COURSE MANUAL

Statistical Inference II

STA 221

University of Ibadan Distance Learning Centre
Open and Distance Learning Course Series Development

ISBN 978-021-270-1

General Editor: Prof. Bayo Okunade

University of Ibadan Distance Learning Centre
University of Ibadan,
Nigeria

Telex: 31128NG

Tel: +234 (80775935727)
E-mail: ssu@dlc.ui.edu.ng
Website: www.dlc.ui.edu.ng

Vice-Chancellor's Message

The Distance Learning Centre is building on a solid tradition of over two decades of service in the provision of External Studies Programme and now Distance Learning Education in Nigeria and beyond. The Distance Learning mode to which we are committed is providing access to many deserving Nigerians in having access to higher education especially those who by the nature of their engagement do not have the luxury of full time education. Recently, it is contributing in no small measure to providing places for teeming Nigerian youths who for one reason or the other could not get admission into the conventional universities.

These course materials have been written by writers specially trained in ODL course delivery. The writers have made great efforts to provide up to date information, knowledge and skills in the different disciplines and ensure that the materials are user-friendly.

In addition to provision of course materials in print and e-format, a lot of Information Technology input has also gone into the deployment of course materials. Most of them can be downloaded from the DLC website and are available in audio format which you can also download into your mobile phones, IPod, MP3 among other devices to allow you listen to the audio study sessions. Some of the study session materials have been scripted and are being broadcast on the university's Diamond Radio FM 101.1, while others have been delivered and captured in audio-visual format in a classroom environment for use by our students. Detailed information on availability and access is available on the website. We will continue in our efforts to provide and review course materials for our courses.

However, for you to take advantage of these formats, you will need to improve on your I.T. skills and develop requisite distance learning Culture. It is well known that, for efficient and effective provision of Distance learning education, availability of appropriate and relevant course materials is a *sine qua non*. So also, is the availability of multiple plat form for the convenience of our students. It is in fulfilment of this, that series of course materials are being written to enable our students study at their own pace and convenience.

It is our hope that you will put these course materials to the best use.

Prof. Abel Idowu Olayinka
Vice-Chancellor

Foreword

As part of its vision of providing education for "Liberty and Development" for Nigerians and the International Community, the University of Ibadan, Distance Learning Centre has recently embarked on a vigorous repositioning agenda which aimed at embracing a holistic and all encompassing approach to the delivery of its Open Distance Learning (ODL) programmes. Thus we are committed to global best practices in distance learning provision. Apart from providing an efficient administrative and academic support for our students, we are committed to providing educational resource materials for the use of our students. We are convinced that, without an up-to-date, learner-friendly and distance learning compliant course materials, there cannot be any basis to lay claim to being a provider of distance learning education. Indeed, availability of appropriate course materials in multiple formats is the hub of any distance learning provision worldwide.

In view of the above, we are vigorously pursuing as a matter of priority, the provision of credible, learner-friendly and interactive course materials for all our courses. We commissioned the authoring of, and review of course materials to teams of experts and their outputs were subjected to rigorous peer review to ensure standard. The approach not only emphasizes cognitive knowledge, but also skills and humane values which are at the core of education, even in an ICT age.

The development of the materials which is on-going also had input from experienced editors and illustrators who have ensured that they are accurate, current and learner-friendly. They are specially written with distance learners in mind. This is very important because, distance learning involves non-residential students who can often feel isolated from the community of learners.

It is important to note that, for a distance learner to excel there is the need to source and read relevant materials apart from this course material. Therefore, adequate supplementary reading materials as well as other information sources are suggested in the course materials.

Apart from the responsibility for you to read this course material with others, you are also advised to seek assistance from your course facilitators especially academic advisors during your study even before the interactive session which is by design for revision. Your academic advisors will assist you using convenient technology including Google Hang Out, You Tube, Talk Fusion, etc. but you have to take advantage of these. It is also going to be of immense advantage if you complete assignments as at when due so as to have necessary feedbacks as a guide.

The implication of the above is that, a distance learner has a responsibility to develop requisite distance learning culture which includes diligent and disciplined self-study, seeking available administrative and academic support and acquisition of basic information technology skills. This is why you are encouraged to develop your computer skills by availing yourself the opportunity of training that the Centre's provide and put these into use.

In conclusion, it is envisaged that the course materials would also be useful for the regular students of tertiary institutions in Nigeria who are faced with a dearth of high quality textbooks. We are therefore, delighted to present these titles to both our distance learning students and the university's regular students. We are confident that the materials will be an invaluable resource to all.

We would like to thank all our authors, reviewers and production staff for the high quality of work.

Best wishes.

Professor Bayo Okunade
Director

Table of Contents

Study Session 1: Sampling

Introduction

Have you ever tasted a hot soup and decided whether the soup was tasty or not? If yes, then you are a sampler. Sampling is a part of our day-to-day life, which we use either advertently or inadvertently. Another example is a pathologist who takes a few drops of blood and tests for any abnormality in the blood of the whole body.

The process of using information obtained from the smaller quantity to make statement about the larger quantity is called sampling. In this lecture, we shall examine why this process is sometimes necessary and the various techniques for doing it. We shall first learn some fundamental concepts, which are related to sampling.

Learning Outcomes for Study Session 1

When you have studied this session you should be able to:

1.1 Define Sampling

1.2 Discuss the reasons for sampling.

1.3 Discuss the various procedures of sampling.

1.4 Distinguish between census, population and sampl

1.1 Sampling Definition

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed, but may include simple random sampling or systematic sampling.

1.1.1 Sampling and Non-Sampling Errors

Any statistical inference based on sample results may not always be correct. This is because sample results are either based on partial or incomplete analysis of the population characteristics. This error is referred to as the **sampling error** because each sample taken may produce a different estimate of the population characteristic compared to those results that would have been obtained by a complete enumeration of the population.

Non-sampling Errors arise during census as well as sampling surveys because of biases and mistakes such as:

1. Faulty Planning
2. Non-response;
3. Non-random selection of samples;
4. Incompleteness and inaccuracy of returns;
5. Compilation errors.

In-Text Question

__________is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population.

a. Statistics

b. Standard deviation

c. Sampling

d. Sample error

In-Text Answer

c

1.2 Sampling Methods

Several sampling methods are available, which are classified into two categories:

1. Random Sampling Methods
2. Non-Random Sampling Methods

1. **Random Sampling Methods**

A statistic is a measurable characteristic of a sample, such as a mean or standard deviation. A sampling method is a procedure for selecting sample elements from a population. A random number is a number determined totally by chance, with no predictable relationship to any other number. The following are example of random sampling methods:

- Simple random sampling
- Stratified sampling
- Cluster sampling
- Systematic sampling
- Multi-stage sampling
- **Simple Random Sampling (SRS)**

In this type of sampling, each unit of the population has equal chance of being included in the sample. If the units are drawn one by one in such a way that a unit drawn at a time is replaced back into the population before the subsequent draw, it is known as simple random sampling with replacement (srswr).

However, if the unit selected once is not included in the population at any subsequent draw, it is called simple random sampling without replacement (srswor). One disadvantage of this method is that all members of the population have to be available for selection. However, this availability may not be possible in most cases.

- **Stratified Sampling**

This method is useful when the population as a whole consists of a number of heterogeneous groups while the units within each group are relatively homogeneous. Thus, population is divided into distinct groups called *strata*. A simple random sample is drawn from each *stratum* or *group*, in proportion to its size. Individual stratum samples are combined into one to obtain an overall sample for analysis.

- **Cluster Sampling**

This method is also referred to as area sampling method. It is useful when the population consists of a very large number of similar groups which are wide-spreading. It is devised to meet the problem of costs or inadequate sampling *frame* (a complete listing of all members in the population so that each member can be identified by a distinct number).

By the use of map references, the entire area to be analyzed is divided into smaller areas and a sample of the desired number of areas is selected by a SRS method. Such groups are termed as *clusters*. The members of the clusters are called **elementary units**. From each cluster, we may select a random sample of the desired size.

- **Systematic Sampling**

This method is useful when units in the population are physically arranged in some sequence, and every k^{th} unit is included in the sample after the first has been randomly selected. The value k is called the *sampling interval.*

A systematic sample has the advantage of being quick and easy to use. However, we need to be careful lest there be a cyclic variation in the frame, and this is picked up in the sample because it is organized cyclically.

- **Multi stage Sampling**

In multistage sampling, the whole population is divided into a number of primary units called *stages*, each of which is composed of second stage units. A series of samples are then taken at successive stages. The sample size at each stage is determined by the relative population size at each stage.

Non-Random Sampling Methods

It results in a biased sample, a non-random sample of a population (or non-human factors) in which all individuals, or instances, were not equally likely to have been selected. If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling.

- Quota sampling
- Judgment or Purposive sampling
- Convenience sampling

1. Quota Sampling

This method is often used in market and social surveys. The selection of respondents lies with the investigator's discretion; although, care must be taken to ensure that each respondent satisfies certain criteria, which are essential for the study. Because quota sampling is non-random, it leads to substantial complications in the statistical analysis of the survey results.

2. Judgment or Purposive Sampling

In this method, the investigator selects units of the sample that he/she feels are most representative of the population with respect to the population characteristics under study. Great precaution is needed in drawing conclusions based on judgment samples to make inferences about a population.

3. Convenience Sampling

Perhaps this is the easiest method for collecting data on a particular issue. The investigator simply selects units to be included in the sample at his/her convenience, rather than following a pre specified rule. Precautions are also needed in interpreting the results of convenience sampling that are used to make inferences about a population.

1.3 Some Basic Concepts of Sampling

The following are some of the basic concept of sampling

- **Census**
- **Sample**

1. Census

A census involves a complete count (or a complete enumeration) of every individual member of the population of interest, such as persons in a country, households in a town, shops in a city, students in a college, and so on.

Apart from the cost and the large amount of resources (such as enumerators, clerical assistance, etc.) that are required, the main problem is the time required to process the data. Thus, the results are not known immediately.

2. Population

In statistical sense, population is a group of items, units or subjects, which is under reference of study. It is often referred to as *universe* by a number of statisticians and scientists. The inhabitants of a region, number of cars in a city, workers in a factory, students in a university, insects in a field, etc., are few examples of populations. Generally, populations or universe is classified into four categories:

- **Finite population**- the number of items or units is fixed, limited and countable, e.g. workers in a factory.

- **Infinite population**- the number of items or units is uncountable, e.g. stars in the sky.
- **Real population-** the items or units in the population are all physically present or visible.
- **Hypothetical population**- the population results from repeated trials, e.g. the tossing of a coin repeatedly results into a hypothetical population of heads and tails, rolling of a die again and again gives rise to a hypothetical population of numbers from 1 to 6, etc.

3. Sample

A sample is a part or fraction of a population selected on some basis. In principle, a sample should be such that it is a true representative of the population. The process of selecting a sample from the population is called ***sampling***, and the manner or scheme through which the required number of units is selected is called the ***sampling method***.

The foremost purpose of sampling is to gather maximum information about the population under consideration at minimum cost, time and resources. Precisely, sampling is inevitable in the following situations:

- When population is infinite
- When the item or unit is destroyed under investigation;
- When the results are required in a short time
- When resources are limited particularly in respect of money and trained persons
- When population is either constantly changing or in a state of movement;
- When the items or units are scattered.

Summary

In this study session, you have learnt the following:

1. Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a

larger population depends on the type of analysis being performed, but may include simple random sampling or systematic sampling.

2. Several sampling methods are available, which are classified into two categories:
   - Random Sampling Methods
   - Non-Random Sampling Methods
3. A census involves a complete count (or a complete enumeration) of every individual member of the population of interest, such as persons in a country, households in a town, shops in a city, students in a college, and so on.

