MTMS.01.099 Mathematical Statistics

Lecture 5. Statistical inference. Population and Sample

Tõnu Kollo

Fall 2017

Introduction

An investigator wants to study a (some) problem(s).
Population – the complete set of individuals, objects or scores of interest.

- Often too large to analyse it
- It may be real or hypothetical

For example, a researcher may be interested in the relation between class size (variable 1) and academic performance (variable 2) for the population of third-grade children.
Data is needed

Sample

Usually populations are too large to examine the entire group. Therefore, a sample is selected to represent the population in a research study. The goal is to use the results obtained from the sample to help answer questions about the population.
Sample – A subset of the population.

- A sample may be classified as random (each member has equal chance of being selected from a population) or convenience (what's available).
- Random selection attempts to ensure the sample is representative of the population.

- We have problem and data. What to do?
- **Sampling and experimentation:** Clarify the question, decide on methods of collection and analysing data to produce valid information.
- **Exploring data:** Using graphical and numerical techniques to study patterns and departures from patterns in data;
- **Describe probabilistic behaviour** Exploring random phenomena using probability and simulation to find out the proper distribution.
- **Statistical Inference:** Finding estimates of interest and testing hypothesis.

Statistical Investigation

A *statistical investigation* usually consists of four parts – planning, collection of data, analysis and presentation:

- **Planning** consists of all sorts of preparations.
- **Collection of data** is a general term that may mean, for example, an interviewer asking people about their opinions.
- The **analysis** can assume very different forms. In simple cases it may consist only of condensation of data in a table or a diagram. For this purpose we use *descriptive statistics*. For a more detailed analysis of sampling investigations we use some form of *statistical analysis*.
- The **presentation** may consist of graphical illustrations, a summary of results,conclusions and practical recommendations.

Descriptive Statistics

Descriptive statistics includes methods for organizing and summarizing data.

For example:

- tables or graphs are used to organize data,
- quantities and techniques are used to describe a sample by characteristics to illustrate the sample data e.g. mean, standard deviation.

