- Predicting the values of the dependent variable.
- Controlling the values of the dependent variable.

# Simple linear regression

We begin by discussing the simplest (but still extremely useful!) form of regression, linear regression, starting with the case of single explanatory variable.

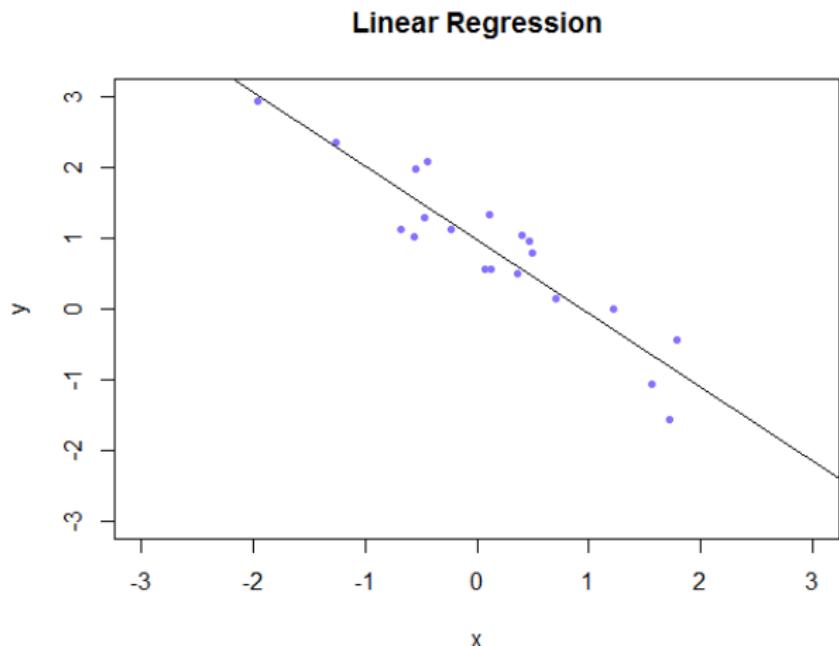
## Simple linear regression, assumptions

- Consider  $n$  observations (pairs)  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  of  $(x, y)$ . Assume, for simplicity, that the values  $x_i$  are non-random (otherwise we need an assumption of *exogeneity*).
- Assume that the values  $y_i$  depend linearly on the value  $x_i$ :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where the **regression coefficients**  $\beta_0$  and  $\beta_1$  are unknown constants.

# Simple linear model



**Figure:** When the values of the variable  $x$  increase, the values of the variable  $y$  decrease linearly.

# Simple linear regression

The simple linear regression model is usually coupled with the following additional assumptions.

## Simple linear regression, assumptions, continued

- The expected value of the errors is  $E[\varepsilon_i] = 0$  for all  $i = 1, \dots, n$ .
- The errors have the same variance  $\text{Var}[\varepsilon_i] = \sigma^2$ .
- The errors are uncorrelated i.e.  $\rho(\varepsilon_i, \varepsilon_j) = 0$ ,  $i \neq j$ .
- The errors are i.i.d. (*a stronger version of the previous two assumptions*).

## Simple linear regression

Under the previous assumptions, the random variables  $y_i$  have the following properties:

- Expected value:  $E[y_i] = \beta_0 + \beta_1 x_i, \quad i = 1, \dots, n,$
- Variance:  $\text{Var}(y_i) = \text{Var}(\varepsilon_i) = \sigma^2.$
- Correlation:  $\rho(y_i, y_j) = 0, \quad i \neq j.$
- If we chose to assume that the errors are i.i.d., then  $y_i$  are independent of each other.

## Simple linear regression, parameters

The linear model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

has three unknown parameters: regression coefficients  $\beta_0$ ,  $\beta_1$  and the error variance  $\text{Var}(\varepsilon_i) = \sigma^2$ .

These parameters are usually unknown and have to be *estimated* from the observations.

Under the assumption that  $E[\varepsilon_i] = 0$ , for all  $i = 1, \dots, n$ , the simple linear model can be given as

$$y_i = E[y_i] + \varepsilon_i, \quad i = 1, \dots, n,$$

where  $E[y_i] = \beta_0 + \beta_1 x_i$  is the **systematic part** and  $\varepsilon_i$  is the **random part** of the model.

## Simple linear regression, parameter interpretation

The systematic part

$$E[y_i] = \beta_0 + \beta_1 x_i$$

of the linear model defines the **regression line**

$$y = \beta_0 + \beta_1 x,$$

where  $\beta_0$  (**intercept**) is the intersection of the regression line and the  $y$ -axis and  $\beta_1$  is the **slope** of the regression line.

- The intercept  $\beta_0$  tells the expected value of the response when the explanatory variable  $x$  has the value zero.
- The slope  $\beta_1$  tells how much the expected value of the response variable  $y$  changes when the explanatory variable  $x$  grows by one unit.
- The error variance  $\text{Var}(\varepsilon_i) = \sigma^2$  describes the magnitude of the random deviations of the observed values from the regression line.

# Contents

- 1 Linear regression model
- 2 Parameter estimation
- 3 Assessing model fit
- 4 Inference for model parameters

## Simple linear regression, objective

The aim in (simple) linear regression analysis is to find **estimates** for the regression coefficients  $\beta_0$  and  $\beta_1$ .

The estimates  $\hat{\beta}_0, \hat{\beta}_1$  should be chosen such that the **fitted values/predictions**,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

best match the observations in some suitable sense.

Numerous ways of choosing the “best” estimates exist and the most popular of these is the **method of least squares**.

## The method of least squares

- In the method of least squares we choose the estimates by minimizing the sum of squared differences between the observations  $y_i$  and the fitted values  $\hat{y}_i$ ,

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2.$$

- The solutions are

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \hat{\rho}(x, y) \frac{s_y}{s_x} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where  $s_x, s_y, s_{xy}, \hat{\rho}(x, y)$  are the sample standard deviations, the sample covariance and the sample correlation of  $x$  and  $y$ .

## The estimated regression line

- The least squares estimates give an estimated regression line

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\&= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\rho}(x, y) \frac{s_y}{s_x} x_i \\&= \bar{y} + \hat{\rho}(x, y) \frac{s_y}{s_x} (x_i - \bar{x})\end{aligned}$$

- The slope (up or down) of the line is determined by the correlation between the two variables:
  - ▶ If  $\hat{\rho}(x, y) > 0$ , the line is increasing.
  - ▶ If  $\hat{\rho}(x, y) < 0$ , the line is decreasing.
  - ▶ If  $\hat{\rho}(x, y) = 0$ , the line is horizontal.

# Contents

- 1 Linear regression model
- 2 Parameter estimation
- 3 Assessing model fit
- 4 Inference for model parameters

## Fitted values and residuals

- Recall that the fitted value of the variable  $y_i$ , i.e. the value given to the variable  $y$  by the regression line at points  $x_i$ , is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \dots, n.$$

- The **residual**  $\hat{\varepsilon}_i$  of the estimated model is the difference

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \dots, n$$

between the observed value  $y_i$  (of the variable  $y$ ) and fitted value  $\hat{y}_i$ .

- The smaller the residuals of the estimated model are, the better the regression model explains the observed values of the response variable.

## Example

### Linear Regression

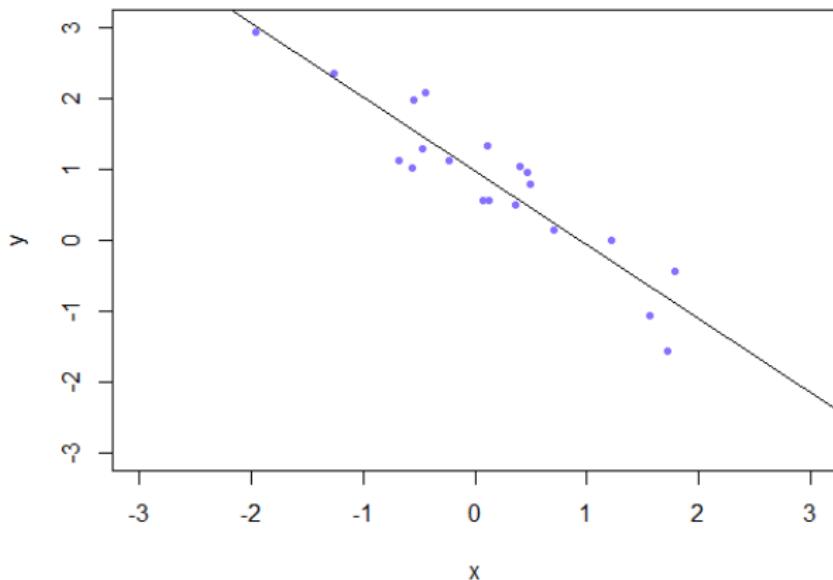


Figure: Estimated regression line minimizes the squared sum of the residuals.

## Residual mean square estimation

Under the regression assumptions, an unbiased estimate for the error variance  $\text{Var}(\varepsilon_i) = \sigma^2$  is given by

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

## Coefficient of determination

- **Coefficient of determination** (also known as “R-squared”) gives a single number with which to assess the accuracy of the model fit.
- Coefficient of determination is defined as

$$R^2 = 1 - \frac{SSE}{SST} = (\hat{\rho}(y, \hat{y}))^2,$$

where

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{and} \quad SSE = \sum_{i=1}^n \hat{\varepsilon}_i^2$$

measure the variation of the data “before” and “after” fitting the model.

- If SSE is small compared to SST, the model has managed to *explain* a large proportion of the variance in the data.
- We always have  $0 \leq R^2 \leq 1$ .

## Properties of the coefficient of determination

The following conditions are equivalent:

- The coefficient of determination  $R^2 = 1$ .
- All the residuals vanish,  $\hat{\varepsilon}_i = 0$ ,  $i = 1, \dots, n$ .
- All the observations  $(x_i, y_i)$  lie on the same line.
- The sample correlation coefficient  $\hat{\rho}(x, y) = \pm 1$ .
- The regression model explains the variation of the observed values of the response  $y$  completely.

The following conditions are equivalent:

- The coefficient of determination  $R^2 = 0$ .
- The regression coefficient  $\hat{\beta}_1 = 0$ .
- The sample correlation coefficient  $\hat{\rho}(x, y) = 0$ .
- The regression model completely fails in explaining the variation of the observed values of the dependent variable  $y$ .

## Model diagnostics

We will discuss the checking of the model assumptions ("linear model diagnostics") next time, when we have first defined multiple linear regression.

