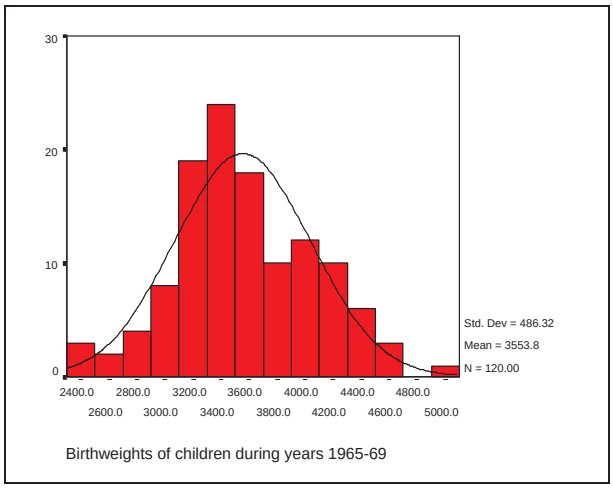
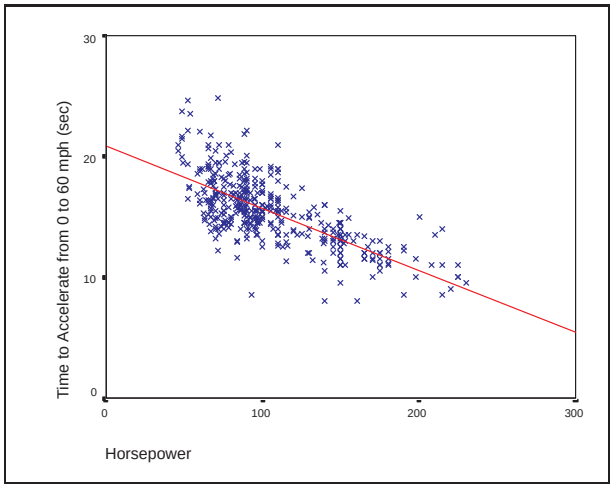
Basics of Statistics

Jarkko Isotalo



Birthweights of children during years 1965-69



Preface

These lecture notes have been used at Basics of Statistics course held in University of Tampere, Finland. These notes are heavily based on the following books.

Agresti, A. & Finlay, B., *Statistical Methods for the Social Sciences*, 3th Edition. Prentice Hall, 1997.

Anderson, T. W. & Sclove, S. L., *Introductory Statistical Analysis.* Houghton Mifflin Company, 1974.

Clarke, G.M. & Cooke, D., *A Basic course in Statistics.* Arnold, 1998.

Electronic Statistics Textbook,
http://www.statsoftinc.com/textbook/stathome.html.

Freund, J.E.,*Modern elementary statistics.* Prentice-Hall, 2001.

Johnson, R.A. & Bhattacharyya, G.K., *Statistics: Principles and Methods,* 2nd Edition. Wiley, 1992.

Leppälä, R., *Ohjeita tilastollisen tutkimuksen toteuttamiseksi SPSS for Windows -ohjelmiston avulla*, Tampereen yliopisto, Matematiikan, tilastotieteen ja filosofian laitos, B53, 2000.

Moore, D., *The Basic Practice of Statistics.* Freeman, 1997.

Moore, D. & McCabe G., *Introduction to the Practice of Statistics*, 3th Edition. Freeman, 1998.

Newbold, P., *Statistics for Business and Econometrics.* Prentice Hall, 1995.

Weiss, N.A., *Introductory Statistics.* Addison Wesley, 1999.

Please, do yourself a favor and go find originals!

1 The Nature of Statistics

[Agresti & Finlay (1997), Johnson & Bhattacharyya (1992), Weiss (1999), Anderson & Sclove (1974) and Freund (2001)]

1.1 What is statistics?

Statistics is a very broad subject, with applications in a vast number of different fields. In generally one can say that statistics is the methodology for collecting, analyzing, interpreting and drawing conclusions from information. Putting it in other words, statistics is the methodology which scientists and mathematicians have developed for interpreting and drawing conclusions from collected **data**. Everything that deals even remotely with the collection, processing, interpretation and presentation of data belongs to the domain of statistics, and so does the detailed planning of that precedes all these activities.

