Introduction to Statistical Inference
Introduction

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Virtanen/Ailus

Contents

Practical Things

This Course

This course is an introduction to statistical analysis and statistical inference. Course topics include estimation, simple parametric and nonparametric tests, statistical dependence and correlation, linear regression analysis and analysis of variance. Software R is used in the exercises of this course.

Lectures

- Lecturer: Pauliina Ilmonen, pauliina.ilmonen(a)aalto.fi
- Lectures: Thursdays 12.15-14.00 Hall C.

Lectures

Lectures are an important part of this course. Attendance is not compulsory, but highly recommended. If you are unable to attend the lectures, you are expected to ask for notes from the other students.

Exercises

Exercises are another very important part of this course. Attendance is not compulsory, but again highly recommended. You do get points by attending the exercises and by doing your homework assignments. There are several exercise groups. Please attend one of them. If none of the exercise group times is suitable for you, but you would still like to get exercise points, you may contact the course head assistant (matias.heikkila(a)aalto.fi). Note that you need a super good reason for getting points without attendance!

Exercises

Students should be present and ready with their computers turned on at the beginning of the exercise session. Arriving late is not allowed. The first problem of each week's exercise session is a homework assignment. The homework assignment should be solved at home before the start of the session, using only pen, paper, and if necessary, a pocket calculator (no computers). Active participation is required at the exercise sessions. Students can obtain points from homework assignments and participation as follows: Active attendance plus homework assignment –> 0.5 points. Active attendance only –> 0.25 p. (As total one can obtain max 12*0.5=6 p.) Note that one can not get homework points without attending the corresponding exercise session! Note also that one can get homework assignment points even if the answer is not correct/complete. Trying your very best is enough! The exercise points are valid until the end of 2017.

How to pass this course?

You are expected to

- Attend the lectures and be active - not compulsory, no points, but highly recommended. The first lecture is on Thursday 5.1.
- Participate to exercises and solve your homework problems - not compulsory, but highly recommended - max 6 points.
- Take the exam - max 24 points.

Max total points = 24 + 6 = 30: You need at least 15 points (or 12 points from the exam) in order to pass the course.

How to get a good grade?

- Attend the lectures and be active!
- Work hard on your homework assignments.
- Be active in the exercises!
- Study for the exam!

Grading is based on the total points as follows: at least 15p (or 12p from the exam) -> 1, at least 16p -> 2, at least 19p -> 3, at least 22p -> 4, at least 25p -> 5. Please note that you can not get grade 5 without attending the exercises. The reason for this is that learning to use R, and learning to conduct statistical analysis in practice, are crucial parts of the course.

Exam

- In the exam, the focus is on the lecture material and on the lecture discussions.
- In the exam you may have your pens and pencils, a ruler and an eraser. On top of that you may have one A4 of notes. The rules for the note are: size A4, text on one side only, it must be hand-written, your name has to be on the top right corner of the note. Please take an id-card with you to the exam. The exam organizers do not know you.
- The first exam is on 05.04.2017 from 13.00 to 16.00 and the second one is on 23.05.2017 from 09.00 to 12.00.

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Virtanen/Ailus

Introduction

Statistical research

Statistical research is collecting, organizing, analysing and interpreting data.

Statistical models are mathematical and are based on probability theory.

Research in statistics

Research in statistics is proposing and developing new methodology and studying the existing methods. The goal in research in statistics is to develop methods for statistical research.

Statistical research

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

Population and sample

- In statistical analysis, a population is a collection of all the people, items, or events about which one wants to make inferences. (For example university students in Finland.)
- In statistical analysis, a sample, is a subset of the population (i.e. the people, items, or events) that one collects and analyzes to make inferences. (For example 200 randomly chosen university students.)
- In statistical analysis, an observation is an elements of the sample. (For example Helena, a student at Aalto University.)

Variables and data

In Statistical research, data consists of the values of selected variables that describe the observations. The data points (the values of the selected variables) can also be called observations.

Examples:

- temperature, height, blood pressure (continuous quantitative variables)
- gender, eye colour (categorical qualitative variables)
- clothing size (s, m, l), (ordinal quantitative variable)

Statistical research projects

Statistical research projects can usually be conducted in the following steps:

1. Setting of the research topic and the relevant research questions. Research questions should be defined precisely.
2. Defining of the population and interesting variables.
3. Planning of the sample collection. Collected sample must represent the population!
4. Collection of the sample.
5. Organization of the sample.
6. Description of the variables and the sample, descriptive statistics and visualization.
7. Inference based on statistical analysis. Model assumptions have to be tested separately!
8. Critical evaluation of the analysis. Possible errors and weaknesses have to be reported.
9. Communication of the research and findings.

Think about all the steps in statistical research. In each step — What can go wrong?

