One Sample test - using R

Hypothesis testing - 1

S. Devi Yamini

Overview

1. Basics
   - Hypothesis Testing
   - Formulation of Hypotheses
   - Types of errors
   - Steps in Hypothesis Testing

Overview

1. Basics
   - Hypothesis Testing
   - Formulation of Hypotheses
   - Types of errors
   - Steps in Hypothesis Testing
2. R coding for One sample tests

Introduction

Population is the set of individuals, items or data points sharing some common characteristics under study.

Introduction

Population is the set of individuals, items or data points sharing some common characteristics under study.

A sample is a subset of the population that is selected for an analysis and is used to draw conclusions about the entire population.

Introduction

Population is the set of individuals, items or data points sharing some common characteristics under study.

A sample is a subset of the population that is selected for an analysis and is used to draw conclusions about the entire population.

Example

1. Population: All APC's in VIT;
Sample: The set of all APC's in this workshop to anlayze the effectiveness of your teaching skills.

Introduction

Population is the set of individuals, items or data points sharing some common characteristics under study.

A sample is a subset of the population that is selected for an analysis and is used to draw conclusions about the entire population.

Example

1. Population: All APC's in VIT;
Sample: The set of all APC's in this workshop to anlayze the effectiveness of your teaching skills.
2. Population: All patients visiting a hospital in a month;
Sample: 60 patients selected randomly to study the waiting time in the hospital

Notations

Population parameters

Population mean (μ)
Population standard deviation (σ)
Population size (N)
Population proportion (P)

Sample statistic

Sample mean ($\bar{x}$)
Sample standard deviation (s)
Sample size (n)
Sample proportion (p)

- When we draw many random samples of the same size from a population and compute a statistic (like mean or standard deviation) for each sample, the collection of these values forms a sampling distribution.
- The statistic (e.g., sample mean $\bar{X}$) behaves as a random variable because its value changes from sample to sample.
- If $X_1, X_2, \ldots, X_n$ are values of the statistic X obtained from independent random samples of fixed size from a population, then

$$\bar{X} = \frac{\sum X_i}{n}$$

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Sampling distribution is the probability distribution of a statistic obtained from all possible samples of a fixed size drawn from a population.

Hypothesis Testing

What is Hypothesis?

Problem: Does studying longer improve the test scores?
Hypothesis: "If students study for more than two hours, then their exam scores will increase".

Hypothesis Testing

What is Hypothesis?

Problem: Does studying longer improve the test scores?
Hypothesis: "If students study for more than two hours, then their exam scores will increase".

Statistical Hypothesis

A statement or a claim about one or more population parameters.
Eg. $\mu = 100$, $\sigma = 10$, $N = 500$

Hypothesis Testing

What is Hypothesis?

Problem: Does studying longer improve the test scores?
Hypothesis: "If students study for more than two hours, then their exam scores will increase".

Statistical Hypothesis

A statement or a claim about one or more population parameters.
Eg. $\mu = 100$, $\sigma = 10$, $N = 500$

Types of Statistical hypothesis

- Alternate hypothesis (H_1 or H_a): An hypothesis to be tested or to answer a question
- Null hypothesis (H_0): An hypothesis of no difference. That is, an hypothesis which opposes H_1.

How to choose the Alternate hypothesis?

- Testing a research hypothesis:

Example

Suppose a particular bike has an average fuel efficiency of 40km/litre. The R& D department of the company developed a new fuel injection system to increase the average fuel efficiency. To test whether the new fuel injection system has an average fuel efficiency exceeding 40km/litre, we have to set up the hypothesis as

How to choose the Alternate hypothesis?

- Testing a research hypothesis:

Example

Suppose a particular bike has an average fuel efficiency of 40km/litre. The R& D department of the company developed a new fuel injection system to increase the average fuel efficiency. To test whether the new fuel injection system has an average fuel efficiency exceeding 40km/litre, we have to set up the hypothesis as

$$H_0 : \mu = 40$$

$$H_1 : \mu > 40$$

Right tailed test

- Testing the validity of a claim:

General rule

- If the claim contains equality $(=, \leq, \geq)$, then it is considered as null hypothesis
- If the claim contains strict inequality $(<, >)$, then it is considered as alternative hypothesis

- Testing the validity of a claim:

General rule

- If the claim contains equality $(=, \leq, \geq)$, then it is considered as null hypothesis
- If the claim contains strict inequality $(<, >)$, then it is considered as alternative hypothesis

Example

Suppose a manufacturer of soft drinks claims that two-litre soft drink bottles has an average of at least 150g sugar. To test the manufacturer's claim, we have the hypothesis as

- Testing the validity of a claim:

General rule

- If the claim contains equality $(=, \leq, \geq)$, then it is considered as null hypothesis
- If the claim contains strict inequality $(<, >)$, then it is considered as alternative hypothesis

Example

Suppose a manufacturer of soft drinks claims that two-litre soft drink bottles has an average of at least 150g sugar. To test the manufacturer's claim, we have the hypothesis as

$$H_0 : \mu \geq 150$$

$$H_1 : \mu < 150$$

Left tailed test

Try these!

