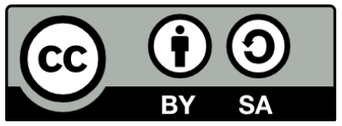
STA 206: Statistics for Physical Sciences and Engineering II

Published by the Centre for Open and Distance Learning,
University of Ilorin, Nigeria

E-mail: codl@unilorin.edu.ng
Website: https://codl.unilorin.edu.ng



Course Development Team

Subject Matter Expert

Adeniyi O.I., Ph.D.

Instructional Designers

Olawale Koledafe

Damilola Adesodun

Ayodele S. Olorunfemi

Hassan Selim Olarewaju

Language Editors

Bankole Ogechi Ijeoma

From the Vice Chancellor

Courseware development for instructional use by the Centre for Open and Distance Learning (CODL) has been achieved through the dedication of authors and the team involved in quality assurance based on the core values of the University of Ilorin. The availability, relevance and use of the courseware cannot be timelier than now that the whole world has to bring online education to the front burner. A necessary equipping for addressing some of the weaknesses of regular classroom teaching and learning has thus been achieved in this effort.

This basic course material is available in different electronic modes to ease access and use for the students. They are available on the University's website for download to students and others who have interest in learning from the contents. This is UNILORIN CODL's way of extending knowledge and promoting skills acquisition as open source to those who are interested. As expected, graduates of the University of Ilorin are equipped with requisite skills and competencies for excellence in life. That same expectation applies to all users of these learning materials.

Needless to say, that availability and delivery of the courseware to achieve expected CODL goals are of essence. Ultimate attention is paid to quality and excellence in these complementary processes of teaching and learning. Students are confident that they have the best available to them in every sense.

It is hoped that students will make the best use of these valuable course materials.

Professor S. A. Abdulkareem
Vice Chancellor

Foreword

Courseware remains the nerve centre of Open and Distance Learning. Whereas some institutions and tutors depend entirely on Open Educational Resources (OER), CODL at the University of Ilorin considers it necessary to develop its own materials. Rich as OERs are and widely as they are deployed for supporting online education, adding to them in content and quality by individuals and institutions guarantees progress. Doing it in-house as we have done at the University of Ilorin has brought the best out of the Course Development Team across Faculties in the University. Credit must be given to the team for prompt completion and delivery of assigned tasks in spite of their very busy schedules.The development of the courseware is similar in many ways to the experience of a pregnant woman eagerly looking forward to the D-day when she will put to bed. It is customary that families waiting for the arrival of a new baby usually do so with high hopes. This is the apt description of the eagerness of the University of Ilorin in seeing that the centre for open and distance learning [CODL] takes off.

The Vice-Chancellor, Prof. Sulyman Age Abdulkareem, deserves every accolade for committing huge financial and material resources to the centre. This commitment, no doubt, boosted the efforts of the team. Careful attention to quality standards, ODL compliance and UNILORIN CODL House Style brought the best out from the course development team. Responses to quality assurance with respect to writing, subject matter content, language and instructional design by authors, reviewers, editors and designers, though painstaking, have yielded the course materials now made available primarily to CODL students as open resources.

Aiming at a parity of standards and esteem with regular university programmes is usually an expectation from students on open and distance education programmes. The reason being that stakeholders hold the view that graduates of face-to-face teaching and learning are superior to those exposed to online education. CODL has the dual-mode mandate. This implies a combination of face-to-face with open and distance education. It is in the light of this that our centre has developed its courseware to combine the strength of both modes to bring out the best from the students. CODL students, other categories of students of the University of Ilorin and similar institutions will find the courseware to be their most dependable companion for the acquisition of knowledge, skills and competences in their respective courses and programmes.

Activities, assessments, assignments, exercises, reports, discussions and projects amongst others at various points in the courseware are targeted at achieving the objectives of teaching and learning. The courseware is interactive and directly points the attention of students and users to key issues helpful to their particular learning. Students' understanding has been viewed as a necessary ingredient at every point. Each course has also been broken into modules and their component units in sequential order.

At this juncture, I must commend past directors of this great centre for their painstaking efforts at ensuring that it sees the light of the day. Prof. M. O. Yusuf, Prof. A. A. Fajonyomi and Prof. H. O. Owolabi shall always be remembered for doing their best during their respective tenures. May God continually be pleased with them, Aameen.

Bashiru, A. Omipidan
Director, CODL

INTRODUCTION

I welcome you to Internet Technology I, a second-semester course. Internet Technology I is a two (2) unit course that provides a general introduction to Internet Technology, covering a brief history of the Internet, how it grew from its humble origins into the worldwide network that is available today, identifying the most popular Internet services such as information retrieval, WWW and communication services.

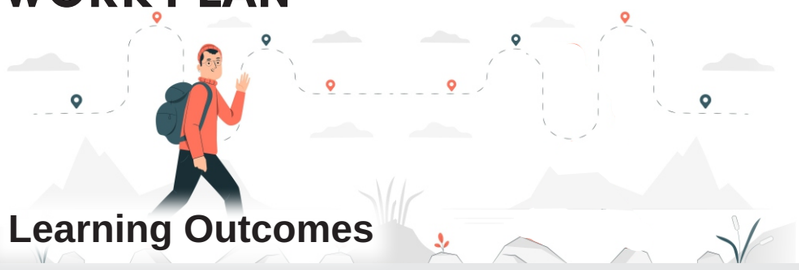
The relationship between the Internet and the World Wide Web is discussed. The course provides you with comprehensive knowledge on the concepts of Internet Technology, which include its internet architecture and internet protocol. The two most important protocols that allow networks to communicate with one another and exchange information, that is the TCP (Transmission Control Protocol) and IP (Internet Protocol), are also discussed.

Also, the functions of each layer at the TCP/IP networking model are covered. The brief history of HTML, XML, XHTML and DHTML is addressed. The course also covers in depth, HTML5, CSS and Javascript. The course also discusses the concept of a markup language and how to create web pages using HTML5 elements, CSS and Javascript. the course also discusses other equally important topics like WYSIWYG, Test Editors, including notepad, notepad++ and others.

Course Goal

The major goal of this course, CSC 224,is to introduce you to the concept of Internet technology and teach you how to develop websites using available technologies.

WORK PLAN



Learning Outcomes

At the end of this course, you should be able to:

- know when to make use of Z-test statistic and t-test statistic.
- fit simple linear regression between two variables
- find the correlation coefficients between two variables
- .carry out tests of independence on two variables and
- make parametric inference about a population whose distribution is unknown

Week 01 | Week 02 | Week 03

Course Infomation

This is a compulsory course for students in the Departments of Chemistry, Industrial Chemistry, Geology, Mathematics, Computer Science. You are expected to participate in all the course activities and have minimum of 75% attendance to be able to write the final examination

Pre-requisite | **NIL**

Course Guide

Module 1

Study Unit 1: Definition of Basic concepts

Study Unit 2: Sampling distribution of Means

Study Unit 3: Sampling distribution of Proportion

Module 2

Study Unit 4: Interval Estimation: Large sample Estimation of a population Mean

Study Unit 5: Interval Estimation: Small sample Estimation of a population Mean

Study Unit 6: Interval Estimation: Large sample Estimation of a population Proportion

Module 3

Study Unit 7: Basic Concepts of Test of Hypothesis

Study Unit 8: Test for a population Mean

Study Unit 9: Test for a population proportion

Module 4

Study Unit 10: Confidence interval for the difference between two population means

Study Unit 11: Hypothesis Testing Concerning difference of two means

Study Unit 12: Inferences about differences between two population proportions

Module 5

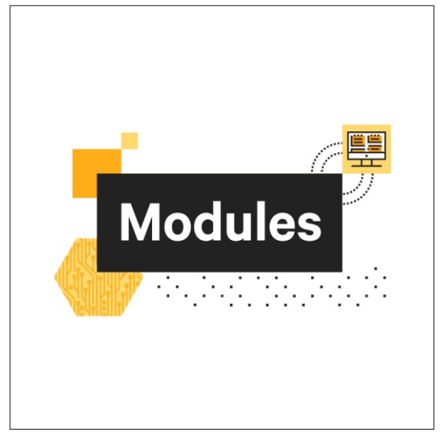
Study Unit 13: Simple Regression Analysis

Study Unit 14: Correlation

Module 4

Study Unit 15: Contingency Table

Study Unit 16: Nonparametric Inference



Course Requirements

Requirements for success

The CODL Programme is designed for learners who are absent from the lecturer in time and space. Therefore, you should refer to your Student Handbook, available on the website and in hard copy form, to get information on the procedure of distance/e-learning. You can contact the CODL helpdesk which is available 24/7 for every of your enquiry.

Visit CODL virtual classroom on http://codllms.unilorin.edu.ng. Then, log in with your credentials and click on STA 206. Download and read through the unit of instruction for each week before the scheduled time of interaction with the course tutor/facilitator. You should also download and watch the relevant video and listen to the podcast so that you will understand and follow the course facilitator.

At the scheduled time, you are expected to log in to the classroom for interaction.

Self-assessment component of the courseware is available as exercises to help you learn and master the content you have gone through.

You are to answer the Tutor Marked Assignment (TMA) for each unit and submit for assessment.

Embedded Support Devices

Support menus for guide and references

Throughout your interaction with this course material, you will notice some set of icons used for easier navigation of this course materials. We advise that you familiarize yourself with each of these icons as they will help you in no small ways in achieving success and easy completion of this course. Find in the table below, the complete icon set and their meaning.

 **Introduction**	 **Learning Outcomes**	 **Main Content**

 **Summary**	 **Tutor Marked Assignment**	 **Self Assessment**
 **Web Resources**	 **Downloadable Resources**	 **Discuss with Colleagues**
 **References**	 **Futher Reading**	 **Self Exploration**

Grading and Assessment

TMA

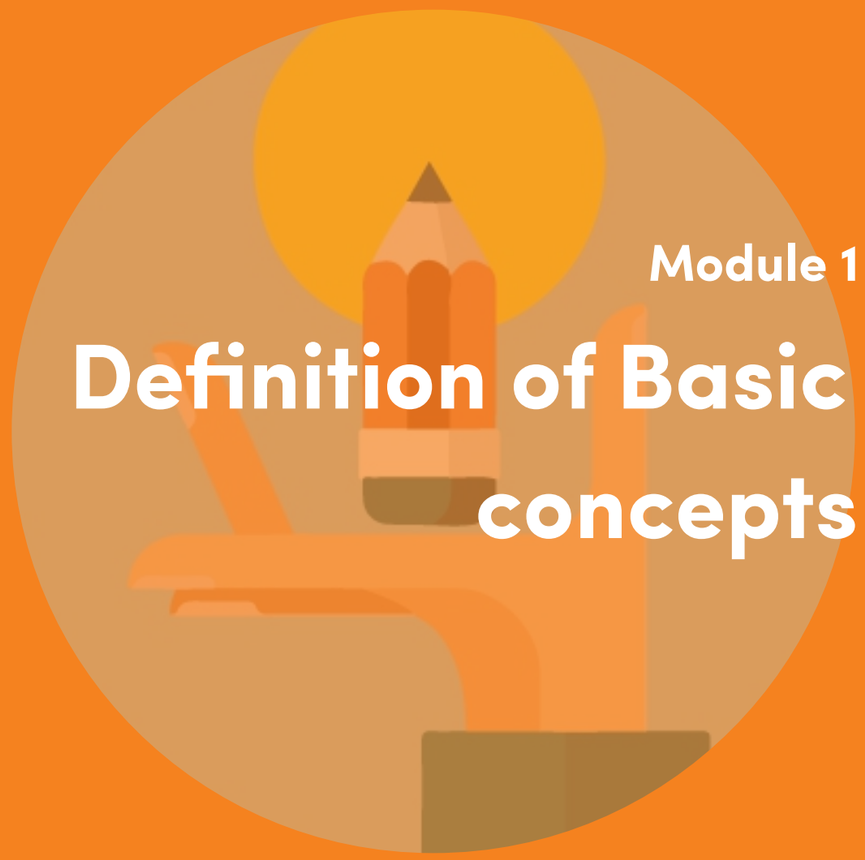
CA

Exam

Total

BASIC

CONCEPTS



Module 1

Definition of Basic concepts

01 | **Picture:** Basic concept

Photo: Wikipedia.com

02 | **Picture:** Basic concept

Photo: Wikipedia.com

UNIT 1

Definition of Basic concepts

Introduction

In this opening unit, I will give the basic definitions of concepts of estimation of a population parameter and also discuss the qualities of an estimator.

Learning Outcomes

At the end of this unit, you should be able to:

1. differentiate between parameter and statistic;
2. define an estimator and estimate;
3. differentiate between point estimation and interval estimation;
4. state the qualities of a good estimator; and
5. identify the qualities of a good estimator when given different estimators.

Main Content

8 mins

Estimation is the branch of Statistics that focuses on making inference about a population. It is a process by which information and a conclusion are drawn about a population based on samples.

Definition of basic concepts

Sampling with replacement: it is the procedure of sampling whereby a selected unit is replaced to the population before the next selection. The probability of selecting a unit remains unchanged after each draw. All the N units in the population have an equal chance of being selected.

The number of possible selection (sample) is ${}_nN$
The probability of selection is 1_nN
E.g. selecting 2 elements from 4, say (ABCD)
The two elements can be drawn with replacement from a population of size 4 in $4_2=16$

i.e AA AB AC AD
BA BB BC BD
CA CB CC CD
DA DB DC DD 16 samples

The probability that sample AB will be selected is 1/16
The probability of selecting element A is 8/16=1/2

Sampling without replacement: it is the procedure of sampling whereby a selected unit is removed from the population for all subsequent selections.

The number of possible selection (sample) is $\binom{N}{n}=\frac{N!}{n!(N-n)!}$

The probability of selection is $\frac{1}{N!/n!(N-n)!}=\frac{n!(N-n)!}{N!}$

E.g. selecting 2 elements from 4, say (ABCD)
The two elements can be drawn with replacement from a population of size 4 in
, i.e. AB AC AD
BC BD
CD 6 samples
The probability that sample AB will be selected is 1/6
The probability of selecting element A is 3/6=1/2

The parameter is a characteristic or function of the population. It is a value or quantity obtained from the information gathered about a population. For example, the mean age of all students taking STA 206. Other examples could be the population mean, population variance, and population proportion.

The statistic is a characteristic or function of the sample taking from the population. It is a value or quantity obtained from the information gathered about a sample taking from the population, which can be used to estimate the population parameter. For example, the mean age of all 50 students taking STA 206 out of all the students. Other examples could be the sample mean, sample variance, and sample proportion.

The estimator is a mathematical rule for obtaining an estimate of a given quantity based on observed data.

The estimate is any quantity computed from the sample data, which is used to give information about an unknown parameter of the population. For example; value obtained in calculating the mean age of 50 students taking STA 206 out of all the students.

The theory of estimation is divided into two parts; point estimation and interval estimation.

Point estimation: A point estimation draws inferences about the population by estimating the value of an unknown population parameter using a single value or point. It uses a single numerical value to provide an estimated value of the population parameter.

Interval estimation: An interval estimation draws an inference about the population by estimating the value of the unknown parameter using an interval the lower values of the interval is refer to as lower limit while the upper value is the upper limit.

Qualities of a good estimator

In general, an estimate must be close to the true value of the population parameter or to vary within only a small range of the true parameter. The closeness is judged based on the following qualities called the qualities or properties of a good estimator.

1. Unbiasedness: if $\hat{\underline{\theta}}$ is an estimator of a population parameter θ, it is said to be unbiased if the expected value of gives the parameter of the population for all samples sizes

Example 1: if x is a binomial random variable with parameter n and p, show that the sample proportion Proof: $\hat{p}=\frac{x}{n}$ is an unbiased estimate of p.

Proof: $\hat{p} = \frac{x}{n}$

$$E(p) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n}(np) = p$$

Example 2: given that $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ are estimators for estimating the population mean such that each of the estimators is a linear combination of five sample observations x_1, x_2, x_3, x_4 and x_5

$$\hat{\theta}_1 = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$\hat{\theta}_2 = \frac{x_1 + x_5}{2}$$

$$\hat{\theta}_3 = \frac{3x_1 + x_2 + 2x_3}{4}$$

Which of the estimators is biased?

Solution

$$E(\theta_1) = E\left[\frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}\right]$$
$$= \frac{1}{5}\left[E(x_1) + E(x_2) + E(x_3) + E(x_4) + E(x_5)\right]$$
$$= \frac{1}{5}[\theta + \theta + \theta + \theta + \theta]$$
$$= \frac{5\theta}{5} = \theta$$

$$E(\hat{\theta}_2) = E\left[\frac{x_1 + x_5}{2}\right]$$
$$= \frac{1}{2}\left[E(x_1) + E(x_2)\right]$$
$$= \frac{1}{2}[\theta + \theta]$$
$$= \frac{2\theta}{2} = \theta$$

$$E(\hat{\theta}_3) = E\left[\frac{3x_1 + x_2 + 2x_3}{4}\right]$$
$$= \frac{1}{4}\left[3E(x_1) + E(x_2) + 2E(x_3)\right]$$
$$= \frac{1}{4}[3\theta + \theta + 2\theta]$$
$$= \frac{1}{4}[6\theta] = \frac{3}{2}\theta$$

The estimator $\hat{\theta}_3$ is biased because it is different from the population parameter θ

2. Consistency: An unbiased estimator is said to be consistent if as the sample size increases the estimator approaches the population parameter being estimated. Variance or standard deviation is usually used to measure how close a sample statistic is to the population parameter.

Example 3: show that the sample means $\bar{x}$ a consistent estimator of the population parameter means .

Proof: μ ere is a need first to show that the sample means $\bar{x}$ an unbiased estimator first.

$$\bar{x} = \frac{\sum x_i}{n}$$
$$= \mathrm{E}(\bar{x}) = E\left[\frac{1}{n}\sum x_i\right]$$
$$= \frac{1}{n}E[x_1 + x_2 + \ldots + x_n]$$
$$= \frac{1}{n}\left[E(x_1) + E(x_2) + \ldots + E(x_n)\right]$$
$$= \frac{1}{n}[\mu + \mu + \ldots + \mu]$$
$$= \frac{n\mu}{n} = \mu$$

$\bar{x}$ is an unbiased estimator of μ

$$V(\bar{x}) = V\left(\frac{1}{n}\sum x\right) = \frac{1}{n^2}V\left(\sum x\right)$$
$$= \frac{1}{n^2}V(x_1 + x_2 + \ldots + x_n)$$
$$= \frac{1}{n^2}\left[V(x_1) + V(x_2) + \ldots + V(x_n)\right]$$
$$= \frac{1}{n^2}\left[\sigma^2 + \sigma^2 + \ldots + \sigma^2\right]$$
$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

As n increases, $V(\bar{x})$ it becomes smaller, hence $\bar{x}$ is a consistent estimator of μ

3. Relative efficiency (Minimum variance): an estimator is said to be efficient it unbiased and has the minimum variance when compared with other unbiased estimators.

Example 4: given the unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ given in example 2, assuming that $V(x_i) = \sigma^2$, which of the estimators is relatively efficient?

$$V(\theta_1) = V\left[\frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}\right]$$
$$= \frac{1}{5^2}V(x_1 + x_2 + x_3 + x_4 + x_5)$$
$$= \frac{1}{25}\left(\sigma^2 + \sigma^2 + \sigma^2 + \sigma^2 + \sigma^2\right)$$
$$= \frac{5\sigma^2}{25} = \frac{\sigma^2}{5}$$

$$V(\theta_2) = V\left[\frac{x_1 + x_5}{2}\right]$$
$$= \frac{1}{2^2}V(x_1 + x_5)$$
$$= \frac{1}{4}\left(\sigma^2 + \sigma^2\right)$$
$$= \frac{2\sigma^2}{4} = \frac{\sigma^2}{2}$$

$\hat{\theta}_1$ has a relatively minimum variance compared to $\hat{\theta}_2$, hence $\hat{\theta}_1$ is relatively efficient to $\hat{\theta}_2$.

4. Sufficiency: an estimator is efficient if it utilizes all the information that the sample contains about the population parameter. It must use all the information about the sample. This implies that no other estimator can add any information about the two population parameter, which has been estimated. For example, the sample mean is a sufficient estimator, while the median is not sufficient.

Summary

The basic definitions have been presented to you in this unit. The properties or qualities of a good estimator were discussed with examples.

Self-Assessment Questions

1. What is the mathematical rule for obtaining a given quantity based on observed data?
2. A characteristic or function of the sample is referred to as_____
3. A characteristic or function of the population is referred to as_____
4. Any quantity computed from the sample data which is used to give information about an unknown quantity in a population is called___ _____
5. The statistical process by which information and conclusion is drawn about a process based on the sample is
6. An estimator that draw inference about a population by estimating the value of an unknown parameter using a single value is called
7. An estimator that draws an inference about a population by estimating the value of an unknown parameter using two values is called
8. The property of an estimator which equates the expected value of the estimator and the parameter to each other is
9. When the sample size is increased, the estimator approaches the population parameter, which of the properties of a good estimator explains this?
10. An estimator that utilizes all the information that a sample container is said to be

Tutor Marked Assessment

- Differentiate between estimator and estimate.
- Differentiate between parameter and statistic.
- What is the difference between point estimation and interval estimation?
- An estimator that utilizes all the information that a sample container is said to be
- given that $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ are estimators for estimating the population mean such that each of the estimators is a linear combination of five sample observations x_1, x_2, x_3, x_4 and x_5

$$\hat{\theta}_1 = \frac{x_1 + x_2 + x_3}{3}$$

$$\hat{\theta}_2 = \frac{x_1 + x_5}{2}$$

$$\hat{\theta}_3 = \frac{x_1 + x_2 + x_3}{4}$$

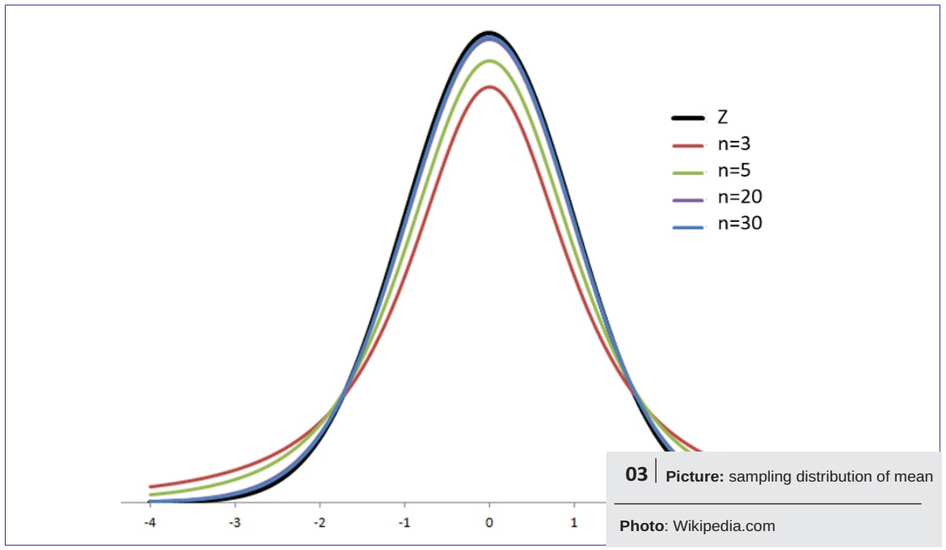
Which of the estimators is biased?