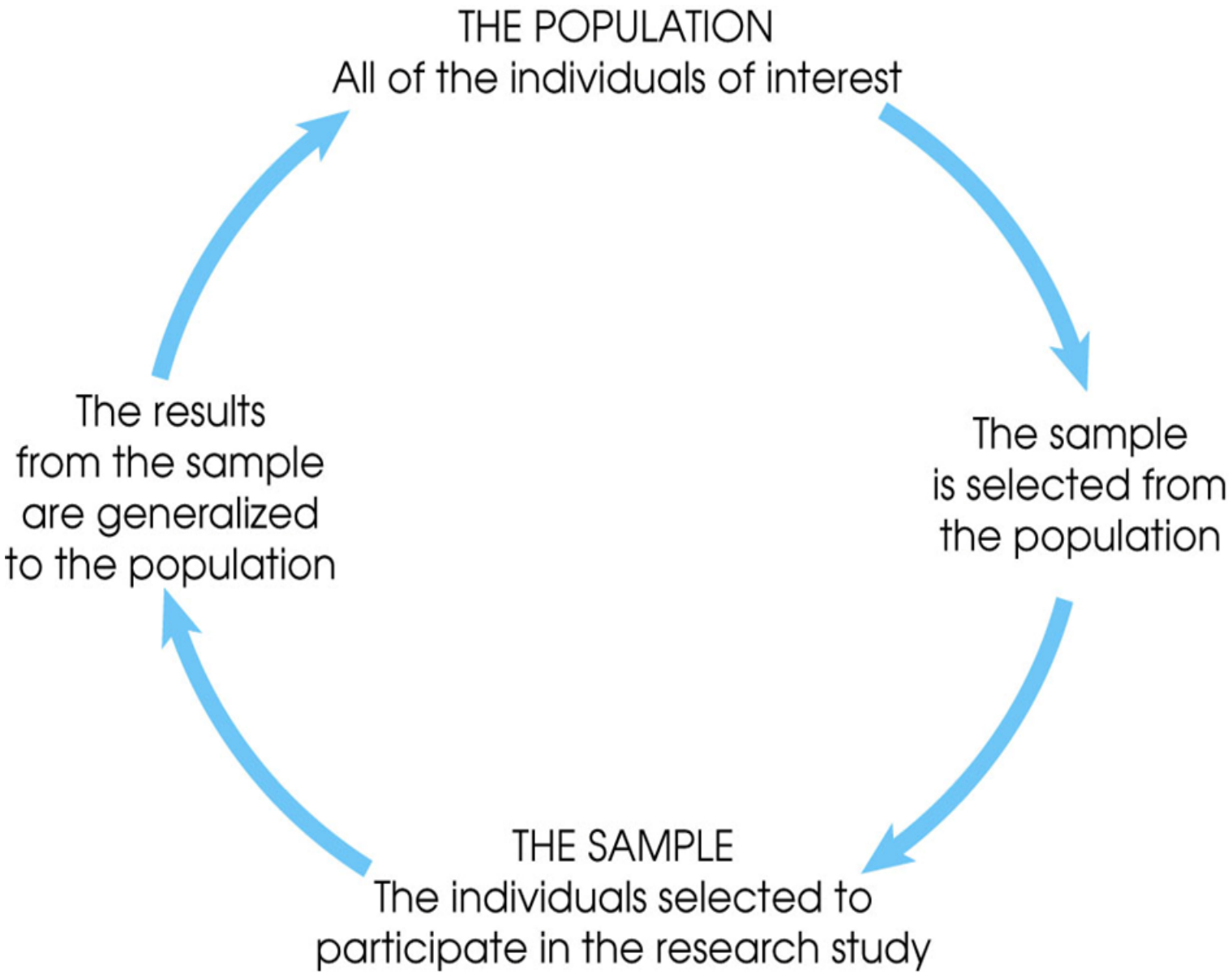
Statistical Inference

Statistical Inference – the process of drawing conclusions about a **population** based on information in a **sample**.
Statistical inference can be divided into three areas: point estimation, interval estimation (confidence intervals) and hypothesis testing.

How this would look like?

„ In hospital a new drug was 10% more effective for reduction in blood pressure than the previous one. The patients were people with high blood pressure (over 160). We can make conclusions for such people but not for all people."

Population and Sample



Describing a population

We assume that population is described by random variables.

- Take for simplicity one variable X.
- Distribution of X has parameters, say $\theta_1, \ldots, \theta_k$
- We have n objects in our sample.
- We measure (observe) variable X on these n sample objects and get values $x_1, x_2, \ldots, x_n$. We call these values 'sample' or 'concrete sample'.

Describing a sample

How to describe a sample probabilistically?

- A sample has to be representative.
- Two requirements: independence of objects, equal possibility for objects to be selected into the sample.
- Assume that all selected objects have the same distribution.
- Mathematically this means, that we have random variables $X_1, X_2, \ldots, X_n$, $X_i \sim X$, variables X_i are independent
- The set $X_1, X_2, \ldots, X_n$ is called *theoretical or random sample*

Describing a population

EXAMPLE
Let X be normally distributed, $N(\mu, \sigma^2)$. The population mean μ and the population standard deviation σ are never known exactly.

- We take a sample
- We find estimates of parameters μ and σ^2 from the sample
- Sample characteristics $\overline{x}$ and s^2 are **estimates** of population characteristics μ and σ^2

Variables. Types of Data

Variables of interest are measured in a sample. They may be classified as:

- **Quantitative** i.e. numerical, represent counts or measurements.
  - *Continuous* (e.g. weight of a person, patient's cholesterol levels)
  - *Discrete* (e.g. number of children in a family)
- **Categorical**
  - *Nominal* (e.g. favourate color, blood group)
  - *Ordinal* (ranked e.g. mild, moderate or severe illness; level of education etc.).

The sample mean and variance

Let $x_1, x_2, \ldots, x_n$ be the realised values of a random variable X, from a sample of size n. The **sample mean** is defined as

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

The **sample variance**, s^2, is the following sum of the squared deviations from the sample mean:

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2$$

The **standard deviation of a sample**, s:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2}$$

Standard Error

Since s is an estimate of σ,
an estimate of $\frac{\sigma}{\sqrt{n}}$ is $SE_{\overline{X}} = \frac{s}{\sqrt{n}}$ – this is known as the **standard error** of the sample mean.
Be careful not to confuse the standard deviation and the standard error!

- Standard deviation describes the variability of the data
- Standard error is the measure of the precision of the sample mean.

Sample

Definition

Let $X \sim F$ describe population. A **sample** $x_1, \ldots, x_n$ from F consists of observations on independent random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$, each with distribution F.

Let F be the distribution of a rv X. The distribution depends on the unknown parameters $\theta_1, \ldots \theta_k$,

$$F = F(\theta_1, \ldots, \theta_k).$$

For simplicity consider one parameter θ, $F = F(\theta)$. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be the sample from F and we use it to estimate θ.