# Contents

- 1 Linear regression model
- 2 Parameter estimation
- 3 Assessing model fit
- 4 Inference for model parameters

## Inference for model parameters

We next go discuss *confidence intervals* and *hypothesis tests* for the intercept  $\beta_0$  and slope  $\beta_1$  of the simple linear regression model.

In addition to our earlier assumptions, the following results assume that

### Simple linear regression, assumptions, continued

- The errors  $\varepsilon_i$  are i.i.d.
- The errors  $\varepsilon_i$  are normally distributed.

The assumption of normality can be replaced by a *large enough* sample size.

## Slope, hypothesis test

The following test is used to test whether the slope parameter  $\beta_1$  of the simple linear model equals a given value (most often zero).

### Slope test, assumptions

(The assumptions of slides 8 and 23.)

### Slope test, hypotheses

$$H_0 : \beta_1 = \beta_1^0 \quad H_1 : \beta_1 \neq \beta_1^0.$$

# Slope, hypothesis test

## Slope test, test statistic

- The  $t$ -test statistic,

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{\frac{\hat{\sigma}}{\sqrt{n-1}s_x}},$$

where  $\hat{\sigma} = \sqrt{\text{Var}(\varepsilon_i)}$  (see slide 19) and  $s_x$  is the sample standard deviation of  $x$ , has under  $H_0$  Student's  $t$ -distribution with  $n - 2$  degrees of freedom.

- Under  $H_0$ , the expected value of  $t$  is 0 and **large absolute values** of the test statistic suggest that the null hypothesis  $H_0$  does not hold.

## Slope, confidence interval

A  $(1 - \alpha)100\%$  confidence interval for the slope  $\beta_1$  of the regression line is given as

$$\left( \hat{\beta}_1 - t_{n-2,\alpha/2} \frac{\hat{\sigma}}{\sqrt{n-1}s_x}, \hat{\beta}_1 + t_{n-2,\alpha/2} \frac{\hat{\sigma}}{\sqrt{n-1}s_x} \right),$$

where  $t_{n-2,\alpha/2}$  is the  $(1 - \alpha/2)$ -quantile of the  $t_{n-2}$ -distribution.

## Intercept, hypothesis test

Testing whether the intercept parameter  $\beta_0$  equals a given value is also sometimes of interest.

### Intercept test, assumptions

(The assumptions of slides 8 and 23.)

### Intercept test, hypotheses

$$H_0 : \beta_0 = \beta_0^0 \quad H_1 : \beta_0 \neq \beta_0^0.$$

# Intercept, hypothesis test

## Intercept test, test statistic

- The  $t$ -test statistic

$$t = \frac{\hat{\beta}_0 - \beta_0^0}{\hat{\sigma} \sqrt{\sum_{i=1}^n x_i^2}} \cdot \frac{1}{\sqrt{n(n-1)s_x}},$$

where  $\hat{\sigma} = \sqrt{\text{Var}(\varepsilon_i)}$  (see slide 19) and  $s_x$  is the sample standard deviation of  $x$ , has under  $H_0$  Student's  $t$ -distribution with  $n - 2$  degrees of freedom.

- Under  $H_0$ , the expected value of  $t$  is 0 and **large absolute values** of the test statistic suggest, that the null hypothesis  $H_0$  does not hold.

## Intercept, confidence interval

A  $(1 - \alpha)100\%$  confidence interval for the intercept  $\beta_0$  of the regression line is given as

$$\left( \hat{\beta}_0 - t_{n-2,\alpha/2} \frac{\hat{\sigma} \sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{n(n-1)s_x}}, \hat{\beta}_0 + t_{n-2,\alpha/2} \frac{\hat{\sigma} \sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{n(n-1)s_x}} \right),$$

where  $t_{n-2,\alpha/2}$  is the  $(1 - \alpha/2)$ -quantile of the  $t_{n-2}$ -distribution.

# MS-C1620 Statistical inference

## 9 Linear regression II

Jukka Kohonen

Department of Mathematics and Systems Analysis  
School of Science  
Aalto University

Academic year 2020–2021  
Period III–IV

# Contents

- 1 Multiple linear regression
- 2 Parameter estimation
- 3 Assessing model fit
- 4 Inference for model parameters
- 5 Extensions

# Multiple linear regression

The simple linear regression can be extended to **multiple linear regression** incorporating several explanatory variables.

## Multiple linear regression, assumptions

- Consider  $n$  observations  $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$  and assume that the  $p$ -dimensional  $\mathbf{x}_i$  are non-random.
- Assume, that  $p < n$ .
- Assume, that the values of the variable  $y$  depend linearly on the values of the variable  $x$ ,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n,$$

where the **regression coefficients**  $\beta_0, \beta_1, \dots, \beta_p$  are unknown constants.

# Multiple linear regression

The multiple linear regression model is usually coupled with the following additional assumptions.

## Multiple linear regression, assumptions, continued

- The expected value of the errors is  $E[\varepsilon_i] = 0$  for all  $i = 1, \dots, n$ .
- The errors have the same variance  $\text{Var}[\varepsilon_i] = \sigma^2$ .
- The errors are uncorrelated i.e.  $\rho(\varepsilon_i, \varepsilon_j) = 0$ ,  $i \neq j$ .
- The errors are i.i.d. (*a stronger version of the previous two assumptions*).

## Multiple linear regression

Under the previous assumptions, the random variables  $y_i$  have the following properties:

- Expected value:  $E[y_i] = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i, \quad i = 1, \dots, n,$
- Variance:  $\text{Var}(y_i) = \text{Var}(\varepsilon_i) = \sigma^2.$
- Correlation:  $\rho(y_i, y_j) = 0, \quad i \neq j.$
- If we chose to assume that the errors are i.i.d., then  $y_i$  are independent of each other.

## Multiple linear regression, parameters

The multiple linear model

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i, \quad i = 1, \dots, n,$$

has the following parameters: regression coefficients  $\beta_0$ ,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$  and the error variance  $\text{Var}(\varepsilon_i) = \sigma^2$ .

These parameters are usually unknown and have to be *estimated* from the observations.

Under the assumption that  $E[\varepsilon_i] = 0$ , for all  $i = 1, \dots, n$ , the simple linear model can be given as

$$y_i = E[y_i] + \varepsilon_i, \quad i = 1, \dots, n,$$

where  $E[y_i] = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i$  is the **systematic part** and  $\varepsilon_i$  is the **random part** of the model.

## Multiple linear regression, parameter interpretation

The systematic part

$$E[y_i] = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i$$

of the linear model defines the **regression plane**

$$\text{" } y = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}. \text{"}$$

- The intercept  $\beta_0$  tells the expected value of the response when the explanatory variable vector  $\mathbf{x}$  is the zero vector.
- The regression coefficient  $\beta_j$  tells how much the expected value of the response variable  $y$  changes when the value of the explanatory variable  $x_j$  grows by one unit *and the other variables stay constant.*
- The error variance  $\text{Var}(\varepsilon_i) = \sigma^2$  describes the magnitude of the random deviations of the observed values from the regression plane.

# Contents

- 1 Multiple linear regression
- 2 Parameter estimation
- 3 Assessing model fit
- 4 Inference for model parameters
- 5 Extensions

## Multiple linear regression, objective

The estimates  $\hat{\beta}_0, \hat{\beta}$  should be chosen such that the **fitted values/predictions**,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}^\top \mathbf{x}_i,$$

best match the observations in some suitable sense.

Again, the most popular solution method is the **method of least squares**.

Let  $\mathbf{X}$  be a  $n \times (p + 1)$  matrix whose first column is full of ones and the remaining columns correspond to the observed values of the  $p$  explanatory variables. Let the  $n$ -vector  $\mathbf{y}$  contain the observed response values.

## The method of least squares

- In the method of least squares we choose the estimates by minimizing the sum of squared differences between the observations  $y_i$  and the fitted values  $\hat{y}_i$ ,

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left( y_i - (\hat{\beta}_0 + \hat{\beta}^\top \mathbf{x}_i) \right)^2.$$

- Denoting  $\hat{\beta}^* = (\hat{\beta}_0, \hat{\beta}^\top)^\top$ , the explicit solution is

$$\hat{\beta}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

## Unstability of the solution

The least squares solution requires inverting the matrix  $\mathbf{X}^\top \mathbf{X}$ . If its rank is smaller than  $p$ , then some of the explanatory variables are fully linearly dependent. In that case, some of the variables can be excluded from the analysis without losing any information.

If  $\mathbf{X}^\top \mathbf{X}$  is full rank, it could still be that it is very close to being singular. This corresponds to **multicollinearity**, the case where the explanatory variables are not necessarily fully linearly dependent but still exhibit large correlations.

Multicollinearity can make the regression coefficients **unstable** and its presence can be investigate using the **variance inflation factors** of the explanatory variables.

## Variance inflation factor

Variance inflation factor (VIF) for an explanatory variable  $x_{ik}$  is defined as:

$$VIF_k = \frac{1}{1 - R_k^2},$$

where  $R_k^2$  is the coefficient of determination for a model where  $x_{ik}$  is the dependent variable and the remaining predictors are used as explanatory variables.

VIF is calculated separately for each explanatory variable  $x_{ik}$ . If the variable  $x_{ik}$  is uncorrelated with the other explanatory variables, then the VIF = 1.

If an explanatory variable has  $VIF \geq 10$ , multicollinearity is likely present, and some of the variables should be dropped from the model.

Roughly, the aim is to select variables such that the coefficient of determination (of the  $(\mathbf{x}_i, y_i)$ -model) is as high as possible and the explanatory variables are as uncorrelated as possible.

# Contents

- 1 Multiple linear regression
- 2 Parameter estimation
- 3 Assessing model fit
- 4 Inference for model parameters
- 5 Extensions

## Fitted values and residuals

- Recall that the fitted value of the variable  $y_i$ , i.e., the value given to the response variable by the estimated regression plane at the point  $\mathbf{x}_i$ , is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i, \quad i = 1, \dots, n.$$

- The **residual**  $\hat{\varepsilon}_i$  of the estimated model is the difference

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \dots, n$$

between the observed value  $y_i$  (of the variable  $y$ ) and fitted value  $\hat{y}_i$ .

- The smaller the residuals of the estimated model are, the better the regression model explains the observed values of the response variable.

## Residual mean square estimation

Under the regression assumptions, an unbiased estimate for the error variance  $\text{Var}(\varepsilon_i) = \sigma^2$  is given by

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \sum_{i=1}^n (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2 = \frac{1}{n - p - 1} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n - p - 1} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Note that the divisor equals the sample size  $n$  minus the number of estimated regression parameters  $p + 1$ .