References

- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA
- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.

Further Reading

- http://www.saylor.org/books
- https://study.com/academy/lesson/point-interval-estimations-definition-differences-quiz.html
- http://www3.govst.edu/kriordan/files/gsu610files/PDF/2006/MGMT610CH07.pdf



03 | **Picture:** sampling distribution of mean

Photo: Wikipedia.com

UNIT 2

Sampling distribution of Means

Introduction

In this study unit I will introduce you to the concepts of the mean, the standard deviation, and the sampling distribution of a sample statistic, with an emphasis on the sample mean.

Learning Outcomes

At the end of this unit, you should be able to:

1. become familiar with the concept of the probability distribution of the sample mean;
2. state the meaning of the formulas for the mean and standard deviation of the sample mean;
3. state what the sampling distribution of x
4. is when the sample size is large; and
5. state what the sampling distribution of is when the population is normal.

Main Content

8 mins

A statistic, such as the sample mean or the sample standard deviation, is a number computed from a sample. Since a sample is random, every statistic is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty. As a random variable it has a mean, a standard deviation, and a probability distribution. The probability distribution of a statistic is called its sampling distribution. Sample statistics are computed in order to estimate the corresponding population parameter.

Suppose we wish to estimate the mean μ of a population. In actual practice we would typically take just one sample. Imagine however that we take sample after sample, all of the same size n, and compute the sample mean $\bar{x}$ of each one. We will likely get a different value of $\bar{x}$ each time. The sample mean $\bar{x}$ is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty. We will write $\bar{x}$ when the sample mean is thought of as a random variable, and write $\bar{x}$ for the values that it takes. The random variable $\bar{x}$ has a mean, denoted μ_x

and a standard deviation, denoted as $\hat{\sigma}_x$.

Notations

	Sampling without replacement	Sampling with replacement
Possible Samples	${}^{N}C_n$	N^n
Expected value	μ	μ
Variance of $\bar{x}$	$\left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$	$\frac{\sigma^2}{n}$
Standard deviation of $\bar{x}$	$\sqrt{\left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}}$	$\sqrt{\frac{\sigma^2}{n}}$

Example 1: A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples without replacement of size two and compute the sample mean for each one.

I. Show that the mean of all the sample means equals the population mean
ii. Compute the variance for the population
iii. Compute the variance of the sample

Solution

The following table shows all possible samples without replacement of size two, along with the mean of each:

Sample	Sample Mean $\bar{x}$	$\left[\bar{x}-E(x)\right]^2$
152, 156	154	16
152, 160	156	4
152, 164	158	0
156, 160	158	0
156, 164	160	4
160, 164	162	16
Total	**948**	**40**

Number of possible samples ${}^{4}C_n = 4^2 = 6$

Population mean $\mu = \frac{152+156+160+164}{4} = 158$

i. Means of means $\bar{\bar{x}} = \frac{154+156+158+158+160+160+162}{6} = 158$

$$\sigma^2 = \frac{\sum\left[\bar{x}-\mu\right]}{N}$$

ii. Population Variance $= \frac{(154-158)^2+(156-158)^2+(160-158)^2+(164-158)^2}{4}$

$= \frac{16+4+4+36}{4} = 15$

iii. Sample variance $V(\bar{x}) = \frac{40}{6} = 6.67$

Example 2: the mean and the standard deviation of the tax value of all vehicles registered in a certain state are #13525 and #4180 respectively. A random sample of size 100 is drawn from the population of vehicles. Find the mean μ_x and standard deviations σ_x of the sample mean $\bar{x}$

Solution

n=100, $\mu = 13525$

The sample mean $\mu_x = \mu = 13525$

The sample standard deviation $= \sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{4180}{\sqrt{100}} = 418$

Central limit Theorem

For sample of size 30 or more, the sample mean is approximately normally distributed with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ where n is the sample size. The greater the sample size the better the approximation.

Example 3: let $\bar{x}$ be the mean of a random sample of size 50 drawn from a population with mean 112 and standard deviation 40.

i. Find the mean and standard deviation of $\bar{x}$
ii. Find the probability that $\bar{x}$ assumes a value between 110 and 114
iii. Find the probability that $\bar{x}$ assumes a value greater than 113

Solution

i. The mean of $\bar{x}$, $\mu_{\bar{x}} = \mu = 112$

The standard deviation of $\bar{x}$, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{50}} = 5.6569$

ii. Since the sample size is at least 30, the central limit theorem applies: $\bar{x}$ is approximately normally distributed.
The probability that $\bar{x}$ assumes a value between 110 and 114

$$P(110 < \bar{x} < 114) = P\left(\frac{110-\mu_{\bar{x}}}{\sigma_{\bar{x}}} < Z < \frac{114-\mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$
$$= P\left(\frac{110-112}{5.6569} < Z < \frac{114-112}{5.6569}\right)$$
$$= P(-0.35 < Z < 0.35)$$
$$= \Phi(-0.35) + \Phi(0.35)$$
$$= 0.1368 + 0.1368 = 0.2736$$

The probability of $\Phi(-0.35)$ and $\Phi(0.35)$ is gotten from the Normal distribution table

iv. The probability that $\bar{x}$ assumes a value greater than 113

$$P(\bar{x} > 113) = P\left(Z > \frac{113-\mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$
$$= P\left(Z > \frac{113-112}{5.6569}\right)$$
$$= P(Z > 0.18)$$
$$= \Phi(0.18)$$
$$= 1 - 0.5714 = 0.4286$$

Example 4: Suppose that in a certain region of the country the mean duration of first marriages that end in divorce is 7.8 years, standard deviation 1.2 years. Find the probability that in a sample of 75 divorces, the mean age of the marriages is at most 8 years.

Solution:

Population Mean duration of marriage $\mu = 7.8$

Population standard deviation of duration of marriage $\sigma = 1.2$

Number of samples divorcees $n = 75$

The mean duration of marriages $\bar{x}$, $\mu_{\bar{x}} = \mu = 7.8$

The standard deviation of $\bar{x}$, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{75}} = 0.1732$

The probability that in a sample of 75 divorces, the mean age of the marriages is at most 8 years.

$$P(\bar{x} \le 8) = P\left(Z \le \frac{7.8-\mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$
$$= P\left(Z \le \frac{7.8-8}{0.1732}\right)$$
$$= P(Z \le -1.15)$$
$$= \Phi(-1.15)$$
$$= 0.1251$$

Example 5: Suppose the mean length of time between submission of a state tax return requesting a refund and the issuance of the refund is 47 days, with standard deviation 6 days. Find the probability that in a sample of 50 returns requesting a refund, the mean such time will be more than 50 days.

Solution:

Population Mean length of time between submission of a state tax return and the issuance of the refund $\mu = 47$

Population standard deviation of time between submission of a state tax return and the issuance of the refund $\sigma = 6$

Number of samples of returns requesting a refund $n = 50$

The mean length of time between submission of a state tax return and the issuance of the refund $\bar{x}$, $\mu_{\bar{x}} = \mu = 47$

The standard deviation of $\bar{x}$, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{50}} = 0.8485$

The probability that in a sample of 50 returns requesting a refund, the mean such time will be more than 50 days.

$$P(\bar{x} > 50) = P\left(Z > \frac{50 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$

$$= P\left(Z > \frac{50 - 47}{0.8485}\right)$$

$$= P(Z > 3.54)$$

$$= \Phi(3.54)$$

$$= 1.0000$$

Summary

In this unit, you have learnt about the concept of the probability distribution of the sample mean and the formulas for the mean and standard deviation of the sample mean. This unit also presented that when the sample size is at least 30, the sample mean is normally distributed and when the population is normal, the sample mean is normally distributed regardless of the sample size.

Self-Assessment Questions

1. The numerical population of grade point averages of University of Ilorin students has a mean of 2.61 and a standard deviation of 0.5. If a random sample of size 100 students is taken from the population of students, what is the probability that the grade point average will be between 2.51 and 2.71?
2. A population has mean 128 and standard deviation 22.
   a. Find the mean and standard deviation of for samples of size 36.
   b. Find the probability that the mean $\bar{x}$ of a sample of size 36 will be within 10 units of the population mean, that is, between 118 and 138.
   3.Suppose the mean length of time that a caller is placed on hold when telephoning a customer service centre is 23.8 seconds, with standard deviation 4.6 seconds. Find the probability that the mean length of time on hold in a sample of 1,200 calls will be within 0.5 second of the population mean.

Tutor Marked Assessment

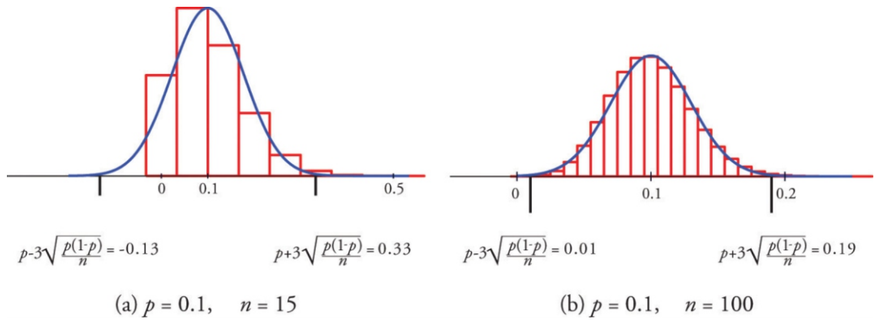
1. A population has mean 73.5 and standard deviation 2.5.
   a. Find the mean and standard deviation of $\bar{x}$ for samples of size 30.
   b. Find the probability that the mean of a sample of size 30 will be less than 72.
2. A normally distributed population has mean 1,214 and standard deviation 122.
   a. Find the probability that a single randomly selected element X of the population is between 1,100 and 1,300.
   b. Find the mean and standard deviation of $\bar{x}$ for samples of size 25.
   c. Find the probability that the mean of a sample of size 25 drawn from this population is between 1,100 and 1,300.
3. Suppose speeds of vehicles on a particular stretch of roadway are normally distributed with mean 36.6 mph and standard deviation 1.7 mph.
   a. Find the probability that the speed of a randomly selected vehicle is between 35 and 40 mph.
   b. Find the probability that the mean speed $\bar{x}$ of 20 randomly selected vehicles is between 35 and 40 mph.
4. Suppose the mean cost across the country of a 30-day supply of a generic drug is #46.58, with standard deviation #4.84. Find the probability that the mean of a sample of 100 prices of 30-day supplies of this drug will be between #45 and #50.

References

- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA
- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.

Further Reading

- https://www.sheffield.ac.uk/polopoly_fs/1.43999!/file/tutorial-10-reading-tables.pdf
- https://stattrek.com/sampling/sampling-distribution.aspx
- http://www.cogsci.ucsd.edu/~nunez/COGS14B_W17/W3b.pdf



04 | **Picture:** Distribution of proportion

Photo: Wikipedia.com

UNIT 3

Sampling distribution of Proportion

Introduction

In this study unit, I will introduce you to the concepts of the proportion, the standard deviation, and the sampling distribution of a sample statistic, with an emphasis on the sample proportion.

Learning Outcomes

At the end of this unit, you should be able to:

1. recognize that the sample proportion $\hat{p}$ is a random variable;
2. state the meaning of the formulas for the mean and standard deviation of the sample proportion;

Main Content

7 mins

Often, sampling is done in order to estimate the proportion of a population that has a specific attribute or characteristic of interest.

For example

Production Plant: the proportion defective items coming from a production plant or the proportion of all.

Bank Transaction: people entering a bank who make a transaction before leaving.

Genetics: to estimate proportion who carry the gene for a particular disease.

Consumer Preferences: to estimate proportion of consumers who prefer hp laptop compared with Dell laptop.

Television Ratings: to estimate proportion of households watching a particular Television program

The sample proportion is a random variable that varies from sample to sample in a way that cannot be predicted with certainty. It has a mean μ_p and a standard deviation σ_p.

The proportion is $p = \dfrac{\text{Number of sample with characteristic of interest}}{\text{Total number of observation}}$

While $q = 1 - \dfrac{\text{Number of sample with characteristic of interest}}{\text{Total number of observation}}$

$=1\text{-}p$

Suppose random samples of size n are drawn from a population in the proportion of the characteristic of interest is p. the mean μ_p and the standard deviation σ_p of the sample proportion is given as

$\mu_p = p$ and $\sigma_p = \sqrt{\dfrac{pq}{n}}$

Sampling distribution of the sample proportion

The sampling distribution of is not exactly normal, but for large samples, the sample proportion is approximately normally distributed with mean $\mu_p = P$ and standard deviation $\sigma_p = \sqrt{\dfrac{pq}{n}}$. A sample is large if the interval $\left[p-3\sigma_p, p+3\sigma_p\right]$ lies wholly within the interval $[0,1]$.

Example 1: An online retailer claims that 95% of all orders are shipped within 24 hours of being received. A consumer group placed 500 orders of different sizes and at different times of the day; 420 orders were shipped within 24 hours.

a. Compute the sample proportion of items shipped within 24 hours.

b. Confirm that the sample is large enough to assume that the sample proportion is normally distributed. Use p = 0.95, corresponding to the assumption that the retailer's claim is valid.

c. Assuming the retailer's claim is true, find the probability that a sample of size 500 would produce a sample proportion as low as was observed in this sample.

Solution

a. The sample proportion $p = \dfrac{\text{Number of orders shipped within 24 hours}}{\text{Total number of number of orders placed}}$

$= \dfrac{420}{500} = 0.84$

b. Since $p = 0.90, q = 1 - p = 0.10$

$\sigma_p = \sqrt{\dfrac{0.90 \times 0.10}{500}} = 0.013$

Hence, $\left[p - 3\sigma_p, p + 3\sigma_p\right] = \left[0.90 - 3 \times 0.013, 0.90 + 3 \times 0.013\right]$

$= \left[0.861, 0.939\right]$

The interval falls within the interval [0, 1]

c. The probability that a sample of size 500 would produce a sample proportion as low as was observed in this sample.

$$P(\hat{p} \le 0.84) = P\left[Z \le \frac{0.84 - \mu_p}{\sigma_p}\right]$$

$$= P\left[Z \le \frac{0.84 - 0.90}{0.013}\right]$$

$$= P\left[Z \le -4.62\right] = 0.0000$$

Example 2: With the advent of cell phones, a survey conducted revealed that 7% of all households have no home telephone but depend completely on cell phones. Find the probability that in a random sample of 450 households, between 25 and 35 will have no home telephone. Assume that the population follows a normal distribution.

Solution

Number of households=450

Proportion with no home phones $p=0.07, q=1, -p=0.93$

$$\sigma_p=\sqrt{\frac{0.07\times 0.93}{450}}=0.012$$

The probability that in a random sample of 450 households, between 25 and 35 will have no home telephone.

$$\hat{p}=\frac{25}{450}=0.06 \quad \text{and} \quad \frac{35}{450}=0.08$$

$$P(0.06\le\hat{p}\le 0.08)=P\left[\frac{0.06-\mu_p}{\sigma_p}\le Z\le\frac{0.08-\mu_p}{\sigma_p}\right]$$

$$=P\left[\frac{0.06-0.07}{0.012}\le Z\le\frac{0.08-0.07}{0.012}\right]$$

$$=\mathrm{P}[-0.83\le Z\le 0.83]$$

$$=0.5934$$

Example 3: Suppose that 2% of all cell phone connections by a certain provider are dropped. Find the probability that in a random sample of 1,500 calls at most 40 will be dropped.

Solution

Number of random sampled calls=1500

Proportion of cell phone connections that were dropped $p=0.02, q=1, -p=0.98$

$$\sigma_p=\sqrt{\frac{0.02\times 0.98}{1500}}=0.0004$$

The probability that in a random sample of 1,500 calls at most 40 will be dropped.

$$\hat{p}=\frac{40}{1500}=0.027$$

$$P(\hat{p}\le 0.027)=P\left[Z\le\frac{0.027-\mu_p}{\sigma_p}\right]$$

$$=P\left[Z\le\frac{0.027-0.02}{0.0004}\right]$$

$$=\mathrm{P}[Z\le 17.5]=0.0000$$

Example 4: In one study it was found that 86% of all homes have a functional smoke detector. Assume that the normal distribution applies. Find the probability that in a random sample of 600 homes,

a. Exactly 80% will have a functional smoke detector

b. between 80% and 90% will have a functional smoke detector.

Solution

Number of homes=600

Proportion of home with functional smoke detector $p=0.86, q=1, -p=0.14$

$$\sigma_p=\sqrt{\frac{0.86\times 0.14}{600}}=0.014$$

a. The probability that exactly 80% will have a functional smoke detector.
80% of 600 is 480

$$\hat{p}=\frac{480}{600}=0.8$$

$$P(\hat{p}=0.8)=P\left[Z=\frac{0.8-\mu_p}{\sigma_p}\right]$$

$$=P\left[Z=\frac{0.8-0.86}{0.014}\right]$$

$$=\mathrm{P}[Z=-4.29]=0.5$$

b. between 80% and 90% will have a functional smoke detector.
90% of 600 is 540, therefore,

$$\hat{p}=\frac{480}{600}=0.8 \quad \text{and} \quad \frac{540}{600}=0.9$$

$$P(0.8\le\hat{p}\le 0.9)=P\left[\frac{0.8-\mu_p}{\sigma_p}\le Z\le\frac{0.9-\mu_p}{\sigma_p}\right]$$

$$=P\left[\frac{0.8-0.86}{0.014}\le Z\le\frac{0.9-0.86}{0.014}\right]$$

$$=\mathrm{P}[-4.29\le Z\le 2.86]$$

$$=0.9979$$

Summary

The concept of the probability distribution of the sample proportion is what we have discussed thus far and the formulas for the proportion and standard deviation of the sample proportion were given. This unit also presented that when the sample size is large, the sample proportion is normally distributed.

Self-Assessment Questions

SAQ

1. A random sample of size 300 was drawn from a population in which the proportion of characteristic of interest is 0.35, decide whether the sample size is large enough to assume that the sample proportion is normally distributed

2. Suppose that 29% of all residents of a community favour a bill passed by the legislative. Find the probability that in a random sample of 50 residents at least 35% will favour the bill. First verify that the sample is sufficiently large to use the normal distribution.

3. Suppose the proportion of all university students who boarded the school bus in the past 6 months is p = .40. For a population of N = 200 students of the university representing all university students, what is the probability that the proportion of students who boarded the school bus in the past 6 months is less than .32?

Tutor Marked Assessment

- Suppose that 8% of all males suffer some form of colour blindness. Find the probability that in a random sample of 250 men at least 10% will suffer some form of colour blindness. First verify that the sample is sufficiently large to use the normal distribution.

- An outside financial auditor has observed that about 4% of all documents he examines contain an error of some sort. Assuming this proportion to be accurate, find the probability that a random sample of 700 documents will contain at least 30 with some sort of error. You may assume that the normal distribution applies.

References

- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.

- Spiegel M.R and Stephens L.J (1999). Schaum's Outlines of Theory and Problems of Statistics. McGraw-Hill, New York

- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA

Further Reading

- https://www.khanacademy.org/math/ap-statistics/sampling-distribution-ap/sampling-distribution-proportion/v/sampling-distribution-of-sample-proportion-part-1?

- https://www.sheffield.ac.uk/polopoly_fs/1.43999!/file/tutorial-10-reading-tables.pdf

Population Mean Formula

$$\mu = \frac{\sum X}{N}$$

05 | **Picture: Population mean**

Photo: wikipedia.com

Module 2

Interval Estimation: Large sample Estimation of a population Mean

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

06 | **Picture:** Population mean

Photo: Wikipedia.com

UNIT 4

Interval Estimation: Large sample Estimation of a population Mean

Introduction

In the last Module, we discussed the sampling distribution of a sample statistic, with emphasis on the sample mean and the sample proportion. In this Module we will discuss the major problem of statistical inference which is the estimation of population parameters.

Learning Outcomes

At the end of this unit, you should be able to:

1. get familiar with the concept of an interval estimate of the population mean; and
2. apply formulas for a confidence interval for a population mean.

Main Content

8 mins

A confidence interval is an interval of values that is likely to include the unknown value of the parameter. **Confidence interval** includes three parts which are: A confidence level, a statistic and a margin of error. The interval estimate of a confidence interval is defined as the sample statistic ± margin of error.

In estimating the population parameter, there are two methods which are point estimate and interval estimate as discussed in unit one. But confidence intervals are preferred to point estimates, because confidence intervals indicate the precision of the estimate and the uncertainty of the estimate.

Confidence level

A confidence level is the probability part of a confidence interval. It describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter. The confidence level is interpreted in percent. For example, a 95% confidence level indicates that 95% of the intervals contain the true population parameter and a 90% confidence level means that 90% of the intervals contain the true population parameter. Note that some confidence intervals may not include the true value of the population parameter.

Margin of Error

The margin of error is the range of values above and below the sample statistic.

Confidence interval for the population mean μ when the population standard deviation σ is known

The basic assumptions for estimating the population mean μ when the population standard deviation σ is known are as follows:

1. A simple random sample of size n is drawn from the population
2. The population standard deviation or variance $(\sigma \text{ or } \sigma^2)$ is known
3. The random variable x is assumed to follow a normal distribution
4. If the distribution of the population from which the random variables are selected is unknown, then a sample size of $n \geq 30$ is required.

By central limit theorem, **the confidence interval** is given as

$$\bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$$

Suppose a confidence level c is desired, the value of c is between 0 and 1 but usually equal to a number such as 0.90, 0.95 or 0.99. The value of Z_c is the number such that the area under the standard normal curve falling between $-Z_c$ and Z_c is equal to c. The value of Z_c is called the critical value for a confidence level c.

Table 4.1: some confidence levels and their corresponding critical values

Confidence level c	Critical value Z_c
0.50	0.6745
0.70	1.04
0.75	1.15
0.80	1.28
0.85	1.44
0.90	1.645
0.95	1.96
0.98	2.33
0.99	2.58

Worked Examples

Example 1: A random sample of size 100 is drawn from a population known to have a standard deviation 11.3. If the mean of the random sample is 105.2,

a. Construct a 90% confidence interval for the sample population mean
b. Construct a 95% confidence interval for the sample population mean
c. Construct a 99% confidence interval for the sample population mean

Solution

n=100, $\bar{x} = 105.2, \sigma = 11.3$

a. 90% confidence interval for the sample population mean

$$CI = \bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$= 105.2 \pm Z_{0.90}\left(\frac{11.3}{\sqrt{100}}\right)$$

$$= 105.2 \pm 1.645(1.13)$$

$$105.2 \pm 1.86$$

The 90% confidence interval of the sample population mean is with the interval $103.34 \leq \mu \leq 107.06$

b. 95% confidence interval for the sample population mean

$$CI = \bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$= 105.2 \pm Z_{0.95}\left(\frac{11.3}{\sqrt{100}}\right)$$

$$= 105.2 \pm 1.96(1.13)$$

$$105.2 \pm 2.22$$

The 95% confidence interval of the sample population mean is with the interval $102.98 \le \mu \le 107.42$

c. 99% confidence interval for the sample population mean

$$CI = \bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$= 105.2 \pm Z_{0.99}\left(\frac{11.3}{\sqrt{100}}\right)$$

$$= 105.2 \pm 2.58(1.13)$$

$$105.2 \pm 2.92$$

The 99% confidence interval of the sample population mean is with the interval $102.28 \le \mu \le 108.12$

Example 2: A brand of light bulb is known to have a life span whose standard deviation is 75 hours. If a random sample of 35 bulbs taken from the brand has mean life span of 850 hours, determine the 95% and 90% confidence interval for the mean life span.

