TRAINING COURSES ON APPLIED ECONOMETRIC ANALYSIS (SUMMER SCHOOL) FOR YOUNG ECONOMISTS / RESEARCHERS 2016-2017

Basics of Probability theory

Farrukh Ataev

15 September 2016

Outline

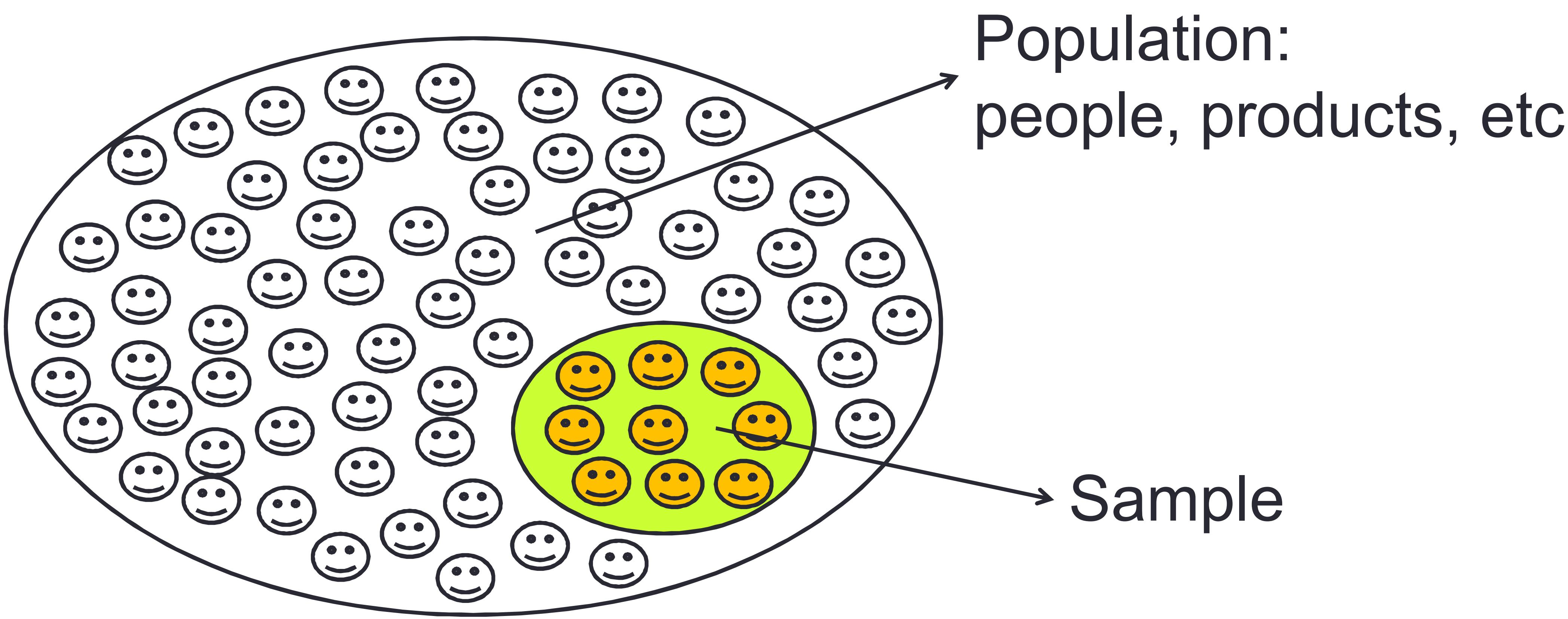
- **Session 3 Sets and Basic probability concepts**
  - (Wed 14 Sept at 13:30-15:00)
- **Session 4 Probability distributions (Discrete)**
  - (Wed 14 Sept at 15:30-17:00)
- **Session 1 Probability distributions (Continuous)**
  - (Thu 15 Sept at 9:00-10:30)
- **Session 2 Inferential statistics and Confidence intervals**
  - (Thu 15 Sept at 11:00-12:30)

Session 2 Statistical inference and Confidence intervals

- Statistics
  - Descriptive and Inferential
- Sampling methods
- Central limit theorem
- Confidence intervals