## Coefficient of determination

- **Coefficient of determination** (also known as “R-squared”) gives a single number with which to assess the accuracy of the model fit.
- Coefficient of determination is defined as

$$R^2 = 1 - \frac{SSE}{SST},$$

where

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{and} \quad SSE = \sum_{i=1}^n \hat{\varepsilon}_i^2$$

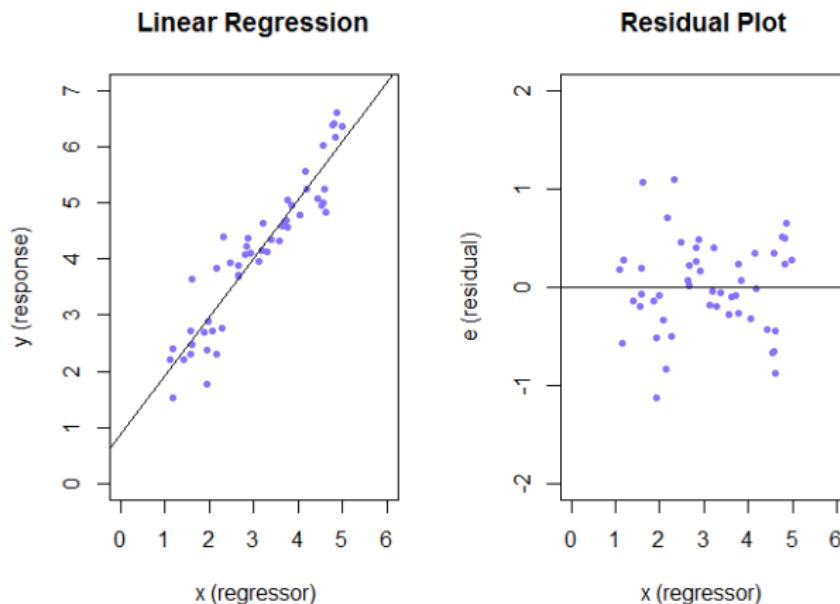
measure the variation of the data “before” and “after” fitting the model.

- If SSE is small compared to SST, the model has managed to *explain* a large proportion of the variance in the data.
- We always have  $0 \leq R^2 \leq 1$ .

# Diagnostics

- The verification of the assumptions of a regression model is called **diagnostics**.
- The diagnostics are usually performed by using the plots of the **residuals versus the fitted values** (or the explanatory variable  $x_i$  if there is only a single one).
- If the model assumptions hold, the residuals,
  - ① are approximately evenly distributed on both sides of zero,
  - ② have constant variance regardless of the value of  $x$  (no heteroscedasticity),
  - ③ exhibit no unusual (non-linear) patterns in general.
- Additionally, if the sample size is small and we cannot rely on the central limit theorem, the normality of the residuals should be tested.

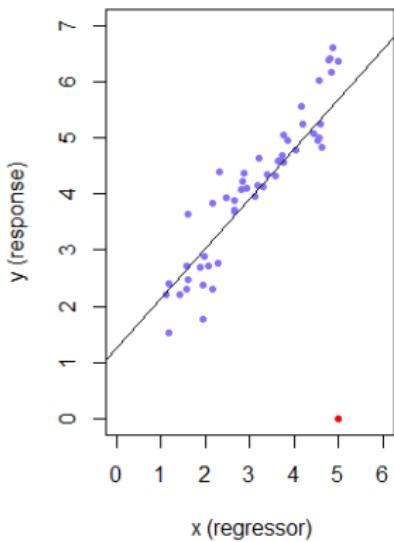
# Example, linear regression



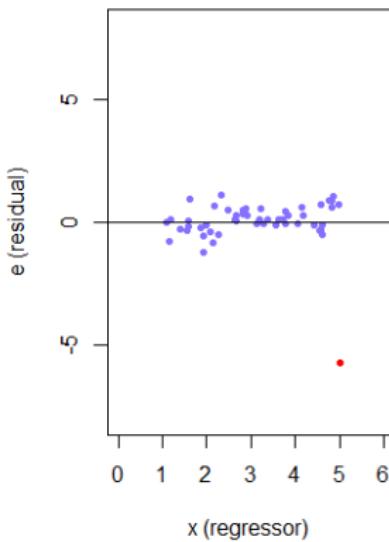
Kuva: Example of mostly satisfactorily looking residuals.

## Example, outlier

Linear Regression



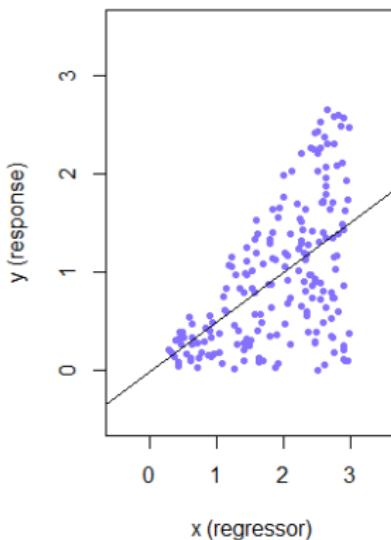
Residual Plot



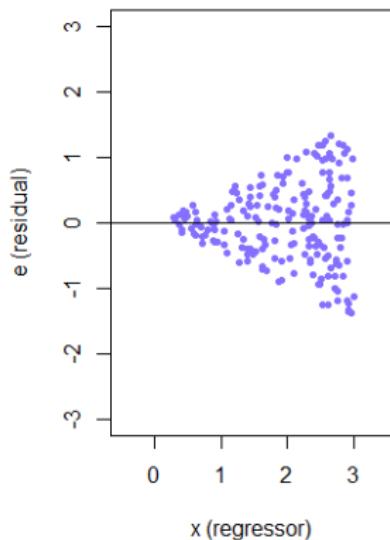
Kuva: Diagnostics can also be used to spot outliers.

## Example, heteroskedasticity

Linear Regression



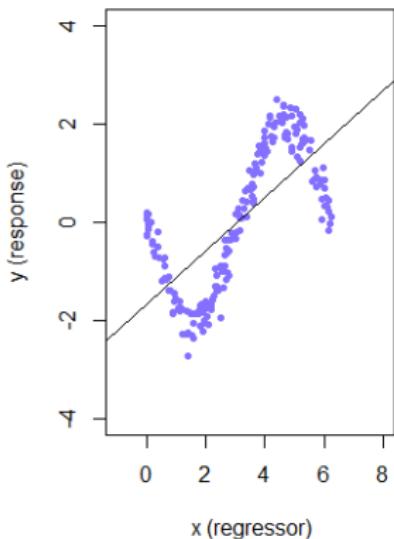
Residual Plot



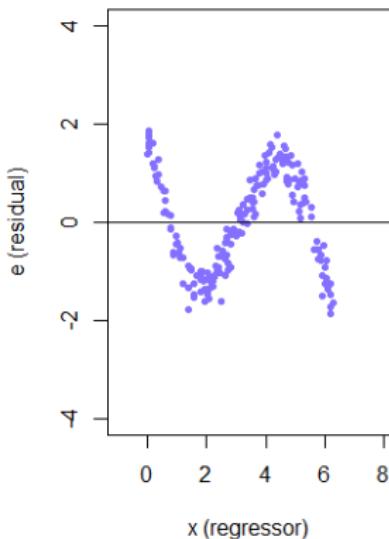
Kuva: The variance of the residuals depends clearly on the value of  $x$  (heteroskedasticity).

## Example, non-linear dependence

Linear Regression



Residual Plot



Kuva: The residuals exhibit clear non-linear dependency on  $x$ .

# Contents

- 1 Multiple linear regression
- 2 Parameter estimation
- 3 Assessing model fit
- 4 Inference for model parameters
- 5 Extensions

## Inference for model parameters

Analogous results as for the simple linear regression (confidence intervals, hypothesis tests) also exist for multiple linear regression under the assumptions that

### Multiple linear regression, assumptions, continued

- The errors  $\varepsilon_i$  are i.i.d.
- The errors  $\varepsilon_i$  are normally distributed.

The assumption of normality can be replaced by a *large enough* sample size.

However, we will not go through the theory behind them here.

## Bootstrap in linear models

In addition to the standard normality/CLT-based inference output by the software, bootstrap can be used to obtain confidence intervals for the model parameters. A bootstrap sample can be created in two ways.

- **Observation resampling:** we simply draw a bootstrap sample of size  $n$  from amongst the observations  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$  and fit a linear model to it (and repeat the same  $B$  times).
- **Residual resampling:** a bootstrap resample is obtained as  $(\mathbf{x}_1, y_1^*), \dots, (\mathbf{x}_n, y_n^*)$ , where  $y_i^* = \hat{y}_i + \hat{\varepsilon}_i^*$  and  $(\hat{y}_1, \dots, \hat{y}_n)$  are the fitted values of the original sample and  $\hat{\varepsilon}_1^*, \dots, \hat{\varepsilon}_n^*$  is a bootstrap sample of the **residuals** of the model for the original sample.

The residual resampling is suited for situations where we want to preserve the explanatory variable structure of the original data also in the bootstrap resamples (e.g. two treatment groups of certain sizes).

# Contents

- 1 Multiple linear regression
- 2 Parameter estimation
- 3 Assessing model fit
- 4 Inference for model parameters
- 5 Extensions

## Extensions of the multiple regression model

The following slides list some of the most common cases where extensions to the standard linear regression methodology are required.

**Problem:** *My data exhibits non-linear dependencies.*

**Solution:**

- Transform your predictors. That is,  $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$  is still a multiple regression model.
- Add **interaction terms**,  $\beta_{12}x_{i1}x_{i2}$  (note: interaction terms make interpreting the coefficients difficult).
- A more automated solution is given by non-linear regression, such as *kernel regression*.

**Problem:** *I have several response variables.*

**Solution:** Model them simultaneously using *multivariate regression*:

$$\mathbf{y}_i = \boldsymbol{\beta}_0 + \mathbf{B}^\top \mathbf{x}_i + \varepsilon_i, \quad i = 1, \dots, n.$$

## Extensions of the multiple regression model

**Problem:** *My data contains a large number of outliers and I would rather not discard them all.*

**Solution:** Use a fitting method which is less sensitive to outlying values such as “the method of least absolute values”, or  $\ell_1$ -regression.