DEFINITION 1.1 (Statistics). *Statistics consists of a body of methods for collecting and analyzing data.* (Agresti & Finlay, 1997)

From above, it should be clear that statistics is much more than just the tabulation of numbers and the graphical presentation of these tabulated numbers. Statistics is the science of gaining information from numerical and categorical[1] data. Statistical methods can be used to find answers to the questions like:

- What kind and how much data need to be collected?
- How should we organize and summarize the data?
- How can we analyse the data and draw conclusions from it?
- How can we assess the strength of the conclusions and evaluate their uncertainty?

[1]Categorical data (or qualitative data) results from descriptions, e.g. the blood type of person, marital status or religious affiliation.

That is, statistics provides methods for

1. Design: Planning and carrying out research studies.
2. Description: Summarizing and exploring data.
3. Inference: Making predictions and generalizing about phenomena represented by the data.

Furthermore, statistics is the science of dealing with uncertain phenomenon and events. Statistics in practice is applied successfully to study the effectiveness of medical treatments, the reaction of consumers to television advertising, the attitudes of young people toward sex and marriage, and much more. It's safe to say that nowadays statistics is used in every field of science.

Example 1.1 (Statistics in practice). Consider the following problems:
–agricultural problem: Is new grain seed or fertilizer more productive?
–medical problem: What is the right amount of dosage of drug to treatment?
–political science: How accurate are the gallups and opinion polls?
–economics: What will be the unemployment rate next year?
–technical problem: How to improve quality of product?

1.2 Population and Sample

Population and sample are two basic concepts of statistics. Population can be characterized as the set of individual persons or objects in which an investigator is primarily interested during his or her research problem. Sometimes wanted measurements for all individuals in the population are obtained, but often only a set of individuals of that population are observed; such a set of individuals constitutes a sample. This gives us the following definitions of population and sample.

Definition 1.2 (Population). *Population is the collection of all individuals or items under consideration in a statistical study.* (Weiss, 1999)

Definition 1.3 (Sample). *Sample is that part of the population from which information is collected.* (Weiss, 1999)

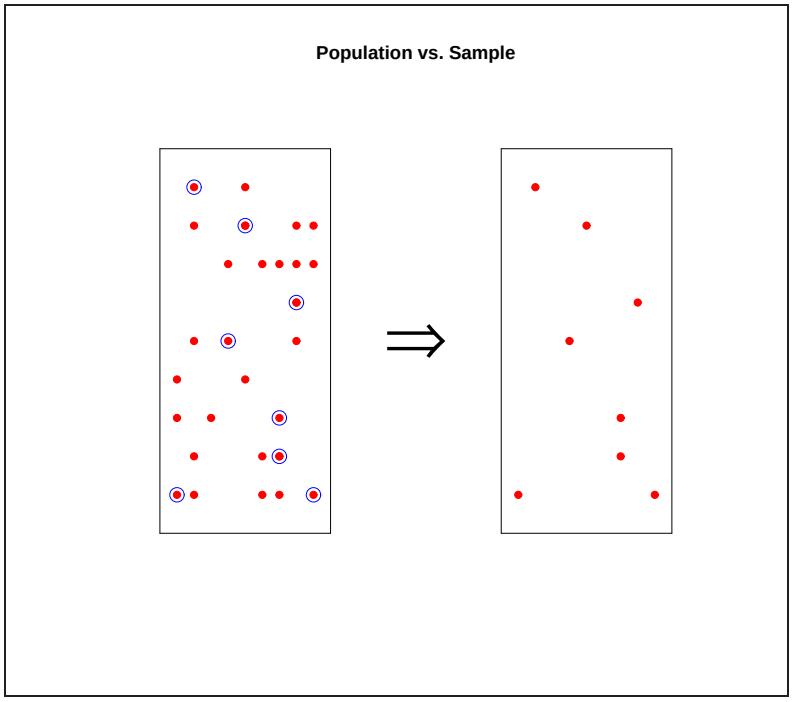


Figure 1: Population and Sample

Always only a certain, relatively few, features of individual person or object are under investigation at the same time. Not all the properties are wanted to be measured from individuals in the population. This observation emphasize the importance of a set of measurements and thus gives us alternative definitions of population and sample.