Solution

$n = 35, \bar{x} = 850, \sigma = 75$

95% confidence interval for the mean life span.

$$CI = \bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$= 850 \pm Z_{0.95}\left(\frac{75}{\sqrt{35}}\right)$$

$$= 850 \pm 1.96(12.68)$$

$$= 850 \pm 24.85$$

The 95% confidence interval for the mean life span is $825.15 \le \mu \le 874.85$

90% confidence interval for the mean life span.

$$CI = \bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$= 850 \pm Z_{0.90}\left(\frac{75}{\sqrt{35}}\right)$$

$$= 850 \pm 1.645(12.68)$$

$$= 850 \pm 20.86$$

The 95% confidence interval for the mean life span is $829.14 \le \mu \le 870.86$

Example 3: The grade point average GPA of students from a university is known to have a standard deviation of 0.51. A random sample of 120 student selected from the University gave a mean GPA of 2.71. Construct a 95 percent confidence interval for the mean GPA for all students in the University.

Solution

$n = 120, \bar{x} = 2.71, \sigma = 0.51$

$$CI = \bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$= 2.71 \pm Z_{0.95}\left(\frac{0.51}{\sqrt{120}}\right)$$

$$= 2.71 \pm 1.96(0.05)$$

$$= 2.71 \pm 0.098$$

The 95% confidence interval for grade point average is $2.61 \le \mu \le 2.81$

Confidence interval for the population mean μ when the population standard deviation σ is unknown

In an attempt to find the confidence interval for the population mean μ using the normal distribution, the population standard deviation must be known as part of the assumption to make the normal distribution a valid distribution for use. However, in most cases when the population mean μ is unknown, the population standard deviation is also not known. In such cases, the sample standard deviation s is used as an approximate of the population standard deviation σ. When s is used as an approximate of σ and the sample size $n \ge 30$, by central limit theorem, **the confidence interval** is given as

$$\bar{x} \pm Z_c\left(\frac{s}{\sqrt{n}}\right)$$

Example 4: The mean and standard deviations of the maximum loads by 60 cables are given by 11.09 and 0.73 respectively. Find the

(a) 95% confidence interval for the mean of the maximum loads of all cables produced by the company

(b) 99% confidence interval for the mean of the maximum loads of all cables produced by the company

Solution

$n = 60, \bar{x} = 11.09, s = 0.73$

a) 95% confidence interval for the mean of the maximum loads of all cables produced by the company

$$CI = \bar{x} \pm Z_c\left(\frac{s}{\sqrt{n}}\right)$$

$$= 11.09 \pm Z_{0.95}\left(\frac{0.73}{\sqrt{60}}\right)$$

$$= 11.09 \pm 1.96(0.09)$$

$$= 11.09 \pm 0.18$$

95% confidence interval for the mean of the maximum loads of all cables produced by the company is $10.91 \le \mu \le 11.27$

b) 99% confidence interval for the mean of the maximum loads of all cables produced by the company

$$CI = \bar{x} \pm Z_c\left(\frac{s}{\sqrt{n}}\right)$$

$$= 11.09 \pm Z_{0.99}\left(\frac{0.73}{\sqrt{60}}\right)$$

$$= 11.09 \pm 2.58(0.09)$$

$$= 11.09 \pm 0.23$$

99% confidence interval for the mean of the maximum loads of all cables produced by the company $10.86 \le \mu \le 11.32$

Summary

The meanings of confidence level, margin of errors and critical values are concepts we have discussed in this unit. Finding the critical value corresponding to a given confidence level was discussed. Computation of confidence intervals for mean when population standard deviation is known and when the population standard deviation is unknown but the sample size is greater or equal to 30, also interpretation of result was done.

Self-Assessment Questions

1. A random sample of size 200 is drawn from a population known to have a standard deviation 5.2. If the mean of the random sample is 345.6.
(a) Construct a 90% confidence interval for the sample population mean
(b) Construct a 95% confidence interval for the sample population mean
(c) Construct a 99% confidence interval for the sample population mean

2. A random sample is drawn from a population of known standard deviation $\sigma = 22.1$. Construct 95% interval for the population mean if

(a) $n = 121$, $\bar{x} = 82.4$

(b) $n = 81$, $\bar{x} = 82.4$

3. A random sample of 40 farms in a zone gave a mean of #7.56 per 100 pounds of oranges. Assume that σ is known to #1.34 per 100 pounds

(a) Find the 99% confidence interval for the population mean price (per 100 pounds) that all farms in the zone get from oranges.

(b) Find the 90% confidence interval for the population mean price (per 100 pounds) that all farms in the zone get from oranges.

Tutor Marked Assessment

- Suppose that 8% of all males suffer some form of colour blindness. Find the probability that in a random sample of 250 men at least 10% will suffer some form of colour blindness. First verify that the sample is sufficiently large to use the normal distribution.

- An outside financial auditor has observed that about 4% of all documents he examines contain an error of some sort. Assuming this proportion to be accurate, find the probability that a random sample of 700 documents will contain at least 30 with some sort of error. You may assume that the normal distribution applies.

1. A random sample of size 400 is drawn from a population known to have a standard deviation 2.2. If the mean of the random sample is 765.6,
(a) Construct a 90% confidence interval for the sample population mean
(b) Construct a 95% confidence interval for the sample population mean
(c) Construct a 99% confidence interval for the sample population mean
2. A random sample of size 256 is drawn from a population with mean and standard deviations unknown. The sample mean $\bar{x} = 1011$ and sample standard deviation $s = 34$
(a) Construct a 90% confidence interval for the sample population mean
(b) Construct a 95% confidence interval for the sample population mean
3. The mean and standard deviation of the diameter of a sample of 250 rivet heads manufactured by a company are 0.72642 and 0.00058 respectively.
(a) Find the 98% confidence interval for the mean diameter of all the rivet heads manufactured by the company.
(b) Find the 80% confidence interval for the mean diameter of all the rivet heads manufactured by the company.

References

- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.

- Spiegel M.R and Stephens L.J (1999). Schaum's Outlines of Theory and Problems of Statistics. McGraw-Hill, New York

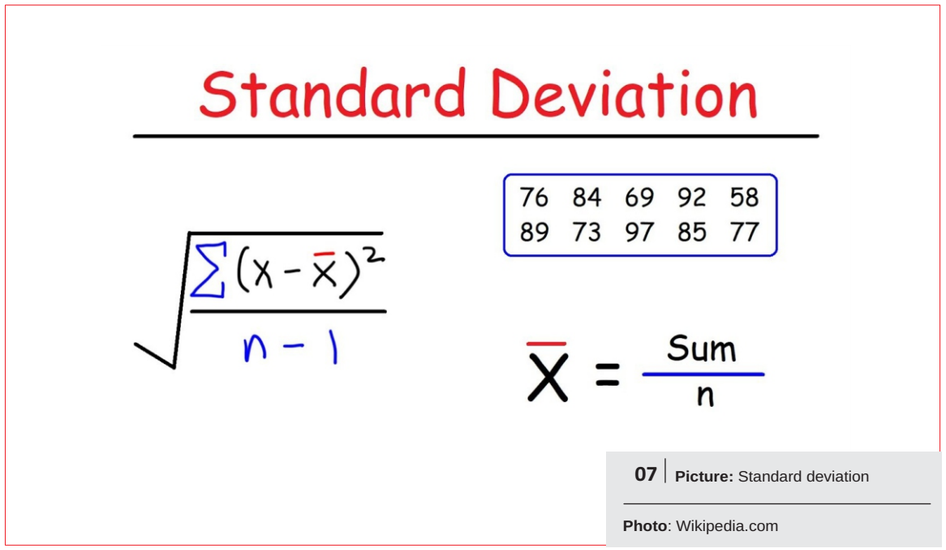
- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA

Further Reading

- https://www.britannica.com/science/statistics/Estimation-of-a-population-mean

- https://saylordotorg.github.io/text_introductory-statistics/s11-01-large-sample-estimation-of-a-p.html

- https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/07%3A_Estimation/7.1%3A_Large_Sample_Estimation_of_a_Population_Mean



07 | **Picture:** Standard deviation

Photo: Wikipedia.com

UNIT 5

Interval Estimation: Small sample Estimation of a population Mean

Introduction

In this study unit, we shall turn our attention to estimating the population mean when the sample size is small (less than 30). Focus will be on situation when the population standard deviation is known and unknown.

Learning Outcomes

At the end of this unit, you should be able to:

1. become familiar with student's t distribution for interval estimation of the population mean with Small sample size.
2. apply formulas for a confidence interval for a population mean for situation of known and unknown population standard deviation

Main Content

5 mins

When the population mean μ is estimated with small sample (less than 30) size, the central limit theorem does not apply. It is assumed that the numerical population from which the sample are taken has a normal distribution if the population standard deviation is known. The formula for the confidence interval $\bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$ can still be used to construct the confidence interval for the population mean μ.

If the population standard deviation is unknown and the sample size n is small (less than 30), then, the sample standard deviation s is substituted for the population standard deviation σ and a student's t distribution with n-1 degree of freedom is used instead of the normal distribution.

The confidence intervals are given as

$\bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$, if population standard deviation σ is known

$\bar{x} \pm t_{c,(n-1)}\left(\frac{s}{\sqrt{n}}\right)$, if population standard deviation σ is unknown

Example 1: A random sample of 12 physician revealed that they have mean of 22.8 rounds of golf play per year. Assume that the number of rounds is normally distributed with a standard deviation of 7. Estimate the 95% confidence interval for the mean number of rounds per year played by physician.

Solution

n=12, $\bar{x} = 22.8, \sigma = 7$

95% confidence interval for the sample population mean

$$CI = \bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$$
$$= 22.8 \pm Z_{0.95}\left(\frac{7}{\sqrt{12}}\right)$$
$$= 22.8 \pm 1.96(2.02)$$
$$= 22.8 \pm 3.9592$$

The 95% confidence interval of the sample population mean is with the interval $18.8408 \le \mu \le 26.7592$

Example 2: It is known that the amount of time needed to change the oil on a car is normally distributed with a standard deviation of 5 minutes. A random sample of 100 oil changes yields a sample mean of 22 minutes. Compute the 99% confidence interval estimate of the population mean.

Solution

n=100, $\bar{x} = 22, \sigma = 5$

99% confidence interval for the sample population mean

$$CI = \bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$$
$$= 22 \pm Z_{0.99}\left(\frac{5}{\sqrt{100}}\right)$$
$$= 22 \pm 2.575(0.5)$$
$$= 22 \pm 1.2875$$

The 95% confidence interval is $20.7125 \le \mu \le 23.2875$

Example 3: The marks 9 students on a statistics quiz gave a mean of 6.9 and standard deviation 2.4. Compute the 90% confidence interval for the mean score.

Solution

n=9, $\bar{x} = 6.9, s = 2.4$

90% confidence interval for the sample population mean

$$CI = \bar{x} \pm t_{0.90,(9-1)}\left(\frac{s}{\sqrt{n}}\right)$$
$$= 6.9 \pm t_{0.90,8}\left(\frac{2.4}{\sqrt{9}}\right)$$
$$= 6.9 \pm 1.397(0.8)$$
$$= 6.9 \pm 1.1176$$

The 90% confidence interval is $5.7824 \le \mu \le 8.0176$

Example 4: The mean number of hours used by 20 professors to attend departmental meetings is 9.85 hours per year. Assuming that the time follows a normal distribution with standard deviation 3 hours, construct the

(a) 90% confidence interval of the population mean

(b) 99% confidence interval of the population mean

Solution

n=20, $\bar{x} = 9.85, \sigma = 3$

(a) 90% confidence interval of the population mean

$$CI = \bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$= 9.85 \pm Z_{0.90}\left(\frac{3}{\sqrt{20}}\right)$$

$$= 9.85 \pm 1.645(0.671)$$

$$= 9.85 \pm 1.104$$

The 90% confidence interval is $8.746 \le \mu \le 10.954$

(b) 99% confidence interval of the population mean

$$CI = \bar{x} \pm Z_c\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$= 9.85 \pm Z_{0.99}\left(\frac{3}{\sqrt{20}}\right)$$

$$= 9.85 \pm 2.575(0.671)$$

$$= 9.85 \pm 1.728$$

The 99% confidence interval is $8.122 \le \mu \le 11.578$

Summary

In this unit, you have learnt about the computation of confidence intervals for mean when the population standard deviation is known and when the population standard deviation is unknown but the sample size is less than 30. The interpretation of the results was also done.

Self-Assessment Questions

1. The ages of 8 men in a bar gave a mean of 43.75. The population standard deviation is 10. Determine the 95% percent confident interval for the population mean.

2. The number of cars sold annually by used car dealer is normally distributed with a standard deviation of 15. A random sample of 400 dealers was taken and the mean number of cars sold annually was found to be 75. Find the 95% estimate of the population mean.

3. A random sample of 400 university of Ilorin students who lives off campus were asked the distance from home to school. The average distance was 8.84 kilometres with standard deviation 2.70 kilometre. Construct a 99% confidence interval for the mean distance from home to school.

4. An Estate manager wishes to estimate the average length of time a tenant remain in the same apartment before moving out. A random sample of 300 rentals gave a mean of length of occupancy of 4.8 years with standard deviation 0.8 years. Construct a 95% confidence interval for the mean length of occupancy of apartment.

5. The number of hours per day that children watch cartoon gave a mean of 6.4 for 150 selected households with a standard deviation of 0.85. Construct a 99% confidence interval for the mean hour of watching cartoon.

Tutor Marked Assessment

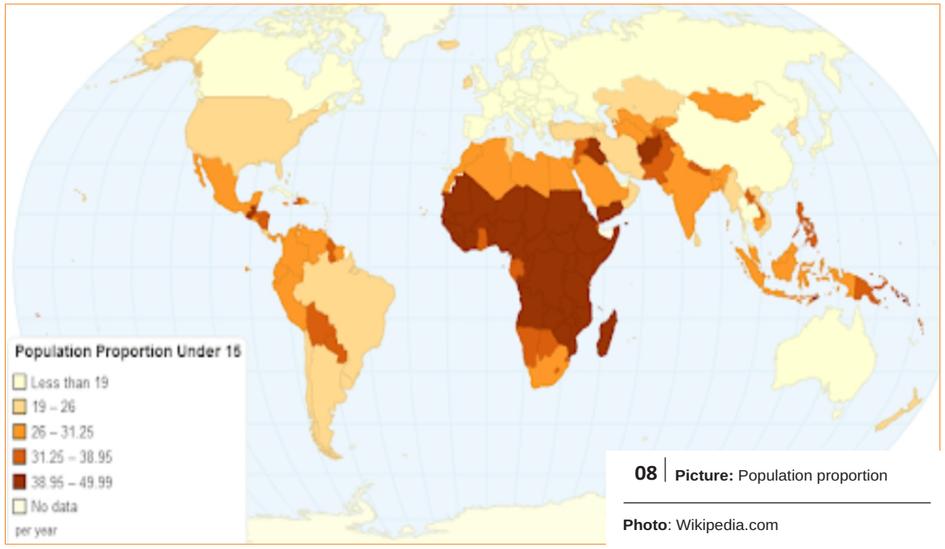
- The ages of 15 fresh students in a department gave a mean of 18.3 and a standard deviation 2.4. Determine the 95% percent confident interval for the population mean.
- The amount of time undergraduate students spend weekly on part time jobs is normally distributed with standard deviation 20 minutes. A random sample of 18 undergraduate students drawn gave a mean of 125 minutes. Find the 90% confidence interval estimate of the population mean.
- A sample of 10 electric bulbs produced by a company gave a mean life length of 1200h and a standard deviation 100h. Find the 90% confidence interval for the population mean
- The mean and standard deviation of the diameter of sample of 250 bowls manufactured by a plastic company are 0.72642 inches and 0.00058 inches respectively. Find (i) 99% (ii) 95% (iii) 90% confidence interval for the mean diameter of all the bowls manufactured by the company.

References

- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.
- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA
- Spiegel M.R and Stephens L.J (1999). Schaum's Outlines of Theory and Problems of Statistics. McGraw-Hill, New York

Further Reading

- https://saylordotorg.github.io/text_introductory-statistics/s11-01-large-sample-estimation-of-a-p.html
- https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/07%3A_Estimation/7.1%3A_Large_Sample_Estimation_of_a_Population_Mean



08 | **Picture:** Population proportion

Photo: Wikipedia.com

UNIT 6

Interval Estimation: Large sample Estimation of a population Proportion

Introduction

In unit three of module 1, I presented the mean, standard deviation and sampling distribution of the sample proportion, the approach of the study four and five can be applied to produce a confidence interval for a population proportion. This will be discussed in this unit.

Learning Outcomes

At the end of this unit, you should be able to:

1. highlight the procedure for applying the formula for a confidence interval for a population proportion; and
2. interpret the confidence interval for a population proportion.

Main Content

5 mins

The approach of the large sample is applied. It is assumed that the population follows a normal distribution. Suppose a large random samples of size n are drawn from a population in the proportion of the characteristic of interest is p. the mean μ_p and the standard deviation σ_p of the sample proportion is given as $\mu_p = P$ and $\sigma_p = \sqrt{\frac{pq}{n}}$, the confidence interval for the population proportion is given by

$$CI = \hat{p} \pm Z_c \sqrt{\frac{pq}{n}}, \text{ w}$$

here p is the proportion of characteristic of interest and $q=1-p$

Example 1: In a random sample of 500 eligible voters in the next election, 275 of them said they will vote the incumbent president. Find the 90% confidence interval for the proportion of eligible voters who will vote the incumbent president.

Solution

$n = 500, x = 275$

$$p = \frac{x}{n} = \frac{275}{500} = 0.55,$$

$$q = 1 - p = 1 - 0.55 = 0.45$$

The confidence interval is

$$CI = \hat{p} \pm Z_c \sqrt{\frac{pq}{n}}$$

$$= 0.55 \pm z_{0.90} \sqrt{\frac{0.55 \times 0.45}{500}}$$

$$= 0.55 \pm 1.645 \times 0.022$$

$$= 0.55 \pm 0.036$$

The 90% confidence interval for the proportion of eligible voters who will vote the incumbent president is $0.514 \le p \le 0.586$

Example 2: A psychologist believes that a large percentage of male drivers do not ask questions when they are lost. A sample of 350 driver was interviewed showed that 80% of male drivers when lost continue to drive hoping to find the location they seek rather than ask question. Find a point estimate of the proportion and the 99% confidence interval for the proportion of drivers.

Solution

$n = 350, p = 0.8, q = 1 - p = 0.2$

The confidence interval is

$$CI = \hat{p} \pm Z_c \sqrt{\frac{pq}{n}}$$

$$= 0.80 \pm z_{0.99} \sqrt{\frac{0.80 \times 0.20}{350}}$$

$$= 0.80 \pm 2.575 \times 0.021$$

$$= 0.80 \pm 0.05$$

The 99% confidence interval for the proportion of drivers who do not ask questions when they are lost is $0.75 \le p \le 0.85$

Example 3: 50% of a total of 600 customers of a bookshop sampled said they are satisfied with their services and prices. Find the 99% percent confidence interval.

Solution

$n = 600, p = 0.5, q = 1 - p = 0.5$

The confidence interval is

$$CI = \hat{p} \pm Z_c \sqrt{\frac{pq}{n}}$$

$$= 0.50 \pm z_{0.99} \sqrt{\frac{0.50 \times 0.50}{600}}$$

$$= 0.50 \pm 2.575 \times 0.02$$

$$= 0.50 \pm 0.05$$

The 99% confidence interval is $0.45 \le p \le 0.55$

Summary

In this unit, we carried out the computation of confidence intervals for population proportion and interpretation of result was done.

Self-Assessment Questions

1. In a random sample of 250 employed people, 61 said that they bring work home with them at least occasionally.

a. Give a point estimate of the proportion of all employed people who bring work home with them at least occasionally.

b. Construct a 99% confidence interval for that proportion.

2. In order to estimate the proportion of students who graduate within four years, the Administration of University of Ilorin examined the records of 600 randomly selected students who entered the university four years ago, and found that 312 had graduated.

a. Give a point estimate of the six-year graduation rate, the proportion of entering students who graduate within six years.

b. Assuming that the sample is sufficiently large, construct a 98% confidence interval for the six-year graduation rate.

3. A popular online electronic retailer claims that 95% of all orders are shipped within 24 hours after the orders are placed. Out of 440 orders placed at different times, 371 orders were shipped with 24 hours. Construct a 90% confidence interval of the proportion of order placed within 24 hours by the online electronic retailer.

4. Samsung is a leading smartphone. In a random sample of 900 adults, 420 of them uses a Samsung smart phone. Give a point estimate of the proportion of adults using Samsung smart phone and construct a 95% confidence interval for the number of adults using Samsung smart phone.

Tutor Marked Assessment

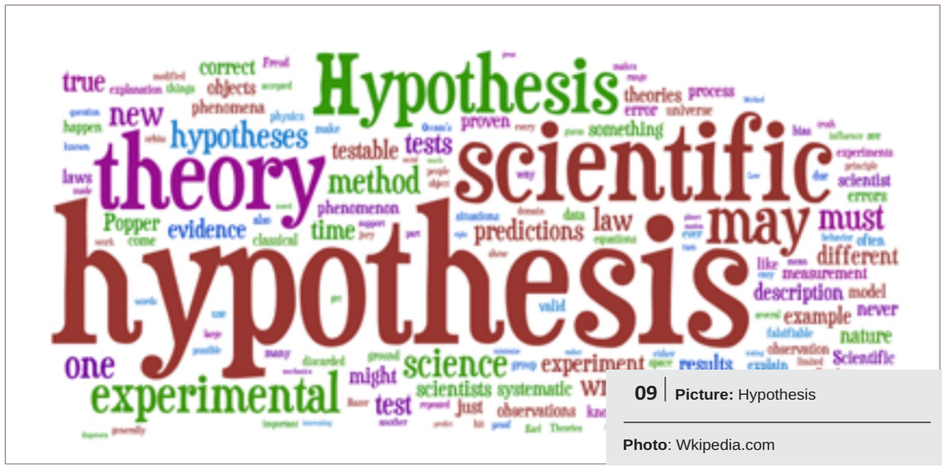
- From a particular brand of paracetamol, 95% of 400 samples of people who used it for relieve of headache said that they had relieve within two minutes of its usage. Find the 99% confidence interval for the proportion of people who get relived from headache within two minutes of its usage.
- 16 out of a sample of 800 components drawn from an assembly line that produces electronic for a missile system was defective. Construct the 95% confidence interval for the population proportion of the defective items.
- 30% of a sample of 300 students attending public secondary school in a city had access to computer at home. Construct a 90% confidence interval for the population proportion of public secondary school who have access to computer at home.

