Statistical Inference

An Introduction

Contents

Basic Terms as Prerequisite

- **Variable** (under study) – What you measure (ex. monthly salary of employees)
- **Population**- Set of all units in the study (all employees in the organization)
- **Sample**- Subset of units selected from population (ex. monthly salary of few selected employees in the organization)
- **Distribution**-How values of variable are distributed in the population (ex. normal distribution)
- **Factor**- Defines subgroups in the study.(ex. Gender, where gender wise salary distribution can be studied.)
- **Descriptive Statistics**- mean, median, standard deviation etc of the variable under study.. (ex. Average salary)

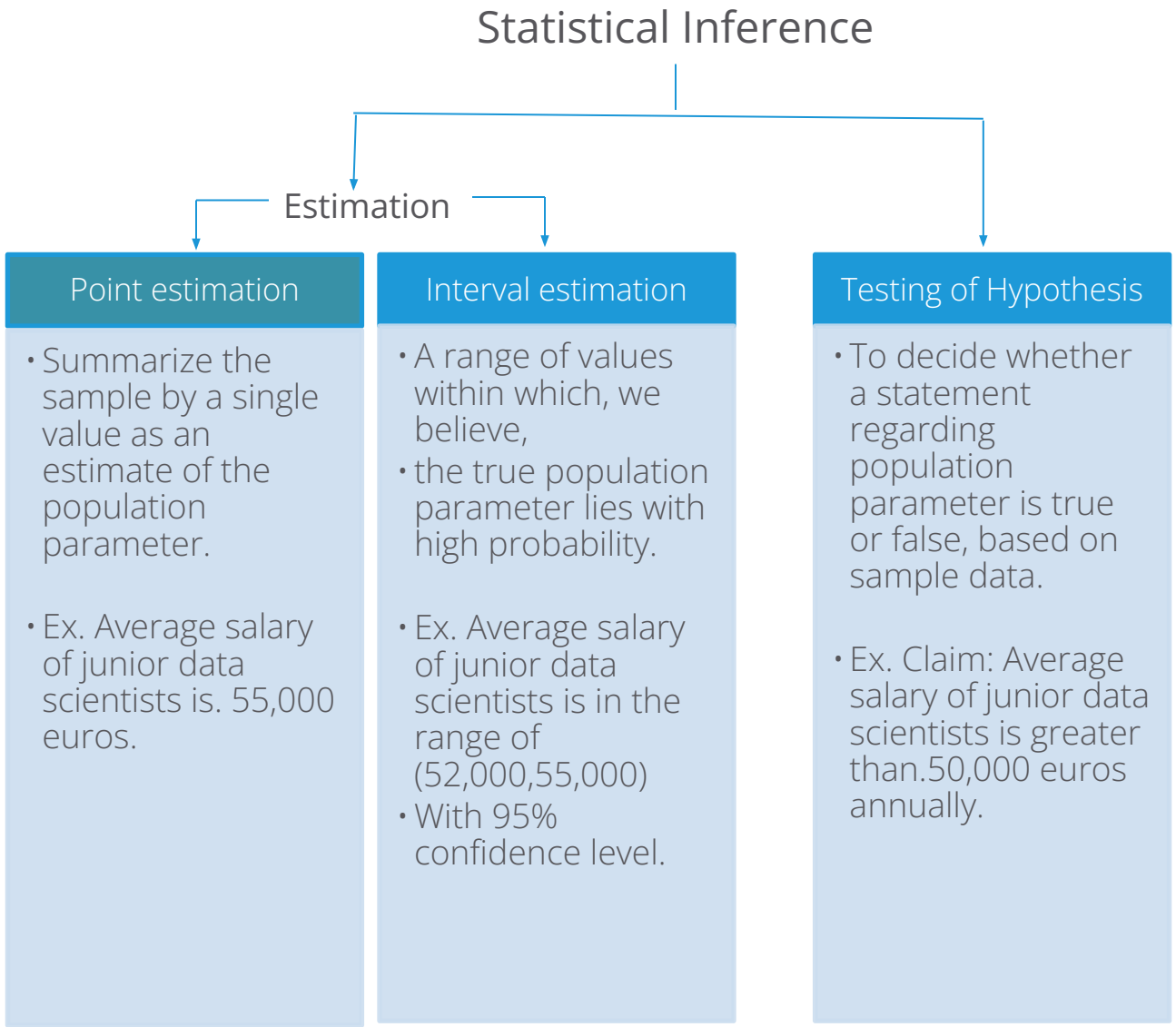
What is Statistical Inference ?

- Statistical inference is the process of drawing conclusion about unknown population properties, using a sample drawn from the population.

- These unknown population properties can be:
  - Mean
  - Proportion
  - Variance etc.

- Such unknown population properties are called as ‘Parameters’.

What is Statistical Inference ?

