INTRODUCTION TO DATA ANALYTICS

Class #10

Statistical Inference

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology

IIIT Sri City

QUOTE OF THE DAY..

Live as if you were to die tomorrow. Learn as if you were to live forever.

IN THIS PRESENTATION…

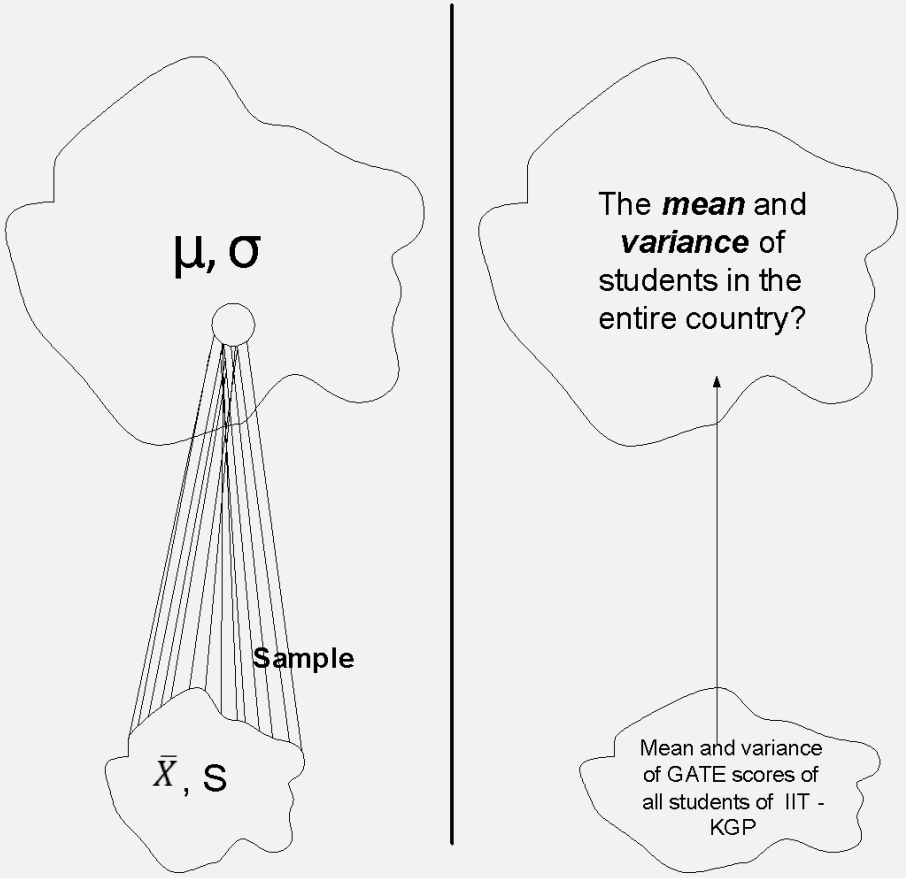
- Principle of Statistical Inference (SI)
- Hypothesis in SI
- Hypotheses testing procedures
- Errors in hypothesis testing
- Case Study 1: Coffee Sale
- Case Study 2: Machine Testing
- Summary of Sampling Distributions in Hypothesis Testing

INTRODUCTION

What do you think about this piece?

INTRODUCTION

The primary objective of statistical analysis is to use data from a sample to make inferences about the population from which the sample was drawn.



This lecture aims to learn the basic procedures for making such inferences.

BASIC APPROACHES

Approach 1: Hypothesis testing

- We conduct **test on hypothesis**.
  - We hypothesize that one (or more) parameter(s) has (have) some specific value(s) or relationship.
- Make our decision about the parameter(s) based on one (or more) sample statistic(s)
- Accuracy of the decision is expressed as the probability that the decision is incorrect.

Approach 2: Confidence interval measurement

- We estimate one (or more) parameter(s) using sample statistics.
  - This estimation usually done in the form of an interval.
- Accuracy of the decision is expressed as the **level of confidence** we have in the interval.

HYPOTHESIS TESTING

Statistical inference

Null hypothesis

Sample

Alternative hypothesis

HYPOTHESIS TESTING

What is Hypothesis?

- "A hypothesis is an educated prediction that can be tested" (study.com).
- "A hypothesis is a proposed explanation for a phenomenon" (Wikipedia).
- "A hypothesis is used to define the relationship between two variables" (Oxford dictionary).
- "A supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation" (Walpole).

STATISTICAL HYPOTHESIS

- If the hypothesis is stated in terms of population parameters (such as mean and variance), the hypothesis is called statistical hypothesis.
- Data from a sample (which may be an experiment) are used to test the validity of the hypothesis.
- A procedure that enables us to agree (or disagree) with the statistical hypothesis is called a **test of the hypothesis.**

Example 6.2:

1. To determine whether the wages of men and women are equal.
2. A product in the market is of standard quality.
3. Whether a particular medicine is effective to cure a disease.

THE HYPOTHESES

- The main purpose of statistical hypothesis testing is to choose between two competing hypotheses.

Example 6.3:

One hypothesis might claim that wages of men and women are equal, while the **alternative** might claim that men make more than women.