Self-Assessment Questions (SAQs) for Study Session 1

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 1.1 (Testing Learning Outcomes 1.1)

Define Sampling

SAQ 1.2 (Testing Learning Outcomes 1.2

Discuss the reasons for sampling

SAQ 1.3 (Testing Learning Outcomes 1.3)

Discuss the various procedures of sampling.

SAQ 1.4 (Testing Learning Outcomes 1.4)

Distinguish between census, population and sample

Study Session 2: Sampling Distribution of a Statistics

Introduction

In the previous study session, population and sampling was discussed. Imagine that you have a large population to study and the description of its characteristics is not possible by census method. Then, in order to make statistical inference, samples of given size are drawn repeatedly from the population and '***statistic***' computed for each sample.

The computed value of a particular statistic will differ from sample to sample. This implies that, if the same statistic is computed for each of the samples, the value is likely to vary from sample to sample.

Learning Outcomes for Study Session 2

When you have studied this session, you should be able to:

2.1 Explain the Concept of a Sampling Distribution

2.2 Explain the Concept of standard error

2.3 Differentiate between Standard error from Standard deviation.

2.1 Concept of a Sampling Distribution

The following are some concepts of sampling distribution;

Sampling Unit

Elementary units or group of such units which besides being clearly defined, identifiable and observable, are convenient for purpose of sampling are called sampling units.

For instance, in a family budget enquiry, usually a family is considered as the sampling unit since it is found to be convenient for sampling and for ascertaining the required information. In a crop survey, a farm or a group of farms owned or operated by a household may be considered as the sampling unit.

Population

The collection of all units of a specified type in a given region at a particular point or period of time is termed as a population or universe. Thus, we may consider a population of persons, families, farms, cattle in a region or a population of trees or birds in a forest or a population of fish in a tank etc. depending on the nature of data required.

Population Distribution

The population distribution is the distribution of values of its members and has mean denoted by μ, variance σ^2 and standard deviation σ. For example, a population consisting of the numbers 0, 2, 4 and 6 has mean $\mu = 3$ and standard deviation $\sigma = \sqrt{5}$.

Definition [All possible samples of size n]

Let a population consist of N elements.

1. If a random sample of size n is selected from the population **with replacement,** then there are N^n possible samples of size n that can be drawn from the population.
2. If a random sample of size n is selected from the population **without replacement**, then there are $N_{C_n} = \frac{N!}{n!(N-n)!}$ possible samples of size n that can be drawn from the population.

Example

A population consists of the numbers 0, 2, 4 and 6, List all possible samples of size 2 that can be drawn

1. with replacement
2. without replacement

Solution

1. The population size N = 4 and sample size n = 2, therefore, 4^2 =16 possible samples can be drawn with replacement. The list of the possible samples is given as follow:

Sample number	Sample elements
1	0, 0
2	0, 2
3	0, 4
4	0, 6
5	2, 0
6	2, 2
7	2, 4
8	2, 6
9	4, 0
10	4, 2
11	4, 4
12	4, 6
13	6, 0
14	6, 2
15	6, 4
16	6, 6

The population size N = 4 and sample size n = 2, therefore, $4_{C_2} = \frac{4!}{2!(4-2)!} = 6$ possible samples can be drawn without replacement. The list of the possible samples is given as follow:

Sample number	**Sample elements**
17	0, 2
18	0, 4
19	0, 6
20	2, 4
21	2, 6
22	4, 6

Sampling Distribution of a sample statistic

If a particular statistic (e.g. sample mean, sample standard deviation, etc.) is computed for each of the possible samples, the value of the statistic will differ from sample to sample. Thus, it would be theoretically possible to construct a frequency table showing the values assumed by the statistic and their frequency of occurrence.

This distribution of values of a statistic is called a sampling distribution. Thus, we see that there would be an overall mean (where it is centered), a standard deviation (representing the spread) and a shape if the histogram is plotted. So, we can talk of the mean of sampling distribution of a statistic (denoted μ_m if m is the statistic), and standard deviation of sampling distribution of a statistic (denoted σ_m if m is the statistic).

These properties help lay down rules for making statistical inferences about a population on the basis of a single sample drawn from it, that is, without even repeating the sampling process.

In-Text Question

Sampling distribution is the probability distribution of all possible values of a given statistic from all the distinct possible samples of equal size drawn from a population

In-Text Answer

True

Statistic, Estimator and Estimate

Suppose a sample of n units is selected from a population of N units according to some probability scheme and let the sample observations be denoted by $y_1, y_2, \ldots, y_n$. Any function of these values which is free from unknown population parameters is called a statistic.

An estimator is a statistic obtained by a specified procedure for estimating a population parameter. The estimator is a random variable and its value differs from sample to sample and the samples are selected with specified probabilities. The particular value, which the estimator takes for a given sample, is known as an estimate

In-Text Question

Population is the collection of all units of a specified type in a given region at a particular point or period of time is termed as a population or universe

In-Text Answer

True

2.2 Standard Error of Statistic

The standard deviation of sampling distribution of a statistic is called the standard error of the statistic. It is clearly different from the population standard deviation (σ). The population standard deviation describes the variation among values of members of the population, whereas the standard deviation of sampling distribution measures the variability among values of the statistic due to sampling error.

Standard error is a measure of a reasonable difference between a particular sample statistic and the population parameter.

It is used in tests of whether a particular sample could have been drawn from a given parent population. It is also used in working out confidence limits and confidence intervals.

2.3 Difference between standard error and standard deviation

The **standard deviation**, or **SD**, measures the amount of variability or dispersion for a subject set of data from the mean, while the **standard error** of the mean, or SEM, measures how far the sample mean of the data is likely to be from the true population mean. The SEM is always smaller than the **SD**.

When dealing with numerical data sets, many people get confused between the standard deviation of the sample and the standard error of the sample mean. We want to stress the difference between these.

Standard deviation (SD)

This describes the spread of values in the sample. The sample standard deviation, *s*, is a random quantity -- it varies from sample to sample -- but it stays the same on average when the sample size increases.

Standard error of the mean (SE)

This is the standard deviation of the sample mean, $\bar{x}$, and describes its accuracy as an estimate of the population mean, μ. When the sample size increases, the estimator is based on more information and becomes more accurate, so its standard error decreases.

In-Text Question

Standard deviation describes the spread of values in the sample.

In-Text Answer

True

Summary

In this study session, you have learnt the following:

1. There are N^n possible samples of size n that can be drawn with replacement from a population having N elements;
2. There are $N_{C_n} = \frac{N!}{n!(N-n)!}$ possible samples of size n that can be drawn without replacement from the population having N elements.;
3. Sampling distribution is the probability distribution of all possible values of a given statistic from all the distinct possible samples of equal size drawn from a population.; and
4. Standard error of statistic measures the amount of chance error in the sampling process.

Self-Assessment Questions (SAQs) for Study Session 2

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 2.1 (Testing Learning Outcomes 2.1)

Explain the Concept of a sampling distribution

SAQ 2.2 (Testing Learning Outcomes 2.2)

Explain the Concept of standard error

SAQ 2.3 (Testing Learning Outcomes 2.3)

Differentiate between standard error from standard deviation.

Study Session 3: Sampling Distribution of Sample Mean

Introduction

The sample mean is referred to as the *point estimate* of the population mean. For example, if you are interested in the mean rent charged for a 2-bedroom apartment in the Bodija area of Ibadan, you may obtain a random sample and from that sample you obtain the sample mean.

This sample mean is one number which estimates the Population mean rent for 2-bedroom apartment in the area. The *sampling distribution of the mean* refers to the distribution of all the possible sample means that could be obtained if you select all possible samples of a given size.

.

Learning Outcomes for Study Session 3

When you have studied this session, you should be able to:

3.1 List the properties of the sampling distribution of the sample mean;

3.2 Determine the sampling distribution of mean when population has normal distribution;

3.3 Determine the sampling distribution of mean when population has non-normal distribution.

3.4 Determine the sampling distribution of the difference between two sample means.

3.1 Properties of the Sampling Distribution of the Sample Mean

There are three very important properties associated with the sampling distribution of the sample mean. These properties are the *centre, spread* and *shape* of the sampling distribution.

1. **Centre: The sample mean is an unbiased estimator**

The arithmetic mean $\mu_{\bar{X}}$ of sampling distribution of mean values (also called mean of means) is equal to the population mean μ regardless of the form of population distribution, that is, $\mu_{\bar{X}} = \mu$.

Example 1

A population consists of the numbers 0, 2, 4 and 6. The population mean μ = 3. Now, consider all possible samples of size 2 without replacement from the population and their means as shown in the following table.

sample number	sample elements	sampling distribution of mean
1	0, 2	1
2	0, 4	2
3	0, 6	3
4	2, 4	3
5	2, 6	4
6	4, 6	5

The arithmetic mean of sampling distribution of mean value is $\mu_{\bar{X}} = \frac{1+2+3+3+4+5}{6} = 3$.

In-text Question

The arithmetic mean $\mu_{\bar{X}}$ of sampling distribution of mean values (also called *mean of means*) is equal to the population mean μ regardless of the form of population distribution, that is, $\mu_{\bar{X}} = \mu$.

TRUE or FALSE

In-text Answer

True

2. **Spread: Standard error of the mean**

The sampling distribution of mean has a standard deviation (also called standard error of the mean) equal to the population standard deviation divided by the square root of the sample size, that is, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. It should be noted from this formula that its value tends to be smaller as the size of sample n increases and vice-versa.

When the standard deviation σ of population is not known, the standard deviation *s* of the sample is used to compute the standard error, that is, $\sigma_{\bar{X}} = \frac{s}{\sqrt{n}}$.

3. **Shape: The shape of the sampling distribution of mean**

The sampling distribution of sample mean values from normally distributed population is the normal distribution for samples of all sizes.

3.2 Sampling Distribution of Mean When Population has Normal Distribution

If all possible samples of size n are drawn *with replacement* from a population having normal distribution with mean μ and standard deviation σ, then it can be shown that the sampling distribution of mean $\bar{X}$ and standard error $\sigma_{\bar{X}}$ will also be normally distributed irrespective of the size of the sample.

The procedure for making statistical inference using sampling distribution about the population mean μ based on $\bar{X}$ of sample means is summarized as follows:

- If the population standard deviation σ value is known and either
    a. population distribution is normal, or
    b. population distribution is not normal, but the sample size n is large ($n \geq 30$), then the sampling distribution of mean $\mu_{\bar{X}}$ = μ. and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, is very close to the standard normal distribution given by $z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

- If the population is finite and the samples of fixed size n are drawn *without replacement*, then $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$, where $\sqrt{\frac{N-n}{N-1}}$ is called the *finite correction factor*.

3.3 Sampling Distribution of Mean When Population has Non-Normal Distribution

- If population is not normally distributed, then we make use of the central limit theorem. This theorem says that:

As the sample size is increased, the sampling distribution of mean can be approximated by the normal distribution, regardless of the population distribution.

As a general rule, if the sample size is at least 30, the sampling distribution of mean $\bar{X}$ is assumed to be normally distributed, regardless of the form of the population distribution.

- If σ is not known, the value of z cannot be calculated for a specific sample. In such a case, the standard deviation of population σ must be estimated using the sample standard deviation s. Thus $\sigma_{\bar{X}} = \frac{s}{\sqrt{n}}$ and consequently $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a distribution called student's t distribution.

3.4 Sampling Distribution of Difference between Two Sample Means

The concept of sampling distribution of sample mean can also be used to compare a population of size N_1 having μ_1 and standard deviation σ_1 with another similar type of population of size N_2 having μ_2 and standard deviation σ_2. Let $\bar{X}_1$ and $\bar{X}_2$ be the mean of sampling distribution of mean of the two populations, respectively.