**Problem:** *My data points exhibit significant correlations.*

**Solution:** Depending on the nature of the correlations, use either *mixed models* (correlation within groups of subjects), *time series* methods (temporal correlation) or something else.

## Extensions of the multiple regression model

**Problem:** *My response variable takes values in some specific subset of  $\mathbb{R}$ . That is, the ranges of the two sides of the model equation  $E[y_i] = \beta_0 + \beta^\top \mathbf{x}_i$  do not match.*

**Solution:** Use a suitable *link function* to transform your model equation,  $g(E[y_i]) = \beta_0 + \beta^\top \mathbf{x}_i$ . This leads, e.g., to *logistic regression* and *log-linear models*.

**Problem:** *Too many variables,  $p \geq n$ .*

**Solution:** This makes the matrix  $\mathbf{X}^\top \mathbf{X}$  singular, meaning that no least squares solution exist. This problem can be solved by regularized regression estimates discussed briefly next time.

# MS-C1620 Statistical inference

## 10 Linear regression III

Jukka Kohonen

Department of Mathematics and Systems Analysis  
School of Science  
Aalto University

Academic year 2020–2021  
Period III–IV

# Contents

1 Variable selection

2 Shrinkage methods

## Variable selection

In modern data analysis it is common to encounter data sets with extremely **large numbers of predictors/explanatory variables**.

It is, however, possible that not all of the predictors are actually related to the response variable (maybe we did not have a clear idea what would make for a good predictor and measured a large number of predictor, “just in case” ).

Methods which aim to identify the relevant variables are known as **variable selection** methods

## Reasons for variable selection

Possible reasons for trying to narrow down the number of predictors in a regression model:

- ① Helps us **interpret** the model better (understand the phenomenon underlying the data).
- ② Predictors not related to the response add extra noise in prediction.
- ③ Avoiding collinearity.
- ④ Cost issues, it might be cheaper to observe only a subset of the variables.

## Backward selection

Two most basic methods of variable selection are **backward selection** and **forward selection** based on  $p$ -values.

The backward selection works by selecting a  $p$ -value cutoff  $\alpha_0$  (e.g. 0.05) and proceeding as follows:

- ① Estimate the model using all predictors.
- ② Remove the predictor with the highest  $p$ -value greater than or equal to  $\alpha_0$  and estimate the new model.
- ③ Repeat step 2 until all predictors have  $p$ -values less than  $\alpha_0$ .

That is, backward selection begins with a full model and one-by-one removes the variables that are the least “important”, until we are left with the subset of “most important” variables.

## Forward selection

The forward selection works by selecting a  $p$ -value cutoff  $\alpha_0$  (e.g. 0.05) and proceeding as follows:

- ① Start with a model with no predictors at all.
- ② For each predictor one at a time, check what their  $p$ -value would be if they were added to the model and add the one with the smallest  $p$ -value below  $\alpha_0$  to the model.
- ③ Repeat step 2 until no new predictors with  $p$ -values less than  $\alpha_0$  can be added.

That is, forward selection begins with an empty model and one-by-one adds the variables that are the most “important”, until no more “important” variables are left to be added.

## Backward and forward selection

While the backward and forward selection are natural and simple to use, they have some drawbacks:

- ① It is possible to miss the optimal model as not all possible combinations of the predictors are considered during the process. (a combination of the backward and forward selection, *stepwise selection*, would avoid this)
- ② The more predictors we are left with, the higher is the probability of encountering at least one type I error. That is, it could be that not all retained predictors are actually statistically significant.
- ③ The absolute  $p$ -value cut-off might miss some “almost significant” predictors which are actually relevant.

Note: both backward and forward selection get increasingly complex if one allows for interaction terms between the predictors (e.g.  $\text{age} \times \text{sex}$ ).

## Alternative method

Alternative to the backward and forward methods is to go through all possible models and choose in some sense the *best* one. E.g., if one has  $d$  variables and uses only “main effects” (no interactions), there are a total  $2^d$  models to choose from.

The *best* model should make a compromise between fitting the data well (large enough  $R^2$  to be useful) and the number of variables (few enough variables to be interpretable).

Instead of  $R^2$ , it is common to measure the model’s goodness of fit using *log-likelihood*,

$$\ell = -\frac{n}{2} (\log(2\pi) + \log(\hat{\sigma}^2) + 1).$$

The larger  $\ell$  is (the smaller the residual variance  $\hat{\sigma}^2$  is), the better the model explains the behavior of the response variable.

## Akaike information criterion

Both  $R^2$  and  $\ell$  never decrease when we add predictors to the model. As such they cannot be used on variable selection on their own (that is, both  $R^2$  and  $\ell$  would always be in favor of adding more variables to the model).

One of the most common metrics for model selection is known as *Akaike information criterion* and it compares the models based on their log-likelihoods but **penalizes** for a large number of variables.

$$\text{AIC} = -2\ell + 2k.$$

where  $k$  is #parameters in model ( $\approx$  #variables used).

Generally, AIC is

- **Small** for simple models (using few variables) that explain the response well (large  $\ell$ ).
- **Large** for complex models (using many variables) that fail to explain the response (small  $\ell$ ).

Choose the model with the smallest AIC, out of all  $2^d$  models (if  $d$  variables available).

## Alternative criteria

Multiple criteria having the same idea as AIC (reward for explaining the response, penalize for using many variables) exist:

- *Bayesian information criterion,*

$$\text{BIC} = -2\ell + k \log(n),$$

smaller is better (again  $k = \#\text{parameters estimated}$ )

- *Adjusted R<sup>2</sup>,*

$$R_A^2 = 1 - \frac{n-1}{n-k}(1 - R^2),$$

larger is better ( $p = \#\text{variables}$ )

## Drawbacks of the criteria

Also the criteria-based variable selection methods have their drawbacks:

- AIC and BIC assume normally distributed errors.
- Even though we are sure to find the optimal model, going through all  $2^d$  of them is computationally costly.
  - ▶ One solution to this is to combine the criteria with the backward/forward selection. That is, always include/drop the variable which most improves the criterion value. For AIC this can be done with the function `step` in R.
- It is still possible to miss *almost significant* predictors if they do not improve the fit enough.

# Contents

1 Variable selection

2 Shrinkage methods

## Constraint for the model coefficients

Our next tools for variable selection, **shrinkage methods**, allow “continuous” variable selection. That is, the output shows us how close the model is to including specific variables.

Shrinkage methods conduct variable selection by limiting the *size* of the estimated coefficients in the model,

$$\|\hat{\beta}\| \leq \text{some limit.}$$

Idea:

- *Unconstrained model*: Unlimited amount of “money” to “spend” on the coefficients/predictors. Randomness of the data can cause the model to make some “bad purchases”.
- *Constrained model*: With a limited amount of “money” to “spend”, the model must focus on acquiring only the most important variables.

## Vector norms

The size of the estimated coefficients  $\hat{\beta}$  can be measured using **vector norms**. Most commonly used are the norms  $\|\cdot\|_r$ ,  $1 \leq r < \infty$ ,

$$\|\mathbf{v}\|_r = (|v_1|^r + |v_2|^r + \cdots + |v_p|^r)^{1/r}, \quad \text{where } \mathbf{v} = (v_1, v_2, \dots, v_p).$$

Two particular choices include,

- $r = 1$ , leading to a method known as *LASSO*.
- $r = 2$ , leading to a method known as *ridge regression*.

We start with the latter one.

## Ridge regression

Ridge regression has the same assumptions as regular multiple regression (excluding the normality assumption as no inference is made in ridge regression) and it minimizes the least squares criterion,

$$\sum_{i=1}^n \left( y_i - (\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i) \right)^2,$$

under the constraint that  $\|\boldsymbol{\beta}\|_2^2 \leq s$ , for some  $s$ .

This problem can be shown to be equivalent to minimizing,

$$\sum_{i=1}^n \left( y_i - (\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i) \right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2,$$

where there is one-to-one correspondence between  $\lambda$  and  $s$ .

## The parameter $\lambda$

The parameter  $\lambda \geq 0$  is a so-called “tuning parameter” and it controls how much the coefficients are penalized.

- If  $\lambda = 0$ , there is no penalization and the estimates are simply the usual least squares estimates (we have an unlimited amount of “money”).
- The larger the value of  $\lambda$ , the more the coefficients are penalized (“shrunk” towards zero), making only the important variables stand out (we have less “money” at our disposal and have to make informed purchases).

We will discuss the optimal choice of  $\lambda$  later.

## Ridge solution

The ridge regression solution can be expressed analytically if we first center our data.

That is, replace each predictor  $x_{ij}$  by  $x_{ij} - \bar{x}_j$  where  $\bar{x}_j$  is the sample mean of the  $j$ th predictor. **The centering eliminates the need for the intercept term** in the model and the least squares criterion is then

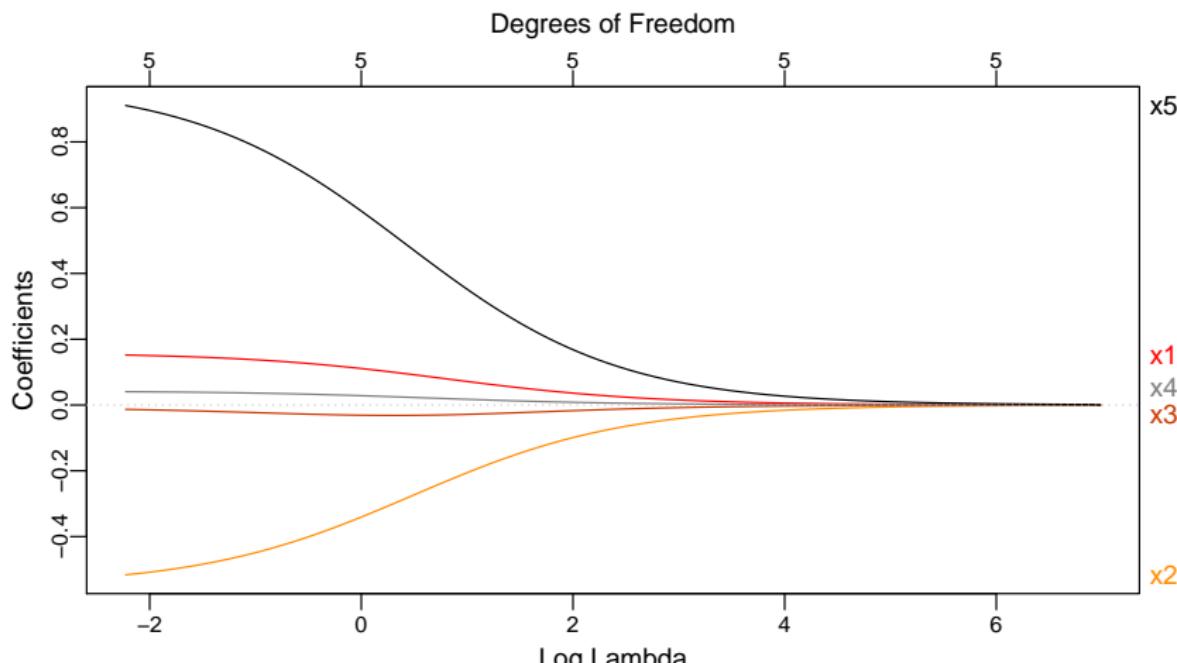
$$\sum_{i=1}^n \left( y_i - \boldsymbol{\beta}^\top \mathbf{x}_i \right)^2 + \lambda \cdot \boldsymbol{\beta}^\top \boldsymbol{\beta},$$

where  $\boldsymbol{\beta}^\top \boldsymbol{\beta} = \|\boldsymbol{\beta}\|_2^2$ . Setting the derivative of the function to zero shows that it is minimized by the **ridge solution**,

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.$$

## Ridge coefficient profiles

The results of ridge regression are best visualized by computing them for a sequence of values of  $\lambda$  and plotting the coefficient values versus  $\lambda$  (or  $\log(\lambda)$ ).