Different statistical studies

Statistical research projects can be conducted in several different ways. Research questions, population, goals and resources all have an effect on the choice of the methods.

Different statistical studies

- Observational research (In observational research, observations are made without changing any existing conditions. For example, temperature is measured or the lung cancer risk of smokers is compared to the lung cancer risk on non-smokers.)
- Controlled experiments (In controlled experiments, the effect of one variable to another is examined by controlling existing conditions. For example, the effect of allergy medicine is compared to the effect of placebo by randomizing patients to two groups.)

Different statistical studies

- Simulations (In simulations, mathematical modeling is used in order to mimic natural conditions or processes. For example, the spread of the Ebola virus is predicted by applying computer simulations, or the safety of a new car model is tested using crash test dummies.)
- Survey (In surveys, the goal is to find a representative sample of the population and get answers to some particular questions. For example, opinion polls are used in order to predict election results, or health related questionnaires are used in assessing the health of the university students.)

Problems in statistics

Can one trust statistical analyses? Why do my results look weird?

Some probability theory

Definition, A sample space

The fundamental object in probability theory is a nonempty sample space Ω.

An event is a subset $A \subseteq \Omega$.

Definition, a σ-algebra

A family of subsets of Ω, denoted by F, is a σ-algebra on Ω if

1. $\emptyset \in F$.
2. If $A \in F$, then $A^c \in F$.
3. If $A_1, A_2, .., A_i, ... \in F$, then $\cup_i A_i \in F$.

Definition, a probability measure

Let Ω be nonempty, and let F be a σ-algebra on Ω. The mapping $P : F \to [0, 1]$ is a probability measure, if

1. $P(A) \in [0, 1]$, for all $A \in F$,
2. $P(\Omega) = 1$
3. For all $A_1, A_2, .., A_i, ... \in F$, with $A_j \cap A_i = \emptyset, i \neq j$, it holds that $P(\cup_i A_i) = \sum_i P(A_i)$.

Corollaries

- $P(A) = 1 - P(A^c)$.
- $P(B \cup C) = P(B) + P(C) - P(B \cap C)$.

Example

Consider rolling two dice. The corresponding sample space $\Omega = \{(1,1),(1,2),...,(6,6)\}$.

The event "both dice > 2" is
$A = \{\omega = (\omega_1, \omega_2) \in \Omega \mid \omega_1 > 2, \omega_2 > 2\}$.

In this example, $P(\{\omega\}) = 1/36$, for all $\omega \in \Omega$.

Definition, conditional probability

Let $P(B) \neq 0$. The probability of an event A given B, $P(A|B)$, is the probability of A under the assumption that B has already occurred. Conditional probability, A given B is

$$P(A|B) = \frac{P(B \cap A)}{P(B)}.$$

Example

The probability of getting 3 (event A) when rolling the first dice, given that the other dice gave 4 (event B):
$P(A|B) = P(B \cap A)/P(B) = (1/36)/(6 * 1/36) = 1/6.$

Definition, Independence

The events $A_1, \ldots, A_n$ are independent if for all $1 \leq i_1 < i_2 < \cdots < i_k \leq n$

$$P(A_{i_1} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k}).$$

Example

In the dice example $P(A \cap B) = 1/36$ and on the other hand $P(A)P(B) = 1/36$, for all A, B, $(A, B) \in \Omega$.

Definition, random variable

A real-valued random variable X is a mapping from the sample spare to the real line, i.e. $X = X(\omega) : \Omega \to \mathbb{R}$.

More precisely: Let Ω be nonempty and let F be a σ-algebra on Ω. Let $X = X(\omega) : \Omega \to \mathbb{R}$ be function. If $\{\omega | X(\omega) \leq r\} \in F$ for all $r \in \mathbb{R}$ (i.e. X is F measurable), then X is a random variable.

Example, two dice

As an example of a random variable, consider the sum:
$X : \{(1,1), ..., (6,6)\} \to \{2, ..., 12\}$, $X(\omega) = \omega_1 + \omega_2$.

Note, however, that the identity function $Y(\omega_1, \omega_2) = (\omega_1, \omega_2)$ also defines a random variable. Since $Y : \Omega \to \mathbb{R}^2$, this random variable is vector valued.

Definition, probability function

The probability function of a random variable X, denoted P_X, is

$$P_X(A) = P(\{\omega : X(\omega) \in A\}).$$

Definition, cumulative distribution function

The cumulative distribution function (cdf) of a random variable X, denoted F_X, is

$$F_X(x) = P(\{\omega \in \Omega : X(\omega) \leq x\}) \text{ (or shortly } = P_X(X \leq x)).$$

Random variable

Usually, in practice, ω is not observed directly and analysis is based on the observed random variable $X(\omega)$. Thus statistical analysis is based on the measure P_X, not on P.