Then the difference between their mean values μ_1 and μ_2 can be estimated by generalizing the formula of standard normal variable as follows:

$$z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_{\bar{X}_1} - \mu_{\bar{X}_2}\right)}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

where

$\mu_{\bar{X}_1-\bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2 \Rightarrow$ mean of sampling distribution of difference of two means

$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \Rightarrow$ Standard error of sampling distribution of difference of two means

n1 and *n2* $\Rightarrow$ independent random samples drawn from first and second population, respectively.

Example

The strength of the wire produced by company A has a mean of 4,500 kg and a standard deviation of 200 kg. Company B has a mean of 4,000 kg and a standard deviation of 300 kg. If 50 wires of company A and 100 wires of company B are selected at random and tested for strength, what is the probability that the sample mean strength of A will at least 600 kg more than that of B?

Solution

We are given the following information:

Company A: $\mu_A = 4{,}500$, $\sigma_A = 200$ and $n_A = 50$

Company B: $\mu_B = 4{,}000$, $\sigma_B = 300$ and $n_B = 100$.

Thus,

$$\mu_{\bar{X}_A-\bar{X}_B} = \mu_{\bar{X}_A} - \mu_{\bar{X}_B} = \mu_A - \mu_B = 4{,}500 - 4000 = 500$$

and

$$\sigma_{\bar{X}_A-\bar{X}_B} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = \sqrt{\frac{40{,}000}{50} + \frac{90{,}000}{100}} = 41.23$$

$$P\left[\left(\bar{X}_A - \bar{X}_B\right) \geq 600\right] = P\left[z \geq \frac{(\bar{X}_A-\bar{X}_B)-(\mu_A-\mu_B)}{\sigma_{\bar{X}_A-\bar{X}_B}}\right] = P\left[z \geq \frac{600-500}{41.23}\right] = 0.0075.$$

Summary

In this study session, you have learnt the following:

1. Properties of the Sampling Distribution of the Sample Mean are:

   a. The sample mean is an unbiased estimator of the population mean.

   b .Standard error of the mean equals to the population standard deviation divided by the square root of the sample size.

   c .The sampling distribution of sample mean values from normally distributed population is the normal distribution for samples of all size.:

2. If the sample size is at least 30, the sampling distribution of mean $\bar{X}$ is assumed to be normally distributed, regardless of the

Self-Assessment Questions (SAQs) for Study Session 3

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 3.1 (Testing Learning Outcomes 3.1)

List the properties of the sampling distribution of the sample mean;

SAQ 3.2 (Testing Learning Outcomes 3.2)

Determine the sampling distribution of mean when population has normal distribution;

SAQ 3.3 (Testing Learning Outcomes 3.3)

Determine the sampling distribution of mean when population has non-normal distribution

SAQ 3.4 (Testing Learning Outcomes 3.4)

Determine the sampling distribution of the difference between two sample means

Study Session 4: Sampling Distribution of Sample Proportion

Introduction

There are many situations in which each individual member of the population can be classified into two mutually exclusive categories, such as success or failure, accept or reject, head or tail of a coin, and so on. For instance, the population could be registered voters living in a city, and the attribute is “plans to vote for party A in presidential elections”.

We take a random sample from the population and observe the number in the sample planning to vote party A in presidential elections. There are two possible outcomes, “success” and “failure.” Success on voter *i* means voter *i* plans to vote party A and failure vice versa. The *sample proportion* can then be defined as the number of successes divided by the sample size.

With the same logic of sampling distribution of mean, the sampling distribution of sample proportion can be derived.

Learning Outcomes for Study Session 4

When you have studied this session, you should be able to:

4.1 Define sample proportion;

4.2 List the properties of the sampling distribution of sample proportion.

4.3 Determine the sampling distribution of the difference of two proportions.

4.1 Sample Proportion

The sample proportion $\overline{p}$ is defined as:

$$\overline{p} = \frac{\text{Number of successes, } x}{\text{Sample size, } n}.$$

The sample proportion $\overline{p}$ having the characteristic of interest is the best statistic to use for statistical inferences about the population proportion parameter p. For example, a company writing industrial accident insurance might estimate as 0.71 the proportion of its policyholders who file at least one claim per year, if a sample check of 200 policies shows that 142 had at least one claim filed during 2006.

4.2 Sampling Distribution of Sample Proportion

With the same logic of sampling distribution of mean, the following properties hold for the sampling distribution of sample proportion:

1. The mean of the sampling distribution of sample proportion, denoted as $\mu_{\overline{p}}$, is equal to the population proportion p. That is,

   $$\mu_{\overline{p}} = p$$

2. The standard deviation of the sampling distribution of sample proportion (also called *standard error of* $\overline{p}$) $\sigma_{\overline{p}}$ is given by:

   $\sigma_{\overline{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$ if sampling is with replacement and

   $\sigma_{\overline{p}} = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}$ if sampling is without replacement.

3. If a large sample size $(n \geq 30)$ satisfies the following two conditions:

   i. $np \geq 5$;

   ii. $n(1-p) \geq 5$;

then the sampling distribution of the sample proportion is approximately normally distributed. Thus, to standardize proportion $\overline{p}$, the standard normal variable

$$z = \frac{\overline{p}-\mu_{\overline{p}}}{\sigma_{\overline{p}}} = \frac{\overline{p}-p}{\sqrt{P(1-p)/n}}$$

is approximately the standard normal distribution.

Example

A manager in the billing section of a mobile phone company checks on the proportion of customers who are paying their bills late. Company policy dictates that this proportion should not exceed 20 per cent. Suppose that the proportion of all invoices that were paid late is 20 per cent. In a random sample of 140 invoices, determine the probability that more than 28 per cent invoices are paid late.

Solution

Given $\mu_{\overline{p}} = p = 0.20$, n = 140;

$$\sigma_{\overline{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.20\times 0.80}{140}} = 0.033$$

$$P\left[\overline{p} \geq 0.28\right] = P\left[z \geq \frac{\overline{p}-p}{\sigma_{\overline{p}}}\right]$$

$$= P\left[z \geq \frac{0.28-0.20}{0.033}\right]$$

$$=P\left[z \geq 2.42\right] = 0.0082$$

In-text Question

The mean of the sampling distribution of sample proportion, denoted as $\mu_{\overline{p}}$, is equal to the population proportion p. That is: $\mu_{\overline{p}} = p$ .true or False

In-text Answer

True

4.3 Sampling Distribution of the Difference of Two Sample Proportions

Suppose two populations of size N_1 and N_2 are given. For each sample of size n_1 from first population, compute sample proportion $\overline{p}_1$ and standard deviation $\sigma_{\overline{p}_1}$. Similarly, for each sample of size n_2 from second population, compute sample proportion $\overline{p}_2$ and standard deviation $\sigma_{\overline{p}_2}$.

For all combinations of these samples from these populations, we can obtain a sampling distribution of the difference $\overline{p}_1$ - $\overline{p}_2$ of samples proportions. Such a distribution is called *sampling distribution of difference of two sample proportions*. The mean and standard deviation of this distribution are given as follows.

Mean: $\mu_{\overline{p}_1-\overline{p}_2} = \mu_{\overline{p}_1} - \mu_{\overline{p}_2} = p_1 - p_2$

Standard Deviation: $\sigma_{\overline{p}_1-\overline{p}_2} = \sqrt{\sigma^2_{\overline{p}_1} + \sigma^2_{\overline{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

If the sample size n_1 and n_2 are large, that is, $n_1 \geq 30$ and $n_2 \geq 30$, then the sampling distribution of difference of proportions is closely approximated by a normal distribution.

Example

Ten per cent of machines produced by company A are defective, and five per cent of those produced by company B are defective. A random sample of 250 machines is taken from company A and a random sample of 300 machines from company B. What is the probability that the difference in sample proportion is less than or equal to 0.02?

Solution

We are given the following information

$\mu_{\overline{p}_A-\overline{p}_B} = \mu_{\overline{p}_A} - \mu_{\overline{p}_B} = p_A - p_B = 0.10 -$ 0.05=0.05; $n_A = 250$ and $n_B = 300$.

Thus standard error of the difference in a sample proportion is given by

$$\sigma_{\bar{p}_A-\bar{p}_B} = \sqrt{\frac{p_A(1-p_A)}{n_A}+\frac{p_B(1-p_B)}{n_B}} = \sqrt{\frac{0.10\times0.90}{250}+\frac{0.50\times0.95}{300}} = 0.0228$$

The desired probability of difference in sample proportions is given by

$$P\left[\left(\bar{p}_A-\bar{p}_B\right)\le 0.02\right] = P\left[z\le\frac{\left(\bar{p}_A-\bar{p}_B\right)-\left(p_A-p_B\right)}{\sigma_{\bar{p}_A-\bar{p}_B}}\right]$$

$$= P\left[z\le\tfrac{0.02-0.05}{0.0228}\right] = P\left[z\le -1.32\right] = 0.0934\,.$$

In-text Question

The formula below is for calculating___________________

$$\sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{\sigma^2_{\bar{p}_1}+\sigma^2_{\bar{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

a. Mean

b. Arithmetic mean

c. Standard deviation

d. Probability

In-text Answer

c. Standard deviation

Summary

In this study session, you have learnt the following:

1. The sample proportion is defined as the number of units in the sample having the characteristic of interest divided by the sample size.
2. The sampling distribution of sample proportion has the following properties:

1. $\mu_{\bar{p}} = p$.

2. $\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$

3. $z = \frac{\bar{p}-\mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{\bar{p}-p}{\sqrt{P(1-p)/n}}$ is approximately the standard normal.

3. The mean and standard deviation of the sampling distribution of the difference of two sample proportions are given as follows:

Mean: $\mu_{\bar{p}_1-\bar{p}_2} = \mu_{\bar{p}_1} - \mu_{\bar{p}_2} = p_1 - p_2$

Standard Deviation: $\sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{\sigma^2_{\bar{p}_1} + \sigma^2_{\bar{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Self-Assessment Questions (SAQs) for Study Session 4

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 4.1 (Testing Learning Outcomes 4.1)

Define sample proportion

SAQ 4.2 (Testing Learning Outcomes 4.2

List the properties of the sampling distribution of sample proportion

SAQ 4.3 (Testing Learning Outcomes 4.3

Determine the sampling distribution of the difference of two proportions

Study Session 5: Estimation and Confidence Intervals

Introduction

Statistical methods of estimation are applicable almost anywhere; in science, in business, as well as in other areas of life. In science, a biologist may wish to estimate what proportion of a certain kind of insect is born physically defective. In business, a finance company may wish to estimate what proportion of its customers plan to buy a new car within the next year.

Furthermore, in driving, we may want to estimate what percentage of car accidents is due to faulty brakes. Basically, one may be dealing with the estimation of a percentage or proportion, average or mean and measure of variation. These (and particularly the first two) are the parameters with which we are concerned in most problems of estimation.

You will learn about them in the next study session but in this session you will learn the general principles of estimation.

Learning Outcomes for Study Session 5

When you have studied this session, you should be able to:

5.1 Define Estimation

5.2 Distinguish clearly between point and interval estimation;

5.3 List the properties of a point estimator

.

5.1 Estimation:

A point **estimate** of a population parameter is a single value of a **statistic**. For example, the sample mean x is a point **estimate** of the population mean μ. Similarly, the sample proportion p is a point **estimate** of the population proportion P. Interval **estimate**.

5.1.1 Types of Estimation

There are two types of estimation for a population parameter:

- Point estimation
- Interval estimation

Point Estimation

In point estimation, a single sample statistic (such as $\overline{X}, s, \ or\ \overline{p}$) is calculated from the sample to provide a best estimate of the true value of the corresponding population parameter (such as $\mu, \sigma, \ or\ p$). Such a single relevant statistic is termed as point estimator, and the value of the statistic is termed as point estimate.