## Variable selection with ridge regression

Variable selection with ridge regression is tricky as

- The coefficient sizes are not a measure of the variables' importance, as they depend on the scales of the variables.
- The coefficients never reach exactly zero. All variables are part of the model for all values of  $\lambda$ .

As a conclusion, ridge regression should not be used for variable selection.

However, it still has other benefits:

- It helps deal with multicollinearity.
- It avoids overfitting to noise by shrinking the coefficients of the noise variables.

A better alternative in terms of variable selection is given by the LASSO estimator.

## LASSO

LASSO (least absolute shrinkage and selection operator) has the same assumptions as ridge regression and it minimizes the least squares criterion,

$$\sum_{i=1}^n \left( y_i - (\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i) \right)^2,$$

under the constraint that  $\|\boldsymbol{\beta}\|_1 \leq s$ , for some  $s$ .

This problem can be shown to be equivalent to minimizing,

$$\sum_{i=1}^n \left( y_i - (\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i) \right)^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

where there is one-to-one correspondence between  $\lambda$  and  $s$ .

## LASSO vs. ridge regression

The formulations of LASSO and ridge regression look very similar, differing only in their choice of norm,  $\|\cdot\|_1$  for LASSO and  $\|\cdot\|_2$  (squared) for ridge.

However, this difference plays a big role in the methods' results. The geometry induced by the norm  $\|\cdot\|_1$  is such that it can **force coefficients to equal exactly zero** for an appropriate choice of the tuning parameter  $\lambda \geq 0$ .

The parameter  $\lambda$  again controls how much the coefficients are penalized/shrunk towards zero with the same interpretations as in ridge regression.

## LASSO solution

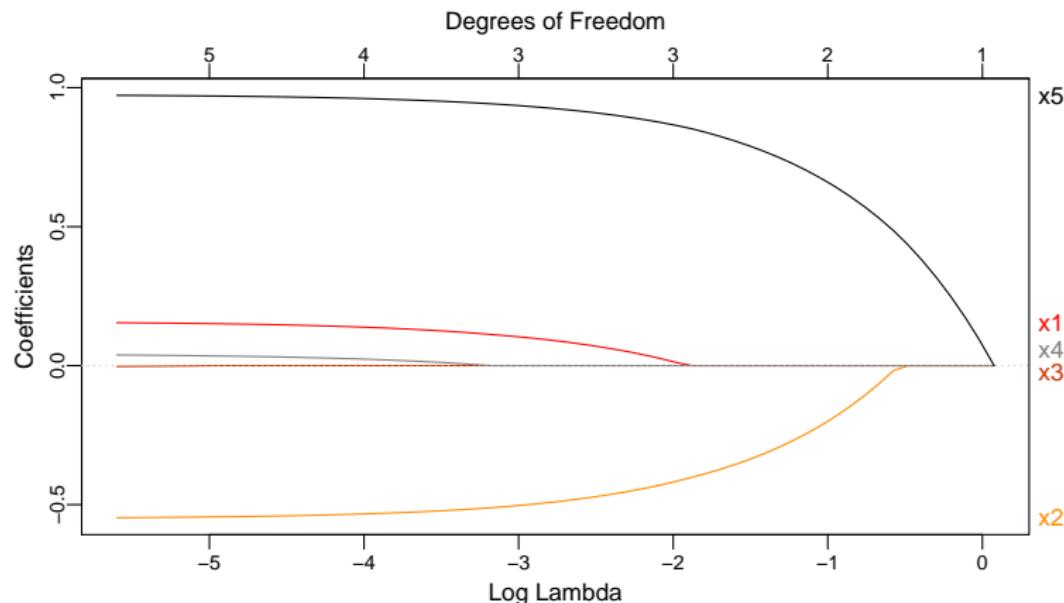
The non-differentiable absolute values in the LASSO penalty,  $\|\beta\|_1 = |\beta_1| + |\beta_2| + \dots + |\beta_d|$ , mean that the LASSO solution  $\hat{\beta}_{LASSO}$  cannot be obtained by standard means.

Statistical software computes the solution (which has no closed form) numerically.

As with ridge regression, it is standard to compute the LASSO solution for multiple values of  $\lambda$  and plot the resulting coefficients as functions of  $\lambda$  (or  $\log(\lambda)$ ).

## LASSO coefficient profiles

The LASSO coefficient profiles show that after particular values of  $\lambda$  each coefficient hits zero and stays there.



## LASSO coefficient profiles

The plot on the previous slide shows that

- the black variable is the most important (we “buy” it first),
- the orange variable is the second most important,
- the red variable is the third most important,
- and so on...

## Choosing the value of $\lambda$

But how should one choose which  $\lambda$  ("budget") to pick?

It is standard to select  $\lambda$  in LASSO using **cross-validation**, by choosing the value which makes the best predictions.

- Too small value of  $\lambda$  (too much "money" at our disposal) makes us include also irrelevant variables (noise) in the model and this makes prediction difficult.
- Too large value of  $\lambda$  (too little "money" at our disposal) makes us leave important variables out of the model, again making prediction difficult.

The best choice is usually in between the above two.

## Training, validation and test data sets

In modern study of prediction methods (especially in machine learning), it is common to divide the data into three **disjoint** sets, training, validation and test data.

- **Training data** is used to fit the model (estimate the parameters), possibly for multiple values of a tuning parameter  $\lambda$ .
- **Validation data** is used to choose the value of the tuning parameter such that the obtained model makes the smallest average squared error in predicting the response values in the validation data.
- **Test data** is used to evaluate the performance of the obtained model. The smaller the prediction error on the test data, the better the method is.

The data sets are kept disjoint so that no step influences another, and that we get fully objective results in the testing step.

## Cross-validation

Cross-validation is a modification of the previous scheme including only the training and validation steps.

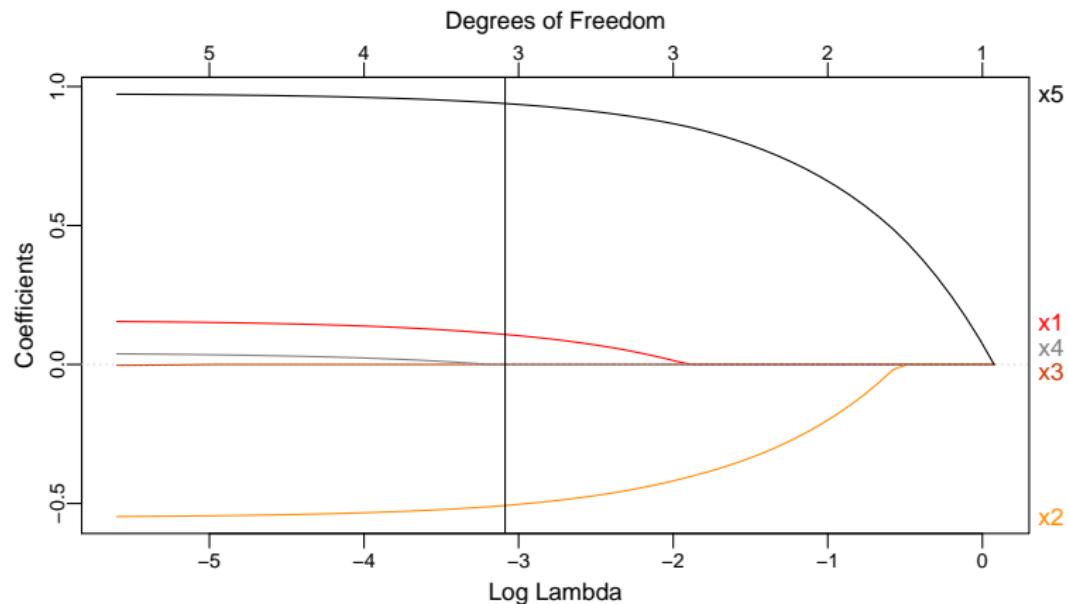
$k$ -fold cross validation proceeds as follows,

- ① Split the data into  $k$  groups of as equal size as possible.
- ② For each of the  $k$  groups:
  - ▶ Use the remaining  $k - 1$  groups together to fit the model for several values of  $\lambda$ .
  - ▶ Compute the average squared prediction errors of the fitted models on the left-out set.
- ③ For each used value of  $\lambda$ , average the obtained  $k$  average squared prediction errors.
- ④ Choose the  $\lambda$  with the smallest average.

Being based on multiple evaluations of the models, cross-validation leads to a more “robust” choice of the tuning parameters than a single validation would (“majority vote vs. single person deciding”).

## Cross-validation in LASSO

The tuning parameter  $\lambda$  is in LASSO usually selected in the manner described in the previous slide, e.g. using 10-fold cross validation. The vertical line below shows the optimal value for the example data.