Remember that the parameter θ is an abstract quantity which often has a concrete interpretation in the real world.

Statistic, Estimator

A function g of theoretical (random) sample $g = g(X_1, \ldots, X_n) = g(\boldsymbol{X})$ is called **statistic**. If values of a statistic can be considered as estimates of θ, the function is called an **estimator of** θ and denoted by $\hat{\theta}(\boldsymbol{X})$.

Definition

Given a sample of realized observations, the number $\hat{\theta}(\boldsymbol{x})$ is called a **point estimate** of θ.

Definition

A **point estimator** of θ is a function $\hat{\theta}(\boldsymbol{X})$.

In statistical theory, a point estimate $\hat{\theta}(\boldsymbol{x})$ is regarded as a value of a random variable $\hat{\theta}(\boldsymbol{X})$.
It is important to distinguish between $\hat{\theta}(\boldsymbol{x})$, which is a *numerical value* computed from the sample, and $\hat{\theta}(\boldsymbol{X})$ which is a *random variable*.

The distribution of an estimator

The properties of a point estimator are described with its distribution Finding the distribution of a point estimator is an important and sometimes difficult task. It can be solved in different ways.

1. *Analytic method.* The distribution is derived by means of probability theory, exactly or approximately.

2. *Simulations.* The sampling procedure is repeated a large number of times and an approximation to the distribution is then obtained by tabulating all values of $\hat{\theta}(\boldsymbol{X})$ in the spirit of descriptive statistics.

Point estimation: properties

Let θ be a unknown population parameter, for example, the mean of population and there are 3 possible estimates:

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \hat{\theta}_2 = \frac{x_1 + x_n}{2} \quad \hat{\theta}_3 = \frac{x_{(1)} + x_{(n)}}{2}$$

Which estimate one should use?
Which of these estimates are good estimates?

Point estimation: properties (2)

Definition

A point estimate $\hat{\theta}(\boldsymbol{x})$ is said to be **unbiased** if the corresponding estimator has expectation θ, that is, for each $\theta \in A$

$$E[\hat{\theta}(\boldsymbol{X})] = \theta,$$

where A is a set of possible parameter values.

If the expectation is different from θ the estimate is said to be **biased**. Bias is denoted by B and is computed by $B = E\hat{\boldsymbol{\theta}} - \theta$.

Point estimation: properties (3)

Definition

If two estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased and

$$Var[\hat{\theta}_1(\boldsymbol{X})] \leq Var[\hat{\theta}_2(\boldsymbol{X})]$$

for all $\theta \in A$ (with strict inequality for some $\theta \in A$), then $\hat{\theta}_1$ is said to be **more efficient** than $\hat{\theta}_2$.

Definition

If, for any fixed $\theta \in A$ and for any given $\varepsilon > 0$,

$$P(|\hat{\theta}(\boldsymbol{X}) - \theta| > \varepsilon) \to 0$$

as the sample size n goes to infinity, then the point estimate $\hat{\theta}(\boldsymbol{x})$ is said to be **consistent.**

Point estimation: properties (4)

Definition

The Mean Square Error (MSE) of an estimator $\hat{\theta}(\boldsymbol{X})$ for estimating θ is

$$MSE(\hat{\theta}) = E(\hat{\theta}(\boldsymbol{X}) - \theta)^2 = Var(\hat{\theta}(\boldsymbol{X})) + B^2$$

Theorem

$\hat{\theta}$ is a consistent iff

$$(1) \qquad lim_{n\to\infty} E(\hat{\theta}(\boldsymbol{X})) = \theta$$

$$(2) \qquad lim_{n\to\infty} Var(\hat{\theta}(\boldsymbol{X})) = 0.$$

Usually the fact stated above is taken as a definition of consistency.

Example continues

Let $X \sim N(\theta, \sigma^2)$. We had

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1), \hat{\theta}_2 = \frac{x_1 + x_n}{2} \quad (2).$$

$$E(\hat{\theta}_1(\boldsymbol{X})) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n} nE(X) = \theta.$$

Also $E(\hat{\theta}_2(\boldsymbol{X})) = \theta$. Both estimates (1) and (2) are unbiased. But

$$Var(\hat{\theta}_1(\boldsymbol{X})) = \frac{1}{n}\sigma^2, \quad Var(\hat{\theta}_2(\boldsymbol{X})) = \frac{1}{2}\sigma^2.$$

Estimate (1) is more efficient than (2).