DEFINITION 1.4 (Population). *A (statistical) population is the set of measurements (or record of some qualitive trait) corresponding to the entire collection of units for which inferences are to be made.* (Johnson & Bhattacharyya, 1992)

DEFINITION 1.5 (Sample). *A sample from statistical population is the set of measurements that are actually collected in the course of an investigation.* (Johnson & Bhattacharyya, 1992)

When population and sample is defined in a way of Johnson & Bhattacharyya, then it's useful to define the source of each measurement as **sampling unit**, or simply, a **unit**.

The population always represents the target of an investigation. We learn about the population by sampling from the collection. There can be many

different populations, following examples demonstrates possible discrepancies on populations.

EXAMPLE 1.2 (Finite population). In many cases the population under consideration is one which could be physically listed. For example:
–The students of the University of Tampere,
–The books in a library.

EXAMPLE 1.3 (Hypothetical population). Also in many cases the population is much more abstract and may arise from the phenomenon under consideration. Consider e.g. a factory producing light bulbs. If the factory keeps using the same equipment, raw materials and methods of production also in future then the bulbs that will be produced in factory constitute a hypothetical population. That is, sample of light bulbs taken from current production line can be used to make inference about qualities of light bulbs produced in future.

1.3 Descriptive and Inferential Statistics

There are two major types of statistics. The branch of statistics devoted to the summarization and description of data is called *descriptive statistics* and the branch of statistics concerned with using sample data to make an inference about a population of data is called *inferential statistics*.

DEFINITION 1.6 (Descriptive Statistics). *Descriptive statistics consist of methods for organizing and summarizing information* (Weiss, 1999)

DEFINITION 1.7 (Inferential Statistics). *Inferential statistics consist of methods for drawing and measuring the reliability of conclusions about population based on information obtained from a sample of the population.* (Weiss, 1999)

Descriptive statistics includes the construction of graphs, charts, and tables, and the calculation of various descriptive measures such as averages, measures of variation, and percentiles. In fact, the most part of this course deals with descriptive statistics.

Inferential statistics includes methods like point estimation, interval estimation and hypothesis testing which are all based on probability theory.

EXAMPLE 1.4 (Descriptive and Inferential Statistics). Consider event of tossing dice. The dice is rolled 100 times and the results are forming the sample data. Descriptive statistics is used to grouping the sample data to the following table

Outcome of the roll	Frequencies in the sample data
1	10
2	20
3	18
4	16
5	11
6	25

Inferential statistics can now be used to verify whether the dice is a fair or not.

Descriptive and inferential statistics are interrelated. It is almost always necessary to use methods of descriptive statistics to organize and summarize the information obtained from a sample before methods of inferential statistics can be used to make more thorough analysis of the subject under investigation. Furthermore, the preliminary descriptive analysis of a sample often reveals features that lead to the choice of the appropriate inferential method to be later used.

Sometimes it is possible to collect the data from the whole population. In that case it is possible to perform a descriptive study on the population as well as usually on the sample. Only when an inference is made about the population based on information obtained from the sample does the study become inferential.

1.4 Parameters and Statistics

Usually the features of the population under investigation can be summarized by numerical *parameters*. Hence the research problem usually becomes as on investigation of the values of parameters. These population parameters are unknown and sample *statistics* are used to make inference about them. That is, a statistic describes a characteristic of the sample which can then be used to make inference about unknown parameters.