## Cross-validation in LASSO

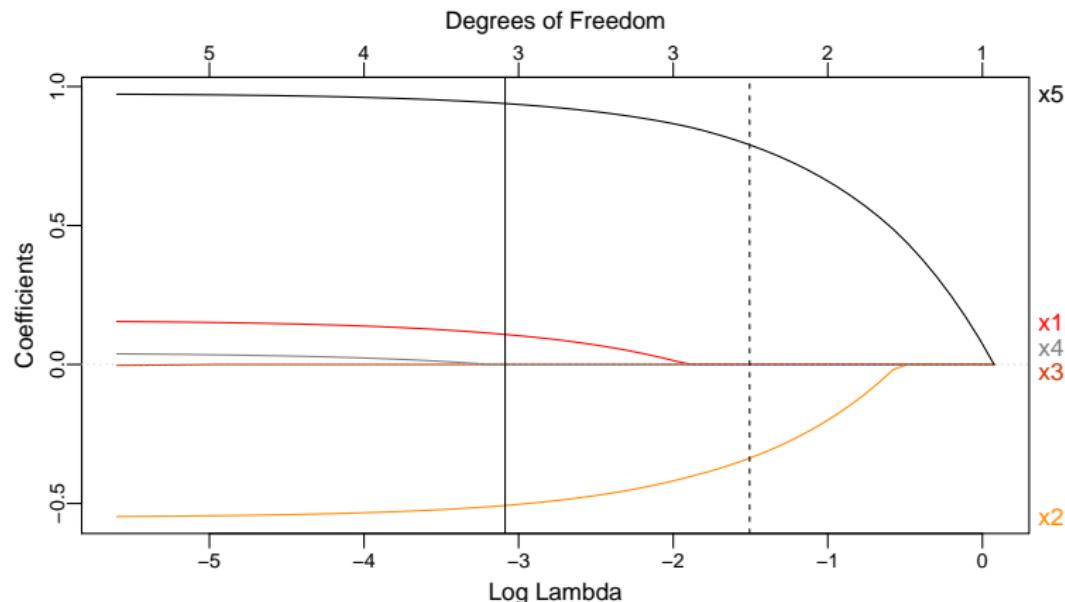
Usually the choice of  $\lambda$  which yields the lowest average prediction error produces a model which still contains too many variables from a practical point of view.

To obtain a more *sparse* model (one with less variables), we choose the simplest model which still explains the response “almost as well as” the optimal model.

The standard choice is to pick the largest value of  $\lambda$  which has a prediction error still within one standard deviation of the optimal prediction error.

## Cross-validation in LASSO

The below plot shows both the optimal value (solid line) and the “one-standard-error” value (dashed line) of  $\lambda$ . The latter has selected the variables  $x_2, x_5$  in the model.



## Other methods

Besides those that we covered, numerous methods exist for variable selection. For example,

- Algorithms such as Branch-and-Bound can be used to speed-up the criteria-based variable selection methods.
- Elastic net is a combination of ridge regression and LASSO.
- LARS (least angle regression) is combination of forward selection and LASSO.

## Key references

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Vol. 1. No. 10. New York, NY, USA: Springer series in statistics, 2001.

# MS-C1620 Statistical inference

## 8 Kernel regression

Jukka Kohonen

Department of Mathematics and Systems Analysis  
School of Science  
Aalto University

Academic year 2020–2021  
Period III–IV

# Contents

1 Kernel regression

2 Kernel density estimation

3 More information

# Regression function

Simple linear regression (Lecture 7) is a special case of fitting a **regression function** to the data.

$$y_i = g(x_i) + \epsilon_i$$

Linear model  $g(x) = \beta_0 + \beta_1 x$  has two parameters.

Many other functional forms of  $g$  could be used, e.g.

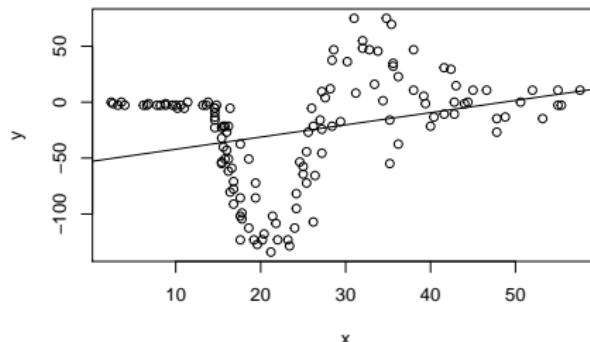
- higher order polynomials
- multiple explanatory variables (Lecture 9)
- piecewise regression
- **kernel regression (this lecture)**

## Example: Motorcycle data (testing crash helmets)

Head acceleration in a simulated motorcycle accident.

$x$  = time, explanatory variable (unit=ms)

$y$  = acceleration, response variable (unit=g)



Linear model fits badly, but would a polynomial be any better?

Instead of trying to fit a “global” model to all of the data, let’s try to understand its behaviour “locally”.

## General idea of local regression

Key idea: At any given point  $x$ , the value of the regression function  $g(x)$  is calculated **from nearby data points** (not all data points).

Some variants of the idea:

- KNN regression: Average the  $k$  nearest data points.
- Sliding window: Average all data points that are within  $h$  units of  $x$ .
- Kernel regression (kernel smoothing): Average nearby data points, giving **bigger weight** to data points that are very near. A **kernel function**  $K$  maps distances to weights.

## Simple kernel regression

Nadaraya-Watson regression: At any point  $x$ , define regression function value as a weighted average of data points

$$g(x) = \sum_{i=1}^n w_i y_i,$$

where the weights are calculated as

$$w_i = \frac{K(x - x_i)}{\sum_{j=1}^n K(x - x_j)},$$

and  $K$  (kernel function) is some nice function that gives big values when  $x_i$  is near  $x$ . The divisor just makes sure that the weights sum to one.

## Choice of kernel function

The kernel function is typically defined in two steps:

- ① Choose a **shape**, such as a triangular function, parabola, or the density function of standard normal distribution
- ② Choose a **bandwidth** that **scales** the shape to desired width = how far datapoints are used in the averaging

Example: parabolic (Epanechnikov) kernel

$$K_1(u) = \frac{3}{4}(1 - u^2)$$

for  $-1 \leq u \leq 1$ , and zero outside that interval.

Then scaled to bandwidth  $h$  with

$$K_h(u) = K_1(u/h).$$

This is positive for  $-h \leq u \leq h$ .

See [https://en.wikipedia.org/wiki/Kernel\\_\(statistics\)](https://en.wikipedia.org/wiki/Kernel_(statistics)) for many other kernel shapes.

## Choice of bandwidth

Large bandwidth = averaging many datapoints = very smooth regression function that only shows “large scale” features of the data. Also efficiently smoothes small errors away.

Small bandwidth = averaging few datapoints = very wild regression function that follows the data very closely. But also retains its errors.

Many methods exist for choosing the “best” bandwidth (see literature), but for exploratory analysis you could just experiment with different values. There are also “adaptive” methods which use smaller bandwidth if there are many data points nearby.

## Local linear regression

Instead of taking the **average** of the nearby points, we can also fit a **straight line** to them. This is called **local linear regression**.

In other words, we do a linear regression, but only on the data points  $x_i$  that are near  $x$ , and weighted by a kernel function. Then define  $g(x)$  to be the value of that regression line **at**  $x$ .

Note that for each value  $x$  where we are evaluating the regression function, we look at *different* “nearby” datapoints or use different weights, so the regression function  $g(x)$  that we obtain need not be “linear” at all.

Nadaraya-Watson (local constant) and local linear regression usually produce similar results, except at **edges of the data**. (Consider what happens in time series prediction.)

Just like in “global” regression, in local regression we can also use higher degree polynomials (e.g. parabolic).

# Contents

- 1 Kernel regression
- 2 Kernel density estimation
- 3 More information

## Kernel density estimation

The same idea, “looking at nearby datapoints”, can be used to estimate the density function of a distribution, if we have a sample  $x_1, x_2, \dots, x_n$  from it.

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i)$$

You can think of the datapoints  $x_i$  as representing point masses  $1/n$  each, then doing kernel smoothing to distribute those masses around  $x_i$  over some distance (by the kernel function).

This gives often a nicer, smoother estimate of the unknown density than a histogram.

# Contents

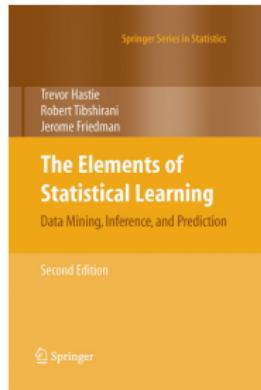
1 Kernel regression

2 Kernel density estimation

3 More information

## More information

You can learn more about local / kernel regression from e.g. the freely available book <https://web.stanford.edu/~hastie/ElemStatLearn/> (Chapter 6: Kernel smoothing methods)



# MS-C1620 Statistical inference

## 11 Analysis of variance

Jukka Kohonen

Department of Mathematics and Systems Analysis  
School of Science  
Aalto University

Academic year 2020–2021  
Period III–IV

# Contents

1 Analysis of Variance

2 Kruskal-Wallis test

# ANOVA

Analysis of variance (ANOVA) is the generalization of the two sample *t*-test for more than two populations.

In analysis of variance the population consists of two or more independent groups. Observations are assumed to follow normal distribution. Each group is independently sampled.

ANOVA tests the equality of the expected values of the groups.

- Is there a difference in mean salary in the 10 largest cities in Finland?
- Is there a difference in the average lengths of products made in different production lines?

# ANOVA

## ANOVA, assumptions

- Let  $x_{1j}, x_{2j}, \dots, x_{nj}$  be i.i.d. observed values of a  $\mathcal{N}(\mu_j, \sigma^2)$ -distributed random variable  $x_j$ ,  $j = 1, \dots, k$ . Assume that the  $k$  samples are independent.

We thus have  $k$  independent random samples from univariate normal distributions. The variances of all  $k$  distributions are assumed to be the same.

## ANOVA, hypotheses

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k.$$

$$H_1 : \mu_h \neq \mu_j \text{ for some groups } h \neq j.$$

## ANOVA, the basic idea

In analysis of variance, the total variance is divided into two parts. The first part measures the **variation between the group means** and the second part measures **variation within the groups**.

If the first part is much larger than the second part, there is evidence against the null hypothesis and we reject it. Otherwise, it is plausible that the group means are equal.

Hence, the test of the equality of the expected values is based on the comparison of between-groups variance and within-groups variance (giving the name of the method).