References

- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.
- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA
- Spiegel M.R and Stephens L.J (1999). Schaum's Outlines of Theory and Problems of Statistics. McGraw-Hill, New York

Further Reading

- https://saylordotorg.github.io/text_introductory-statistics/s11-01-large-sample-estimation-of-a-p.html
- https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/07%3A_Estimation/7.3%3A_Large_Sample_Estimation_of_a_Population_Proportion
- https://courses.lumenlearning.com/odessa-introstats1-1/chapter/a-population-proportion/



09 | **Picture:** Hypothesis

Photo: Wkipedia.com

UNIT 7

Basic Concepts of Test of Hypothesis

Introduction

In the preceding lessons, you have been able to learn that a major focus of statistics is to draw inference about a population based on information from samples taken from the population. Specifically, drawing inference about the population parameter was discussed in module 2. However, decision concerning the parameter of the population on the basis of the sample taken has to be made. Such decisions are called statistical decisions. In order to take the decision, the test of hypothesis is carried out on the estimated value of the population parameter from the sample. This is what we shall be discussing in this unit.

Learning Outcomes

At the end of this unit, you should be able to:

1. ihighlight the rationale of tests of hypotheses;
2. idiscuss the basic concepts in relation with hypothesis testing;
3. identify the null and alternative hypothesis;
4. identify one-tailed and two-tailed test;
5. identify the types of errors in hypothesis testing; and
6. discuss the steps in carrying out the test of hypothesis.

Main Content

Definition of concepts

8 mins

A hypothesis is a statement about the value of a population parameter which may be true or not.

Null Hypothesis

It is denoted by H_0, it is a hypothesis formulated about the population parameter for the purpose of rejecting, it is often called the working hypothesis. The statement about the null hypothesis is assumed to be true of the population parameter unless there is a convincing evidence to the statement. For example, a pharmaceutical company that produce a type of paracetamol claims that the average time it takes a person suffering from headache to recover after taking the type of paracetamol is 3 minutes, the null hypothesis is stated as $H_0: \mu = 3$.

A commercial for a household appliance manufacturer claims that 5% of all its products requires a service call in the first year of purchase, the null hypothesis is stated as $H_0: p = 0.05$.

Alternative Hypothesis

It is denoted by H_1, it is a statement about the population parameter that contradicts the null hypothesis. It is accepted as true if there is convincing evidence in favour of it which leads to the rejection of the null hypothesis called the working hypothesis. From the examples mentioned under the null hypothesis, if it is believed that the average time it takes a person suffering from headache to recover after taking the type of paracetamol is above 3 minutes, then, the alternative hypothesis is stated as $H_1: \mu > 3$. Also, if the claim of the household appliance manufacturer is believed to be less than 5%, the alternative hypothesis is stated as $H_1: \mu < 0.05$.

Test of hypothesis

It is a statistical procedure in which a choice is made between the null hypothesis and the alternative hypothesis based on information from a sample. The result of hypothesis test could be either to reject the null hypothesis or fail to reject the null hypothesis.

Types of tests in hypothesis testing

The null hypothesis always states that the population parameter is equal to a specified value while the alternative hypothesis can either states that the population parameter is less than, greater than or not equal to the specified value. The direction of the alternative hypothesis gives the type of the hypothesis being tested. The types of hypothesis test is categorised into two:

One-Tailed Test

If the alternative hypothesis is directional such that it has the form $H_1: \mu < \mu_0$, it is called **left tailed** test, but if the alternative hypothesis is directional such that it has the form $H_1: \mu > \mu_0$, it is called **right tailed** test. Any of the two forms; left or right tailed is referred to as one tailed test.

Two-Tailed Test

If the alternative hypothesis is directional such that it has the form $H_1: \mu \neq \mu_0$, it is called two tailed test.

Types of errors in Hypothesis Testing

There are two types of errors that could arise in taking decision about the population parameter using the hypothesis testing

Type I Error: it is the error committed in rejecting the null hypothesis when in fact it is true.

Type II Error: it is the error committed in failing to reject the null hypothesis when in fact it is not true.

Table 7: Type of error against decision taken

Decision Taken		True state of nature	
		H_0 is true	H_0 is not true
	Do not reject H_0	Correct Decision	Type II Error
	Reject H_0	Type I Error	Correct Decision

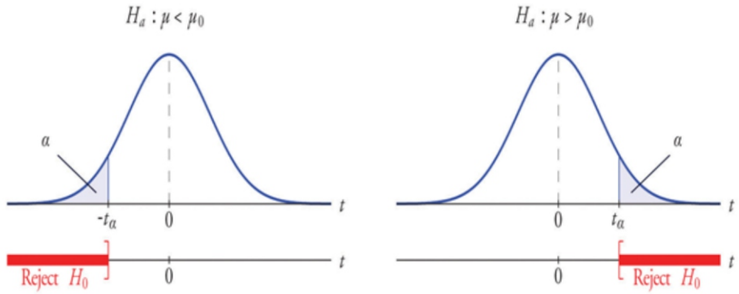
Standardized Test Statistic

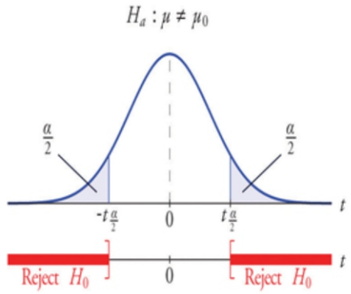
It is the statistic formed by subtracting from the statistic of interest its mean and dividing the difference by its standard deviation. It is a function of the sample data on which decision is taken.

Critical/Rejection Region

It is the set of all test statistic values for which the null hypothesis H_0 will be rejected.

Critical value(s) of a test of hypothesis are the number or numbers that determine the rejection region. Figure 1 shows the probability that the estimate of the population parameter takes a value in an interval is the area under its density curve and above that interval.





Level of Significance α

It is the value used to determine the rejection region. It is the probability that the test procedure will result in Type I error. The significance level is often determined before the samples are drawn so that the results obtained are not influenced by its choice.

Procedure for Testing Hypothesis

1) State the null and alternative hypothesis
2) Identify the relevant statistic and its distribution
3) Compute from the value of the test statistic from the available data
4) Construct the rejection region
5) Compare the value computed in step 3 to the rejection region constructed in step 4
6) Make a decision and interpretation of the result.

Example 1: The manufacturer of Five Alive fruit juice claims that the average quantity of juice inside a pack is 1 litre. Set up the statement of hypothesis to test the claim of the manufacturer.

Solution

The statement for the null hypothesis is what is manufacturer is trying to test, therefore, if μ is the average quantity of juice inside a pack, the manufacturer want to test whether $\mu = 1\ ltr$, the null hypothesis will take the form

$H_0 : \mu = 1$

The statement for the alternative hypothesis. The manufacturer is testing whether the average quantity of juice inside a pack is less or greater than 1 litre. Therefore, alternative hypothesis would be of the form

$H_1 : \mu \neq 1$

Summary

The rationale of tests of hypotheses and the basic concepts in relation with hypothesis testing are the concepts I presented to you in this unit. We were also able to achieve the identification of both the null and alternative hypothesis, one-tailed and two-tailed test and the types of errors in hypothesis testing. Lastly, I taught you the steps in carrying out the test of hypothesis were given. You can therefore assess yourself with the following questions.

Self-Assessment Questions

1. Which of the types of hypothesis determines the choice of One-Tailed or Two-tailed test?

2. If the null hypothesis is rejected, does it mean that the null hypothesis is false beyond doubt?

3. If the null hypothesis is not rejected, does it mean that the null hypothesis is true beyond doubt?

4. An industry association asserts that the average age of all self-described fly fishermen is 42.8 years. A sociologist suspects that it is higher.

Tutor Marked Assessment

State the null and the alternative hypothesis for the following situations.

- The average weight of a female airline passenger with luggage was 145 pounds ten years ago. The FAA believes it to be higher now.
- The average yield per acre for all types of corn in a recent year was 161.9 bushels. An economist believes that the average yield per acre is different this year.
- The average farm size in a predominately rural state was 69.4 acres. The secretary of agriculture of that state asserts that it is less today.
- The administration believes that average age of fresh students is 17 years. A professor of statistics believes that the average age is less in the new session.

References

- Brase C. H & Brase C.P (2007) Understanding Basic Statistics. Houghton Mifflin, USA
- Spiegel M.R and Stephens L.J (1999). Schaum's Outlines of Theory and Problems of Statistics. McGraw-Hill, New York

Further Reading

- https://en.wikipedia.org/wiki/Statistical_hypothesis_testing
- https://www.wisdomjobs.com/e-university/research-methodology-tutorial-355/basic-concepts-concerning-testing-of-hypotheses-11524.html
- https://www.scribd.com/document/29584617/Basic-Concepts-of-Hypothesis-Testing-1

Population Mean

$$\mu = \frac{\sum x}{N}$$

Sample Mean

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

10 | **Picture:** Population mean

Photo: Wikipedia .com

UNIT 8

Test for a population Mean

Introduction

This unit we will be focusing on applying the procedures for testing hypothesis about the population mean when the sample size is at least 30 (large sample size) and below 30 (small sample). Interpretation of the test of hypothesis will also be done.

Learning Outcomes

At the end of this unit, you should be able to:

1. apply the five-step test procedure for a test of hypotheses concerning a population mean when the sample size is large;
2. apply the five-step test procedure for test of hypotheses concerning a population mean when the sample size is small; and
3. interpret the result of a test of hypotheses.

Main Content

Large sample Tests for a population mean | 8 mins

Conducting a test of hypothesis about the mean of a population when the sample size is at least 30$(n \geq 30)$. There are two cases that could arise which are:

Testing the population mean when the population standard deviation σ is known

If x(variable of interest) is normally distributed, the central limit theorem states that the $\bar{x}$ has a mean $\mu_x = \bar{x}$ and a standard deviation $\sigma_x = \frac{\sigma}{\sqrt{n}}$ where μ_x and σ_x are the population mean and standard deviation respectively. Applying the procedure for hypothesis testing, then the test of the population mean when the population standard deviation σ is known follows as:

1) State the null and the alternative hypothesis and set the significance level α
2) The relevant test statistic is given as

$$z = \frac{\bar{x} - \mu_x}{\sigma/\sqrt{n}}$$

3) Using the test statistic, compute its value
4) Use the standard normal distribution to construct the rejection region
5) Compare the value computed in step 3 to the rejection region constructed in step 4
6) Make a decision and interpretation of the result.

Example 1: It has been found out from experience that the mean breaking strength of a particular brand of thread is 10.72 oz with a standard deviation of 1.4 oz. A sample of 36 pieces of thread gave a mean of 9.93 oz. Test at (i) 5% and (ii) 1% level of significance that the thread breaking strength has reduced.

Solution

Following the procedure for Test of hypothesis
The null and alternative hypothesis is

$H_0: \mu = 10.72$

$H_1: \mu < 10.72$

The population standard deviation is known and the sample is large, therefore, the test statistic is given as

$$z = \frac{\bar{x} - \mu_x}{\sigma/\sqrt{n}}$$

$$= \frac{9.93 - 10.72}{1.4/\sqrt{36}} = -3.3862$$

(i) 5% level of significance that the thread breaking strength has reduced.

$z_{0.05} = -1.645$

Decision: Since computed value of the test statistic is less than the critical value, the null hypothesis is rejected.

Conclusion: It is therefore concluded that the thread breaking strength has reduced at 5% level of significance

(ii) 1% level of significance that the thread breaking strength has reduced.

$Z_{0.01} = -2.33$

Decision: Since computed value of the test statistic is less than the critical value, the null hypothesis is rejected.

Conclusion: It is therefore concluded that the thread breaking strength has reduced at 1% level of significance

Example 2: The mean glucose level (in mg/100mh) taken from a horse for 8 different weeks was 93.8. If it is known from past experiences that the standard deviation $\sigma = 12.5$ and mean $\mu = 85$. Test at 5% level of significance if the mean glucose level for the horse is higher than 85 mg/100mh.

Solution

The null and alternative hypothesis is

$H_0: \mu = 85$

$H_1: \mu > 85$

The population standard deviation is known and the sample is large, therefore, the test statistic is given as

$$z = \frac{\bar{x} - \mu_x}{\sigma/\sqrt{n}}$$

$$= \frac{93.8 - 85}{12.5/\sqrt{8}} = 4.3806$$

5% level of significance that the mean glucose level for the horse is higher than 85 mg/100mh.

$$z_{0.05} = 1.645$$

Decision: Since computed absolute value of the test statistic is greater than the critical value, the null hypothesis is rejected.

Conclusion: It is therefore concluded that the mean glucose level for the horse is higher than 85 mg/100mh at 5% level of significance.

Testing the population mean when the population standard deviation σ is unknown

If x(variable of interest) is normally distributed, the central limit theorem states that the $\bar{x}$ has a mean $\mu_x = \bar{x}$ but the population standard deviation σ is unknown but can be computed from the samples given, then the computed standard deviation s can be used. Applying the procedure for hypothesis testing, then the test of the population mean when the population standard deviation σ is known follows as:

1) State the null and the alternative hypothesis and set the significance level α
2) The relevant test statistic is given as

$$z = \frac{\bar{x} - \mu_x}{s/\sqrt{n}}$$

3) Using the test statistic, compute its value
4) Use the standard normal distribution to construct the rejection region
5) Compare the value computed in step 3 to the rejection region constructed in step 4
6) Make a decision and interpretation of the result.

Example 3: It is known that the mean life span of the resident of a state is 77 years. A random sample of 20 obituary notices was selected, the mean $\bar{x}$ age at death was 71.4 and the standard deviation s=20.65. Test at 1% level of significance that the mean life span has reduced from 77 years.

Solution

The null and alternative hypothesis is

$H_0: \mu = 77$

$H_1: \mu < 77$

The population standard deviation is unknown, therefore, the test statistic is given as

$$z = \frac{\bar{x} - \mu_x}{s/\sqrt{n}}$$

$$= \frac{71.4 - 77}{20.65/\sqrt{20}} = -10.217$$

1% level of significance that the mean life span has reduced from 77 years.

$$z_{0.01} = -2.33$$

Decision: Since computed value of the test statistic is less than the critical value, the null hypothesis is rejected.

Conclusion: It is therefore concluded that the mean life span has reduced from 77 years.

Example 4: A lecturer reports that the mean score of the student in a particular course is 76. A random sample of 20 students score gave a mean of 72 and standard deviation of 6. Test at 5% level of significance if there is enough evidence to support the lecturer's claim that the mean score is 76.

Solution

The null and alternative hypothesis is

$H_0: \mu = 76$

$H_1: \mu \neq 76$

The population standard deviation is unknown, therefore, the test statistic is given as

$$z = \frac{\bar{x} - \mu_x}{s/\sqrt{n}}$$

$$= \frac{72 - 76}{6/\sqrt{20}} = -2.9814$$

5% level of significance that the mean score is 76.

$$z_{0.025} = 1.96$$

Decision: Since computed the absolute value of the test statistic is greater than the critical value, the null hypothesis is rejected.

Conclusion: It is therefore concluded that the mean score is significantly different from 76

Summary

We applied the application of the procedure for carrying out statistical test of hypothesis when the number of sample size is at least 30 and below 30. Also, I made efforts to look at situations where the population standard deviation is known and unknown.

Self-Assessment Questions

1. A lecturer reports that the mean score of the student in a particular course is 76. A random sample of 20 students score gave a mean of 72 and standard deviation of 6. Test at 1% level of significance if there is enough evidence to support the lecturer's claim that the mean score is 76.

2. It has been found out from experience that the mean breaking strength of a particular brand of thread is 10.72 oz with a standard deviation of 1.4 oz. A sample of 72 pieces of thread gave a mean of 9.93 oz. Test at 5% level of significance that the thread breaking strength is not 10.72.

3. A soft drink company produces a 35cl bottle of a particular brand of soft drink. A random sample of 100 bottles of the brand of soft drink gave a mean of 34.8cl and standard deviation of 0.4cl.

(I). Can we infer at 1% level of significance that the population mean soft drink is less than 35cl?

(ii). Can we infer at 5% level of significance that the population mean soft drink is less than 35cl?

Tutor Marked Assessment

- It is reported that the mean and standard deviation of systolic blood pressure (SBP) of men is 125 and 14 respectively. A study of 100 sample of mean yield a mean systolic blood pressure (SBP) of 130. Test at 5% level of significance if the data suggest that the mean SBP for men is

  (I). Higher than 125.

  (ii). Not 125

- Quality control test is being conducted on drug. It is known that a single dosage of drug should contain 8mg of active ingredient. A random sample of 20 dosages gave a mean 7.7mg of active ingredient and standard deviation 2mg. Test at 5% level of significance if the data suggest that the mean of active ingredient in all dosage produced is different from 8mg

- A university claim that the average age of newly admitted students into the undergraduate programme is 18 years. The age of a random sample of 200 newly admitted students gave a mean of 17.8 and standard deviation 0.8. Can we conclude at 5% level of significance that the population mean age of the newly admitted students is less than 18 years?

References

- Brase C. H & Brase C.P (2007) Understanding Basic Statistics. Houghton Mifflin, USA

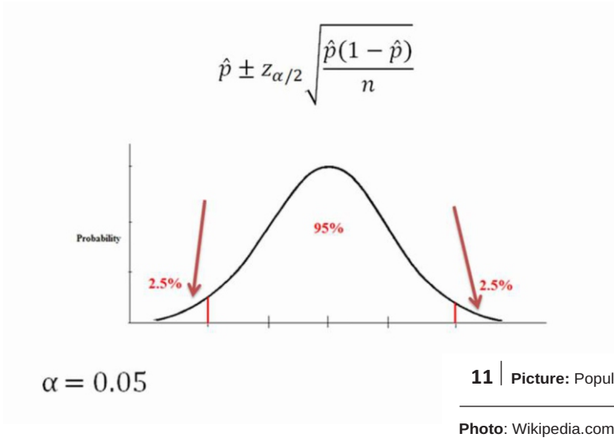
- Spiegel M.R and Stephens L.J (1999). Schaum's Outlines of Theory and Problems of Statistics. McGraw-Hill, New York

Further Reading

- https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/hypothesis-test-for-a-population-mean-1-of-5/

- https://stattrek.com/hypothesis-test/mean.aspx

https://www.econometrics-with-r.org/3-3-hypothesis-tests-concerning-the-population-mean.html



11 | **Picture:** Population proportion

Photo: Wikipedia.com

UNIT 9

Test for a population proportion

Introduction

In this unit, you will learn how to apply the procedures for testing hypothesis about the population proportion when the sample size is large enough.

Learning Outcomes

At the end of this unit, you should be able to:

1. to apply the five-step test procedure for a test of hypotheses concerning a population proportion.
2. interpret the result of a test of hypotheses

Main Content

6 mins

In many situations of practical experience, the interest is to test the proportions or percentages rather than the means.

Testing the proportion

If p is the proportion of characteristic of interest from a sample size of n taking from a population. Applying the procedure for hypothesis testing, then the test of the population proportion is known follows as:

1) State the null and the alternative hypothesis and set the significance level
2) The relevant test statistic is given as

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

where p is the value specified in null hypothesis and $q=1-p$

3) Using the test statistic, compute its value
4) Use the standard normal distribution to construct the rejection region
5) Compare the value computed in step 3 to the rejection region constructed in step 4
6) Make a decision and interpretation of the result.

Example 1: A manufacturer of soft drink claims that majority of adults prefers its leading beverages over that of its main competitors'. To test this claim, 500 randomly selected people were given the two beverages in random order to taste. Among them, 270 preferred the soft drink makers brand, 211 preferred the competitors' brand and 19 could not make up their mind. Determine whether there is sufficient evidence at 5% level of significance to support the soft drink maker's claim against the default that the proportion is evenly split in its preference.

Solution

$$p = \frac{270}{500} = 0.54, q = 1 - p = 0.46$$

The null and the alternative hypothesis is stated as follows

$H_0 : p = 0.5$

$H_1 : p > 0.5$

Significance level $\alpha = 0.05$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

$$= \frac{0.54 - 0.50}{\sqrt{\frac{0.5 \times 0.5}{500}}} = 1.789$$

The Critical value $z_{0.05} = 1.96$

Decision Rule: reject the null hypothesis if the value of the test statistic is greater than the critical value.

Decision: since 1.789 is less than 1.96, the null hypothesis is not rejected, therefore, it is concluded that the data provide sufficient evidence at 5% level of significance that a majority of adult do not prefer the company's beverage to their competitors'.

Example 2: A pharmaceutical company claims that 98% of people who take her newly produced drug will be relieved from headache after five minutes of taking the drug. Out of a random sample of 600 people suffering from headache that took the drug, 580 reported that they were relieved of headache within five minutes of administering it. Test the claim of the pharmaceutical company at 1% level of significance.

Solution

$$\hat{p} = \frac{580}{600} = 0.97, q = 1 - p = 0.02$$

The null and the alternative hypothesis is stated as follows

$H_0 : p = 0.98$

$H_1 : p \neq 0.98$

Significance level $\alpha = 0.01$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

$$= \frac{0.97 - 0.98}{\sqrt{\frac{0.98 \times 0.02}{500}}} = 1.67$$

The Critical value $z_{0.05} = 1.645$

Decision Rule: reject the null hypothesis if the value of the test statistic is greater than the critical value.

Decision: since 1.67 is greater than 1.645, the null hypothesis is rejected, therefore, it is the proportion that recover from headache within five minute of administration is not the same as the manufacturer claim

Summary

The application of the procedure for carrying out statistical test of hypothesis to test for the population proportion when the sample is large enough and interpretation of result is what I taught you in this unit, with the examples and explanations I gave, you should be able to asses yourself with the questions below.

Self-Assessment Questions

SAQ

1. The provost of a college claim that 86% of their final year students eventually graduate at the end of the session. A random sample of 200 final year student in the last academic session showed that 167 eventually graduated. Test at 5% level of significance if the proportion of students that graduated from the college is now less than 86%

2. A university security at the entrance of the computer-based test hall claims that 90% of the students usually throws away the GNS textbook after their examination. To test the security claim, 340 of 400 randomly selected GNS students said they threw away their textbook immediately after their examination. Test at 5% level of significance if the proportion of students that throw away their GNS textbook has reduced.

3. Just as you know that at present, the All progressive Congress (APC) is the leading party in Nigeria. The party leadership believed that 5% of the party member will be in favour of open ballot for the selection of candidate. A random sample of voters who identify with party in Kogi state were selected and asked if they favour an open ballot for the selection of candidate to represent their political party in the next gubernatorial election.

	APC
Sample size *n*	200
Number in favour *x*	14

Test at 5% level of significance, the hypothesis that the proportion of all members of APC who are in favour of open ballot is less than

Tutor Marked Assessment

- A popular shopping mall in a major city claims that 80% of their customer are within the age group 16-45 years. A random sample of 115 customer reveals that 88 are within the age group. Test at 5% level of significance if this indicate that less than 80% of the customers that patronise the shopping mall is within the age group 16-45?