DEFINITION 1.8 (Parameters and Statistics). *A parameter is an unknown numerical summary of the population. A statistic is a known numerical summary of the sample which can be used to make inference about parameters.* (Agresti & Finlay, 1997)

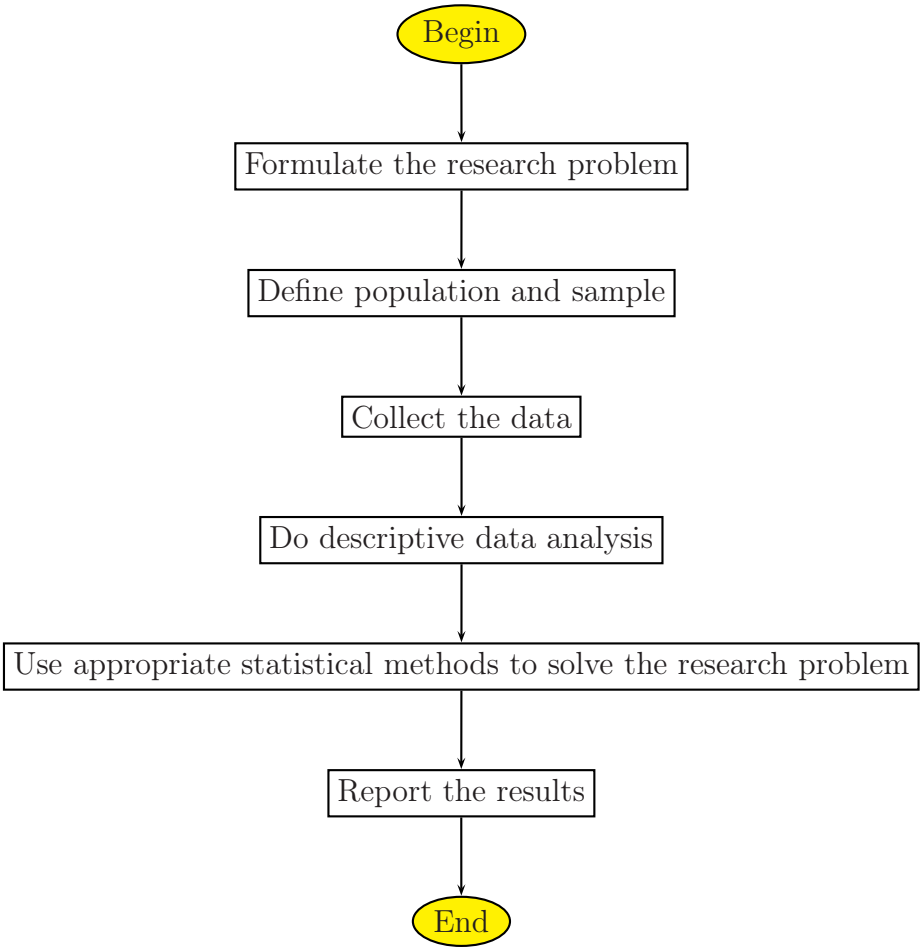
So the inference about some specific unknown parameter is based on a statistic. We use known sample statistics in making inferences about unknown population parameters. The primary focus of most research studies is the parameters of the population, not statistics calculated for the particular sample selected. The sample and statistics describing it are important only insofar as they provide information about the unknown parameters.

EXAMPLE 1.5 (Parameters and Statistics). Consider the research problem of finding out what percentage of 18-30 year-olds are going to movies at least once a month.

- Parameter: The proportion p of 18-30 year-olds going to movies at least once a month.
- Statistic: The proportion $\hat{p}$ of 18-30 year-olds going to movies at least once a month calculated from the sample of 18-30 year-olds.

1.5 Statistical data analysis

The goal of statistics is to gain understanding from data. Any data analysis should contain following steps:



To conclude this section, we can note that the major objective of statistics is to make inferences about population from an analysis of information contained in sample data. This includes assessments of the extent of uncertainty involved in these inferences.

2 Variables and organization of the data

[Weiss (1999), Anderson & Sclove (1974) and Freund (2001)]

2.1 Variables

A characteristic that varies from one person or thing to another is called a **variable**, i.e, a variable is any characteristic that varies from one individual member of the population to another. Examples of variables for humans are height, weight, number of siblings, sex, marital status, and eye color. The first three of these variables yield numerical information (yield numerical measurements) and are examples of **quantitative (or numerical) variables**, last three yield non-numerical information (yield non-numerical measurements) and are examples of **qualitative (or categorical) variables**.