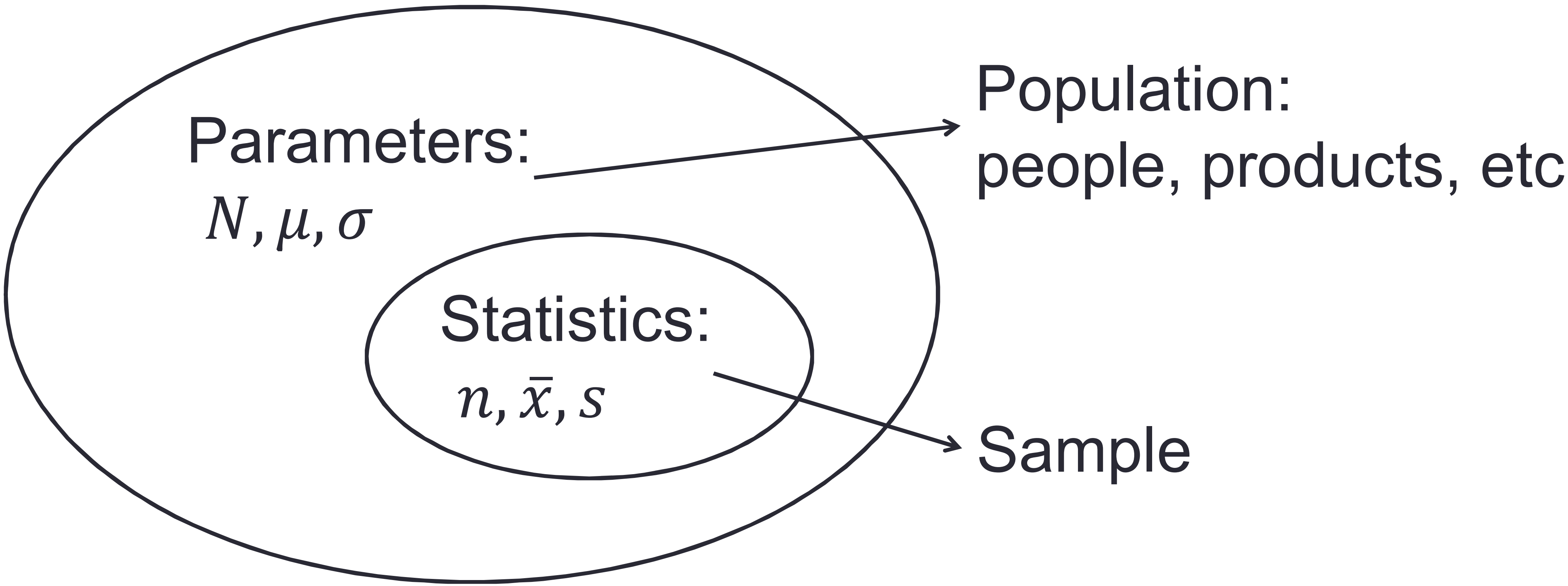
Statistics

- Science of collecting, organizing, presenting and interpreting data
- Population is a collection of all elements in a study
- Sample is a subset of the population



Statistics

- Descriptive (tabular, graphical and numerical methods to summarize data)
- Inferential (using sample data to make general conclusions (inferences) about population)



Sampling criteria

- Representative of population
- The goal is to use the results from sample to estimate the population, that is, the results of the sample are generalized to the population.

Sampling methods

- Probability (random) (every unit of population has a certain probability of being selected)
  - **Simple random** (number all and select from a hat)
  - **Systematic** (select a starting point and every *k*th, sample size is n = N/k)
  - **Stratified** (divide into strata (groups), randomly select proportionally from **each** stratum and combine them to form a full sample, each stratum is homogenous) (divide students into strata by gender, age, level)
  - **Cluster** (divide into clusters (groups), randomly select **some** groups and all their members, each cluster is heterogenous)
- Non-probability (non-random) (some elements of population have no chance of selection, thus could be subject to bias)
  - **Convenience** (select from population easier to reach) (select the first n people who enter a store, people in a street, volunteers for a study)
  - **Quota** (divide into strata, non-randomly select proportionally from each stratum and combine them to form a full sample)
  - **Purposive/Judgemental** (select based on an expert's opinion)

Confidence interval

- The interval (margin) at which the population parameter is located around the sample statistic.
- The larger the sample size the more precise the estimate.

Example

A class consists of 50 students, of whom 20 are foreigners. The teacher held a questionnaire on study hours of a sample of 36 students (9 are foreigners) and obtained the following data:

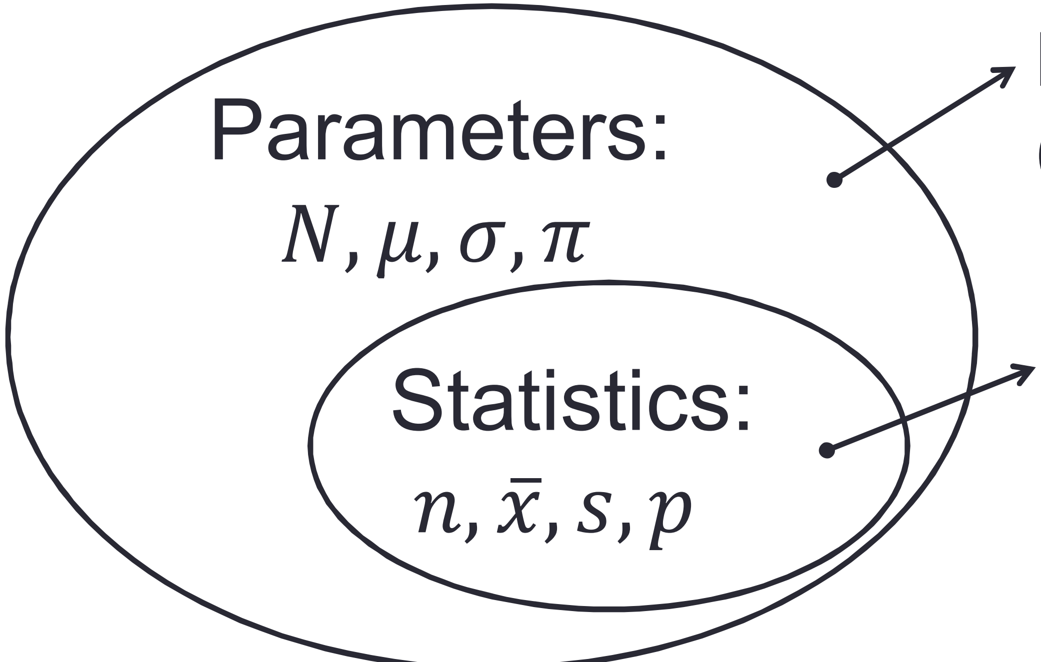
3.2	**2.5**	**2.7**	**5.1**	**4.1**	**3.5**	**2.8**	**2.5**	**3.3**	**3.9**	**4.4**	**3.6**
3.0	**2.9**	**2.6**	**3.1**	**3.2**	**3.4**	**2.4**	**2.5**	**2.8**	**2.7**	**3.8**	**6.2**
4.3	**1.5**	**1.8**	**2.8**	**3.2**	**3.3**	**3.7**	**2.9**	**2.8**	**4.8**	**5.2**	**3.0**

From past research the teacher knows the standard deviation as 1.1.

- What is the size of the population (or sample)?
- What is the proportion of foreigners in population (or sample)?
- What is the population mean (or SD) study time? $\mu = ?\ \ (\sigma = 1.1)$ (Descriptive statistics)
- What is the sample mean (or SD) study time? $\bar{x} = 3.3\ \ (s = 1)$
- Can the sample mean time be generalized to the whole class? (Inferential statistics)
- Find a) 90% b) 95% confidence interval of the population mean.

Example (cont.)

A class consists of 50 students, of whom 20 are foreigners. The teacher held a questionnaire on study hours of a sample of 36 students (9 are foreigners) and obtained the data above:



Population:
$(N = 50, \mu = ?, \sigma = 1.1, \pi = 0.4)$

Sample:
$(n = 36, \bar{x} = 3.3, s = 1, p = 0.25)$

Confidence interval: $\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

$$3.3 - ME \leq \mu \leq 3.3 + ME$$

Example (cont.)

Find a) 90% b) 95% conf. interval of the population mean.

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Answer: $\bar{x} = 3.3,\ \sigma = 1.1,\ n = 36$

a) $1 - \alpha = 0.9 \Rightarrow \alpha = 0.1 \Rightarrow \alpha/2 = 0.05.$

$$z_{0.05} = NORM.S.INV(0.05) = -1.64.$$

$$3.3 - 1.64 \cdot \frac{1.1}{6} \leq \mu \leq 3.3 + 1.64 \cdot \frac{1.1}{6} \Rightarrow 3 \leq \mu \leq 3.6$$

b) $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025.$

$$z_{0.025} = NORM.S.INV(0.025) = -1.96.$$

$$3.3 - 1.96 \cdot \frac{1.1}{6} \leq \mu \leq 3.3 + 1.96 \cdot \frac{1.1}{6} \Rightarrow 2.94 \leq \mu \leq 3.66$$

Central Limit Theorem (for mean)

If a random sample of size *n* is drawn from a population with mean μ and standard deviation σ, the distribution of the sample mean $\bar{x}$ approaches a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as the sample size increases.

2 properties:

1) If the population is **normal**, the sample mean has a normal distribution centered at μ, with a standard error equal to $\sigma/\sqrt{n}$.
2) If the population is **not normal**, the sample mean has more normal distribution centered at μ, with a standard error equal to $\sigma/\sqrt{n}$ as the sample size increases ($n \geq 30$).

Exercise

We would like to estimate the mean amount of cash carried by executives in the hotel industry. We have a sample of size n=36, drawn from a normally distributed population with a known standard deviation $\sigma = 4$. The sample mean is 72.

a) For a confidence level of 95%, calculate the margin of error.

b) Calculate the width and the limits of the confidence interval for the unknown population mean.

c) Interpret the interval.

Exercise solution

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Answer: $\bar{x} = 72,\ \sigma = 4,\ n = 36$

a) For a conf. level of 95%, calculate the margin of error.

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025.$$

$$z_{0.025} = NORM.S.INV(0.025) = -1.96.$$

$$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{4}{\sqrt{36}} = 1.31.$$

b) Calculate the width and the limits of the confidence interval for the unknown population mean.

$$72 - 1.31 \leq \mu \leq 72 + 1.31 \Rightarrow$$
$$70.69 \leq \mu \leq 73.31$$

c) Interpret the interval.

Thank you