## ANOVA, components

To conduct ANOVA we calculate,

- ① the group means

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij},$$

where  $n_j$  is the group size of the  $j$ th group,

- ② the combined sample mean

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij},$$

where  $n = \sum_{j=1}^k n_j$ ,

## ANOVA, components

- ③ the variance between groups (group sum of squares)

$$SSG = \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2 = \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2 \quad \text{and}$$

- ④ the variance within groups (error sum of squares)

$$SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2.$$

# ANOVA

## ANOVA, test statistic

- The  $F$ -test statistic,

$$F = \frac{SSG / (k - 1)}{SSE / (n - k)},$$

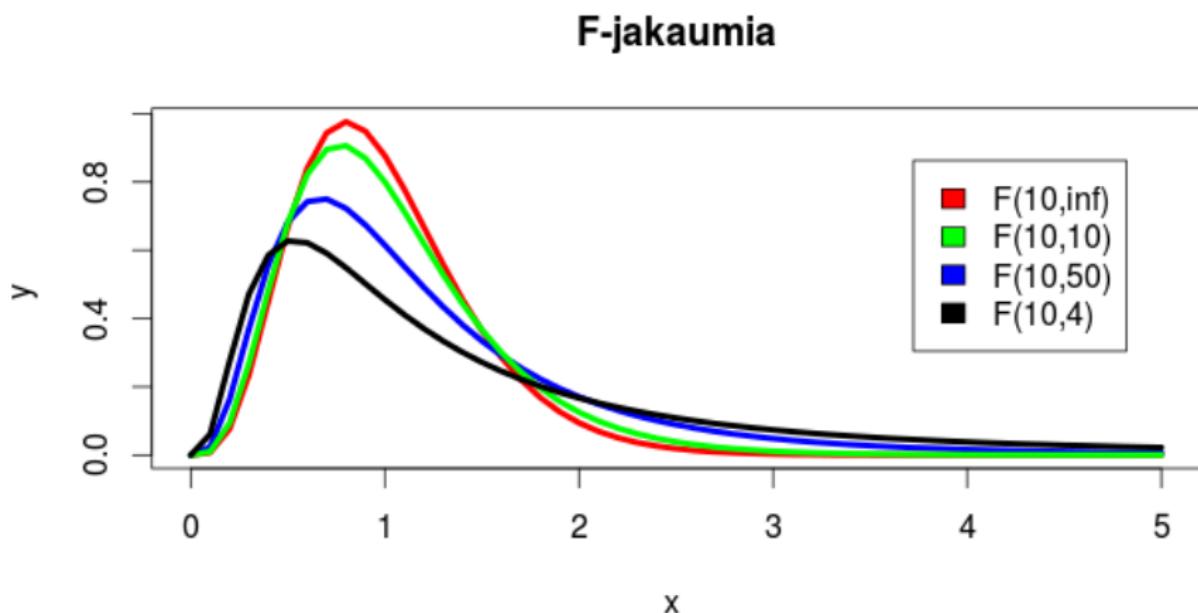
follows under  $H_0$  the  $F$ -distribution with the parameters  $(k - 1)$  and  $(n - k)$ .

- The expected value of the test statistic under  $H_0$  is  $\frac{n-k}{n-k-2}$  and **large** values of the test statistic give evidence against the null hypothesis.

## *F*-distribution

*F*-distribution is a family of distributions indexed by two parameters.

It is rarely encountered outside of theoretical results.



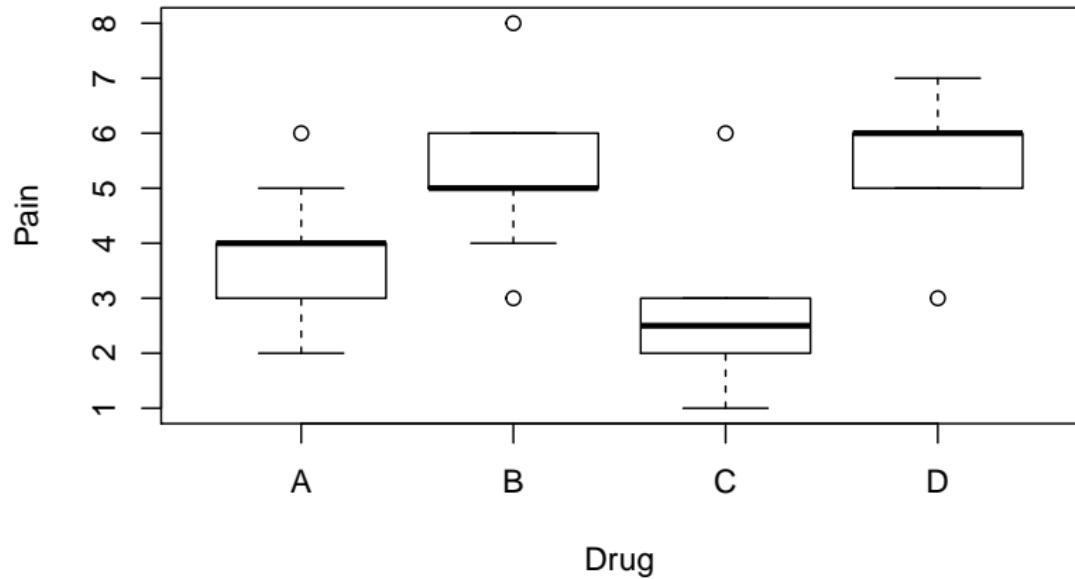
## Example

- A drug company wants to test the effectiveness of 4 drugs in relieving headache. The company recruited 40 volunteers and randomized them to take one of the four drugs during their next headache. The subjects reported their pain after one hour of taking the drug on a scale 1-10 (1 = no pain).

A	4	3	2	4	6	5	4	3	4	4
B	6	5	4	5	8	5	3	5	5	6
C	2	3	1	3	2	6	3	3	2	2
D	6	7	6	5	5	5	6	3	6	7

- We are interested in studying whether the drugs differ in their effectiveness on the significance level  $\alpha = 0.05$ .

## Example



Kuva: Boxplots of pain by drug.

- The group means are,

$$\bar{x}_A = 3.9 \quad \bar{x}_B = 5.2 \quad \bar{x}_C = 2.7 \quad \bar{x}_D = 5.6,$$

and the combined sample mean is  $\bar{x} = 4.35$ .

- The variance between groups is,

$$SSG = 52.1,$$

and the variance within groups is,

$$SSE = 55.$$

- This gives the F-statistic value,

$$F = \frac{(40 - 4)52.1}{(4 - 1)55} = 11.367,$$

and the  $p$ -value of 0.0000216, indicating that at least one of the treatments differs in effectiveness from the others.

## Multiple testing problem

If the null hypothesis is rejected based on the  $F$ -test, then we know that **at least two of the groups differ** (but we do not know which ones!).

The next step in the analysis is usually to find out the groups with statistically significant differences in expected values.

A simple way to do this is to analyze the groups in pairs of two with  $t$ -test.

There are  $C = \frac{k(k-1)}{2}$  pairs in total to compare and conducting all possible comparisons has the side effect that the **probability of type I error is inflated greatly above its set level**.

This is called the *multiple testing problem*.

## Bonferroni correction

Let  $\beta$  be the significance level of the  $C$  pair-wise comparisons, i.e., the (upper bound for the) probability that  $H_0$  is incorrectly rejected in a single comparison, i.e., the probability of type I error in a single comparison.

Let  $\gamma$  be the probability that  $H_0$  is incorrectly rejected in at least one test, when the test is repeated  $C$  times, i.e., the probability of making at least type I error during the  $C$  tests.

Probability theory shows that,

- if the tests are independent (which they most likely are not), then  $\gamma = 1 - (1 - \beta)^C$ .
- in the general case, we have the bound  $\gamma \leq C\beta$ .

Thus, to be absolutely sure that the probability of making at least one type I error during the  $C$  tests is at most some  $\alpha$ , the individual comparisons must be done on significance level  $\beta = \frac{\alpha}{C}$ .

# Bonferroni correction

## Bonferroni correction

That is, for each pair  $\mu_j, \mu_k$  we conduct a  $t$ -test to test for their equality and reject the null hypothesis  $H_0 : \mu_j = \mu_k$  if the corresponding  $p$ -value satisfies

- $p < \frac{\alpha}{C}$ , or equivalently
- $pC < \alpha$  (each  $p$ -value is magnified  $C$ -fold).

This is known as the *Bonferroni correction*.

## Example, continued

- The drug company wants to know exactly which of the drugs differ from each other in effectiveness.
- To keep the 5 % Type I error rate, the pairwise comparisons are done on a significance level  $\beta = \alpha/C = 0.05/6 = 0.0083$ .

Pair	$H_0$
AB	Not rejected
AC	Not rejected
AD	Rejected
BC	Rejected
BD	Not rejected
CD	Rejected

- Drug D is statistically significantly different from drugs A and C and furthermore drug B is statistically significantly different from drug C.

## Assumptions of ANOVA

ANOVA makes two key assumptions:

- ① The groups are **normally distributed**.
- ② The groups have **equal variances**.

As usual, the first of the assumptions can (by central limit theorem) be replaced with a large enough sample size  $n$ .

The second one is required also for large samples. However, **ANOVA is robust to moderate violations from it**. As a rule of thumb, the largest group variance should be at most 4 times the smallest group variance.

The variance assumption can also be tested using *Bartlett's test*.

## Bartlett's test for equality of variances

### Bartlett's test, assumptions

Let  $x_{1j}, x_{2j}, \dots, x_{nj}$  be i.i.d. observed values of a  $\mathcal{N}(\mu_j, \sigma_j^2)$ -distributed random variable  $x_j$ ,  $j = 1, \dots, k$ . Assume that the  $k$  samples are independent.

### Bartlett's test, hypotheses

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2.$$

$$H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ for some } i \neq j.$$

## Bartlett's test for equality of variances

To conduct Bartlett's test for equality of variances we calculate,

- ① The individual variance estimates

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2,$$

- ② The pooled variance estimate

$$s^2 = \frac{1}{n - k} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2,$$

- ③ and the components of the test statistic,

$$Q = (n - k) \ln s^2 - \sum_{j=1}^k (n_j - 1) \ln s_j^2$$
$$h = 1 + \frac{1}{3(k-1)} \left( \left( \sum_{j=1}^k \frac{1}{n_j - 1} \right) - \frac{1}{n - k} \right).$$

# Bartlett's test for equality of variances

## Bartlett's test, test statistic

- Bartlett's test statistic,

$$B = \frac{Q}{h},$$

follows, for large  $n$ , under  $H_0$  approximately the  $\chi^2$ -distribution with  $k - 1$  degrees of freedom.

- The expected value of the test statistic under  $H_0$  is approximately  $k - 1$  and **large** values of the test statistic suggest that the null hypothesis  $H_0$  is false.

## ANOVA and linear regression

ANOVA is closely related to multiple linear regression.