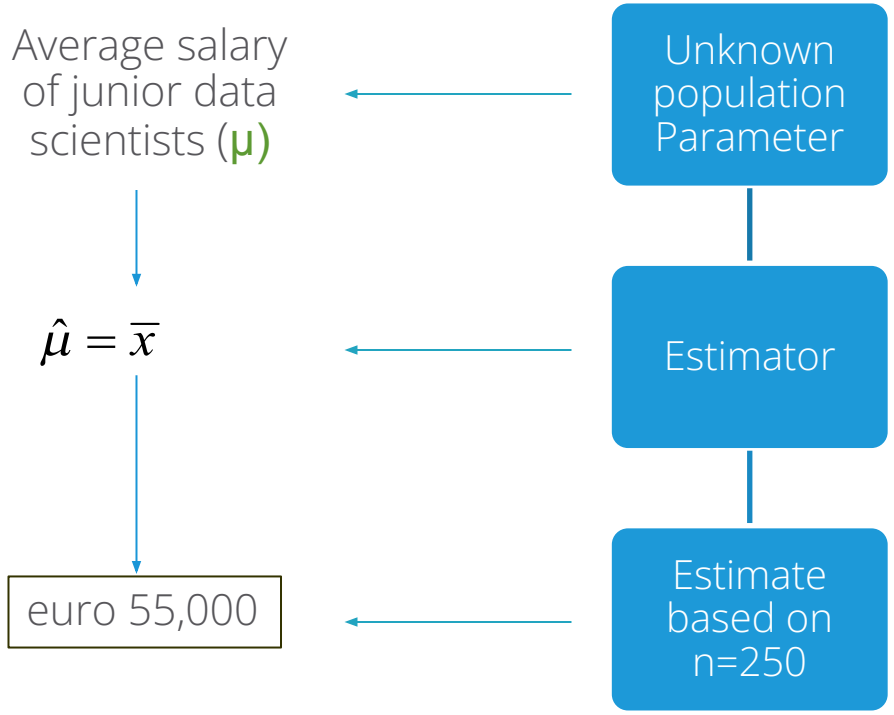


Parameter, Estimator, Estimate

- **Parameter**: Unknown property or characteristic of population
  - (population mean (μ), variance (σ^2), proportion (P))
- **Estimator**: A rule or function based on sample observations which is used to estimate the parameter
  - (sample mean, sample variance, sample proportion)
- **Estimate:** A particular value computed by substituting the sample observations into an Estimator.

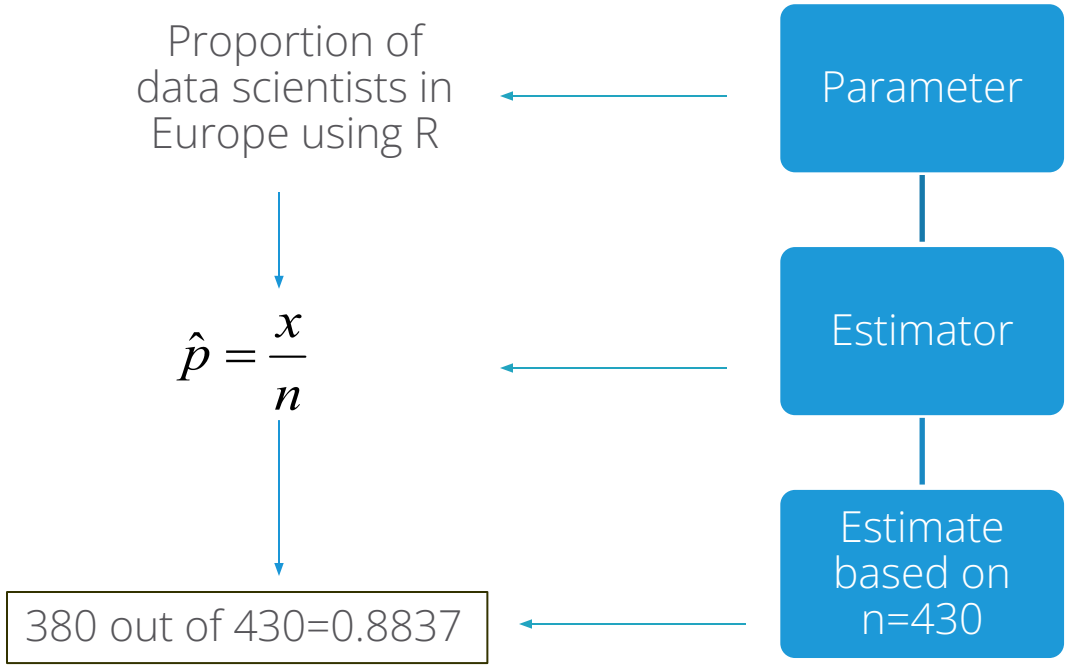
Parameter, Estimator, Estimate

- Research Question: What is the average salary of junior data scientists in Europe?
  - Average salary of junior data scientists in Europe is Population **Parameter.**
  - Sample of 250 junior data scientists is observed and Sample mean is computed.
  - Sample mean is used as **Estimator** of Population Mean.
  - Sample mean “55,000” which is calculated from sample of 250 is the **Estimate.**