- It is believed that about 82% of college students union leader are extroverts. A random sample of 73 student union leader reveals that 56 were extroverts. Test at 1% level of significance if this indicate that the proportion of student union leader that are extroverts is different?

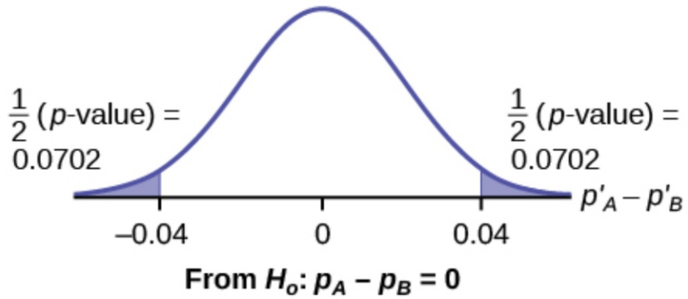
References

- Brase C. H & Brase C.P (2007) Understanding Basic Statistics. Houghton Mifflin, USA
- Spiegel M.R and Stephens L.J (1999). Schaum's Outlines of Theory and Problems of Statistics. McGraw-Hill, New York

Further Reading

- https://stattrek.com/hypothesis-test/proportion.aspx
- https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/hypothesis-test-for-a-population-proportion-3-of-3/
- http://www.stat.ucla.edu/~magtira/XL10/chapter8.pdf

Unit 9 | Module 2 | CSC 206 Test for a population proportion



12 | **Picture:** comparing two independent population proportion

Photo: Wikipedia.com

UNIT 10

Comparison between two independent population means

Introduction

In estimation and making inferences, the interest may be to consider that parameter of two populations. For instance, comparison may be made on the average income of all adults in one region of the country with the average income of all adults in another region of the same country. Our focus in this unit will centre on two population problems.

Learning Outcomes

At the end of this unit, you should be able to:

1. understand the logical framework for estimating the difference between the means of two independent populations and perform tests of hypotheses concerning those means.
2. learn to construct a confidence interval for the difference in the means of two independent populations.
3. learn to perform a test of hypotheses concerning the difference between the means of two independent populations.

Main Content

Comparison of two independent population means

8 mins

Suppose the interest is to compare the means of two independent population taken large samples from each of the population; say population 1 and population 2.

Note: Two population are said to independent of each other if samples are drawn from each population without reference to the other population.

Table 1: Parameter and statistic of the two populations

Parameter/Statistic	Population 1	Population 2
Population Mean	μ_1	μ_2
Population Standard Deviation	σ_1	σ_2
Sample size	n_1	n_2
Sample Mean	$\bar{x}_1$	$\bar{x}_2$
Sample Standard Deviation	s_1	s_2

Confidence interval for the difference between two independent populations (large sample)

The sample mean $\bar{x}_1$ is a good estimator of the population mean μ_1 and the sample mean $\bar{x}_2$ is a good estimator of population mean μ_2, the point estimate of the difference between the two population $(\mu_1 - \mu_2)$ is $(\bar{x}_1 - \bar{x}_2)$. The samples drawn from the two population are large, that is, $n_1, n_2 \geq 30$. If the population standard deviation for the two population are known, then, the confidence interval for the difference between the two-population means is given as

$$(\bar{x}_1 - \bar{x}_2) \pm z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

If the population standard deviation for the two population are unknown, then, the confidence interval for the difference between the two population means is given as

$$(\bar{x}_1 - \bar{x}_2) \pm z_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example 1: To compare customer satisfaction level of two competing cable television companies, 174 customers of company 1 and 335 customers of company 2 were randomly selected and were asked to rate their cable companies on a five-point scale, with 1 being satisfied and 5 most satisfied. The survey result is summarized below:

Company 1: $n_1 = 174, \bar{x}_1 = 3.51, s_1 = 0.51$

Company 2: $n_2 = 355, \bar{x}_2 = 3.24, s_2 = 0.52$

Construct a point estimate and a 95% confidence Interval for $\mu_1 - \mu_2$, the difference in average satisfaction level of customers of the two companies as measured on this five point scale.

Solution

The point estimate of $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 = 3.51 - 3.24 = 0.27$$

The average customer satisfaction for company 1 is 0.27 points higher on the five point scale than it is for company 2.

The confidence interval

$$CI = (\bar{x}_1 - \bar{x}_2) \pm z_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= 0.27 \pm z_{0.05} \sqrt{\frac{0.51^2}{174} + \frac{0.52^2}{355}}$$

$$= 0.27 \pm 0.12$$

$$= (0.15 \leq \mu_1 - \mu_2 \leq 0.39)$$

There is 99% confidence that the average level of customer satisfaction for company 1 is between 0.15 and 0.39 points higher on the five-point scale than that of company 2.

Confidence interval for the difference between two independent populations (small sample)

If samples taken from one or the two population is small $(n < 30)$, then the confidence interval is given as

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(n_1+n_2-2)\sigma/2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

where $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ is a pooled sample variance. The number of degree of freedom is $n_1 + n_2 - 2$.

Example 2: A software company markets a new computer game with two experimental packaging designs. Design 1 is sent to 11 stores; their average sales in the first month is 52 units with sample standard deviation is 12 unit. Design 2 is sent to 6 stores; their average for the first month is 46 with sample standard deviation 10 units. Construct a point estimate and a 95% confidence interval for the difference in average monthly sales between the two package designs.

Solution

The point estimate of $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 = 52 - 46 = 6$$

The average monthly sales for Design 1 is 6 units more per month than the average monthly sales for Design 2.

The degree of freedom $df = n_1 + n_2 - 2 = 11 + 6 - 2 = 15$

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$
$$= \frac{(10)(12)^2 + (5)(10)^2}{15} = 129.3$$

Thus,

$$CI = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$
$$= 6 \pm t_{0.025}\sqrt{129.3\left(\frac{1}{11}+\frac{1}{6}\right)}$$
$$= 6 \pm 12.3$$
$$= (-6.3, 18.3)$$

There is 95% confidence that the average for the first month for design 1 is between -6.3 and 18.3 the first month for design 2.

Comparison between two dependent population means (paired samples)

Two samples are said to be dependent of each other if each data value in one sample can be paired with a corresponding data value in the other sample. Its confidence interval is given as

$$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

Where there are n pairs, $\bar{d}$ is the mean and s_d is the standard deviation of the differences. The degree of freedom is $n-1$

Example 3: A random sample of 12 students were given a diagnostic test before teaching a topic and another test after teaching the topic. The scores of the students are given below. Construct a point estimate and a 95% confidence interval for the difference in the two scores.

	1	2	3	4	5	6	7	8	9	10	11	12
Before	68	44	30	58	35	33	52	69	23	69	48	30
After	59	42	20	62	25	30	56	62	25	75	40	26

SOLUTION

	1	2	3	4	5	6	7	8	9	10	11	12
Before	68	44	30	58	35	33	52	69	23	69	48	30
After	59	42	20	62	25	30	56	62	25	75	40	26
Difference d	9	2	10	-4	10	3	-4	7	-2	-6	8	4
d^2	81	4	100	16	100	9	16	49	4	36	64	16

The point estimate of the difference is

$$\sum d = 37,\ \bar{d} = \frac{37}{12} = 3.0833$$

$$s_d^2 = \frac{\sum d^2 - \left(\sum d\right)^2/n}{n-1}$$
$$= \frac{495 - 37^2/12}{11} = 34.6288$$

$$s_d = \sqrt{34.6288} = 5.8846$$

The confidence interval

$$CI = \bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

$$= 3.0833 \pm 2.2010 \frac{5.8846}{\sqrt{12}}$$

$$= 3.0833 \pm 3.7391$$

$$= \left(-0.6558 \leq \bar{d} \leq 6.8224\right)$$

Summary

The purpose of what you learnt in this unit is to use information samples to estimate difference in two population mean with emphasis on two independent population with large samples and small samples. We also looked at confidence interval for means of paired samples when the samples taken are not independent of each other.

Self-Assessment Questions

SAQ

1. The scores of 16 students from the Department of Mathematics for a particular course shows a 71 and a standard deviation of 8, while the scores of 14 of students from Department of Computer Science gave a mean of 67 and a standard deviation of 5. Construct a 99% confidence interval for the difference in the mean scores of the students from the two departments

2. A sample of 100 electric bulbs produced by manufacturer A showed a mean lifetime of 1190h and a standard deviation of 90h. A sample of 75 electric bulbs produced by manufacturer B showed a mean lifetime of 1230h and a standard deviation of 120h. Construct a 99% confidence interval for the difference in the mean lifetimes of the bulbs produced by the two manufacturer.

Tutor Marked Assessment

- A sample of 100 electric bulbs produced by manufacturer A showed a mean lifetime of 1190h and a standard deviation of 90h. A sample of 75 electric bulbs produced by manufacturer B showed a mean lifetime of 1230h and a standard deviation of 120h. Construct a 95% confidence interval for the difference in the mean lifetimes of the bulbs produced by the two manufacturer.

- The scores of 16 students from the Department of Mathematics for a particular course shows a 71 and a standard deviation of 8, while the scores of 14 of students from Department of Computer Science gave a mean of 67 and a standard deviation of 5. Construct a 90% confidence interval for the difference in the mean scores of the students from the two departments

- A random sample of 10 students were examined on a particular course through the manual based test (MBT) and computer based test (CBT). The scores of the students are given below. Construct a point estimate and a 95% confidence interval for the difference in the two scores.

	1	2	3	4	5	6	7	8	9	10
MBT	66	78	82	53	54	78	62	56	68	45
CBT	70	74	89	45	67	87	78	60	56	50

References

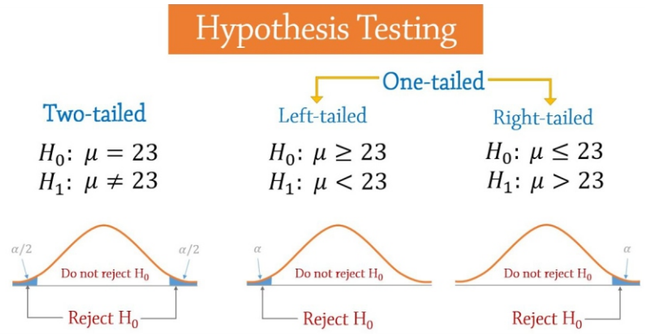
- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.
- Spiegel M.R and Stephens L.J (1999). Schaum's Outlines of Theory and Problems of Statistics. McGraw-Hill, New York

  Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA

Further Reading

- https://opentextbc.ca/introbusinessstatopenstax/chapter/comparing-two-independent-population-means/
- https://openstax.org/books/introductory-business-statistics/pages/10-1-comparing-two-independent-population-means

Unit 10 | Module 2 | CSC 206 Comparison between two independent population means



13 | **Picture:** Hypothesis test

Photo: Wikipedia.com

UNIT 11

Hypothesis Testing Concerning difference of two means

Introduction

In this study unit, we will be looking at tests of hypotheses concerning two independent populations (with large and small samples), paired observations will also be part of our focus.

Learning Outcomes

At the end of this unit, you should be able to:

1. perform a test of hypotheses concerning the difference between the means of two independent populations when the sample size is large
2. perform a test of hypotheses concerning the difference between the means of two independent populations when the sample size is small
3. perform a test of hypotheses concerning the difference between the means for paired sample.

Main Content

Hypothesis testing concerning difference between two independent populations (Large sample)

8 mins

In order for us to test for the significance difference between the two independent population means, the null hypothesis will be written in the form

$$H_0: \mu_1 - \mu_2 = \mu_d$$

where μ_d is a number deduced from the statement of the situation.

The alternative hypothesis can take one of the three forms depending on the statement of the situation.

$H_1: \mu_1 - \mu_2 < \mu_d$ Left-Tailed Test

$H_1: \mu_1 - \mu_2 > \mu_d$ Right-Tailed Test

$H_1: \mu_1 - \mu_2 \neq \mu_d$ Two-Tailed Test

Test statistic for the Hypothesis Testing concerning difference between two independent populations (Large sample)

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$ If population standard deviation for the two population is unknown

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ If population standard deviation for the two population is known

Example 1: To compare customer satisfaction level of two competing cable television companies, 174 customers of company 1 and 335 customers of company 2 were randomly selected and were asked to rate their cable companies on a five-point scale, with 1 being satisfied and 5 most satisfied. The survey result is summarized below:

Company 1: $n_1 = 174, \bar{x}_1 = 3.51, s_1 = 0.51$

Company 2: $n_2 = 355, \bar{x}_2 = 3.24, s_2 = 0.52$

Test at 1% level of significance whether the data provide sufficient evidence to conclude that company 1 has higher mean satisfaction rating than company 2.

Solution

To say that the mean customer satisfaction for company 1 is higher than that for company 2 means that $\mu_1 > \mu_2$, which in terms of their difference is $\mu_1 - \mu_2 > 0$, therefore the statement of hypotheses are stated below

The null hypothesis

$$H_0: \mu_1 - \mu_2 = 0$$

The alternative hypothesis

$$H_0: \mu_1 - \mu_2 > 0$$

The test statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{(3.51 - 3.24) - 0}{\sqrt{\frac{0.51^2}{174} + \frac{0.52^2}{355}}} = 5.684$$

The critical value $z_c = z_{0.01} = 2.326$

Decision Rule: reject the null hypothesis if the value of the test statistic is greater than the value of the critical region

Decision: since the value of the test statistic is greater than the value of the critical region, the null is rejected.

Conclusion: the data provide sufficient evidence at 1% level of significance to conclude that company 1 has higher mean satisfaction rating than company 2.

Test statistic for the Hypothesis Testing concerning difference between two independent populations (Small sample)

$$t=\frac{(\bar{x}_1-\bar{x}_2)-\mu_d}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where $s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ is a pooled sample variance. The number of degree of freedom is n_1+n_2-2.

Example 2: A software company markets a new computer game with two experimental packaging designs. Design 1 is sent to 11 stores; their average sales in the first month is 52 units with sample standard deviation is 12 unit. Design 2 is sent to 6 stores; their average for the first month is 46 with sample standard deviation 10 units. Test at 1% level of significance whether the data provide sufficient evidence to conclude that mean sales per month of the two design are different.

Solution

The statement of hypotheses are

The null hypothesis

$H_0: \mu_1-\mu_2=0$

The alternative hypothesis

$H_0: \mu_1-\mu_2\neq 0$

The degree of freedom $df=n_1+n_2-2=11+6-2=15$

$$s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$
$$=\frac{(10)(12)^2+(5)(10)^2}{15}=129.3$$

The test statistic

$$t=\frac{(\bar{x}_1-\bar{x}_2)-\mu_d}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$
$$=\frac{(52-46)-0}{\sqrt{129.3\left(\frac{1}{11}+\frac{1}{6}\right)}}=1.040$$

The critical value $t_{0.025}=2.131$

Decision: since the value of the test statistic is less than the critical value, the null hypothesis is not rejected.

Conclusion: the data do not provide sufficient evidence to conclude that mean sales per month of the two design are different at 1% level of significance

Hypothesis testing concerning difference between paired samples

Testing hypothesis concerning the difference of two population using paired difference sample is done as it is done for independent samples, the null hypothesis is expressed in terms of μ_d instead of $\mu_1-\mu_2$.

$H_0: \mu_d=d_0$

where d_0 is a number deduced from the statement of the situation.

The alternative hypothesis can take any of the form;

$H_1=\mu_d<d_0$ Left-Tailed Test

$H_1=\mu_d>d_0$ Right-Tailed Test

$H_1=\mu_d\neq d_0$ Two-Tailed Test

Test statistic for the Hypothesis Testing concerning difference between two paired samples

$$t=\frac{\bar{d}-d_0}{\frac{s_d}{\sqrt{n}}}$$

Example 3: A random sample of 12 students were given a diagnostic test before teaching a topic and another test after teaching the topic. The scores of the students are given below. Test a 5% level of significance that the efficiency of the new drug method of treating the particular topic.

	1	2	3	4	5	6	7	8	9	10	11	12
Before	68	44	30	58	35	33	52	69	23	69	48	30
After	59	42	20	62	25	30	56	62	25	75	40	26

Solution

	1	2	3	4	5	6	7	8	9	10	11	12
Before	68	44	30	58	35	33	52	69	23	69	48	30
After	59	42	20	62	25	30	56	62	25	75	40	26
Difference d	9	2	10	-4	10	3	-4	7	-2	-6	8	4
d^2	81	4	100	16	100	9	16	49	4	36	64	16

The point estimate of the difference is

$$\sum d = 37,\ \bar{d} = \frac{37}{12} = 3.0833$$

The statement of hypothesis

The null hypothesis

$$H_0 : \mu_d = d_0$$

The alternative hypothesis

$$H_1 = \mu_d \neq d_0$$

$$s_d^2 = \frac{\sum d^2 - \left(\sum d\right)^2 / n}{n-1}$$

$$= \frac{495 - 37^2/12}{11} = 34.6288$$

$$s_d = \sqrt{34.6288} = 5.8846$$

The test statistic

$$t = \frac{\bar{d} - d_0}{\frac{s_d}{\sqrt{n}}}$$

$$= \frac{3.0833 - 0}{\frac{5.8846}{\sqrt{12}}} = 1.8150$$

The critical value $t_{0.025,11} = 2.2010$

Decision: since the value of the test statistic is less than the critical value, the null hypothesis is not rejected.

Conclusion: the data does not provide sufficient evidence at the 5% level of significance that the scores are different

Summary

In this unit, I have taught you the application of the procedure of carrying out statistical test of hypothesis and also its application to difference of two populations with emphasis on situation when the population are independent of each other (large and small samples) and situation when the samples are dependent on each other (paired samples).

Self-Assessment Questions

1. The mean number of miles per gallon used by five Toyota corolla of Brand A gasoline is 22.6 with 0.48 standard deviation. Using Brand B gasoline, the number of gasoline used by the five Toyota corolla gave a mean of 21.4 and 0.54 standard deviation. Test at 5% level of significance if brand A gasoline is better than brand B gasoline.

2. Two types of chemical solutions, 1 and 2 were tested for their PH level. 60 samples from chemical solution 1 gave mean of 7.52 with standard deviation of 0.024. Analysis of 50 samples from chemical solution 2 gave a mean of 7.49 with a standard deviation of 0.032. Determine whether the two type of solutions have different PH values.

3. A random sample 12 chicks were given a particular type of feed, their weight (in g) before and after the administration of feed are given in the table below. Test at 1%level of significance if there is difference in the weight of the chicks.

	1	2	3	4	5	6	7	8	9	10	11	11
Before	73	68	73	71	71	72	68	58	74	79	64	76
After	70	74	78	76	67	81	67	63	70	72	89	79

Tutor Marked Assessment

- A random sample of 10 students were examined on a particular course through the manual based test (MBT) and computer based test (CBT). The scores of the students are given below. Test at 5% level of significance that there is difference in the scores of the students.

	1	2	3	4	5	6	7	8	9	10
MBT	66	78	82	53	54	78	62	56	68	45
CBT	70	74	89	45	67	87	78	60	56	50

- The scores of 16 students from the Department of Mathematics for a particular course shows a mean of 71 and a standard deviation of 8, while the scores of 14 of students from Department of Computer Science gave a mean of 67 and a standard deviation of 5. Test at 5% level of significance that the mean score of students in Mathematics Department is greater than that of students from Computer science Department.

- A sample of 100 electric bulbs produced by manufacturer A showed a mean lifetime of 1190h and a standard deviation of 90h. A sample of 75 electric bulbs produced by manufacturer B showed a mean lifetime of 1230h and a standard deviation of 120h. Test at 5% level of significance that there is difference between the lifetimes of the bulbs produced by the two manufacturers.

References

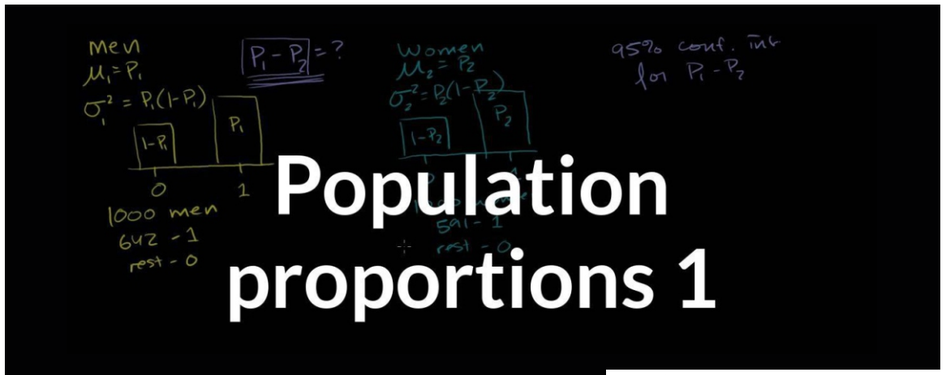
- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.

- Spiegel M.R and Stephens L.J (1999). Schaum's Outlines of Theory and Problems of Statistics. McGraw-Hill, New York

- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA

Further Reading

- https://stattrek.com/hypothesis-test/difference-in-means.aspx
- https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/hypothesis-test-for-a-difference-in-two-population-means-1-of-2/
- http://www.stat.yale.edu/Courses/1997-98/101/meancomp.htm



14 | **Picture:** Population proportion

Photo: Wikipedia.com

UNIT 12

Inferences about differences between two population proportions

Introduction

In estimation and making inferences, our interest may be to consider two populations have a specific characteristic of interest. For instance, the proportion of male lecturer to female lecturer. Our focus in this unit will be on inferences about differences between two population proportions.

Learning Outcomes

At the end of this unit, you should be able to:

1. learn how to construct a confidence interval for the difference in the proportions of two distinct populations that have a particular characteristic of interest.
2. learn how to perform a test of hypotheses concerning the difference in the proportions of two distinct populations that have a particular characteristic of interest.

Main Content

Comparison of two independent population proportions

6 mins

Suppose our interest is to compare the proportion of two independent population taken a large sample from each of the population; say population 1 and population 2.

Table 1: Parameter and statistic of the two populations

Parameter/Statistic	Population 1	Population 2
Population Proportion	P_1	P_2
Sample size	n_1	n_2
Sample proportion	p_1	p_2

Confidence interval for the difference between two independent population proportions

The sample proportion p_1 is a good estimator of the population proportion P_1 and the sample proportion p_2 is a good estimator of population proportion P_2, the point estimate of the difference between the two population $(P_1 - P_2)$ is $(p_1 - p_2)$. The samples must be large and independent of each other. The confidence interval for the difference between the two population proportions is given as

$$(p_1 - p_2) \pm z_c \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}},$$

Hypothesis testing concerning difference between population proportions

Testing hypothesis concerning the difference of two population proportions is done as it is done for independent samples for population means, the null hypothesis is expressed as;

$$H_0 : P_1 - P_2 = P_0$$

where P_0 is a number deduced from the statement of the situation.
The alternative hypothesis can take any of the form;

$H_1 = P_1 - P_2 < P_0$ Left-Tailed Test

$H_1 = P_1 - P_2 > P_0$ Right-Tailed Test

$H_1 = P_1 - P_2 \neq d_0$ Two-Tailed Test

Test statistic for the Hypothesis Testing concerning difference between two population proportions

$$z = \frac{(p_1 - p_2) - p_0}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

Example 1: There are two major parties in Nigeria, the All progressive Congress (APC) and the People Democratic Party (PDP). A random sample of voters who identify with one of the two parties in Kogi state were selected and asked if they favour an open ballot for the selection of candidate to represent their political party in the next gubernatorial election.