1. A computer chip manufacturer claims that no more than 2 percent of the chips it sends out are defective. An electronics company, impressed with this claim, has purchased a large quantity of such chips. To determine if the manufacturer's claim can be taken literally, the company has decided to test a sample of 300 of these chips. If 10 of these 300 chips are found to be defective, should the manufacturer's claim be rejected?

Try these!

1. A computer chip manufacturer claims that no more than 2 percent of the chips it sends out are defective. An electronics company, impressed with this claim, has purchased a large quantity of such chips. To determine if the manufacturer's claim can be taken literally, the company has decided to test a sample of 300 of these chips. If 10 of these 300 chips are found to be defective, should the manufacturer's claim be rejected?

2. All cigarettes presently on the market have an average nicotine content of at least 1.6 mg per cigarette. A firm that produces cigarettes claims that it has discovered a new way to cure tobacco leaves that will result in the average nicotine content of a cigarette being less than 1.6 mg. To test this claim, a sample of 20 of the firm's cigarettes were analyzed. If it is known that the standard deviation of a cigarette's nicotine content is .8 mg, what conclusions can be drawn, at the 5 percent level of significance, if the average nicotine content of the 20 cigarettes is 1.54?

How to choose the Alternate hypothesis?

- Testing in decision making situations:

Example

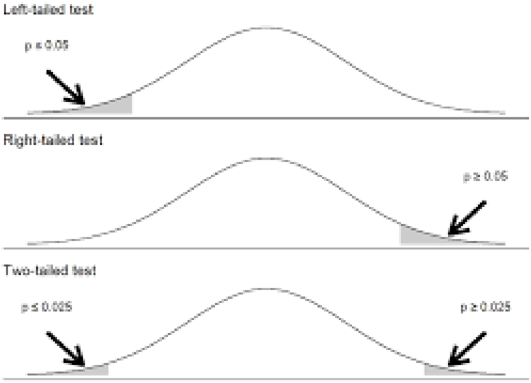
Suppose a quality control inspector has to decide whether to accept or return the shipment to the supplier based on whether it meets the specifications, say the mean length of three inches per part, the hypothesis must be

$$H_0 : \mu = 3$$

$$H_1 : \mu \neq 30$$

Two-tailed test

Graphical intepretation of one-tailed and two-tailed test



Types of errors

	H_0 is true	H_0 is false
Reject H_0	Type I error	Correct decision
Accept H_0	Correct decision	Type II error

P(Type I error) $= \alpha =$ level of significance (LOS)
P(Type II error) $= \beta$

Types of errors

- Type I error: Rejecting H_0 when H_0 is true. Probability of Type I error is known as significance level (α)
- Example:
  Criminal conviction: An innocent person is convicted for a crime.
  Medical diagnosis: A person is diagnosed with a medical condition when he does not have.
- Type II error: Accepting H_0 when H_1 is true.
- A disease test reports a negative result when the patient is actually infected.)

Types of errors

Scenario	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)	Error Type
Criminal trial	Person is innocent	Person is guilty	Type I
Medical diagnosis (false +)	Person is healthy	Person is diseased	Type I
Disease test (false −)	Person is healthy	Person is diseased	Type II

Critical values

Test LOS	1% (0.01)	5% (0.05)	10% (0.1)
Two tailed	$\lvert Z_{\frac{\alpha}{2}} \rvert = 2.58$	$\lvert Z_{\frac{\alpha}{2}} \rvert = 1.96$	$\lvert Z_{\frac{\alpha}{2}} \rvert = 1.645$
Right tailed	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left tailed	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

Steps involved in Hypothesis testing

1. Formulate Null and Alternate hypothesis
2. Identify the level of significance
3. Calculate the test statistic / p-value
4. Identify the critical or rejection region
5. Conclusion: Reject H_0 or Fail to reject H_0

R coding for One sample tests

1. Interpretation of test objects (p-value, estimate, CI)

2. Confidence interval :

A confidence interval (CI) is a range of values calculated from sample data that is used to estimate an unknown population parameter (such as a mean or a proportion) with a stated level of confidence.
Example: $100(1-\alpha)\%$ confidence interval for population mean is

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

3. Assumption checking using plots

4. Writing reusable testing functions