Parameter, Estimator, Estimate

- Research Question: What is the proportion of data scientists in Europe who use R for data analysis?
  - Proportion of data scientists in Europe who use R for data analysis is population parameter.
  - Sample of 430 data scientists observed and proportion (or percentage) is calculated.
  - Sample proportion is used as an estimator of population proportion.
  - 380 out of 430 which is calculated from sample is **Estimate.**



Point Estimation vs. Interval Estimation

- In both the previous examples, (estimation of average salary of junior data scientists and proportion of data scientists using R) estimator is a single value estimating unknown population parameter.
- A confidence interval gives an estimated range of values which is likely to include an unknown population parameter with some probability, the estimated range being calculated from a given set of sample data.
- Generally, 95% or 90% Confidence Intervals are used.
- 95% confidence interval is a range estimate within which the true value of the parameter lies with probability 0.95.

Sampling distribution and Sampling error

- Research Question: What is the average salary of junior data scientists in Europe?

  - 50 samples, each of size 250 junior data scientists are observed and sample mean for each of theses 50 samples are computed. Here, sample mean will vary based on sample values.

- A probability distribution of all these means of the sample is called the **sampling distribution** of mean.

- **Standard error** is standard deviation of the these mean values.

Hypothesis Testing

- **Hypothesis**: An assertion about the distribution / parameter of the distribution of one or more random variables.
- **Null Hypothesis (H_0)**: An assertion which is generally believed to be true until researcher rejects it with evidence.
- **Alternative Hypothesis (H_1)**: A researcher's claim which contradicts null hypothesis.
- In simple words, testing of hypothesis is to decide whether a statement regarding population parameter is true or false, based on sample data.
- **Test Statistic**: The statistic on which decision rule of rejection of null hypothesis is defined.
- **Critical region or Rejection region**: the region, in which, if the value of test statistic falls, the null hypothesis is rejected.

Hypothesis Testing : Example

Objective	A consumer protection agency wants to test a Paint Manufacturer's claim, that average drying time of their new paint is less than 20 minutes.

- Sample: n=36 boards were painted from 36 different cans and the drying time was observed.
- Estimator of mean drying time is sample mean $\hat{\mu} = \bar{x}$

Null Hypothesis (H_0): $\mu = 20$

Alternate Hypothesis (H_1): $\mu < 20$

Test Statistic	In this case the test statistic is based on $\bar{x}$
Decision Criteria	Reject null hypothesis if test statistic based on sample mean is less than critical value.

Two types of error

- While testing the hypothesis using any decision rule, one of the following scenario might occur.

Decision	Reality	
	Ho is true	Ho is false
Reject Ho	Type I error	Correct
Do Not Reject Ho	Correct	Type II error

- **For example**, in legal system,
  H_0: person is not guilty H_1: person is guilty

Decision	Reality	
	Not Guilty	Guilty
Guilty	Type I Error -- Innocent person goes to jail	Correct
Not Guilty	Correct	Type II error Guilty person is set free

Two Types of error

- **Level of significance (LOS):** Probability of Type I error is called as ‘Level of Significance (α)’
  generally set as 5% (los=0.05) and null hypothesis is rejected if observed risk(p value) is less
  than 0.05

- **p-value:** is the smallest level of significance that would lead to rejection of the null hypothesis (generally if p <0.05, we reject the null hypothesis).

- **α =** Probability [Type I Error] = Probability [Reject H_0 | H_0 is True]

- **β =** Probability [Type II Error] = Probability [Do not reject H_0 | H_0 is not True]

- **Power of the test** is given by: (1 – β)

One tailed and two tailed tests

- Hypothesis test where the alternative hypothesis is one-tailed (right-tailed or left-tailed), is called a **one-tailed test**.

  $H_0: \mu = \mu_0$
  $H_1: \mu > \mu_0$ (Right-tailed) or $H_1: \mu < \mu_0$ (left-tailed)

- Hypothesis test where the alternative hypothesis is two-tailed is called **two-tailed test**.

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

Quick Recap

Statistical Inference	• It is the process of drawing conclusion about unknown population properties, using a sample drawn from the population.
Point Estimation	• Summarize the sample by a single value as an estimate of the population parameter.
Interval Estimation	• A range of values within which, we believe, the true population parameter lies with high probability.
Testing of Hypothesis	• To decide whether a statement regarding population parameter is true or false
Type I error	• α = Probability [Type I Error] = Probability [Reject H0 | H0 is True]
Type II error	• β = Probability [Type II Error] = Probability [Do not reject H0 | H0 is not True] • Power of the test is given by: $(1 - \beta)$