	PDP	APC
Sample size *n*	150	200
Number in favour *x*	90	14

(a) Construct the 95% confidence interval for the difference in their proportion

(b) Test at 5% level of significance, the hypothesis that the proportion of all members of PDP who are in favour of open ballot is less than that of all members if APC in favour of it.

Solution

$$p_1 = \frac{90}{150} = 0.6, q_1 = 1 - p_1 = 0.4$$

$$p_2 = \frac{14}{200} = 0.07, q_2 = 1 - p_2 = 0.93$$

(a) The 95% confidence interval

$$CI = (p_1 - p_2) \pm z_c \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$= (0.6 - 0.07) \pm Z_{0.05} \sqrt{\frac{0.6 \times 0.4}{150} + \frac{0.07 \times 0.93}{200}}$$

$$= 0.53 \pm 1.96 \times 0.044$$

$$= 0.53 \pm 0.086$$

$$= (0.444 \leq (p_1 - p_2) \leq 0.616)$$

(b) The statement of hypothesis is states as follows
The null hypothesis is

$$H_0: p_1 - p_2 = 0$$

The alternative hypothesis is

$$H_1: p_1 - p_2 < 0$$

The test statistic

$$z = \frac{(p_1 - p_2) - p_0}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

$$= \frac{0.53 - 0}{0.04} = 13.25$$

The critical value $z_{0.05} = 1.645$

Decision: Reject the null hypothesis if the value of the test statistic is greater than the critical value

Decision: Since the value of the test statistic is greater than the critical value, the null hypothesis is rejected.

Conclusion: The proportion of all members of PDP who are in favour of open ballot is less than that of all members if APC in favour of it

Example 2: A random sample of 500 jam keys manufactured by Machine 1 and 300 jam keys manufactured by Machine 2 showed that 48 and 15 were defective respectively.

(a) Construct the 95% confidence interval for the difference in their proportion
(b) Test at 1% level of significance, that the two machines show different qualities of performance.
(C) Test at 5% level of significance that machine 2 performs better than machine 1.

Solution

$$p_1 = \frac{48}{500} = 0.096, q_1 = 1 - p_1 = 0.904$$

$$p_2 = \frac{15}{300} = 0.05, q_2 = 1 - p_2 = 0.95$$

(a) The 95% confidence interval

$$CI = (p_1 - p_2) \pm z_c \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$= (0.096 - 0.05) \pm Z_{0.05} \sqrt{\frac{0.096 \times 0.904}{500} + \frac{0.05 \times 0.95}{300}}$$

$$= 0.046 \pm 1.96 \times 0.019$$

$$= 0.046 \pm 0.037$$

$$= (0.009 \le (p_1 - p_2) \le 0.083)$$

(b) The statement of hypothesis is states as follows
The null hypothesis is

$$H_0: p_1 - p_2 = 0$$

The alternative hypothesis is

$$H_1: p_1 - p_2 \neq 0$$

The test statistic

$$z = \frac{(p_1 - p_2) - p_0}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

$$= \frac{0.046 - 0}{0.019} = 2.421$$

The critical value $z_{0.05} = 1.96$

Decision: Reject the null hypothesis if the value of the test statistic is greater than the critical value

Decision: Since the value of the test statistic is greater than the critical value, the null hypothesis is rejected.

Conclusion: The two machines show different qualities of performance.

(c) The statement of hypothesis is states as follows

The null hypothesis is

$H_0: p_1 - p_2 = 0$

The alternative hypothesis is

$H_1: p_1 - p_2 < 0$

The test statistic

$$z = \frac{(p_1 - p_2) - p_0}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

$$= \frac{0.046 - 0}{0.019} = 2.421$$

The critical value $z_{0.05} = 1.645$

Decision: Reject the null hypothesis if the value of the test statistic is greater than the critical value

Decision: Since the value of the test statistic is greater than the critical value, the null hypothesis is rejected.

Conclusion: The machines 2 performs better than machine 1.

Summary

So far, in this unit we have computed a confidence interval for the difference in two population proportions using a formula provided. The procedure used to test hypotheses concerning a single population proportion was what we used to test hypotheses concerning the difference between two population proportions.

Self-Assessment Questions

1. A random sample of 78 women entering a major bank for transaction showed that 23 of them have Bachelor's degree while a random sample of 73 men entering the bank showed that 20 have Bachelor's degree. Does this indicate that the population proportion of women having Bachelor's degree is different from that of men?

2. Of the total number of 378 students admitted into a distant learning programme last session, 194 were males. From the total number of 516 students admitted for this same programme this academic session, 320 are males. Does this indicates that the proportion of males admitted has increased? Use 5% level of significance.

3. Out of 1100 voters, 56% indicated to vote for a political party candidate when he first contested for the governorship position of a state. A survey pf 800 voters revealed that 46% showed their support for the same candidate for contesting the second time.

   (a)At 5% significance level, can we infer that the candidate's popularity has decreased.

   (b). At 1% significance level, can we infer that the candidate's popularity has decreased.

Tutor Marked Assessment

- The dress code committee of the University of Ilorin came up with a proposal to check conformity to the approved dresses at the university gate before any student is allowed to enter. Of a random sample of 220 female students, 29 supported the proposal. Another random sample of 175 male students, 56 supported the proposal.

  (a) Find the 95% confidence interval for the difference in their proportion

  (b) Does the information indicate a difference between the population of female and male students that supported the proposal? Use 0.05 as the level of significance.

- Of the total number of 378 students admitted into a distant learning programme last session, 194 were males. From the total number of 516 students admitted for this same programme this academic session, 320 are males. Does this indicates that the proportion of males admitted has increased? Use 1% level of significance.

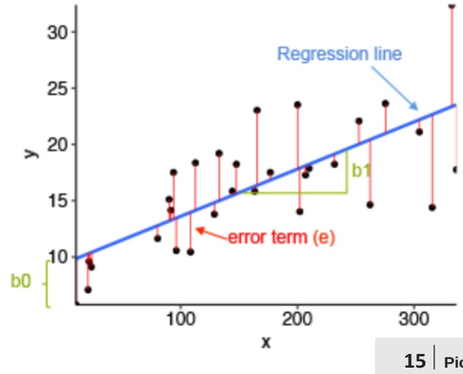
- Of the 100 students graduating from the Department of Statistics in a particular year, 52% graduated with second class upper division while 48% graduated with second class upper division out 100 students that graduated from the Department in the following year. Test at 5% level of significance that the population of students graduating with second class upper division has decreased.

References

- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.
- Spiegel M.R and Stephens L.J (1999). Schaum's Outlines of Theory and Problems of Statistics. McGraw-Hill, New York
- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA

Further Reading

- https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/09%3A_Two-Sample_Problems/9.4%3A_Comparison_of_Two_Population_Proportions
- https://faculty.elgin.edu/dkernler/statistics/ch11/11-1.html



15 | **Picture:** Simple regression analysis

Photo: Wikipedia.com

UNIT 13

Simple Regression Analysis

Introduction

Often, our interest will be to estimate the value of a variable say Y on the basis of sample data corresponding to a given variable say X. In this study unit, we will be considering statistical methods for analysing the relationship between variables x and y will be considered.

Learning Outcomes

At the end of this unit, you should be able to:

1. learn what it means for two variables to exhibit a relationship
2. fit a linear regression model
3. construct confidence interval for the slope of the regression model
4. test for the significance of the slope of the regression model

Main Content

7 mins

Simple regression is concerned with measuring the relationship between two variables. One of the variables is the independent (explanatory) variable while the other variable is called the dependent (response) variable. The dependent variable is usually denoted as Y while the independent variable is denoted as X.

Table 1: Examples of independent and explanatory variable

Independent Variable	Dependent Variable
Height	Weight
Farm Size	Yield
Income	Expenditure
Size of a house	Value of a house

Scatter Diagram: The scatter diagram graphs the pairs of numerical data, with one variable at the y-axis and the other variable at the x-axis. Fig 1 shows the scatter diagram of weight and height of children. Weight is on the y-axis while Height is on the x-axis

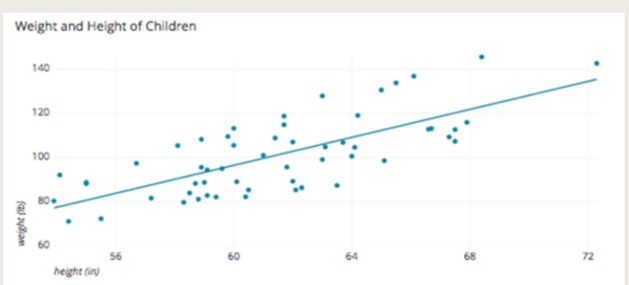


Fig 1: Typical example of a scatter diagram

The simple regression equation/model

$$y = \beta_0 + \beta_1 x_i + e_i,$$

where β_0 is the intercept on the y-axis

β_1 is the slope or the regression coefficient

e_i is the error term

β_0 and β_1 are parameters whose values are to be determined.

The formula for the estimates of the parameters are given as

$$\beta_1 = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} \text{ and } \beta_0 = \frac{\sum y}{n} - \beta_1 \frac{\sum x}{n} \text{ or } \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Confidence interval for the slope of the regression model

The confidence interval for the slope β_1 is given by the following formula.

$$\beta_1 \pm t_{\alpha/2} \frac{s_e}{\sqrt{s_{xx}}}$$

Where, n is the number of paired observation (data)

$$s_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}, \qquad s_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} \qquad s_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}$$

$$s_e = \frac{SSE}{\sqrt{n-2}}, \qquad SSE = s_{yy} - \beta_1 s_{xy}$$

Hypothesis Test of the slope of the regression model

Hypothesis testing for the slope β_1 of the regression model involves the procedures under the test of hypothesis. The null hypothesis take the form

$H_0 : \beta_1 = 0$

The alternative hypothesis is a two tailed test in the form

$H_1 : \beta_1 \neq 0$

Test statistic concerning the slope β_1

$$t = \frac{\beta_1}{s_e / \sqrt{s_{xx}}}$$

The test statistic follows a Student's t distribution with n-2 degree of freedom

Decision Rule: reject the null hypothesis if the absolute value of the t-statistic is greater than the tabulated value of the Student T table value.

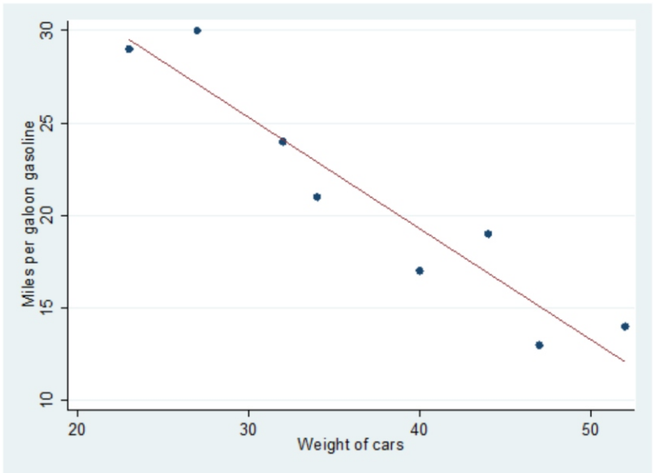
Example 1: to investigate if heavier cars use more gasoline, 8 cars were randomly selected, the weight x (in pounds) of the car and miles per gallon (mpg) of gasoline y were measured. The table below fives the weight and the miles per gallon.

Weight x	27	44	32	47	23	40	34	52
MPG y	30	19	24	13	29	17	21	14

I. Draw the scatter diagram of weight of a car against the miles per gallon of gasoline
ii. Fit a linear regression model to the data
iii. Test for the significance of the slope of the regression model fitted
iv. Suppose that a car weighed 40 pounds, what is the miles per gallon of gasoline?

Solution

I. Draw the scatter diagram of weight of a car against the miles per gallon of gasoline



$n=8, \sum x=299, \sum y=167, \sum xy=5814, \sum x^2=11887, \sum y^2=3773$

ii. Fit a linear regression model to the data

$$\beta_1=\frac{n\sum xy-\left(\sum x\right)\left(\sum y\right)}{n\sum x^2-\left(\sum x\right)^2}$$
$$=\frac{8\times5814-299\times167}{8\times11887-(299)^2}$$
$$=\frac{-3421}{5695}=-0.6$$

$$\beta_0=\frac{\sum y}{n}-\beta_1\frac{\sum x}{n}$$
$$=\frac{167}{8}-(-0.6)\frac{299}{8}$$
$$=43.3$$

The linear regression model is

$$y=43.3-0.6x$$

iii Test for the significance of the slope of the regression model fitted the null hypothesis take the form

$$H_0:\beta_1=0$$

The alternative hypothesis is a two tailed test in the form

$$s_{xx}=\sum x^2-\frac{\left(\sum x\right)^2}{n}$$
$$=11887-\frac{299^2}{8}=711.875$$

$$s_{yy}=\sum y^2-\frac{\left(\sum y\right)^2}{n}$$
$$=3773-\frac{167^2}{8}=286.875$$

$$s_{xy}=\sum xy-\frac{\left(\sum x\right)\left(\sum y\right)}{n}$$
$$=5814-\frac{299\times167}{8}=-427.625$$

$$SSE = s_{yy} - \beta_1 s_{xy}$$
$$= 286.875 - (-0.6 \times -427.625)$$
$$= 30.3$$

$$s_e = \frac{SSE}{\sqrt{n-2}} = \frac{30.3}{\sqrt{8-2}} = 12.3699$$

Test statistic concerning the slope β_1

$$T = \frac{\beta_1}{s_e / \sqrt{s_{xx}}}$$
$$= \frac{-0.6}{12.3699 / \sqrt{711.875}} = -1.2942$$

The test statistic follows a Student's t distribution with n-2 degree of freedom

$$t_{0.025,(8-2)} = t_{0.025,6} = 0.718$$

Decision Rule: reject the null hypothesis if the absolute value of the t-statistic is greater than the tabulated value of the Student T table value.

Decision: Since the absolute value of the t-statistic is greater than the tabulated value of the Student T table value, the null hypothesis is rejected.

Conclusion: The slope of the fitted regression model is significantly different from zero.

iv. Suppose that a car weighed 40 pounds, what is the miles per gallon of gasoline?

The miles per gallon of gasoline is 19.3

Summary

In this study unit, I explained what it means for two variables to exhibit a relationship. Formula for estimating the parameters of the regression model were given. We also computed the confidence interval of the slope parameter and test concerning the slope of the regression model was also carried out using the procedures for testing hypothesis. With the above knowledge, self asses yourself with the following questions.

Self-Assessment Questions

1. Two tests were conducted for the students taking a statistics course, the following are scores of 10 students in the tests.

Test 1 x	75	80	93	65	87	71	98	68	84	77
Test 2 y	82	78	86	72	91	80	95	72	89	74

I. Draw the scatter diagram of test 2 against test 1
ii. Fit a linear regression model to the data
iii. Test for the significance of the slope of the regression model fitted
iv. Suppose that a student score 81 in the test 1, what will be the estimated score in the test 2

Tutor Marked Assessment

- The yield per farm is known to be approximately linearly related to the area of the farm. The yield of maize y in tons and the area of the farm x in hectare. The summary statistic are given as follows:

$$n = 15, \sum x = 249.8, \sum y = 1200.6, \sum xy = 20127.47, \sum x^2 = 4200.56, \sum y^2 = 96725$$

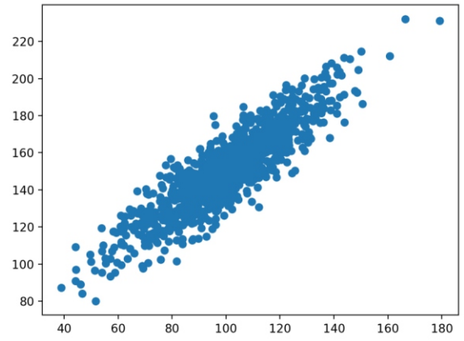
I. Fit a linear regression model to the data
ii. Test for the significance of the slope of the regression model fitted
iii. Suppose that the area of the farm is 56.2, find the yield of maize.

References

- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.
- Jaisingh, L. R. (2000). Statistics for the Utterly Confuse, McGraw-Hill, USA.
- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA

Further Reading

- http://www.cimt.org.uk/projects/mepres/alevel/stats_ch12.pdf
- http://websupport1.citytech.cuny.edu/Faculty/mbessonov/MAT1272/Worksheet%20November%2014%20Solutions.pdf



16 | **Picture:** correlation

Photo: Wikipedia.com

UNIT 14

Correlation

Introduction

A closely related problem of measuring the relationship between two variables is the measures of strength or degree of relationship between them. This unit we will focus on correlation between variables.

Learning Outcomes

At the end of this unit, you should be able to:

1. explain what the linear correlation coefficient is
2. compute linear correlation coefficient and interpret it
3. carry out test concerning the correlation coefficient
4. describe what the coefficient of determination is, how to compute it, and its interpretation

Main Content

8 mins

In order for us to establish the relationship between two variables, the scatter diagram will give us a better description. The typical scatter diagram is shown in figure 1 of study unit 13. The measure of sample correlation coefficient describes the strength/degree of linear relationship between two variables. The Pearson production correlation coefficient is used to measure the degree of relationship.

Properties of correlation coefficient *r*

i. It does have unit of measurement
ii. It lies between -1 and 1 $(-1 \le r \le 1)$
iii. The sign of r indicates the direction of the relationship. For example, for two variables X and Y, if $r < 0$, it means, as X increases, Y tends to decrease. If $r > 0$, it means as X increases, Y tends to increase.
iv. If $r = 1$, then, there is perfect positive relationship between the two variable X and Y
v. If $r = 0$, then, there is no linear relationship between the two variable X and Y
vi. The value of $|r|$ indicates the strength of relationship between the variables. If $|r|$ is close to 1, it means a strong positive relationship, if $|r|$ is close to -1, it means a strong negative relationship.
vii. The value of r does not change regardless of which variable is the independent or dependent variable.

The Pearson production correlation coefficient formula is

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$$

n is the number of random paired sample.

Hypothesis Test concerning the correlation coefficient r

Hypothesis testing for the correlation coefficient r involves the procedures under the test of hypothesis. The null hypothesis take the form

$H_0 : r = 0$

The alternative hypothesis is a two tailed test in the form

$H_1 : r \neq 0$

Test statistic concerning the correlation coefficient r

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The test statistic follows a Student's t distribution with n-2 degree of freedom

Decision Rule: reject the null hypothesis if the absolute value of the t-statistic is greater than the tabulated value of the Student T table value.

Coefficient of Determination r^2

Coefficient of Determination measures how good the least square line is as an instrument of regression. It is the square of the correlation coefficient. It measures the proportion if variation in the dependent variable that is explained by the independent variable. For example, if the correlation coefficient r is 0.986, then the coefficient of determination r^2 is 0.986^2=0.972, it means that the independent variable explained 97.2% of the variation in the dependent variable, the remaining 2.8% of the variation in the dependent variable is due to random chance.

Example 1: to investigate if heavier cars use more gasoline, 8 cars were randomly selected, the weight x (in pounds) of the car and miles per gallon (mpg) of gasoline y were measured. The table below fives the weight and the miles per gallon.

Weight *x*	27	44	32	47	23	40	34	52
MPG *y*	30	19	24	13	29	17	21	14

I. Find the correlation coefficient and interpret it.

ii. Test at 5% level of significance if the correlation coefficient is significantly different from zero.

iii. Find the coefficient of determination.

iv. What percentage of the variation in miles per gallon of the gasoline can be explained by the weight of the car?

v. What percentage of the variation in miles per gallon of the gasoline cannot be explained by the weight of the car?

Solution

$$n=8, \sum x=299, \sum y=167, \sum xy=5814, \sum x^2=11887, \sum y^2=3773$$

i. The correlation coefficient

$$r=\frac{n\sum xy-\left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2-\left(\sum x\right)^2\right]\left[n\sum y^2-\left(\sum y\right)^2\right]}}$$

$$=\frac{8\times5814-299\times167}{\sqrt{\left[8\times11887-299^2\right]\left[8\times3773-167^2\right]}}$$

$$=\frac{-3421}{\sqrt{5692\times2295}}=-0.9463$$

Interpretation: The correlation coefficient is -0.9463, it means that the relationship between weight of a car and the miles per gallon of gasoline is negatively strong, which indicates that as the weight of car increases, the consumption of gasoline decreases.

ii. Test if the correlation coefficient is significantly different from zero.

The statement of hypothesis

$H_0: r=0$

The alternative hypothesis is a two tailed test in the form

$H_1: r\neq 0$

The Test statistic

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$=\frac{-0.9463\times\sqrt{8-2}}{\sqrt{1-(-0.9463)^2}}$$

$$=\frac{-2.3180}{0.3233}=7.1698$$

The $t_{0.025,(n-2)}=t_{0.025,6}=0.718$

Decision: Since the absolute value of the t-statistic is greater than the tabulated value of the Student T table value, therefore the null hypothesis is rejected.

Conclusion: It is concluded that the correlation coefficient is significantly different from zero.

iii. The coefficient of determination.

The coefficient of determination n is the square of the correlation coefficient r^2, therefore the value of the coefficient of determination is 0.8954

iv. What percentage of the variation in miles per gallon of the gasoline can be explained by the weight of the car?

$0.8954\text{x}100=89.54\%$

v. Therefore, 89.54% of the variation in miles per gallon of the gasoline can be explained by the weight of the car

What percentage of the variation in miles per gallon of the gasoline cannot be explained by the weight of the car?

$1-0.8954=0.1046\text{ x }100=10.46\%$

Therefore, 10.46% of the variation in miles per gallon of the gasoline cannot be explained by the weight of the car

Example 2: Two tests were conducted for the students taking a statistics course, the following are scores of 10 students in the tests.

Test 1 x	75	80	93	65	87	71	98	68	84	77
Test 2 y	82	78	86	72	91	80	95	72	89	74

(a) Find the correlation correlating coefficient between the two test and interpret it
(b) Carry out a test of significance of the correlation coefficient

Solution

$$n=10, \sum x=798, \sum y=819, \sum xy=66045, \sum x^2=64722, \sum y^2=67675$$

(a) The correlation coefficient

$$r=\frac{n\sum xy-\left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2-\left(\sum x\right)^2\right]\left[n\sum y^2-\left(\sum y\right)^2\right]}}$$

$$= \frac{10\times 66045 - 798\times 819}{\sqrt{\left[10\times 64722 - 798^2\right]\left[10\times 67675 - 819^2\right]}}$$

$$= \frac{6888}{\sqrt{10416\times 5989}} = 0.8721$$

Interpretation: The correlation coefficient is 0.8721, it means that the relationship between test 1 and test 2 score is positively strong, which indicates that as the score of test 1 increases, the score of test 2also increases.