Quantitative variables can be classified as either **discrete** or **continuous**.

Discrete variables. Some variables, such as the numbers of children in family, the numbers of car accident on the certain road on different days, or the numbers of students taking basics of statistics course are the results of counting and thus these are discrete variables. Typically, a discrete variable is a variable whose possible values are some or all of the ordinary counting numbers like $0, 1, 2, 3, \ldots$. As a definition, we can say that a variable is discrete if it has only a countable number of distinct possible values. That is, a variable is is discrete if it can assume only a finite numbers of values or as many values as there are integers.

Continuous variables. Quantities such as length, weight, or temperature can in principle be measured arbitrarily accurately. There is no indivible unit. Weight may be measured to the nearest gram, but it could be measured more accurately, say to the tenth of a gram. Such a variable, called continuous, is intrinsically different from a discrete variable.

2.1.1 Scales

Scales for Qualitative Variables. Besides being classified as either qualitative or quantitative, variables can be described according to the **scale** on which they are defined. The scale of the variable gives certain structure to the variable and also defines the meaning of the variable.

The categories into which a qualitative variable falls may or may not have a natural ordering. For example, occupational categories have no natural ordering. If the categories of a qualitative variable are unordered, then the qualitative variable is said to be defined on a **nominal scale**, the word nominal referring to the fact that the categories are merely names. If the categories can be put in order, the scale is called an **ordinal scale**. Based on what scale a qualitative variable is defined, the variable can be called as a nominal variable or an ordinal variable. Examples of ordinal variables are education (classified e.g. as low, high) and "strength of opinion" on some proposal (classified according to whether the individual favors the proposal, is indifferent towards it, or opposites it), and position at the end of race (first, second, etc.).

Scales for Quantitative Variables. Quantitative variables, whether discrete or continuos, are defined either on an **interval scale** or on a **ratio scale**. If one can compare the differences between measurements of the variable meaningfully, but not the ratio of the measurements, then the quantitative variable is defined on interval scale. If, on the other hand, one can compare both the differences between measurements of the variable and the ratio of the measurements meaningfully, then the quantitative variable is defined on ratio scale. In order to the ratio of the measurements being meaningful, the variable must have natural meaningful absolute zero point, i.e, a ratio scale is an interval scale with a meaningful absolute zero point. For example, temperature measured on the Certigrade system is a interval variable and the height of person is a ratio variable.

2.2 Organization of the data

Observing the values of the variables for one or more people or things yield **data**. Each individual piece of data is called an **observation** and the collection of all observations for particular variables is called a **data set** or **data matrix**. Data set are the values of variables recorded for a set of sampling units.

For ease in manipulating (recording and sorting) the values of the qualitative variable, they are often **coded** by assigning numbers to the different categories, and thus converting the categorical data to numerical data in a trivial sense. For example, marital status might be coded by letting 1,2,3, and 4 denote a person's being single, married, widowed, or divorced but still coded

data still continues to be nominal data. Coded numerical data do not share any of the properties of the numbers we deal with ordinary arithmetic. With recards to the codes for marital status, we cannot write $3 > 1$ or $2 < 4$, and we cannot write $2 - 1 = 4 - 3$ or $1 + 3 = 4$. This illustrates how important it is always check whether the mathematical treatment of statistical data is really legimatite.

Data is presented in a matrix form (data matrix). All the values of particular variable is organized to the same column; the values of variable forms the column in a data matrix. Observation, i.e. measurements collected from sampling unit, forms a row in a data matrix. Consider the situation where there are k numbers of variables and n numbers of observations (sample size is n). Then the data set should look like

$$\begin{array}{c} \\ \text{Sampling units} \end{array} \begin{array}{c} \text{Variables} \\ \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2k} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3k} \\ & \vdots & & \ddots & \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nk} \end{pmatrix} \end{array}$$

where x_{ij} is a value of the j:th variable collected from i:th observation, $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, k$.