For example, we may calculate that 10 per cent of the items in a random sample taken from a day's production are defective. The result '10 per cent' is a point estimate of the percentage of items in the whole lot that are defective. Thus, until the next sample of items is drawn and examined, we may proceed on manufacturing on the assumption that the day's production contains 10 per cent defective items.

One serious shortcoming of point estimates is that they do not tell us how close we can expect them to be to the quantities they are supposed to estimate. In other words, *point* estimates do not tell us anything about the intrinsic reliability or precision of the method of estimation which is being used.

Therefore, point estimates should always be accompanied by some information, which makes it possible to judge their merits. How this is done is the main concern in interval estimation.

Interval Estimation

For estimating a parameter value, it is important to know, (i) a point estimate, (ii) the amount of possible error in the point estimate, that is, an interval likely to contain the parameter value, and (iii) the statement or degree of confidence that the interval contains the parameter value.

The knowledge of these three pieces of information is called a confidence interval or interval estimation. Thus, an interval estimate of a population parameter is therefore a confidence interval with a statement of confidence that the interval contains the parameter value.

The confidence interval estimate of a population parameter is obtained by applying the formula:

Point estimate ± Margin of error

Where

Margin of error = Z_c x Standard error of a particular statistic

Z_c = critical value of standard normal variable that represents confidence level (probability of being correct) such as 0.90, 0.95, 0.99, and so on.

An interesting feature of the formula for the margin of error is that it can also be used to determine the sample size that is required to attain a desired degree of precision. You will learn about this in study session 7

In-Text Question

There are two types of estimation for a population parameter:

- Point estimation
- Interval estimation

In-Text Answer

True

5.2 Properties of a Point Estimator

The sample statistic value varies from sample to sample. Thus, the accuracy of a given estimator also varies from sample to sample. This means that there is no certainty of the accuracy achieved for the sample one happens to draw.

Although in practice, only one sample is selected at any given time, we should judge the accuracy of an estimator based on its average value over all possible samples of equal size. Therefore, we prefer to choose an estimator whose 'average accuracy' is close to the value of population parameter being estimated. The criteria for selecting an estimator are:

1. Unbiasedness

The value of a statistic measured from a given sample is likely to be above or below the actual value of population parameter of interest due to sampling error. Thus, it is desirable that the expected value or mean of all possible values of a statistic from the estimates over all possible random samples is equal to the population parameter being estimated.

If this is true, then the sample statistic is said to be an unbiased estimator of the population parameter. It can be shown that both $\bar{\mathrm{X}}$ and $\bar{p}$ are unbiased estimators of the corresponding population parameters μ and p respectively. However, sample standard deviation s is not an unbiased estimator of σ. But this bias reduces as sample size increases.

2. Consistency

A point estimator is said to be consistent if its value tends to become closer to the population parameter as the sample size increases. For example, the standard error of sampling distribution of the mean, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, tends to become smaller as sample size n increases.

Thus, the sample mean $\bar{X}$ is a consistent estimator of the population mean μ. Similarly, the sample proportion $\bar{p}$ is a consistent estimator of the population proportion p because $\sigma_{\bar{p}} = \frac{\sigma}{\sqrt{n}}$.

3. Efficiency

For the same population, out of two unbiased point estimators, an unbiased estimator with smaller standard deviation is said to be efficient when it provides an estimate closer to the population parameter. It is for this reason that there is less variation in the sampling distribution of the statistic.

Summary

In this study session, you have learnt the following:

1. A point **estimate** of a population parameter is a single value of a **statistic**. For example, the sample mean x is a point **estimate** of the population mean μ. Similarly, the sample proportion p is a point **estimate** of the population proportion P. Interval **estimate**.

2. Two types of estimation:
   - Point estimation
   - Interval Estimation

3. The confidence interval estimate of a population parameter is obtained by applying the formula:

   Point estimate ± Margin of error

4. Properties of a point estimator:

- Unbiasedness
- Consistency
- Efficiency

Self-Assessment Questions (SAQs) for Study Session 5

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 5.1 (Testing Learning Outcomes 5.1)

Define estimation

SAQ 5.2 (Testing Learning Outcomes 5.2)

Distinguish clearly between point and interval estimation;

SAQ 5.3 (Testing Learning Outcomes 5.3)

List the properties of a point estimator

Study Session 6: Interval Estimation of Population Mean and Proportion

Introduction

The need to use the sample statistic to draw conclusions about the population characteristic is one of the fundamental applications of statistical inference in business, economics, etc. In the previous lecture, you learnt about the general principles of estimation. The focus in this study session is to apply the principles to mean and proportion, which are the most common problems in estimation.

Learning Outcomes for Study Session 6

When you have studied this session, you should be able to:

6.1 Construct and interpret confidence interval estimates for population mean.

6.2 Construct and interpret confidence interval estimates for population proportion.

6.1 Interval Estimation of Population Mean (σ known)

Suppose the population mean μ is unknown and the true population standard deviation σ is known. Then for a large sample size $(n \geq 30)$, the interval estimation of population mean μ is given by

$$\bar{X} \pm Z_{\alpha/2}\sigma_{\bar{X}}$$

or

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Where $Z_{\alpha/2}$ is the *Z*-value representing an area $\alpha/2$ in the right and left tails of the standard normal probability distribution, and (1 – α) is the *level of confidence.*

The values of $Z_{\alpha/2}$ for most commonly-used as well as the other confidence levels can be seen from standard normal probability table below.

Confidence level (1 – α) (%)	*Acceptable* *Error Level,* α	α/2	$Z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.960
99%	0.01	0.005	2.576

In general, a 95 per cent confidence interval estimate implies that, if all possible samples of the same size were drawn, then 95 per cent of them would include the true population mean somewhere within the interval around their sample mean and only 5 per cent of them would not.

Example

The average monthly electricity consumption for a sample of 100 families is 1250 units. Assuming the standard deviation of electric consumption of all families is 150 units, construct a 95 per cent confidence interval estimate of the actual mean electric consumption.

Solution

$\bar{X}$ = 1250, σ = 150, n = 100 and confidence level (1 – α) = 95 per cent.

Using the standard normal probability table $Z_{\alpha/2}$ = 1.96. Therefore,

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1250 \pm 1.96\frac{150}{\sqrt{100}} = 1250 \pm 29.40 \text{ units.}$$

Thus, for 95 per cent level of confidence, the population mean μ is likely to fall between 1220.60 units and 1274.40 units, that is, 1220.60 ≤ μ ≤ 1274.40.

In-text Questions

The interval estimation of population mean μ is given by

$$\bar{X} \pm Z_{\alpha/2}\sigma_{\bar{X}}$$

or

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

True or False

In-text Answers

True

6.2 Interval Estimation of Population Mean (σ unknown)

In practice, the standard deviation of a population σ, is not likely to be known. When σ is unknown and *n* is 30 or more, we proceed as before and estimate σ with the sample standard deviation *s*. The resulting 1 – α *large sample confidence interval* for μ becomes

$$\bar{X} - Z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\alpha/2}\frac{s}{\sqrt{n}}$$

When the σ is not known and the sample size is small, the procedure for interval estimation of population mean is based on a probability distribution known as the *Student-t distribution.* The shape of the distribution is very much like that of the normal curve; it is symmetrical with zero mean, but there is slightly higher probability of getting values falling into the two tails.

Actually, the shape of the *t* distribution depends on the size of the sample or, better, on the quantity *n* – 1, which in this connection is called the *number of degrees of freedom.* As the number of degrees of freedom increases, *t* distribution gradually approaches the normal distribution, and the sample standard deviation *s* becomes a better estimate of population standard deviation σ.

The interval estimate of a population mean when the sample size is small $(n \leq 30)$ with confidence coefficient (1 – α), is given by

$$\bar{X} \pm t_{\alpha/2} \tfrac{s}{\sqrt{n}}$$

or

$$\bar{X} - t_{\alpha/2} \tfrac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2} \tfrac{s}{\sqrt{n}}$$

The critical values of *t* for the given degrees of freedom can be obtained from the table of *t* distribution.

Example

The personnel department of an organization would like to estimate the family dental expenses of its employees to determine the feasibility of providing a dental insurance plan. A random sample of 10 employees reveals the following family dental expenses (in thousands of Naira) in the previous year: 11, 37, 25, 62, 51, 21, 18, 43, 32, 20. Set up a 99 per cent confidence interval of the average family dental expenses for the employees of this organization.

Solution

From the data, the sample mean $\bar{X} = \frac{\sum X}{n} = \frac{320}{10} = 32$ and the sample standard deviation $s = \sqrt{\sum \left(X - \bar{X} \right)^2 / n - 1} = \sqrt{\frac{2358}{9}} = 5.11$

Using this information and $t_{\alpha/2}$ = 1.833 at degree of freedom (*df*) = 9,

We have

$$\bar{X} \pm t_{\alpha/2} \tfrac{s}{\sqrt{n}} = 32 \pm 1.833 \tfrac{5.11}{\sqrt{10}} = 32 \pm 2.962$$

Therefore, the mean expenses per family are likely to fall between ₦29.038 and ₦34.962, that is, $29.038 \leq \mu \leq 34.962$.

6.3 Interval Estimation for Population Proportion

You will recall that normal distribution as an approximation of the sampling distribution of sample proportion is based on large sample conditions. Therefore, the confidence interval estimate for a population proportion at $1 - \alpha$ confidence coefficient is given by

$$\bar{p} \pm Z_{\alpha/2} \sigma_{\bar{p}}$$

or

$$\bar{p} \pm Z_{\alpha/2} \sqrt{\tfrac{\bar{p}(1-\bar{p})}{n}}$$

or

$$\bar{p} - Z_{\alpha/2} \sigma_{\bar{p}} \leq \mu \leq \bar{p} + Z_{\alpha/2} \sigma_{\bar{p}}$$

Example

Suppose we want to estimate the proportion of families, which have two or more children in Agbowo area of Ibadan. If a random sample of 144 families shows that 48 families have two or more children, set up a 95 per cent confidence interval estimate of the population proportion of families having two or more children.

Solution

The sample proportion is: $\bar{p} = \tfrac{x}{n} = \tfrac{48}{144} = \tfrac{1}{3}$

Using the information, n = 144, $\bar{p}$ = 1/3 and $Z_{\alpha/2}$ = 1.96 at 95 per cent confidence coefficient, we have

$$\overline{p}\pm Z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}=\frac{1}{3}\pm 1.96\sqrt{\frac{\frac{1}{3}\left(\frac{2}{3}\right)}{144}}=0.333\pm 0.077$$

Hence the population proportion of families, who have two or more children is likely to be between 25.6 to 41 per cent, that is, $0.256\leq p\leq 0.410$

Summary

In this study session, you have learnt the following:

1. **Confidence Interval for** p

   Based on the large sample condition

   $$\overline{p}\pm Z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

2. **Confidence Interval for μ**

Sample size	Interval Estimate of
Population Mean μ	
Large	
➢ σ assumed known	$\overline{X}\pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
➢ σ estimated by *s*	$\overline{X}\pm Z_{\alpha/2}\frac{s}{\sqrt{n}}$
Small	
➢ σ assumed known	$\overline{X}\pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
➢ σ estimated by *s*	$\overline{X}\pm t_{\alpha/2}\frac{s}{\sqrt{n}}$

Self-Assessment Questions (SAQs) for Study Session 6

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 6.1 (Testing Learning Outcomes 6.1)

Construct and interpret confidence interval estimates for population mean

SAQ 6.2 (Testing Learning Outcomes 6.2

Construct and interpret confidence interval estimates for population proportion

Study Session 7: Sample Size Determination

Introduction

In the previous study session, we noted that the standard error of sampling distribution of sample mean and proportion are both inversely proportional to the sample size n, which is also related to the width of their confidence intervals. The width or range of the confidence interval can be decreased by increasing the sample size n.