Definition, density function and probability mass function

The probability density function (pdf) $f_X(x)$ of a continuous random variable X is the derivative of its cumulative distribution function

$$f_X(x) = \frac{d}{dx} F_X(x).$$

(Note that the density function does not always exist.)

In the case of a discrete random variable X, the analogue of a probability density function is a probability mass function (pmf)

$$p_X(x) = P(X = x),$$

which corresponds to the probability of the event $X = x$.

Random variables are often defined by giving their cumulative distribution functions and/or density functions.

Examples

- discrete X: for example Binomial or Poisson distribution
- continuous X: for example uniform, normal or exponential distribution

Multivariate distributions

Very often, instead of dealing with one random variable X only, we are interested in several random variables $X_1, ..., X_k$.

Joint cumulative distribution function

Let $X_1, ..., X_k$ be random variables. Then the joint cumulative distribution function of $X_1, ..., X_k$ is given by

$$F_{X_1,...,X_k}(x_1, ..., x_k) = P(X_1 \leq x_1, ..., X_k \leq x_k).$$

Joint probability density function

Let $X_1, ..., X_k$ be continuous random variables. Then the joint probability density function of $X_1, ..., X_k$ (if it exists) is given by

$$f_{X_1,...,X_k}(x_1, ..., x_k) = \frac{d^n}{dx_1 \cdots dx_k} F_{X_1,...,X_k}(x_1, ..., x_k).$$

Joint probability mass function

Let $X_1, ..., X_k$ be discrete random variables. Then the joint probability mass function of $X_1, ..., X_k$ is given by

$$p_{X_1,...,X_k}(x_1, ..., x_k) = P(X_1 = x_1, ..., X_k = x_k).$$

Let $X_1, ..., X_k$ be continuous random variables. Assume that the joint probability density function $f_{X_1,...,X_k}(x_1, ..., x_k)$ exists. Then

$$P((X_1, ..., X_k) \in A) = \int_{(x_1,...,x_k)\in A} f_{X_1,...,X_k}(x_1, ..., x_k)dx_1...dx_k.$$

For discrete variables

$$P((X_1, ..., X_n) \in A) = \sum_{c, c \in C, (X_1(c),...,X_n(c)) \in A} p_{X_1,...,X_k}(x_1, ..., x_k).$$

(In the formula above C is the sample space of the random event.)

Marginal distributions

Let $Z_1, ..., Z_h$ and $Y_1, ..., Y_l$ be continuous random variables with joint probability density functions $f_{Z_1,...,Z_h}(z_1, ..., z_h)$, $f_{Y_1,....,Y_l}(y_1, ..., y_l)$ and $f_{Z_1,...,Z_h,Y_1,....,Y_l}(z_1, ..., z_h, y_1, ..., y_l)$. Then

$$f_{Z_1,....,Z_h}(z_1, ..., z_h) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{Z_1,...,Z_h,Y_1,....,Y_l}(z_1, ..., z_h, y_1, ..., y_l) dy_1 ... dy_l.$$

For discrete variables

$$p_{Z_1,....,Z_h}(z_1, ..., z_h) = \sum_{y_1<\infty,...,y_l<\infty} p_{Z_1,...,Z_h,Y_1,....,Y_l}(z_1, ..., z_h, y_1, ..., y_l).$$

Definition, independence

Let $X_1, \ldots, X_n$ be continuous random variables with probability density functions $f_{X_1}(x_1) \ldots f_{X_n}(x_n)$ and a joint probability density function $f_{X_1,\ldots,X_n}(x_1, ..., x_n)$. If

$$f_{X_1,\ldots,X_n}(x_1, ..., x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n),$$

the random variables $X_1, ..., X_n$ are independent.

Discrete random variables are independent, if

$$p_{X_1,\ldots,X_n}(x_1, ..., x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n).$$

Example, independence

Let X and Y have the joint pdf

$$f(x,y) = \begin{cases} x+y, & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & otherwise \end{cases}.$$

Are the variables X and Y independent?
Now

$$f(x) = \int_0^1 (x+y)dy = x + \frac{1}{2}, \;\; 0 < x < 1$$

and

$$f(y) = \int_0^1 (x+y)dy = y + \frac{1}{2}, \;\; 0 < y < 1.$$

If the random variables are independent, then $f(x,y) = f(x) \cdot f(y)$. Let x=1/3 and y=1/3. Now

$$f(x,y) = x + y = 1/3 + 1/3 = 2/3.$$

On the other hand,

$$f(x) \cdot f(y) = (x + 1/2) \cdot (y + 1/2) = 5/6 \cdot 5/6 = 25/36 \neq 2/3.$$

Thus X and Y are not independent.