In fact, it is equivalent to regressing the  $x$ -variable on the indicator variables of the groups,

$$x_{ij} = \beta_0 + \beta_1 I(j=1) + \beta_2 I(j=2) + \cdots + \beta_{k-1} I(j=k-1) + \epsilon_{ij},$$

where e.g.  $I(j=1) = 1$  if  $j = 1$  and  $I(j=1) = 0$  otherwise, and the independent errors  $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ .

The above model includes only  $k - 1$  indicators as adding the  $k$ th would make the model parameters *unidentifiable* (no unique solution).

## ANOVA and linear regression

Parameter interpretation:

- $\beta_0$  is the expected value  $\mu_k$  of the  $k$ th group (*the reference group*).
- $\beta_\ell$ ,  $\ell = 1, \dots, k - 1$  are the differences  $\mu_\ell - \mu_k$  between the expected values of the other groups and the group  $k$ .

Statistical software usually sets the first or last group as the reference group.

The ANOVA test on slide 8 is equivalent to testing the null hypothesis  $\beta_1 = \dots = \beta_{k-1} = 0$  in the regression formulation for ANOVA.

Similar tests for testing simultaneously whether multiple regression coefficient are zero exist also for standard linear regression. However, they are out of our scope.

# Contents

1 Analysis of Variance

2 Kruskal-Wallis test

## Kruskal-Wallis test

Kruskal-Wallis test is a non-parametric alternative to the analysis of variance. That is, it avoids the need for the normality assumption.

It is a generalization of the two-sample rank test/Wilcoxon rank sum test to more than two groups.

Kruskal-Wallis test tests the null hypothesis that *k* independent samples all come from the same distribution.

## Kruskal-Wallis test

### Kruskal-Wallis test, assumptions

- Let  $x_{1j}, x_{2j}, \dots, x_{nj}$  be an i.i.d random sample from the distribution  $F_j$ ,  $j = 1, \dots, k$ , and let the  $k$  samples be independent.
- Assume further that the groups distributions  $F_1, \dots, F_k$  are equal up to location shifts (i.e. the distribution have the same “shape” but possibly different medians/locations) and denote the distributions’ medians by  $m_j$ ,  $j = 1, \dots, k$ .

### Kruskal-Wallis test, hypotheses

$$H_0 : m_1 = m_2 = \dots = m_k.$$

$$H_1 : m_j \neq m_k \text{ for some } j, k.$$

## Kruskal-Wallis test

To compute the Kruskal-Wallis test,

- ① Combine the groups  $x_{1j}, x_{2j}, \dots, x_{nj}$ ,  $j = 1, \dots, k$ , into one larger sample  $z_1, z_2, \dots, z_n$ , where  $n = \sum_{j=1}^k n_j$ .
- ② Order the observations  $z_s$  from the smallest to the largest and let  $R(z_s)$  be the rank of the observation  $z_s$  in the combined sample  $z_1, z_2, \dots, z_n$ .
- ③ Calculate the group means of the ranks

$$\bar{r}_j = \frac{1}{n_j} \sum_{\substack{s=1 \\ z_s=x_{ij}, i=1}}^{n_j} R(z_s)$$

and the mean rank of the combined sample

$$\bar{r} = \frac{1}{n} \sum_{s=1}^n R(z_s).$$

## Kruskal-Wallis test

- ④ Compute the group sum of squares, which describes the variance of the ranks between the groups

$$\sum_{j=1}^k n_j(\bar{r}_j - \bar{r})^2,$$

- ⑤ and the total sum of squares, which describes the variance of the ranks in the combined sample

$$\sum_{s=1}^n(R(z_s) - \bar{r})^2.$$

# Kruskal-Wallis test

## Kruskal-Wallis test, assumptions

- Kruskal-Wallis test statistic,

$$K = (n - 1) \frac{\sum_{j=1}^k n_j (\bar{r}_j - \bar{r})^2}{\sum_{s=1}^n (R(z_s) - \bar{r})^2},$$

follows, for large  $n$ , under  $H_0$  approximately the  $\chi_{k-1}^2$ -distribution.

- Under  $H_0$ , the expected value of the test statistic is approximately  $k - 1$  and **large** values of the test statistic suggest that the null hypothesis  $H_0$  is false.

## Kruskal-Wallis test, some notes

- Statistical software can often calculate exact  $p$ -values of the Kruskal-Wallis test when the sample size is small.
- With large sample sizes, calculation of the exact  $p$ -values would require large amounts of computing and in these cases the asymptotic  $p$ -values (based on the above-mentioned  $\chi^2$ -distribution) are used.
- We assumed above that the observations follow a continuous distribution. However, Kruskal-Wallis test can be used for discrete observations as well. Then it is possible that some of the observations have the same rank. In that case, all those observations are assigned to have the median of the corresponding ranks.
- For example, if two observations have the same rank corresponding to ranks 7 and 8, then both are assigned to have rank 7.5. If three observations have the same ranks corresponding to ranks 3, 4, and 5, then each is assigned to have rank 4.

## Numerical example

Consider three student groups and their statistics exam scores. The table below displays the scores and the corresponding ranks (in parenthesis).

Group 1	Group 2	Group 3
18.0 (14)	16.5 (11)	23 (22)
11.0 (4.5)	10.0 (3)	22 (20)
17.0 (12)	15.0 (8.5)	23 (22)
14.0 (7)	15.0 (8.5)	24 (24)
11.0 (4.5)	20.5 (17)	21 (18)
9.5 (2)	8.0 (1)	21.5 (19)
16.0 (10)	12.0 (6)	23 (22)
		20.0 (16)
		17.5 (13)
		19.0 (15)

## Numerical example

- Calculate the rank means within the groups

$$\bar{r}_1 = \frac{1}{7}(14 + 4.5 + 12 + 7 + 4.5 + 2 + 10) = \frac{54}{7} = 7.714286,$$

$$\bar{r}_2 = \frac{1}{7}(11 + 3 + 8.5 + 8.5 + 17 + 1 + 6) = \frac{55}{7} = 7.857143,$$

$$\bar{r}_3 = \frac{1}{10}(22 + 20 + 22 + 24 + 18 + 19 + 22 + 16 + 13 + 15) = \frac{191}{10} = 19.1,$$

- and the mean rank of the combined sample

$$\bar{r} = \frac{1}{24}(54 + 55 + 191) = \frac{300}{24} = 12.5.$$

## Numerical example

- Calculate the group sum of squares

$$\sum_{j=1}^k n_j (\bar{r}_j - \bar{r})^2 = 7 * (7.714286 - 12.5)^2 + 7 * (7.857143 - 12.5)^2 \\ + 10 * (19.1 - 12.5)^2 = 746.8143,$$

- and the total sum of squares

$$\sum_{s=1}^n (R(z_s) - \bar{r})^2 \\ = (14 - 12.5)^2 + (4.5 - 12.5)^2 + (12 - 12.5)^2 + (7 - 12.5)^2 \\ + (4.5 - 12.5)^2 + (2 - 12.5)^2 + (10 - 12.5)^2 + (11 - 12.5)^2 \\ + (3 - 12.5)^2 + (8.5 - 12.5)^2 + (8.5 - 12.5)^2 \\ + (17 - 12.5)^2 + (1 - 12.5)^2 + (6 - 12.5)^2 \\ + (22 - 12.5)^2 + (20 - 12.5)^2 + (22 - 12.5)^2 + (24 - 12.5)^2 + (18 - 12.5)^2 \\ + (19 - 12.5)^2 + (22 - 12.5)^2 + (16 - 12.5)^2 + (13 - 12.5)^2 + (15 - 12.5)^2 \\ = 1147$$

## Numerical example

- Now

$$K = (n - 1) \frac{\sum_{j=1}^k n_j (\bar{r}_j - \bar{r})^2}{\sum_{s=1}^n (R(z_s) - \bar{r})^2} = (24 - 1) \frac{746.8143}{1147} = 14.97535.$$

- $p$ -value of the test is 0.00056 so the null hypothesis of equal medians is rejected.
- We conclude that there is a statistically significant difference in the exam scores between the groups.

## Bonferroni's method for pairwise comparison of the medians

- If the null hypothesis of the Kruskal-Wallis test is rejected, then the analysis can be continued by finding the groups with statistically significant differences in medians.
- A simple idea is to apply the two-sample rank test for pairwise comparisons, with a total of  $C = \frac{k(k-1)}{2}$  pairs to compare.
- If the combined comparison is to be done on significance level  $\alpha$ , then the pairwise comparisons should be done on significance level  $\beta = \frac{\alpha}{C}$  (as in ANOVA).
- For example, if significance level 0.05 is desired for the combined comparison, then the pairwise comparisons reject the corresponding null hypotheses if the  $p$ -value is smaller than  $\frac{0.05}{C}$ .

# MS-C1620 Statistical inference

## 12 Reflection

Jukka Kohonen

Department of Mathematics and Systems Analysis  
School of Science  
Aalto University

Academic year 2020–2021  
Period III–IV

## Reflection session

The aim of this session is to reflect on Lectures 1–11 of the course.

You will do this in groups. Group  $i$  reflects on Lecture  $i$  ( $1 \leq i \leq 11$ ).

Timeline:

- |                   |  |
|-------------------|--|
| 8:30 – 9:00       | <b>Discuss</b> your lecture in breakout rooms (instructions on next slide). <b>Collect</b> your findings on the Miro board <a href="https://bit.ly/3sYQ10k">https://bit.ly/3sYQ10k</a> |
| 9:00 – 9:15       | Take a break   |
| 9:15 – 10:00      | Each group <b>presents</b> their findings to others, about 3+1 minutes each. Leave 1 minute for questions and comments.  |
| (if time remains) | Free Q&A   |

## Instructions

One student in your group can share the screen and open the lecture slides. Then browse through the slides, discuss and reflect. Use Miro board: <https://bit.ly/3sYQ10k> (navigate to “your” lecture)

Consider the following questions, or others that you find more relevant.

- ① **What is the key idea of the lecture?** Its title is about 2–5 words. Can you expand it to 10–20 words? Can you explain the idea in 50 words to your fellow student?
- ② What is the **one most important slide** of the lecture?
- ③ How does it compare to the **other lectures** or your other studies?
- ④ **What kind of mathematics** is involved? Probability? Computation?
- ⑤ In your field of study, when would you **use these methods**? What problems do you see in applying them in practice?
- ⑥ What was new/familiar/surprising? What was easy/difficult? What further questions do you have? Feel free to express your questions!