The question then is: 'how do we determine the sample size that is required to attain a desired degree of precision?' Interestingly, one feature of the formula for the margin of error is that it can be used to answer the important question.

Learning Outcomes for Study Session 7

When you have studied this session, you should be able to:

7.1 List factors that are responsible for the size of a sample

7.2 Determine sample size for estimating population mean

7.3 Determine sample size for estimating population proportion.

7.1 Factors Determining the Size of a Sample

The following are the factors for determining the size of a sample

- The size of a sample depends upon the following factors:
- The purpose for which the sample is drawn.

- The heterogeneity of the sampling units in the population. The more heterogeneous, the larger is the size of the sample.
- Resources available for the study in terms of time and money.
- Number of technical persons and/or equipment available.
- Precision of estimate required. For a greater precision, a large sample is usually preferred.

7.2 Sample Size for Estimating Population Mean

In order to estimate the population mean, μ, with a condition that the error in its estimation should not exceed a fixed value, say E, we require that the sample mean $\overline{X}$ should fall within range, $\mu \pm E$ with a specified probability. Thus, the margin of error acceptable (i.e., maximum tolerable difference between unknown population mean μ and the sample estimate at a particular level of confidence) can be written as:

$$\overline{X} - \mu = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

or

$$E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

or

$$\sqrt{n} = \frac{Z_{\alpha/2}\sigma}{E}$$

or

$$n = \frac{\left(Z_{\alpha/2}\right)^2 \sigma^2}{E^2}$$

Note that if the population standard deviation σ is not known, then sample standard deviation *s* can be used to determine the sample size n.

Example

Given a population with a standard deviation of 8.6, what sample size is needed to estimate the mean of population within ± 0.5 with 99 per cent confidence?

Solution

Given σ = 8.6, E = ± 0.5 and $Z_{\alpha/2}$ = 2.576 at 99 per cent confidence level.

$$n = \frac{\left(Z_{\alpha/2}\right)^2 \sigma^2}{E^2} = \frac{(2.576)^2 (8.6)^2}{(0.5)^2} \cong 1964$$

7.3 Sample Size for Estimating Population Proportion

The method for determining a sample size for estimating the population proportion is similar to that used for population mean. We require that the sample proportion $\overline{p}$ should fall within the range $\overline{p} \pm E$, with a specified probability.

$$E = Z_{\alpha/2}\sigma_{\overline{p}}$$

$$or$$

$$E = Z_{\alpha/2}\sqrt{\tfrac{pq}{n}}; q = 1 - p$$

$$or$$

$$E^2 = \frac{(Z_{\alpha/2})^2\, pq}{n}$$

$$or$$

$$n = \frac{\left(Z_{\alpha/2}\right)^2 pq}{E^2}$$

Example

A car manufacturing company received a shipment of petrol filters. These filters are to be sampled to estimate the proportion that is unusable. From past experience, the proportion of unusable filter is estimated to be 10 per cent. How large a random sample should be taken to estimate the true proportion of unusable filters to within 0.07 with 99 per cent confidence?

Solution

Given: E = 0.07, $p = 0.10$, and $Z_{\alpha/2} = 2.576$ at 99 per cent confidence level.

Using the formula $$n = \frac{\left(Z_{\alpha/2}\right)^2 pq}{E^2} = \frac{(2.576)^2(0.10\times0.90)}{(0.07)^2} = 121.88$$

Therefore, a slightly larger sample size of n = 122 filters should be used.

Summary

In this study session, you have learnt the following:

1. The size of a sample depends on a number of factors but, most importantly, the precision of estimate is required.
2. Sample size for estimating population mean:

   i. $$n = \frac{\left(Z_{\alpha/2}\right)^2 \sigma^2}{E^2}$$

3. Sample size for estimating population proportion:

$$n = \frac{\left(Z_{\alpha/2}\right)^2 pq}{E^2}; \quad q = 1 - p$$

Self-Assessment Questions (SAQs) for Study Session 7

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 7.1 (Testing Learning Outcomes 7.1)

List factors that are responsible for the size of a sample

.SAQ 7.2 (Testing Learning Outcomes 7.2)

Determine sample size for estimating population mean

SAQ 7.3 (Testing Learning Outcomes 7.3)

Determine sample size for estimating population proportion.

Study Session 8: Hypothesis Testing

Introduction

There is always some disagreement about the value(s) of a parameter or the relationship between parameters. When parametric values are unknown, we estimate them through sample values. If the sample value is exactly the same as we claim, there is no problem in accepting it.

In this study session you will be introduced to statistical hypothesis and the steps involved and also typed one and two errors among others.

Learning Outcomes for Study Session 8

When you have studied this session, you should be able to:

8.1 Define statistical hypothesis;

8.2 List the basic steps involved in hypothesis testing

8.3 Distinguish between one-tailed and two-tailed tests

8.4 Distinguish between Type I and Type II errors.

8.1 Concepts of a Hypothesis

The concepts of a hypothesis can be summarized below

- A statistical hypothesis is a claim (belief or assumption) about an unknown population parameter value.

- The methodology that enables a decision-maker to draw inferences about population characteristics by analyzing the difference between the value of sample statistic and the corresponding hypothesized parameter value is called *hypothesis testing*.
- The statistic used to test a hypothesis is called a test statistic. The greater the difference between the value of the statistic and hypothesized parameter, the more doubt there is about the correctness of the hypothesis.
- The probability level at which the decision-maker concludes that observed difference between the value of the test statistic and hypothesized parameter value cannot be due to chance is called the level of significance *of the test.*

  For example, a decision-maker may feel that a difference that could occur by chance 5 per cent of the time is not significant even if the hypothesis is correct.

8.2 General Procedure for Hypothesis Testing

To test the validity of the claim or assumption about the population parameter, a sample is drawn from the population and analyzed. The results of the analysis are used to decide whether the claim is true or not. The steps involved in hypothesis testing are summarized below:

Step 1: State the null hypothesis (H_0) and alternative hypothesis (H_1)

The null hypothesis refers to a hypothesized numerical value or range of values of the population parameter. It is usually expressed in the form of an equation, which makes a claim, regarding the specific value of the population parameter.

For example,

H_0: $\mu = \mu_0$

Where μ is a population mean and μ_0 represents hypothesized parameter value.

On the other hand, the alternative hypothesis is the logical opposite of the null hypothesis. It states that the specific population parameter value is not equal to the value stated in the null hypothesis.

For example,

$H_1: \mu \neq \mu_0$

Consequently, $H_1: \mu < \mu_0$ or $H_1: \mu > \mu_0$

Step 2: State the level of significance, α (alpha) for the test

The level of significance, usually denoted by α (alpha), is specified before the samples are drawn, so that the results obtained should not influence the choice of the decision-maker. It is the probability of rejecting the null hypothesis when it is true. This probability is usually very small, say 0.01, 0.05, 0.10, etc.

Step 3: Establish critical or rejection region

Sample space of the experiment, which corresponds to the area under the sampling distribution curve of the test statistic is divided into two mutually exclusive regions. These regions are called the *acceptance region* and the *rejection* or *critical region.* If the value of the test statistic falls into the acceptance region, the null hypothesis is accepted; if it is otherwise, it is rejected. The value of the sample statistic that separates the regions of acceptance and rejection is called *critical value.*

Step 4: Calculate the suitable test statistic

The value of test statistic is calculated from the distribution of sample statistic by using the following formula.

$$Test\ statistic = \frac{Value\ of\ sample\ statistic - Value\ of\ hypothesized\ population\ parameter}{S\tan dard\ error\ of\ the\ sample\ statistic}$$

Step 5: Reach a conclusion

Compare the calculated value of the test statistic with the critical value (also called *standard table value* of test statistic). If the calculated absolute value of a test statistic is more than or equal to its critical (or table) value, then reject the null hypothesis; if it is otherwise, accept it.

8.3 One-Tailed and Two-Tailed Test

If an alternative hypothesis is such that it leads to two-sided alternatives to the null hypothesis, it is said to be a *two-tailed test*. For instance, testing

$$H_0: \mu = 20 \text{ vs } H_1: \mu \neq 20$$

Leads to two-sided test as μ can be greater than 20 or less than 20. In this situation, half of the area of critical region lies on the left tail and half on the right. If α is the area of the critical region, then α/2 is the area on both the tails.

Again, if the alternative hypothesis provides one-sided alternative to H_0, e.g.,

$$H_0: \mu = 20 \text{ vs. } H_1: \mu < 20 \text{ (Left-tailed Test)}$$

or

$$H_0: \mu = 20 \text{ vs. } H_1: \mu > 20 \text{ (Right-tailed Test)}$$

the critical region of size α lies only on one tail.

8.4 Errors in Hypothesis Testing

There is a probability of committing an error in making a decision about a hypothesis. Hence, two types of errors are defined as follows:

- Error of rejecting H_0 when it is true – *Type I error.*
- Error of accepting H_0 when it is false – *Type II error.*

Probability of Type I error is denoted by α and probability of Type II error by β. That is:

$$P(\text{reject } H_0/H_0 \text{ is true}) = \alpha$$

$$P(\text{accept } H_0/H_1 \text{ is true}) = \beta$$

The complement (1 – α) of the probability of Type I error measures the probability level of not rejecting a true null hypothesis. It is also referred to as *confidence level.*

Similarly, the complement (1 – β) of the probability of Type II error measures the probability of rejecting the false null hypothesis. It is also called the *power of a statistical test.*

For a given level of significance (α), an increase in sample size will decrease β value. Consequently, this increases the power of the test to detect that the null hypothesis is false. In other words, the level of Type II error can be reduced either by considering a higher value of the level of significance or by establishing the desired trade-off between these two types of errors.

Summary

In this study session, you have learnt the following:

1. A *statistical hypothesis* is a claim (belief or assumption) about an unknown population parametric value.
2. Steps involved in hypothesis testing are the following:
   - State the null and alternative hypotheses
   - State the level of significance for the test
   - Establish critical or rejection region
   - Calculate the suitable test statistic
   - Reach a conclusion
3. If an alternative hypothesis is such that it leads to two-sided alternatives to the null hypothesis, it is said to be a *two-tailed test,* whereas it is *one-tailed* if the alternative hypothesis provides one-sided alternative.

4. Two types of error can be committed in making a decision about a hypothesis.

- **Type I error –** Error of rejecting a true null hypothesis
- **Type II error-** Error of accepting a false null hypothesis

Self-Assessment Questions (SAQs) for Study Session 8

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 8.1 (Testing Learning Outcomes 8.1)

Define statistical hypothesis

SAQ 8.2 (Testing Learning Outcomes 8.2)

List the basic steps involved in hypothesis testing

SAQ 8.3 (Testing Learning Outcomes 8.3)

Distinguish between one-tailed and two-tailed tests

SAQ 8.4 (Testing Learning Outcomes 8.4)

Distinguish between Type I and Type II errors.

Study Session 9: Hypothesis Testing for Population Parameters with Large Samples

Introduction

Hypothesis testing, which involves large samples $(n \geq 30)$ is based on the assumption that the population from which the sample is drawn has a normal distribution. Consequently, the sampling distribution of the sample statistic is also normal.

Even if the population does not have a normal distribution, the sampling distribution of the sample statistic is assumed to be normal due to the central limit theorem because the sample size is large.