Example, independence

Let X and Y have the joint pmf

$$p(x,y) = \begin{cases} \frac{1}{4}, & x \in \{1,2\}, y \in \{1,2\} \\ 0, & \textit{otherwise} \end{cases}.$$

Now $p(x) = \sum_{y \in \{1,2\}} p(x,y) = 1/4 + 1/4 = \frac{1}{2},\ x \in \{1,2\}$, and otherwise $p(x) = 0$, and
$p(y) = \sum_{x \in \{1,2\}} p(x,y) = 1/4 + 1/4 = \frac{1}{2},\ y \in \{1,2\}$, and otherwise $p(y) = 0$.

If $p(x, y) = p(x) \cdot p(y)$, then X and Y are independent. Now

$$p(x) \cdot p(y) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = p(x, y), \; x \in \{1, 2\}, y \in \{1, 2\}$$

and

$$p(x) \cdot p(y) = 0 = p(x, y),$$

otherwise. The random variables are independent!

Conditional distribution

Let $Z_1, ..., Z_n$ and $Y_1, ..., Y_m$ continuous random variables with joint probability density functions $f_{Z_1,...,Z_n}(z_1, ..., z_n)$, $f_{Y_1,....,Y_m}(y_1, ..., y_m)$ and $f_{Z_1,...,Z_n,Y_1,....,Y_m}(z_1, ..., z_n, y_1, ..., y_m)$. Then

$$f_{Y_1,....,Y_m|Z_1,....,Z_n}(y_1, ..., y_m|z_1, ..., z_n) = \frac{f_{Z_1,...,Z_n,Y_1,....,Y_m}(z_1, ..., z_n, y_1, ..., y_m)}{f_{Z_1,...,Z_n}(z_1, ..., z_n)},$$

for $f_{Z_1,...,Z_n}(z_1, ..., z_n) > 0$. For discrete random variables

$$p_{Y_1,....,Y_m|Z_1,....,Z_n}(y_1, ..., y_m|z_1, ..., z_n) = \frac{p_{Z_1,...,Z_n,Y_1,....,Y_m}(z_1, ..., z_n, y_1, ..., y_m)}{p_{Z_1,...,Z_n}(z_1, ..., z_n)},$$

for $p_{Z_1,...,Z_n}(z_1, ..., z_n) > 0$.

Definition, Expected value

Let X be a continuous random variable. If $\int_{-\infty}^{\infty} |h(x)| f_X(x) dx < \infty$, then the expected value of a random variable $h(X)$ is (the real number)

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) dx.$$

Let X be a discrete random variable with the domain I. If $\sum_{x \in I} |h(x)| p_X(x) < \infty$, then the expected value of $h(X)$ is

$$E[h(X)] = \sum_{x \in I} h(x) p_X(x).$$

Example

- The expected value of X, $E[X]$, is obtained by setting $h(X) = X$.
- The variance of X, $\text{var}[X]$, is obtained by setting $h(X) = E[(X - E[X])^2]$.
- The kth moment of X, $E[X^k]$, is obtained by setting $h(X) = X^k$.

Numerical example, expected value

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Virtanen/Ailus

Let X be a continuous random variable with the pdf

$$f_X(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & otherwise \end{cases}.$$

Now

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)dx = \int_0^1 x \cdot 1dx = \frac{1}{2}.$$

Let X be a discrete random variable with the pmf

$$p_X(x) = P(X = x) = \frac{1}{30} \cdot x^2, x = \{1, 2, 3, 4\}.$$

Now

$$E[X] = \sum x \cdot p_X(x) = 1 \cdot \frac{1}{30} + 2 \cdot \frac{4}{30} + 3 \cdot \frac{9}{30} + 4 \cdot \frac{16}{30} = \frac{10}{3}.$$

Theorems, formulae for expectation and variance

Let $X_1, \ldots, X_n$ be random variables with finite expectations and variances. Let $a, b \in \mathbb{R}$. Then

- $E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i]$
- $E[aX_i + b] = aE[X_i] + b$
- $\text{var}[aX_i + b] = a^2\text{var}[X_i]$

Let $X_1, \ldots, X_n$ be independent. Then

- $E[X_1 X_2 \cdots X_n] = E[X_1]E[X_2] \cdots E[X_n]$
- $\text{var}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \text{var}[X_i]$

References

- J. S. Milton, J. C. Arnold: Introduction to Probability and Statistics, McGraw-Hill Inc 1995.
- J. Crawshaw, J. Chambers: A Concise Course in Advanced Level Statistics, Nelson Thornes Ltd 2013.
- R. V. Hogg, J. W. McKean, A. T. Craig: Introduction to Mathematical Statistics, Pearson Education 2005.
- Pertti Laininen: Todennäköisyys ja sen tilastollinen soveltaminen, Otatieto 1998, numero 586.
- Ilkka Mellin: Tilastolliset menetelmät, http://math.aalto.fi/opetus/sovtoda/materiaali.html.