- Hypothesis testing start by making a set of two statements about the parameter(s) in question.
- The hypothesis actually to be tested is usually given the symbol H_0 and is commonly referred as the **null hypothesis.**
- The other hypothesis, which is assumed to be true when null hypothesis is false, is referred as the **alternate hypothesis** and is often symbolized by H_1. It is also called as Research hypothesis.
- The two hypotheses are exclusive and exhaustive.

THE HYPOTHESES

Example 6.4:

Ministry of Human Resource Development (MHRD), Government of India takes an initiative to improve the country's human resources and hence set up **23 IIT's** in the country.

To measure the engineering aptitudes of graduates, MHRD conducts GATE examination for a mark of 1000 in every year. A sample of 300 students who gave GATE examination in 2018 were collected and the mean is observed as 220.

In this context, statistical hypothesis testing is to determine the mean mark of the all GATE-2018 examinee.

The two hypotheses in this context are:

$$H_0: \mu = 220$$

$$H_1: \mu < 220$$

THE HYPOTHESES

Note:

1. As null hypothesis, we could choose $H_0: \mu \leq 220$ or $H_0: \mu \geq 220$

2. It is customary to always have the null hypothesis with an equal sign.

3. As an alternative hypothesis there are many options available with us.

Examples 6.5:

I. $H_1: \mu > 220$

II. $H_1: \mu < 220$

III. $H_1: \mu \neq 220$

4. The two hypothesis should be chosen in such a way that they are **exclusive** and **exhaustive**.
   - One or other must be true, but they cannot both be true.

THE HYPOTHESES

One-tailed test

- A statistical test in which the alternative hypothesis specifies that the population parameter lies entirely above or below the value specified in H_0 is called a one-sided (or one-tailed) test.

Example.

$$H_0: \mu = 100 \qquad H_1: \mu > 100$$

Two-tailed test

- An alternative hypothesis that specifies that the parameter can lie on their sides of the value specified by H_0 is called a two-sided (or two-tailed) test.

Example.

$$H_0: \mu = 100 \qquad H_1: \mu <> 100$$

THE HYPOTHESES

Note:

In fact, a 1-tailed test such as:

$$H_0: \mu = 100$$

$$H_1: \mu > 100$$

is same as

$$H_0: \mu \leq 100$$

$$H_1: \mu > 100$$

In essence, $\mu > 100$, it does not imply that $\mu > 80, \mu > 90$, etc.

HYPOTHESIS TESTING PROCEDURES

- The following five steps are followed when testing hypothesis

1. Specify H_0 and H_1, the null and alternate hypothesis, and an **acceptable level of α.**
2. Determine an appropriate sample-based test statistics and the **rejection region** for the specified H_0.
3. Collect the sample data and calculate the test statistics.
4. Make a decision to either reject or fail to reject H_0.
5. Interpret the result in common language suitable for practitioners.

HYPOTHESIS TESTING PROCEDURE

- In summary, we have to choose between H_0 and H_1
- The standard procedure is to assume H_0 is true.

  (Just we presume innocent until proven guilty)
- Using statistical test, we try to determine whether there is sufficient evidence to declare H_0 true.
- We reject H_0 only when the **chance is small** that H_0 is false.
- The procedure is based on probability theory, that is, there is a chance that we can make errors.

Errors in Hypothesis Testing

In hypothesis testing, there are two types of errors.

Type I error: A type I error occurs when we incorrectly reject H_0 (i.e., we reject the null hypothesis, when H_0 is true).

Type II error: A type II error occurs when we incorrectly fail to reject H_0 (i.e., we accept H_0 when it is not true).

Decision	Observation	
	H_0 is true	H_0 is false
H_0 is accepted	Decision is correct	Type II error
H_0 is rejected	Type I error	Decision is correct

PROBABILITIES OF MAKING ERRORS

Type I error calculation

α: denotes the probability of making a Type I error

$$\alpha = \mathbf{P}(\text{Rejecting } H_0 | H_0 \text{ is true})$$

Type II error calculation

β: denotes the probability of making a Type II error

$$\beta = \mathbf{P}(\text{Accepting } H_0 | H_0 \text{ is false})$$

Note:

- α and β are not independent of each other as one increases, the other decreases
- When the sample size increases, both to decrease since sampling error is reduced.
- In general, we focus on Type I error, but Type II error is also important, particularly when sample size is small.

REFERENCE

- The detail material related to this lecture can be found in

  Probability and Statistics for Engineers and Scientists (8th Ed.) by Ronald E. Walpole, Sharon L. Myers, Keying Ye (Pearson), 2013.

Any question?

QUESTIONS OF THE DAY…

1. In a hypothesis testing, suppose H_0 is rejected. Does it mean that H_1 is accepted? Justify your answer.

2. Give the expressions for z, t and χ^2 in terms of population and sample parameters, whichever is applicable to each. Signifies these values in terms of the respective distributions.

3. How can you obtain the value say $P(z = a)$? What this values signifies?

4. On what occasion, you should consider z-distribution but not t-distribution and vice-versa?

5. Give a situation when you should consider χ^2 distribution but neither z- nor t-distribution.