Learning Outcomes for Study Session 9

When you have studied this session, you should be able to

9.1 Test hypothesis on single population mean with a large sample

9.2 Test hypothesis on difference between mean values of two populations with large samples

9.3 Test hypothesis on single population proportion with a large sample

9.1 Hypothesis Testing for Single Population Mean with Large Sample

Let μ_0 be the hypothesized value of the population mean to be tested.

a. Two-tailed test:

Hypotheses

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

Test statistic

- If σ is known, then based on the central limit theorem, the sampling distribution of mean would follow the standard normal distribution for a large sample size. The z-test statistic is given by

  $$z = \frac{\bar{X}-\mu_0}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$$

- If σ is not known, then a sample standard deviation *s* is used to estimate σ.

Decision rule

- Reject H_0 if $z_{calculated} < - z_{\alpha/2}$ or $z_{calculated} > z_{\alpha/2}$
- Otherwise accept H_0

Where $z_{\alpha/2}$ is the table value (also called critical value) of z at a chosen level of significance α.

b. Left-tailed test

Hypotheses

- $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$

Test statistic

- $z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$

Decision rule

- Reject H_0 if $z_{calculated} < - z_\alpha$
- Otherwise accept H_0

c. Right-tailed test

Hypotheses

- $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$

Test statistic

- $z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$

Decision rule

- Reject H_0 if $z_{calculated} > z_\alpha$
- Otherwise accept H_0

Example

The mean life of a sample of 400 fluorescent bulbs produced by a company is found to be 1570 hours with a standard deviation of 150 hours. Test the hypothesis that the mean life time of the bulbs produced by the company is 1600 hours at 1 per cent level of significance.

Solution

Hypotheses

$H_0 : \mu = 1600$ against $H_1 : \mu \neq 1600$ (Two-tailed test)

Test statistic

Given n=400, $\bar{X}$ = 1570 hours, s = 150 hours and α = 1 per cent. Using the test statistic,

$$z = \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{1570-1600}{150/\sqrt{400}} = -\frac{30}{7.5} = -4$$

Decision

Since $z_{calculated}$ = -4 is less than the critical value $z_{\alpha/2}$ = ±2.58, the H_0 is rejected. Hence we conclude that the mean lifetime of bulbs produced by the company may not be 1600 hours.

9.2 Hypothesis Testing for Difference between Mean Values of Two Populations with Large Samples

For two populations, say population- 1 and population- 2, let the means of a common variable be μ_1 and μ_2 respectively. The corresponding population standard deviations, σ_1 and σ_2 respectively, are also known. Let two independent random samples of large size n_1 and n_2 be drawn from the first and second population, respectively. Let the sample means so calculated be $\bar{X}_1$ and $\bar{X}_2$.

Hypotheses

- $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

Test statistic

- The test statistic will follow the normal distribution for a large sample due to the central limit theorem. The z-test statistic is

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where

$\sigma_{\bar{X}_1 - \bar{X}_2}$ = standard error of the statistic ($\bar{X}_1 - \bar{X}_2$)

$\bar{X}_1 - \bar{X}_2$ = difference between two sample means, that is, sample statistic

$\mu_1 - \mu_2$ = difference between population means, that is, hypothesized population parameter.

If σ_1 and σ_2 are not known, then we may estimate them by using the sample standard deviation s_1 and s_2 respectively and then substitute in the formula for standard error of $\bar{X}_1 - \bar{X}_2$.

Decision Rule

- Reject H_0 if $z_{calculated} > z_{\alpha/2}$ or $z_{calculated} < - z_{\alpha/2}$
- Otherwise accept H_0

Example

A firm believes that the tyres produced by process A on an average last longer than tyres produced by process B. To test this belief, random samples of tyres produced by the two processes were tested and the results are:

Process	*Sample size*	*Average Lifetime (in Km)*	*Standard Deviation (in km)*
A	50	22,400	1000
B	50	21,800	1000

Is there evidence at a 5 per cent level of significance that the firm is correct in its belief?

Solution

Hypotheses

$H_0 : \mu_A = \mu_B$ against $H_1 : \mu_A \neq \mu_B$

Test statistic

Given $\bar{X}_A = 22{,}400$ km, $\bar{X}_B = 21{,}800$ km, $\sigma_A = \sigma_B = 1000$, and $n_A = n_B = 50$

Using the z-test statistic

$$z = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \frac{(22{,}400 - 21{,}800) - (0)}{\sqrt{\frac{(1000)^2}{50} + \frac{(1000)^2}{50}}} = 3$$

Decision

Since $z_{\text{calculated}} = 3$ is more than the critical value $z_{\alpha/2} = \pm 1.645$ at 5 per cent level of significance, therefore H_0 is rejected.

9.3 Hypothesis Testing for Population Proportion with Large Sample

Consider a population where p represents the proportion of individuals or items possessing a certain characteristic. A random sample of size n is selected to compute the sample proportion $\bar{p}$. The three forms of null hypothesis and alternative hypothesis pertaining to the hypothesized population proportion p_0 are as follows:

Hypotheses

(i) H$_0$: $p = p_0$ against H$_1$: $p \neq p_0$ (Two-tailed test)

(ii) H$_0$: $p = p_0$ against H$_1$: $p < p_0$ (Left-tailed test)

(iii) H$_0$: $p = p_0$ against H$_1$: $p > p_0$ (Right-tailed test)

Test statistic

To conduct a test of hypothesis, it is assumed that the sampling distribution of a proportion follows a standardized normal distribution.

We compute a value for the z-test statistic as follows:

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} = \frac{\bar{p} - p_0}{\sqrt{P_0(1 - p_0)/n}}$$

Decision Rule

The rules of rejecting or accepting a null hypothesis, pertaining to the population mean for both two-tailed and one-tailed tests are also applicable for hypothesis testing of population proportion. (see section A above)

9.4 Hypothesis Testing for Difference between Proportions of Two Populations with Large Samples

For two populations, say population- 1 and population- 2, let p_1 and p_2 be the proportions of individuals or items possessing a particular attribute in each population respectively. The corresponding population standard deviations, σ_{p_1} and σ_{p_2} respectively, are unknown.

Let two independent random samples of large size n$_1$ and n$_2$ be drawn from the first and second populations, respectively. Let the sample proportions so calculated be $\bar{p}_1$ and $\bar{p}_2$.

Hypotheses

H$_0$: $p_1 = p_2$ against H$_1$: $p_1 \neq p_2$

Test statistic

The z-test statistic is stated as:

$$z = \frac{(\overline{p}_1 - \overline{p}_2) - (p_1 - p_2)}{\sigma_{\overline{p}_1 - \overline{p}_2}}$$

Invariably, the standard error $\sigma_{\overline{p}_1 - \overline{p}_2}$ of difference between sample proportions is not known. Thus when a null hypothesis states that there is no difference between the population proportions, we combine the two sample proportions $\overline{p}_1$ and $\overline{p}_2$ to get one unbiased estimate of population proportion as follows:

$$\text{Pooled estimate } \overline{p} = \frac{n_1\overline{p}_1 + n_2\overline{p}_2}{n_1 + n_2}$$

The z-test statistic is then restated as:

$$z = \frac{(\overline{p}_1 - \overline{p}_2) - (p_1 - p_2)}{s_{\overline{p}_1 - \overline{p}_2}}$$

Where $s_{\overline{p}_1 - \overline{p}_2} = \sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

Example

In a simple random sample of 600 students taken from a university, 400 were found to be smokers. In another simple random sample of 900 students from another university, 450 were found to be smokers. Do the data indicate that there is a significant difference in the habit of smoking in the two universities?

Solution

Let p_1 and p_2 be the proportion of student smokers in the two universities

Hypotheses

$$H_0 : p_1 = p_2 \text{ against } H_1 : p_1 \neq p_2$$

Test statistic

Given, n1 = 600, $\overline{p}_1$ = 400/600=0.667; n2 = 900, $\overline{p}_2$ = 450/900 = 0.50 and level of significance α = 5 per cent.

$$\overline{p} = \frac{n_1\overline{p}_1 + n_2\overline{p}_2}{n_1 + n_2} = \frac{600(400/600)+900(450/900)}{600+900} = 0.567$$

$$s_{\overline{p}_1-\overline{p}_2} = \sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{0.567\times0.433(\frac{1}{600}+\frac{1}{900})} = 0.026$$

Substituting the values in z-test statistic, we have

$$z = \frac{(\overline{p}_1-\overline{p}_2)-(p_1-p_2)}{s_{\overline{p}_1-\overline{p}_2}} = \frac{(0.667-0.500)-(0)}{0.026} = 6.423$$

Decision

Since $z_{\text{calculated}}$ = 6.423 is greater than the critical value $z_{\alpha/2}$ = ±2.58, H_0 is rejected. Thus, we conclude that there is a significant difference in the habit of smoking in the two universities.

Summary

In this study session, you have learnt the following:

1. Hypothesis testing for a population mean with a large sample $(n \geq 30)$

   Test statistic for a population mean μ

   - σ is assumed known $z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
   - σ is estimated by *s* $z = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$

2. Hypothesis testing for difference between mean values of two populations with large samples $(n \geq 30)$

   Test statistic for $\mu_1 = \mu_2$

   - σ_1 and σ_2 assumed known $z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$
   - σ_1 and σ_2 estimated by s_1 and s_2 respectively $z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$

3. Hypothesis testing for population proportion with a large sample $\left(n \geq 30\right)$

Test statistic for population proportion p

- $z = \frac{\bar{p} - p}{\sqrt{P(1-p)/n}}$

4. Hypothesis testing for difference between proportions of two populations with large samples $\left(n \geq 30\right)$

Test statistic for $p_1 = p_2$

- $z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{s_{\bar{p}_1 - \bar{p}_2}}$

where $s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ and

$\bar{p} = \dfrac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2}$ is the pooled estimator of the population proportion.

Self-Assessment Questions (SAQs) for Study Session 9

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 9.1 (Testing Learning Outcomes 9.1)

Test hypothesis on single population mean with a large sample

SAQ 9.2 (Testing Learning Outcomes 9.2)

Test hypothesis on difference between mean values of two populations with large

SAQ 9.3 (Testing Learning Outcomes 9.3)

Test hypothesis on single population proportion with a large sample

Study Session 10: Hypothesis Testing for Population Parameters with Small Samples

Introduction

When a sample size is small (i.e., less than 30), the central limit theorem does not guarantee us to assume that the sampling distribution of a sample statistic is normal. Consequently, when testing a hypothesis with small samples, we must assume that the samples come from a normally or approximately normally distributed population.

Under these conditions, the sampling distribution of the sample statistic is normal but the critical values depend on whether or not the population standard deviation σ is known. When σ is not known, its value is estimated by computing the sample standard deviation *s*.

Learning Outcomes for Study Session 10

When you have studied this session, you should be able to:

10.1 List the properties and uses of *student's t-distribution*;

10.2 Test hypothesis on single population mean with a small sample; and

10.3 Test hypothesis on difference of two population mean values with a small sample.

10.1 Properties of t-Distribution

1. The shape of this distribution is very much like that of the normal curve; it is symmetrical with mean zero.
2. For large values of degrees of freedom (i.e., as sample size increases), the *t*-distribution tends to a standard normal distribution. This implies that for different degrees of freedom, the shape of the *t*-distribution also differs.
3. The *t*-distribution is less peaked than normal distribution at the centre and higher in the tails.
4. The *t*-distribution has greater dispersion than standard normal distribution. The variance of *t*-distribution is defined only when degree of freedom is greater or equal to 3.
5. The *t*-curve attains its maximum at $t = 0$ so that the mode coincides with the mean. It can be shown analytically that the *t*-curve approaches the standard normal curve as the degree of freedom gets larger.