(b) Test of significance of the correlation coefficient

$H_0 : r = 0$

The alternative hypothesis is a two tailed test in the form

$H_1 : r \neq 0$

The Test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{0.8721\times\sqrt{10-2}}{\sqrt{1-(0.8721)^2}}$$

$$= \frac{2.4667}{0.4893} = 5.0413$$

The $t_{0.025,(n-2)} = t_{0.025,8} = 0.706$

Decision: Since the absolute value of the t-statistic is greater than the tabulated value of the Student T table value, therefore the null hypothesis is rejected.

Conclusion: It is concluded that the correlation coefficient is significantly different from zero.

Summary

In this study unit we have been able to present the formula for computing the Pearson correlation coefficient and, also give the properties of correlation coefficient and interpret its meaning and carry out a test concerning the validity of the correlation coefficient. Likewise, the coefficient of determination between two variables was examined and interpreted.

Self-Assessment Questions

1. The data below gives the information on a manufacturing company man-hour lost due to sickness in the last year and the current year. A simple random sample of 10 employees were taken from the company.

Employee	Man-hour lost in the previous year x	Man-hour lost in the current year y
1	12	13
2	24	25
3	15	15
4	30	32
5	32	36
6	26	24
7	10	12
8	15	16
9	0	2
10	4	12

I. Find the correlation coefficient and interpret it.

ii. Test at 5% level of significance if the correlation coefficient is significantly different from zero.

iii. Find the coefficient of determination.

iv. What percentage of the variation in in the man-hour lost in the current year can be explained by the man-hour lost in the last year?

v. What percentage of the variation in the man-hour lost in the current year cannot be explained by the man-hour lost in the last year?

Tutor Marked Assessment

- The yield per farm is known to be approximately linearly related to the area of the farm. The yield of maize y in tons and the area of the farm x in hectare. The summary statistic are given as follows:

$$n = 15, \sum x = 249.8, \sum y = 1200.6, \sum xy = 20127.47, \sum x^2 = 4200.56, \sum y^2 = 96725$$

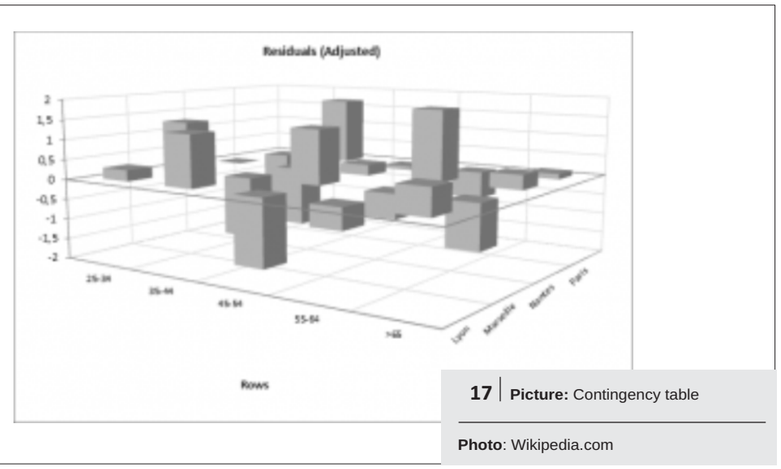
- Find the correlation coefficient and interpret it.
- Test at 5% level of significance if the correlation coefficient is significantly different from zero.
- Find the coefficient of determination.
- What percentage of the variation in the yield of maize can be explained by the area of the farm?
- What percentage of the variation in the yield of maize cannot be explained by the area of the farm?

References

- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.
- Jaisingh, L. R. (2000). Statistics for the Utterly Confuse, McGraw-Hill, USA.
- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA

Further Reading

- http://www.cimt.org.uk/projects/mepres/alevel/stats_ch12.pdf
- http://websupport1.citytech.cuny.edu/Faculty/mbessonov/MAT1272/Worksheet%20November%2014%20Solutions.pdf



17 | **Picture:** Contingency table

Photo: Wikipedia.com

UNIT 15

Contingency Table

Introduction

Previous study units, we focused on the comparison of numerical values of two population parameters but this study unit is going to be different. In what sense? in that we will focus on investigating whether or not two random variables take their values independently, or whether the value of one has a relation to the value of the other. Thus, we will express the hypothesis in words and not mathematically stated as the previous study units.

Learning Outcomes

At the end of this unit, you should be able to:

1. explain what contingency table is and use it compute the Chi-square statistic
2. explain how to use a chi-square test to investigate whether two factors are independent.

Main Content

Contingency Table

8 mins

A contingency Table is a two-way classification table in matrix format that displays the frequency distribution of the variables. It is useful for investigating the relationships between two categorical variables. Suppose there are n objects classified according to the levels of two variables, the typical r by c contingency

Table 1: Typical Contingency Table

		Variable 2		
		I	II	Row Total
Variable 1	A	a_1	a_2	$n_{1.}$
	B	b_1	b_2	$n_{2.}$
Column Total		$n_{.1}$	$n_{.2}$	n

Statement of Hypothesis

H_0: Variable 1 and Variable 2 are independent of each other

H_1: Variable 1 and Variable 2 are dependent on each other

Procedure

1) Display the observed frequencies (observed counts) in the R by C contingency table
2) Calculate the expected frequency (counts) using $e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$

   eij is the expected frequency (count) for the ith row and jth column
3) In order to investigate the relationship between the two categorical variable, the Chi-square test statistic is given as

$$X^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Where oij is the observed frequency and eij is the expected frequency for ith row jth column

4) Obtain the critical value from the Chi-square distribution table $\chi^2_{(r-1)(c-1),\alpha}$
5) Compare the calculated X^2 value with the critical value obtained
6) Reject the Null hypothesis if $X^2(calculated) > \chi^2(tabulated)$, otherwise, do not reject and draw conclusion.

Example 1: There is a theory that the gender of a baby in the womb is related to the baby's heart rate; baby girls are tends to have higher heart rates. To test this theory, the heart rate records of 40 babies were taken during their mother's last prenatal check-up before delivery. The results is given in the table below. Test at 5% level of significance if the baby's heart rate and baby's gender are independent of each other.

Heart Rate		High	low	Row Total
Gender	Girl	11	7	18
	Boy	17	5	22
Column Total		28	12	40

Solution

Statement of hypothesis

H_0 : Baby's gender and baby's heart rate are independent of each other

H_1 : Baby's gender and baby's heart rate are not independent of each other

Calculation of the expected frequencies

Cell 1; $e_{11} = \frac{n_{1.} \times n_{.1}}{n_{..}} = \frac{18 \times 28}{40} = 12.6$

Cell 2; $e_{12} = \frac{n_{1.} \times n_{.2}}{n_{..}} = \frac{18 \times 12}{40} = 5.4$

Cell 3; $e_{21} = \frac{n_{2.} \times n_{.1}}{n_{..}} = \frac{22 \times 28}{40} = 15.4$

Cell 4; $e_{22} = \frac{n_{2.} \times n_{.2}}{n_{..}} = \frac{22 \times 12}{40} = 6.6$

The Chi-square statistic

$$X^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$= \frac{(11-12.6)^2}{12.6} + \frac{(7-5.4)^2}{5.4} + \frac{(17-15.4)^2}{15.4} + \frac{(5-6.6)^2}{6.6}$$

$$= 1.231$$

The critical value from the Chi-square distribution table

$\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{(1)(1),0.05} = 3.841$

Decision rule: Reject the Null hypothesis if $X^2(calculated) > \chi^2(tabulated)$, otherwise, do not reject

Decision: Since the Chi-square value is less than the critical value, we fail to reject the null hypothesis and therefore conclude that Baby's gender and baby's heart rate are independent of each other.

Example 2: To test if being left-handed is hereditary, 250 adults were randomly selected and their handedness and their parent's handedness are noted. Test at 1% level of significance whether there is sufficient evidence in the data to conclude that there is a hereditary element in handedness.

		Number of parents left handed			Total
		0	1	2	
Handedness	Left	8	10	12	30
	Right	178	21	21	220
Total		186	31	33	250

Solution

Statement of hypothesis

H_0 : Being left handed is not hereditary

H_1 : Being left handed is hereditary

Calculation of the expected frequencies

Cell 1; $e_{11} = \frac{n_{1.} \times n_{.1}}{n_{..}} = \frac{30 \times 186}{250} = 22.32$

Cell 2; $e_{12} = \frac{n_{1.} \times n_{.2}}{n_{..}} = \frac{30 \times 31}{250} = 3.72$

Cell 3; $e_{13} = \frac{n_{1.} \times n_{.3}}{n} = \frac{30 \times 33}{250} = 3.96$

Cell 4; $e_{21} = \frac{n_{2.} \times n_{.1}}{n_{..}} = \frac{220 \times 31}{250} = 163.68$

Cell 5; $e_{22} = \frac{n_{2.} \times n_{.2}}{n_{..}} = \frac{220 \times 31}{250} = 27.28$

Cell 6; $e_{22} = \frac{n_{3.} \times n_{.3}}{n_{..}} = \frac{220 \times 33}{250} = 29.04$

The Chi-square statistic

$$X^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$= \frac{(8-22.32)^2}{22.32} + \frac{(10-3.72)^2}{3.72} + \frac{(12-3.96)^2}{3.96} + \frac{(178-163.68)^2}{163.68} + \frac{(21-27.28)^2}{27.28} + \frac{(21-29.04)^2}{29.04}$$

$$= 41.038$$

The critical value from the Chi-square distribution table

$\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{(1)(2),0.01} = 9.210$

Decision rule: Reject the Null hypothesis if $X^2(calculated) > \chi^2(tabulated)$, otherwise, do not reject

Decision: Since the Chi-square value is greater than the critical value, the null hypothesis is rejected and therefore conclude that being left handed is hereditary

Example 3: A large middle school administrator wishes to use celebrity influence to encourage students to make healthier choices in the school cafeteria. The cafeteria is situated at the centre of an open space. Every day at lunch time students get their lunch and a drink in three separate lines leading to three separate serving stations. As an experiment, the school administrator displayed a poster of a popular teen pop star drinking milk at each of the three areas where drinks are provided, except the milk in the poster is different at each location: one shows white milk, one shows strawberry-flavoured pink milk, and one shows chocolate milk. After the first day of the experiment the administrator noted the students' milk choices separately for the three lines. The data are given in the table provided. Test, at the 1% level of significance, whether there is sufficient evidence in the data to conclude that the posters had impact on the students' drink choices.

		Students' Choice			
		Regular	Strawberry	Chocolate	Total
Poster Choice	Regular	38	28	40	106
	Strawberry	18	51	24	93
	Chocolate	32	32	53	117
Total		88	111	117	316

Solution

Statement of hypothesis

H_0: The posters had no impact on the students' drink choices

H_1: The posters had impact on the students' drink choices

Calculation of the expected frequencies

Cell 1; $e_{11} = \frac{n_{1.} \times n_{.1}}{n_{..}} = \frac{106 \times 88}{316} = 29.52$

Cell 2; $e_{12} = \frac{n_{1.} \times n_{2}}{n_{..}} = \frac{106 \times 111}{316} = 37.23$

Cell 3; $e_{13} = \frac{n_{1.} \times n_{3}}{n_{..}} = \frac{106 \times 117}{316} = 39.25$

Cell 4; $e_{21} = \frac{n_{2.} \times n_{.1}}{n_{..}} = \frac{93 \times 88}{316} = 25.9$

Cell 5; $e_{22} = \frac{n_{2.} \times n_{2}}{n_{..}} = \frac{93 \times 111}{316} = 32.67$

Cell 6; $e_{23} = \frac{n_{2.} \times n_{3}}{n} = \frac{93 \times 117}{316} = 34.43$

Cell 7; $e_{31} = \frac{n_{3.} \times n_{.1}}{n_{..}} = \frac{117 \times 88}{316} = 32.58$

Cell 8; $e_{32} = \frac{n_{3.} \times n_{2}}{n_{..}} = \frac{117 \times 111}{316} = 41.1$

Cell 9; $e_{33} = \frac{n_{3.} \times n_{3}}{n} = \frac{117 \times 117}{316} = 38.99$

The Chi-square statistic

$$X^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$= \frac{(38-29.52)^2}{29.52} + \frac{(28-37.23)^2}{37.23} + \frac{(40-39.25)^2}{39.25} + \frac{(18-25.9)^2}{25.9} + \frac{(51-32.67)^2}{32.67} + \frac{(24-34.43)^2}{34.43} + \frac{(32-32.58)^2}{32.58} + \frac{(32-41.1)^2}{41.1} + \frac{(53-38.99)^2}{38.99}$$

$$= 27.651$$

The critical value from the Chi-square distribution table

$\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{(2)(2),0.01} = 13.3$

Decision rule: Reject the Null hypothesis if $X^2(calculated) > \chi^2(tabulated)$, otherwise, do not reject

Decision: Since the Chi-square value is greater than the critical value, the null hypothesis is rejected and therefore conclude that posters had impact on the students' drink choices.

Summary

This study unit I have presented to you what contingency table is and how to use it compute the Chi-square statistic. Also, the use of a chi-square test to investigate whether two factors are independent. Now demonstrate your knowledge with the following questions.

Self-Assessment Questions

1. It is generally believed that children brought up in stable families tends to do well in school. To verify this belief, 290 randomly selected students' record were examined with respect to their family structure and academic status. The data is given below. Test at 1% level of significance there is sufficient evidence in the data to conclude that family structure affect students' performance in school.

		Academic status	
		Graduated	Not Graduated
Family structure	No parent	18	31
	One parent	101	44
	Two parent	70	26

2. A survey was conducted to determine whether the age of driver has effect on the number of automobile accidents in which they are involved. The table below shows that dataset. Test at 5% level of significance that the number of accidents is independent of the driver's age.

		Age of Driver				
		21-30	21-40	41-50	51-60	61-70
Number of Accidents	0	748	821	786	720	672
	1	74	60	51	66	50
	2	31	25	22	16	15
	>2	9	10	6	5	7

Tutor Marked Assessment

- It is generally believed that children brought up in stable families tends to do well in school. To verify this belief, 290 randomly selected students' record were examined with respect to their family structure and academic status. The data is given below. Test at 5% level of significance there is sufficient evidence in the data to conclude that family structure affect students' performance in school.

		Academic status	
		Graduated	**Not Graduated**
Family structure	**No parent**	20	67
	One parent	200	56
	Two parent	100	46

- A survey was conducted to determine whether the age of driver has effect on the number of automobile accidents in which they are involved. The table below shows that dataset. Test at 1% level of significance that the number of accidents is independent of the driver's age.

		Age of Driver				
		21-30	**21-40**	**41-50**	**51-60**	**61-70**
Number of Accidents	**0**	748	821	786	720	672
	1	74	60	51	66	50
	2	31	25	22	16	15
	>2	9	10	6	5	7

References

- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.
- Jaisingh, L. R. (2000). Statistics for the Utterly Confuse, McGraw-Hill, USA.
- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA

Further Reading

- https://en.wikipedia.org/wiki/Contingency_table
- http://mathworld.wolfram.com/ContingencyTable.html
- http://onlinestatbook.com/2/chi_square/contingency.html

18 | **Picture:** Non parametric inference

Photo: Wikipedia.com

UNIT 16

Nonparametric Inference

Introduction

In the previous study unit, the focus was on how to perform the test of hypothesis concerning population parameters such as population mean and population proportion. This study unit we will centre on nonparametric test, its conditions are similar to test of hypothesis carried out in the previous unit but less stringent.

Learning Outcomes

At the end of this unit, you should be able to:

1. explain what nonparametric test is;
2. explain why it is used;
3. perform sign test for a single population and matched-pair data from two dependent samples;
4. perform Wilcoxon signed rank test for matched-pair data from two dependent samples and single population median;
5. perform Wilcoxon rank sum test for difference in population median using two independent samples
6. carry out a rank correlation test

Main Content

17 mins

Test of hypothesis and significance that we discuss in previous study units requires various assumptions about the distribution of the population from which the samples are selected. To carry out a t-test concerning a population mean, the required conditions are that; the sample size must be small and the population must follow a normal distribution. Situation arise when there is need to conduct a t-test with small sample size but the population is not normal. In this case, the nonparametric test is used.

Nonparametric tests are also called distribution-free test because they have fewer conditions. Nonparametric tests do not require the population to follow a particular distribution, such as normal distribution.

Advantages of nonparametric test

i. It can be used on greater variety of data because it requires fewer conditions
ii. It can be applied to categorical (qualitative) data
iii. Its computation tends to be easier compared to parametric test

Disadvantages of nonparametric test

I. It requires a large sample size to reject a null hypothesis given a level of significance
ii. It reduces the actual data to signs or ranks

Sign test for a single population median

In previous study units, we learnt how to perform the one-sample t-test for the population mean; a parametric test requiring the large sample or a normal population. When the population under consideration is not from a normal population and the sample is not large, the Sign test is used.

The Sign test is a nonparametric test that transform the original data into plus or minus signs. It could be conducted for a single population median, matched-paired data from two dependent samples or binomial data. It should be noted that the sigh test is a hypothesis test for the population median and the nit the population mean.

Steps in carrying out Sign **test for a single population median M**

The sample data must be randomly selected.

Step 1: State the Hypotheses. The hypothesis can be any of these three forms:

Table 1: Table of Hypotheses for the sign test for a single population median

Null Hypothesis	Alternative Hypothesis	Type of Test
$H_0: M = M_d$	$H_1: M > M_d$	Right-Tailed Test
$H_0: M = M_d$	$H_1: M < M_d$	Left-Tailed Test
$H_0: M = M_d$	$H_1: M \neq M_d$	Two-Tailed Test

M_d is the value of the population median which is being tested.

Step 2: Find the value of the statistic

- **For small sample size** $(n \leq 25)$, use the Table 2 below

Table 2: Choice of S_M

Type of Test	Test Statistic S_M
Right-Tailed Test	S_M = number of minus signs
Left-Tailed Test	S_M = number of plus signs
Two-Tailed Test	S_M = number of minus signs or plus signs, whichever is smaller

- **For large sample size** $(n > 25)$, first use Table 2 to find the S_M, and then calculate the value of the test statistic Z_M

$$Z_M = \frac{(S_M + 0.5) - \frac{n}{2}}{\frac{\sqrt{n}}{2}}$$

Step 3: Find the critical value and state the rejection rule

- **For small sample size** $(n \le 25)$, use Table 2, choose the column with appropriate level of significance and the applicable One-Tailed or Two-Tailed test. Then, select the row with the appropriate sample size n= number of pluses and minuses. The number that coincide with the row and column is the critical value S_C. The rejection rule to reject the null Hypothesis is reject H_0 if $S_M \le S_C$

- **For large sample size** $(n > 25)$, use the standard normal Table. The critical value Z_C is compared with the value of Z_M of the test statistic. The rejection rule to reject the null Hypothesis is reject H_0 if $Z_M \le Z_C$

Table 3: Sign Table

	alpha values						
n	0.001	0.005	0.01	0.025	0.05	0.10	0.20
5	--	--	--	--	--	0	2
6	--	--	--	--	0	2	3
7	--	--	--	0	2	3	5
8	--	--	0	2	3	5	8
9	--	0	1	3	5	8	10
10	--	1	3	5	8	10	14
11	0	3	5	8	10	13	17
12	1	5	7	10	13	17	21
13	2	7	9	13	17	21	26
14	4	9	12	17	21	25	31
15	6	12	15	20	25	30	36
16	8	15	19	25	29	35	42
17	11	19	23	29	34	41	48
18	14	23	27	34	40	47	55
19	18	27	32	39	46	53	62
20	21	32	37	45	52	60	69
21	25	37	42	51	58	67	77
22	30	42	48	57	65	75	86
23	35	48	54	64	73	83	94
24	40	54	61	72	81	91	104
25	45	60	68	79	89	100	113
26	51	67	75	87	98	110	124
27	57	74	83	96	107	119	134

	alpha values						
n	0.001	0.005	0.01	0.025	0.05	0.10	0.20
28	64	82	91	105	116	130	145
29	71	90	100	114	126	140	157
30	78	98	109	124	137	151	169
31	86	107	118	134	147	163	181
32	94	116	128	144	159	175	194
33	102	126	138	155	170	187	207
34	111	136	148	167	182	200	221
35	120	146	159	178	195	213	235
36	130	157	171	191	208	227	250
37	140	168	182	203	221	241	265
38	150	180	194	216	235	256	281
39	161	192	207	230	249	271	297
40	172	204	220	244	264	286	313
41	183	217	233	258	279	302	330
42	195	230	247	273	294	319	348
43	207	244	261	288	310	336	365
44	220	258	276	303	327	353	384
45	233	272	291	319	343	371	402
46	246	287	307	336	361	389	422
47	260	302	322	353	378	407	441
48	274	318	339	370	396	426	462
49	289	334	355	388	415	446	482
50	304	350	373	406	434	466	503

Step 4: State the conclusion and interpretation: Compare the test statistic with the critical value using the rejection rule.

Example: The monthly number of motor bike accidents in a city given in the table below, test at 5% level of significance whether the population median number of motor bike accidents is less than 20.

Months	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Accidents	23	10	15	20	29	32	26	13	25	19	23	28

Solution

Statement of Hypothesis

$H_0: M = 20$ Versus $H_1: M < 20$

Change each data value that is less than 20 to a minus sign (-), and change each data value that is greater than 20 to a plus sign (+). Ignore any data values that are equal to 20. The sample size ***n*** is the total number of plus signs and minus signs. Discard any data that has the same value as the hypothesized median.

Months	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Accidents	23	10	15	20	29	32	26	13	25	19	23	28
Sign	+	-	-		+	+	+	-	+	-	+	+

To find the critical value and state the rejection rule. The total number of plus signs and minus signs is n= 7 +4=11, which is not greater than 25, so we use the small-sample case. We have a one-tailed test, with α=0.05 and n= 11, the critical value S_C from Table 3 is 10. The rejection rule to reject the null Hypothesis is reject H_0 if $S_M \le S_C$

For small sample size $(n \le 25)$, the test statistic for a Left-Tailed Test, S_M is the number of plus signs, therefore $S_M = 7$

Decision: Since $S_M(7) \le S_C(10)$, the null hypothesis is rejected.

Conclusion: There is evidence that the population median number of motor bike accidents is less than 20at 5% level of significance.

Sign test for Matched-Pair Data from two Dependent samples

In Study unit 11, hypothesis test for the population mean of the difference between two dependent samples was examined. The paired-sample t test we learnt in Study unit 11 required either that the population of differences be normal or that the sample size of the differences be at least 30. The sign test for the population median of the differences, M_d, requires that the sample data be randomly selected only

Steps for carrying out a Sign test for Matched-Pair Data from two Dependent samples

Step 1: For each matched pair, subtract the value of the second variable from the value of the first variable.

Step 2: Interest is only in the sign of the difference found in Step 1, not the difference itself.

Step 3: Exclude ties, that is, omit any matched pairs in which the values for both variables are equal.

Table 4: Hypotheses for the sign test for the population median of the differences M_d

Null Hypothesis	Alternative Hypothesis	Type of Test	Test Statistic S_M
$H_0: M_d = 0$	$H_1: M_d > 0$	Right-Tailed Test	S_M = number of minus signs
$H_0: M_d = 0$	$H_1: M_d < 0$	Left-Tailed Test	S_M = number of plus signs
$H_0: M_d = 0$	$H_1: M_d \neq 0$	Two-Tailed Test	S_M = number of minus signs or plus signs, whichever is smaller

Example: The table below shows the scores of 7 students in a Math course before and after a revision class. Test at 5% level of significance if the mean difference is greater than zero.

Before	20	25	15	10	20	30	15
After	30	30	20	20	25	35	25

Wilcoxon signed rank test

The Wilcoxon signed rank testis a nonparametric hypothesis test in which the original data are transformed into their ranks. The Wilcoxon signed rank test may be conducted for a single population median and matched-pair data from two dependent samples. In the sign test, the data are converted into plus signs or minus signs. The magnitude of the data values is lost, which contributes to the low efficiency of the sign test. The Wilcoxon signed rank test for a single population median, which takes the magnitude of the data into account by ranking the data values.

Wilcoxon signed rank test for Matched-Pair Data from two Dependent samples

The requirements to carry out Wilcoxon signed rank test for Matched-Pair Data from two Dependent samples are that:

(I). the sample data be randomly selected and that the
(ii). the distribution of the differences be symmetric.
(iii). It is not required that the population be normally distributed.

Steps in carrying out Wilcoxon signed rank test for Matched-Pair Data from two Dependent samples

Step 1: State the hypotheses

Table 5: Hypotheses for Wilcoxon signed rank for matched-paired samples

Null Hypothesis	Alternative Hypothesis	Type of Test
$H_0: M_d = 0$	$H_1: M_d > 0$	Right-Tailed Test
$H_0: M_d = 0$	$H_1: M_d < 0$	Left-Tailed Test
$H_0: M_d = 0$	$H_1: M_d \neq 0$	Two-Tailed Test

Step 2: Find the value of the test statistic

First find the signed ranks using the following steps:

a. For each paired data value, find the difference d between each data value and the hypothesized median M_0. Omit data values for which d=0.

b. Find the absolute values of the differences.

c. Rank the absolute values of the differences from smallest to largest. If two or more data values have the same rank, assign to each the mean value of their ranks.

d. Attach to each rank the sign of its corresponding value of d. This is its signed rank. Replace each original data value with its corresponding signed rank.

- **For small sample** $(n \le 30)$: Use Table the Wilcoxon signed Test to find T_D, where T+ is the sum of the positive signed ranks, and |T-| is the absolute value of the sum of the negative signed ranks.

Table 6: Hypothesis test Wilcoxon signed rank for matched-pair data

Type of Test	Test statistic T_{data}
Right-Tailed Test	$T_D = \lvert T^- \rvert$
Left-Tailed Test	$T_D = T^+$
Two-Tailed Test	$T_D = T^+$ or $\lvert T^- \rvert$, whichever is smaller

- **For large sample** $(n > 30)$: Use Table the Wilcoxon signed Test, find T_D, and then the test statistic Z_D

$$Z_D = \frac{T_D - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Table 7: Critical Values of the Wilcoxon Signed Ranks Test

n	Two-Tailed Test		One-Tailed Test	
	α = .05	α = .01	α = .05	α = .01
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43
21	58	42	67	49
22	65	48	75	55
23	73	54	83	62
24	81	61	91	69
25	89	68	100	76
26	98	75	110	84
27	107	83	119	92
28	116	91	130	101
29	126	100	140	110
30	137	109	151	120

Step 3: Find the value of the statistic

- **For small sample size** $(n \le 30)$, use Table 6, choose the column with appropriate level of significance and the applicable One-Tailed or Two-Tailed test. Then, select the row with the appropriate sample size n, n is the number of data for which the difference d is not zero. The number that coincide with the row and column is the critical value T_C. The rejection rule to reject the null Hypothesis is reject H_0 if $T_D \le T_C$

- **For large sample size** $(n > 30)$, use the standard normal Table. The critical value Z_C is compared with the value of Z_D of the test statistic. The rejection rule to reject the null Hypothesis is reject H_0 if $Z_D \le Z_C$

Step 4: State the conclusion and interpret: Using the rejection rule, compare the test statistic with the critical region

Example: The weight (in kg) before and after the introduction of a particular feed to 8 poultry chickens are recorded and presented in the table below. Test, using the Wilcoxon signed rank test, whether there was weight gain in mean weight after the introduction of the feed using level of significance α=0.05.

Before	2.4	2.5	1.2	3.0	2.2	3.2	1.8	1.4
After	2.5	3.0	1.6	3.1	2.1	3.0	1.9	1.6

Solution:

Statement of Hypothesis

$H_0: M_d = 0$ versus $H_1: M_d > 0$

To find the critical value and state the rejection rule. The sample size is the number of data values for which the difference does not equal zero. Because none of the differences equals zero, the sample size is n= 8. Because $(n \le 30)$, we use the small-sample case. To find the critical value, we use Table 7. We have a one-tailed test, with level of significance α=0.05 and n=8, which gives us T_C=3. The rejection rule is to reject H0 if $T_D \le T_C$

To find the value of the test statistic

Before	2.4	2.5	1.2	3.0	2.2	3.2	1.8	1.4
After	2.5	3.0	1.6	3.1	2.1	3.0	1.9	1.6
d=(After-Before)	0.1	0.5	0.4	0.1	-0.1	-0.2	0.1	0.2
|d|	0.1	0.5	0.4	0.1	0.1	0.2	0.1	0.2
Rank of |d|	2.5	8	7	2.5	2.5	5.5	2.5	5.5
Signed Rank	2.5	8	7	2.5	-2.5	-5.5	2.5	5.5

Note: To rank the difference, the there are four 0.1, therefore, the ranks assigned is the average of (1+2+3+4)/4=2.5. The rank assigned to 0.2 is the average of (5+6)/2=5.5 because there are two 0.2.

From the signed ranks Table, we have a right-tailed test, so from Table 9, we have

$T_D = |T^-|$ = the sum of the negative signed ranks = 2.5+5.5 = 8

Decision: reject the null Hypothesis H_0 if $T_D \le T_C$, Since $T_D(8) > T_C(3)$, we fail to reject the null hypothesis.

Conclusion: there was no weight gain in mean weight after the introduction of the feed using level of significance α=0.05.

Wilcoxon signed rank test for a single population median

The same method used for Wilcoxon signed rank test for matched-pair samples can also be used for Wilcoxon signed rank test for a single population median, however, there is no subtraction to find difference because only one sample is used.

Table 8: Hypotheses for Wilcoxon signed rank for single population median

Null Hypothesis	Alternative Hypothesis	Type of Test
$H_0: M = M_0$	$H_1: M > M_0$	Right-Tailed Test
$H_0: M = M_0$	$H_1: M < M_0$	Left-Tailed Test
$H_0: M = M_0$	$H_1: M \neq M_0$	Two-Tailed Test

Example: The breaking strength of a random sample of 50 ropes by a manufacturer are given below. Test at 5% level of significance that the manufacturer's claim that the breaking strength of the rope is 30.

42	28	41	37	38	23	22	36	31	24	23	37	30
25	36	22	41	36	27	28	24	31	32	33	46	41
23	30	25	48	25	34	21	26	29	23	32	41	23
32	35	27	35	28	31	42	38	32	34	32		

Solution:

Statement of Hypothesis

$H_0: M = 30$ versus $H_1: M \neq 30$

To find the critical value and state the rejection rule. There are 50 ropes. Two of these ropes has 30 has their breaking strength, so that d =30-30=0. These 2 ropes are therefore omitted from this hypothesis test. This leaves us with 48 ropes, which is greater than 30, so we use the large-sample case. From the normal Table, the two-tailed test with level of significance α= 0.05 gives us -1.645. The decision rule is reject H_0 if $Z_D \le Z_C$.

Breaking Strength	Difference d	|d|	R|d|	Signed Rank	Breaking Strength	Difference d	|d|	R|d|	Signed Rank
42	12	12	45.5	45.5	41	11	11	42.5	42.5
28	-2	2	8.5	-8.5	23	-7	7	32	-32
41	11	11	42.5	42.5	25	-5	5	21	-21
37	7	7	32	32	48	18	18	48	48
38	8	8	37.5	37.5	25	-5	5	21	-21
23	-7	7	32	-32	34	4	4	17	17
22	-8	8	37.5	-37.5	21	-9	9	40	-40
36	6	6	26	26	26	-4	4	17	-17
31	1	1	2.5	2.5	29	-1	1	2.5	-2.5
24	-6	6	26	-26	23	-7	7	32	-32
23	-7	7	32	-32	32	2	2	8.5	8.5
37	7	7	32	32	41	11	11	42.5	42.5
25	-5	5	21	-21	23	-7	7	32	-32
36	6	6	26	26	32	2	2	8.5	8.5
22	-8	8	37.5	-37.5	35	5	5	21	21
41	11	11	42.5	42.5	27	-3	3	14	-14
36	6	6	26	26	35	5	5	21	21
27	-3	3	14	-14	28	-2	2	8.5	-8.5
28	-2	2	8.5	-8.5	31	1	1	2.5	2.5
24	-6	6	26	-26	42	12	12	45.5	45.5
31	1	1	2.5	2.5	38	8	8	37.5	37.5
32	2	2	8.5	8.5	32	2	2	8.5	8.5
33	3	3	14	14	34	4	4	17	17
46	16	16	47	47	32	2	2	8.5	8.5

To find the value of the test statistic, since the remaining sample is $(n > 30)$, we use the large sample case. Find T_D, and then the test statistic Z_D, the Wilcoxon Statistic is the value of T_D=463 which is the smaller of T^+=713 and $|T^-|=|-463|=463$.

$$Z_D = \frac{T_D - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{463 - \frac{48(48+1)}{4}}{\sqrt{\frac{48(48+1)(96+1)}{24}}} = -1.282$$

Conclusion: Since $Z_D = -1.282 > -1.642$, we do not reject H_0. The manufacturer's claim that the breaking strength of the rope is 30 is uphold.

Wilcoxon Rank Sum Test for Two Independent Samples

The Wilcoxon rank sum test is a nonparametric hypothesis that test whether the two population medians are equal or not. The original data from two independent samples are transformed into their ranks.

The requirements to carry out a Wilcoxon Rank sum are that

(a) the samples are independent random samples,
(b) each sample size is larger than 10
(c) the shapes of the distributions are the same.
(d) It is not required that the populations be normally distributed.

Steps in carrying out Wilcoxon Rank Sum Test for Two Independent Samples

Step 1: State the hypotheses

Table 9: Hypotheses for Wilcoxon signed rank for single population median

Null Hypothesis	Alternative Hypothesis	Type of Test
$H_0 : M_1 = M_2$	$H_1 : M_1 > M_2$	Right-Tailed Test
$H_0 : M_1 = M_2$	$H_1 : M_1 < M_2$	Left-Tailed Test
$H_0 : M_1 = M_2$	$H_1 : M_1 \neq M_2$	Two-Tailed Test

Step 2: Find the value of the test statistic Z_{data}

$$Z_{data} = \frac{R_1 - \mu_R}{\sigma_R}$$

Where, $\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$, $\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$,

R_1 is the sum of the ranks for the first sample

R_2 is the sum of the ranks for the second sample

n_1 and n_2 are sample sizes from the two samples

Step 3: Find the critical value and state the rejection rule.

Table 10 Critical values and rejection rules for the Wilcoxon rank sum test

Significance level	Form of Hypothesis		
	Right-Tailed	**Left-Tailed**	**Two-Tailed**
	$H_0 : M_1 = M_2$ $H_1 : M_1 > M_2$	$H_0 : M_1 = M_2$ $H_1 : M_1 < M_2$	$H_0 : M_1 = M_2$ $H_1 : M_1 \neq M_2$
$\alpha = 0.10$	$Z_C = 1.28$	$Z_C = -1.28$	$Z_C = 1.645$
$\alpha = 0.05$	$Z_C = 1.645$	$Z_C = -1.645$	$Z_C = 1.96$
$\alpha = 0.01$	$Z_C = 2.33$	$Z_C = -2.33$	$Z_C = 2.58$
Rejection Rule: Reject H_0 if:	$Z_{data} \geq Z_C$	$Z_{data} \leq Z_C$	$Z_{data} \leq Z_C$ or if $Z_{data} \geq Z_C$

Step 4: Compare the test statistic with the critical value, using the rejection rule and state the conclusion and the interpretation.

Example: Use the Wilcoxon rank sum test on the data below to determine whether the two population locations differ using 5% significance level.

Sample 1:	15	7	22	20	32	18	26	17	23	30
Sample 2:	8	27	17	25	20	16	21	17	10	18

Solution

Statement of Hypothesis

$H_0 : M_1 = M_2$ versus $H_1 : M_1 \neq M_2$

where M_1 and M_2 represent the population median for the population 1 and population 2 respectively.

To find the critical value and state the rejection rule. The level of significance is α=0.05, so the critical value is Z_C= 1.645, and the rejection rule is to reject H_0 if $Z_{data} \leq Z_C$ or $Z_{data} \geq Z_C$

To find the value of the test statistic. We combine the two samples and arrange in increasing order. We then rank the data values from smallest to largest, as shown in the following table, assigning ties to the mean rank value.

Combined data	7	8	10	15	16	17	17	17	18	18
Rank	1	2	3	4	5	7	7	7	9.5	9.5
Combined data	20	20	21	22	23	25	26	27	30	32
Rank	11.5	11.5	13	14	15	16	17	18	19	20

The sum of the ranks for the sample 1 is

R_1= 4+1+14+11.5+20+9.5+17+7+15+19= 118

The sum of the ranks for the sample 2 is

R_2= 2+18+7+16+11.5+5+13+7+3+9.5= 92

$$\mu_R = \frac{n_1(n_1+n_2+1)}{2} = \frac{10(10+10+1)}{2} = 105$$

$$\sigma_R = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}} = \sqrt{\frac{100(10+10+1)}{12}} = 13.23$$

$$Z_{data} = \frac{R_1 - \mu_R}{\sigma_R} = \frac{118-105}{13.23} = 0.983$$

Conclusion and interpretation. From the rejection rule, since 0.983 is less than 1.645 Therefore, the null hypothesis is rejected. There is insufficient evidence that the population median for the two population are the same.

Rank Correlation Test

The rank correlation test also called Spearman's rank correlation is a nonparametric based on the ranks of matched-pair data. This test may also be applied when the original data are ranks. In the rank correlation test, two variables are tested whether they are related. The sample paired data must be randomly selected. There is no requirement of normality.

Steps in carrying out a Rank correlation Test

Step 1: State the hypotheses

H_0: No rank correlation exists between the two variables

H_1: A rank correlation exists between the two variables.

Step 2: Find the value of the test statistic.

- **For small-sample case** $(n \le 30)$:

  a. Rank the values of the first variable from lowest to highest.

  b. Rank the values of the second variable from lowest to highest.

  c. Find the difference in ranks, ***d***, and square the difference in ranks to get d^2. Add up the d^2 to get $\sum d^2$

  d. Compute the test statistic:

$$r_{data} = 1 - \frac{6\sum d^2}{n^3 - n}$$

  where ***n*** represents the sample size (number of matched pairs)

  Note: Skip a and b if the original data is already ranked

- For large-sample case $(n > 30)$: Use Steps a–d from the small-sample case. However, a normal approximation is used, so the test statistic is called Z_{data}

$$Z_{data} = 1 - \frac{6\sum d^2}{n^3 - n}$$

Step 3: Find the critical value and state the rejection rule.

- **For small-sample case** $(n \le 30)$: Using Table 11, select the column with the appropriate level of significance α and the row with the appropriate sample size n. Reject the H_0 if $r_{data} \ge r_c$ or if $r_{data} \le -r_c$

Table 11: Spearman's Rho Table

n\α	0.2	0.1	0.05	0.02	0.01	0.002
4	1.000	1.000	–	–	–	–
5	0.800	0.900	1.000	1.000	–	–
6	0.657	0.829	0.886	0.943	1.000	–
7	0.571	0.714	0.786	0.893	0.929	1.000
8	0.524	0.643	0.738	0.833	0.881	0.952
9	0.483	0.600	0.700	0.783	0.833	0.917
10	0.455	0.564	0.648	0.745	0.794	0.879
11	0.427	0.536	0.618	0.709	0.755	0.845
12	0.406	0.503	0.587	0.678	0.727	0.818
13	0.385	0.484	0.560	0.648	0.703	0.791
14	0.367	0.464	0.538	0.626	0.679	0.771
15	0.354	0.446	0.521	0.604	0.654	0.750
16	0.341	0.429	0.503	0.582	0.635	0.729
17	0.328	0.414	0.488	0.566	0.618	0.711
18	0.317	0.401	0.472	0.550	0.600	0.692
19	0.309	0.391	0.460	0.535	0.584	0.675
20	0.299	0.380	0.447	0.522	0.570	0.662
21	0.292	0.370	0.436	0.509	0.556	0.647
22	0.284	0.361	0.425	0.497	0.544	0.633
23	0.278	0.353	0.416	0.486	0.532	0.621
24	0.271	0.344	0.407	0.476	0.521	0.609
25	0.265	0.337	0.398	0.466	0.511	0.597
26	0.259	0.331	0.390	0.457	0.501	0.586
27	0.255	0.324	0.383	0.449	0.492	0.576
28	0.250	0.318	0.375	0.441	0.483	0.567
29	0.245	0.312	0.368	0.433	0.475	0.558
30	0.240	0.306	0.362	0.425	0.467	0.549

❖ **For large sample size** $(n>30)$: A normal approximation is used. The critical value Z_c and the rejection rule are given in Table 12.

Table 12 Critical values and rejection rule for the rank correlation test, large-sample case

Level of Significance α	Critical Value Z_c	Rejection Rule
0.10	$\frac{1.645}{\sqrt{n-1}}$	Reject H0 if $Z_{data} \le -Z_c$ or if $Z_{data} \ge Z_c$
0.05	$\frac{1.96}{\sqrt{n-1}}$	
0.01	$\frac{2.58}{\sqrt{n-1}}$	

Step 4: state the conclusion and the interpretation. Compare the test statistic with the critical value, using the rejection rule.

Example: Find and test at 5% level of significance the rank correlation coefficient between the age and weight of 10 individuals given below.

Age:	30	24	18	15	12	10	8	6	4	3
Weight:	394	359	334	319	304	288	276	276	229	194

Solution:

Statement of Hypothesis

H0: No rank correlation exists between age and weight

H1: A rank correlation exists between age and weight

To find the value of the test statistic r_{data}

Age	30	24	18	15	12	10	8	6	4	3
Weight	394	359	334	319	304	288	276	276	229	194
Rank(Age)	10	9	8	7	6	5	4	3	2	1
Rank(Weight)	10	9	8	7	6	5	3.5	3.5	2	1
Difference *d*	0	0	0	0	0	0	-0.5	-0.5	0	0
d^2	0	0	0	0	0	0	0.25	0.25	0	0

$$\sum d^2 = 0.5$$

$$r_{data} = 1 - \frac{6\sum d^2}{n^3 - n} = 1 - \frac{6 \times 0.5}{10^3 - 10} = 0.997$$

To find the critical value and state the rejection rule. Since the sample is $(n \le 30)$, we use the small sample case: Using Table 11, select the column with the appropriate level of significance α=0.05 and the row with the appropriate sample size *n=10*. $r_c = 0.648$. Reject the H0 if $r_{data} \ge r_c$ or if $r_{data} \le -r_c$

Decision: we reject H0 since $r_{data}(0.997) \ge r_c(0.648)$

Conclusion: we conclude that a rank correlation exists between age and weight

Summary

So far, in this study unit I explained what nonparametric test is and why it is used. The sign test for single population and matched-pair data nonparametric tests were described. Wilcoxon signed rank test for matched-pair data from two dependent samples, single population median and Wilcoxon rank sum test for difference in population median using two independent samples were discussed. The rank correlation nonparametric test I also described. With the aid of explanations and examples I have given, test your knowledge with the questions below.

Self-Assessment Questions

(1). Test whether the population median differs from 1000, using level of significance α=0.01

950, 1000, 975, 925, 900, 1000, 1025, 900, 875, 950, 1000, 975, 925, 750, 775, 900

(2). The scores of 12 students in two tests from the same course are given below, test whether the population median score decreased from Test 2 to Test 1 using α= 0.05.

Students	1	2	3	4	5	6	7	8	9	10	11
Test 1	67	89	56	82	46	50	77	76	59	60	57
Test 2	58	77	60	72	57	52	75	60	70	59	82

(3). Using the Wilcoxon signed rank test, test whether the median of body temperature of seven women's differs from 98.6 degrees Fahrenheit. Use level of significance α= 0.05

97.2, 97.8, 98.1, 98.3, 98.7, 98.8, 99.3

(4). The following data represent independent random samples taken from a population of scores of students. Test whether the population median score of population 1 differs from the population median scores of population 2, using level of significance α= 0.05

Population 1: 96, 98, 81, 94, 89, 88, 84, 88, 84, 80, 81, 97

Population 2: 97, 97, 86, 90, 82, 85, 96, 81, 85, 79, 79, 80, 96, 83

(5). The following is the weight and height of 14 children. Assume that the data set represents a random sample. Use the rank correlation test to test for a relationship between age and weight, using level of significance α=0.01.

Weight	60	62	65	70	64	69	58	69	38	59	54	70	54	70
Height	25	21	28	28	19	25	27	20	31	29	24	24	27	30

Tutor Marked Assessment

- What are the advantages and disadvantages of nonparametric tests?
- What is another term for nonparametric tests?
- A sample of 26 scores from a statistics scores is given below, Test the hypothesis at 0.05 significance level that the median score for all the students is 72.

66	74	54	78	58	79	61	62	95	66	60	83	48
70	78	86	52	73	40	46	78	71	55	82	43	78

- Using the Wilcoxon rank sum on the data below, test whether the two population location differ. Use a 5% level of significance.

Sample 1:	15	7	22	20	32	18	26	17	23	30
Sample 2:	8	27	17	25	20	16	21	17	10	18

- The table below shows the rank of scores of 10 students in Mathematics and English. Find the rank correlation coefficient.

Mathematics	8	3	9	2	7	10	4	6	1	5
English	9	5	10	1	8	7	3	4	2	6

References

- Bluman, A. G. (2012). Elementary Statistics: A Step by Step Approach. McGraw-Hill, New York.
- Jaisingh, L. R. (2000). Statistics for the Utterly Confuse, McGraw-Hill, USA.
- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA

Further Reading

- https://statisticsbyjim.com/hypothesis-testing/nonparametric-parametric-tests/?__cf_chl_captcha_tk__=9b5eee0a42b0af7753583e51b0a77b9df415e62b-1593697543-0-
- https://www.statisticshowto.com/parametric-and-non-parametric-data/?__cf_chl_captcha_tk__=55057d76cec2adead76a3661de80b88725fda80e-1593697547-0-
- http://core.ecu.edu/ofe/statisticsresearch/Non-Parametric%20Tests.pdf