10.1.1 Use of t-Distribution

There are various uses of *t*-distribution. A few of them are as follows:

1. Testing hypothesis about the population mean.
2. Testing hypothesis about the difference between two populations' means with independent samples.
3. Testing hypothesis about the difference between two populations' means with dependent samples.
4. Testing hypothesis about an observed coefficient of correlation including partial and rank correlations
5. Testing hypothesis about an observed regression coefficient.

10.2 Hypothesis Testing for Single Population Mean with Small Sample

Let μ_0 be the hypothesized value of the population mean to be tested.

(i) Two-tailed test:

Hypotheses

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

Test statistic

$$t = \frac{\bar{X} - \mu_0}{s_{\bar{X}}} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}; \quad s = \sqrt{\frac{\sum (x - \bar{x})}{n-1}}^{2}$$

Decision rule

1. Reject H_0 if $t_{\text{calculated}} < - t_{\alpha/2}$ or $t_{\text{calculated}} > t_{\alpha/2}$

2. Otherwise accept H_0

Where $t_{\alpha/2}$ is the table value (also called critical value) of t-distribution at a chosen level of significance α and degrees of freedom n - 1.

(ii) Left-tailed test

Hypotheses

$H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$

Test statistic

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Decision rule

1. Reject H_0 if $t_{\text{calculated}} < - t_{\alpha}$

2. Otherwise accept H_0

(iii) Right-tailed test

Hypotheses

$H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$

Test statistic

$$t=\frac{\bar{X}-\mu_0}{s/\sqrt{n}}$$

Decision rule

1. Reject H_0 if $t_{\text{calculated}} > t_\alpha$
2. Otherwise accept H_0

Example

A random sample of 10 electric bulbs from a large consignment gave the following data:

Item :	1	2	3	4	5	6	7	8	9	10
Life in '000 hours	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6

Can we accept the hypothesis that the average life time of the bulbs is 4,000 hours.

Solution

Hypotheses

$H_0 : \mu = 4{,}000$ against $H_1 : \mu \neq 4{,}000$

Calculating $s=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=\sqrt{\frac{3.12}{9}}=0.589$ and $\bar{X}=\frac{\sum x}{n}=\frac{44}{10}=4.4$

Test statistic

$$t=\frac{\bar{X}-\mu_0}{s/\sqrt{n}}=\frac{4.4-4}{0.589/\sqrt{10}}=2.148$$

Decision

Accept H_0 since $t_{\text{calculated}} = 2.148 <$ critical value $t_{\alpha/2} = 2.62$

at $\alpha/2 = 0.025$ and degree freedom = $n - 1 = 9$.

10.3 Hypothesis Testing for Difference of Two Population Means values with Independent Samples

For two normally distributed populations, say population- 1 and population- 2, let the mean values of a common variable be μ_1 and μ_2 respectively. The corresponding population standard deviations are σ_1 and σ_2 respectively. Let two independent random samples of small size n_1 and n_2 be drawn from the first and second populations respectively. Let the sample means so calculated be $\bar{X}_1$ and $\bar{X}_2$.

Hypotheses

$H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

Test statistic

Case 1: Population Standard Deviations are Unknown but Equal (i.e., $\sigma_1 = \sigma_2 = \sigma$)

$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $s_p = \hat{\sigma} = \sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}}$ is called the *pooled sample standard deviation.* The sampling of this *t*-statistic is approximated by the *t*-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Case 2: Population Standard Deviations are Unknown and Unequal

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The sampling of this *t*-statistic is approximated by the *t*-distribution with degrees of freedom given by

$$\text{Degrees of freedom } (df) = \frac{\left[s_1^2 / n_1 + s_2^2 / n_2 \right]^2}{\frac{(s_1^2 / n_1)^2}{n_1 - 1} + \frac{(s_2^2 / n_2)^2}{n_2 - 1}}$$

The number of degrees of freedom in this case is less than that obtained in case 1 above.

Decision Rule

1. Reject H_0 if $t_{\text{calculated}} > t_{\alpha/2}$ or $t_{\text{calculated}} < - t_{\alpha/2}$

2. Otherwise accept H$_0$

Example

The mean life of a sample of 10 electric light bulbs was found to be 1456 hours with standard deviation of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 hours with standard deviation of 398 hours. Is there a significant difference between the means of the two batches?

Solution

Hypotheses

H$_0$: $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$ against H$_1$: $\mu_1 \neq \mu_2$

Given $n_1 = 10$, $\bar{X}_1 = 1456$, $s_1 = 423$; $n_2 = 17$, $\bar{X}_2 = 1280$, $s_2 = 398$ and $\alpha = 0.05$. Thus

Pooled standard deviation, $s_p = \hat{\sigma} = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

$$= \sqrt{\frac{9(423)^2 + 16(398)^2}{10 + 17 - 2}} = 407.18$$

Test Statistic

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(1456 - 1280) - (0)}{407.18 \times \sqrt{\dfrac{1}{10} + \dfrac{1}{17}}} = 1.085$$

Decision

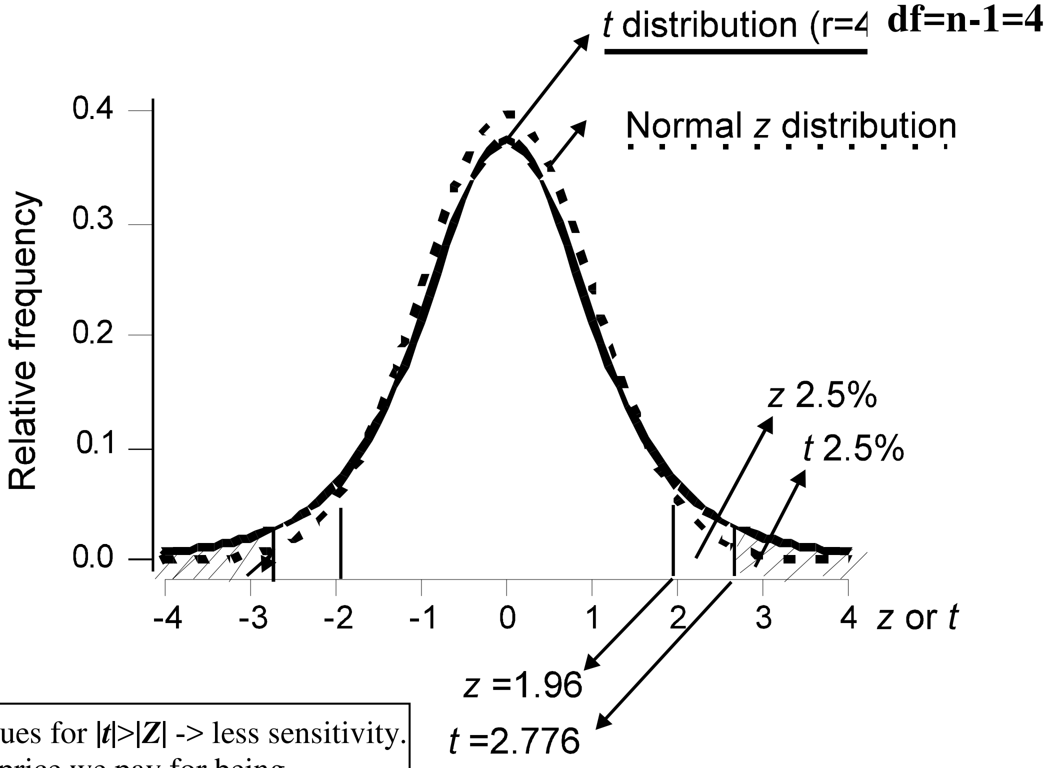
Since the calculated value 1.085 is less than its critical value $t_{\alpha/2}$ = 2.06 at *df* =25, the null hypothesis is accepted. Thus, we conclude that the mean life of electric bulbs of two batches does not differ significantly.

10.4 Distributions, Hypothesis Testing, and Sample Size Determination

Consider a repeated drawing of samples of size n = 5 from a normal distribution. For each sample compute $\overline{Y}$, s, $s_{\overline{Y}}$, and another statistic, *t*:

$$t_{(n-1)} = (\overline{Y} - \mu)/\, s_{\overline{Y}} \quad \text{(Remember } Z = (\overline{Y} - \mu)/\,\sigma_{\overline{Y}}\text{)}$$

The *t* statistics is the number of standard error that separate $\overline{Y}$ and its hypothesized mean μ.



Critical values for |*t*|>|*Z*| -> less sensitivity. This is the price we pay for being uncertain about the population variance

Fig. 1. Distribution of t (df=4) compared to Z. The t distribution is symmetric, but wider and flatter than the Z distribution, lying under it at the center and above it in the tails.

Confidence Limits Based on Sample Statistics

The general formula for any parameter δ is:

Estimated δ ± **Critical value** * Standard error of the estimated δ

So, for a population mean estimated via a sample mean:

$$\mu = \overline{Y} \pm t_{\frac{\alpha}{2},n-1} * s_{\overline{Y}}$$

The statistic $\overline{Y}$ is distributed about μ according to the *t* distribution so it satisfies

$$P\{\overline{Y} - t_{\alpha/2,\, n-1}\; s_{\overline{Y}} \le \mu \le \overline{Y} + t_{\alpha/2,\, n-1}\; s_{\overline{Y}}\} = 1 - \alpha$$

For a confidence interval of size 1-α, use a *t* value corresponding to α /2.

Therefore the confidence interval is

$$\overline{Y} - t_{\alpha/2,\, n-1}\; s_{\overline{Y}} \le \mu \le \overline{Y} + t_{\alpha/2,\, n-1}\; s_{\overline{Y}}$$

These two terms represent the lower and upper 1- α **confidence limits** of the mean. The interval between these terms is called **confidence interval (CI).**

Example: Data Set 1 of Hordeum 14 malt extraction values

$\overline{Y} = 75.94$ $s_{\overline{Y}}$ **=** $1.23 / \sqrt{14} = 0.3279$. A table gives the $\boldsymbol{t}_{0.025,13}$ value of 2.16

95% CI for μ = 75.94 ± 2.160 * 0.3279 ⇒ [75.23- 76.65]

If we repeatedly obtained samples of size 14 from the population and constructed these limits for each, we expect 95% of the intervals to contain the true mean.

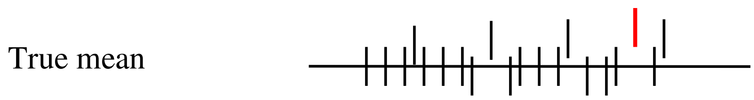


Fig. 2 Twenty 95% confidence intervals. One out of 20 intervals does not include the true mean.

Hypothesis Testing and Power of the Test

Example Barley data. $\bar{Y} = 75.94$, $s_{\bar{Y}} = \sqrt{s^2 / n} = 0.3279$, $t_{0.025,13} = 2.160$, CI:

1) Choose a null hypothesis: Test H_o $\mu = 78$ against the H_1 $\mu \neq 78$.

2) Choose a significance level: Assign $\alpha = 0.05$

3) Calculate the test statistic ***t***:

$$t = \frac{\bar{Y} - \mu}{s_{\bar{Y}}} = \frac{75.94 - 78.00}{0.3279} = -6.28$$

(interpretation: the sample mean is 6.3 SE from the hypothetical mean of 78. Too far!).

4) Compare the absolute value of the test statistic to the critical statistic:

$$|\,-6.28\,| > 2.16$$

5) Since the absolute value of the test statistic is larger, we reject H_0.

This is equivalent to calculate a 95% confidence interval for the mean. Since μ_o (78) is not within the CI [75.23- 76.65] we reject $H_{o.}$

α is called the *significance level* of the test (<0.05): probability of incorrectly rejecting a true H_o, a **Type I** error.

β is the **Type II** error: to incorrectly accept H_o when it is false

		Null hypothesis	
		Accepted	**Rejected**
Null hypothesis	**True**	Correct decision	**Type I error= α**
	False	**Type II error=β**	Correct decision= Power= 1-β

<u>Power of the test</u>: 1-β is the **power** of the test, and represents the probability of correctly rejecting a false null hypothesis. It is a measure of the ability of the test to detect an alternative mean or a significant difference when it is real.

Note that for a given $\overline{Y}$ and s, if 2 of the 3 quantities α, β, and n are specified then the third one can be determined.

Choose the right number of replications to keep **Type I error** α and **Type II error** β under the desired limits (e.g. $\alpha <0.05$ & $\beta<0.20$).

Summary

In this study session, you have learnt the following:

1. Hypothesis testing for single population mean with small sample

   test statistic

   $$t = \frac{\bar{X}-\mu_0}{s_{\bar{X}}} = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}; \quad s = \sqrt{\frac{\sum(x-\bar{x})}{n-1}}$$

2. Hypothesis testing for difference of two population means values with independent samples

Case 1: Population standard deviations are unknown but equal (i.e., $\sigma_1 = \sigma_2 = \sigma$)

$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $s_p = \hat{\sigma} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$ is called the *pooled sample standard deviation.* The sampling of this *t*-statistic is approximated by the *t*-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Case 2: Population standard deviations are unknown and unequal

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The sampling of this *t*-statistic is approximated by the *t*-distribution with degrees of freedom given by

$$\text{Degrees of freedom } (df) = \frac{\left[s_1^2 / n_1 + s_2^2 / n_2\right]^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

Self-Assessment Questions (SAQs) for Study Session 10

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 10.1 (Testing Learning Outcomes 10.1)

List the properties and uses of *student's t-distribution*;

SAQ 10.2 (Testing Learning Outcomes 10.2

Test hypothesis on single population mean with a small sample.

SAQ 10.3 (Testing Learning Outcomes 10.3)

Test hypothesis on difference of two population mean values with a small sample.

Study Session 11: The Chi-Square Distribution

Introduction

All of the inferential statistics we covered in the previous study session are called **parametric statistics.** To use these statistics, you will make some assumptions about the distributions they come from, such as they are normally distributed. With parametric statistics, you will also deal with data for the variable that is at the interval or ratio level of measurement, i.e. test scores, physical measurements, etc.

You shall now consider a widely used *non-parametric* test, **chi-square**, which we can use with data at the nominal level, that is, data that is classificatory. The symbol χ is the Greek letter 'chi'. The sampling distribution of χ^2 is called χ^2-*distribution*.

Learning Outcomes for Study Session 11

When you have studied this session, you should be able to:

11.1 List the properties of the Chi-Square distribution;

11.2 List the conditions for the applications of Chi-Square test;

11.3 List the types of Chi-Square;

11.4 Compute Chi-Square probabilities; and

11.5 List some applications of Chi-Square test.

11.1 The Chi-Square Distribution

The Chi-Square (χ^2) distribution is obtained from the values of the ratio of the sample variance and population variance multiplied by the degrees of freedom, n. This occurs when the population is normally distributed with population variance, σ^2.

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{df.\times s^2}{\sigma^2}$$

11.1.1 Properties of the Chi-Square Distribution

Chi-Square is non-negative. It is the ratio of two non-negative values; therefore it must be non-negative itself. Some of the properties are as follows

- Chi-Square is non-symmetric.
- There are many different Chi-Square distributions, one for each degree of freedom.
- The degree of freedom when working with a single population variance is n – 1.

1. For degrees of freedom greater or equal to 30, the χ^2 curve approximate to the normal curve.
2. The mean of the Chi-Square distribution is n and the variance is 2n.

11.2 Types of Chi-Square

The following are types of chi square:

Pearson's chi-square: It is by far the most common type of chi-square significance test. If simply "chi-square" is mentioned, it is probably Pearson's chi-square. This statistic is used to test the hypothesis of no association of columns and rows in tabular data. It can be used even with nominal data.

Yates' correction: It is an arbitrary, conservative adjustment to chi-square when applied to tables with one or more cells with frequencies less than five. It is only applied to 2 by 2 tables. Some authors also apply it to all 2 by 2 tables since the correction gives a better approximation to the binomial distribution.

Yates' correction is conservative in the sense of making it more difficult to establish significance.

Chi-square goodness-of-fit test: The goodness-of-fit test is simply a different use of Pearsonian chi-square. It is used to test whether or not an observed distribution conforms to any other distribution, such as one based on theory (ex., if the observed distribution is not significantly different from a normal distribution) or one based on some other known distribution (ex., if the observed distribution is not significantly different from a known national distribution based on census data).

Likelihood ratio chi-square: It is an alternative procedure to test the hypothesis of no association of columns and rows in nominal-level tabular data. It is based on maximum likelihood estimation. Though computed differently, likelihood ratio chi-square is interpreted the same way. For large samples,

Likelihood ratio Chi-square will be close in results to Pearson Chi-Square. Even for smaller samples, it rarely leads to different substantive results.

Mantel-Haenszel chi-square: It is also called the *Mantel-Haenszel test for linear association* or *linear by linear association chi-square*. Unlike ordinary and likelihood ratio chi-square, Mantel Haenzel Chi-Square is an ordinal measure of significance. It is preferred when testing the significance of linear relationship between two ordinal variables, since it is more powerful than Pearson Chi-Square (more likely to establish linear association).Mantel- **D.**

11.3 Conditions for the Applications of χ^2 Test

As mentioned before, chi square is a nonparametric test. It does not require the sample data to be more or less normally distributed (as parametric tests like t-tests do); although, it relies on the assumption that the variable is normally distributed in the population from which the sample is drawn. But chi square, in spite of the above, does have some requirements:

1. The sample must be randomly drawn from the population.
2. Data must be reported in raw frequencies (**not percentages or ratios**);
3. Measured variables must be independent;

4. Values/categories on independent and dependent variables must be mutually exclusive and exhaustive;
5. Observed frequencies cannot be too small. It should contain at least 5 observations.

11.3.1 The Chi-Square Test Statistic

The chi-square statistic is used to compare the **observed frequency** of some observation with an **expected frequency**. The comparison of observed and expected frequencies is used to calculate the value of the chi-square statistic, which in turn can be compared with the distribution of chi-square to make an inference about a statistical problem.

The symbol for chi-square and the formula are as follows:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where

O is the observed frequency, and

E is the expected frequency.

The degrees of freedom for the one-dimensional chi-square statistic is *df* **= n - 1**

where **n** is the number of categories or levels of the independent variable.

11.4 Chi-Square Probabilities

Since the Chi-Square distribution is not symmetric, the method for looking up left-tail values is different from the method for looking up right tail values.

- Area to the right – just use the area given.
- Area to the left – the table requires the area to the right; therefore subtract the given area from one and look this area up in the table.

- Area in both tail – divide the area by two. Look up this area for the right critical value and one minus this area for the left critical value.

11.5 Applications of Chi-Square Test

A few important applications of χ^2 test are as follows:

1. test of independence of attributes
2. test of goodness-of-fit
3. Yate's correction continuity
4. test for population variance
5. test for homogeneity

Summary

In this study session, you have learnt the following:

1. The Chi-Square distribution for a normally distributed population with population variance, σ^2 is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{df.\times s^2}{\sigma^2}$$

2. Types of Chi-Square:
   a. Pearson's chi-square
   b. Yates' correction
   c. Chi-square goodness-of-fit test
   d. Likelihood ratio chi-square
   e. Mantel-Haenszel chi-square
3. To compare the observed frequency (**O**) of some observations with an expected frequency (**E**), the Chi-Square statistic used is

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Self-Assessment Questions (SAQs) for Study Session 11

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 11.1 (Testing Learning Outcomes 11.1)

List the properties of the Chi-Square distribution;

SAQ 11.2 (Testing Learning Outcomes 11.2

List the conditions for the applications of Chi-Square test

SAQ 11.3 (Testing Learning Outcomes 11.3

List the types of Chi-Square;

SAQ 11.4 (Testing Learning Outcomes 11.4

Compute Chi-Square probabilities; and

SAQ 11.5 (Testing Learning Outcomes 11.5

List some applications of Chi-Square test.

Study session 12: Applications of Chi-Square Distribution

Introduction

The χ^2 statistic plays an important role in tests dealing with *count data*, or *enumeration data.* This is the data required in dealing with problems where information is obtained by counting rather than measuring. In this study session, you will learn all the five important applications of χ^2-test given in the last study session

Learning Outcomes for Study Session 12

When you have studied this session, you should be able to:

12.1 Test for independence of attributes, that is, analyze contingency tables;

12.2 Test for goodness-of-fit;

12.3 Apply Yate's correction for continuity;

12.4 Test for population variance and Test for homogeneity.

12.1 Contingency Table

A (r x c) contingency table shows the observed frequencies for two categorical variables arranged in a r rows and c columns. The sum of all observed frequencies is n, the sample size. The contingency table format is displayed as follows:

Contingency Table

Variable B	*Variable* A				Total
	A_1	A_2	…	A_c	
B_1	O_{11}	O_{12}	…	O_{1c}	R_1
B_2	O_{21}	O_{22}	…	O_{2c}	R_2
.	.	.	…	.	.
.	.	.	…	.	.
.	.	.	…	.	.
B_r	O_{r1}	O_{r2}	…	O_{rc}	R_r
Total	C_1	C_2	…	C_c	N

Note that:

1. The variables A and B have been classified into mutually exclusive categories.
2. The values O_{ij} in row *i* and column *j* of the table show the frequency of observations fallen in each joint categories *i* and *j*.
3. The row and column totals are the sums of the frequencies.
4. The row and column totals add up to a grand total *n*, which represents the sample size.
5. The expected frequency, E_{ij}, corresponding to the observed frequency in row *i* and *j* in each cell of the contingency table, is calculated as follows:

$$E_{ij} = \frac{\text{Row i total}}{\text{Sample size}} \times \frac{\text{Column j total}}{\text{Sample size}} \times \text{Grand total}$$

$$= \frac{R_i}{n} \times \frac{C_j}{n} \times n = \frac{R_i \times C_j}{n}$$

12.2 Test of Independence of Attributes

The test of independence uses the contingency table format and is also referred to as a *Contingency Table Test.* The analysis of a two-way contingency table helps to answer the question whether two variables are related or independent of each other. Thus, the χ^2-test statistic measures how much the observed frequencies differ from the expected frequencies when variables are independent.

Test Procedure

Step 1: State the null and alternative hypothesis

H_0: No relationship or association exists between two variables, that is, they are independent

H_1: A relationship exists, that is, they are dependent

Step 2: Select a random sample and record the observed frequencies (E values) in each cell of the contingency table and calculate the row, column, and grand totals.

Step 3: Calculate the expected frequencies (E-values) for each cell

Step 4: Compute the value of test-statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Step 5: Calculate the degrees of freedom (*df*).

$df = (r - 1)(c - 1)$, where r = number of rows and c = number of columns.

Step 6: Using a level of significance α and *df*, find the critical (table) value of χ^2_α. This value of χ^2_α corresponds to an area in the right tail of the distribution.

Step 7: Compare the calculated and table values of χ^2. Decide whether the variables are independent or not, using the decision rule:

- Accept H$_0$ if χ^2_{cal} is less than its table value $\chi^2_{\alpha,(r-1)(c-1)}$
- Otherwise reject H$_0$.

Example

Two hundred randomly selected adults were asked whether TV shows as a whole are primarily entertaining, educational, or waste of time (only one answer could be chosen). The respondents were categorized by gender. Their responses are given in the following table:

Gender	*Opinion*			**Total**
	Entertaining	*Educational*	*Waste of time*	
Female	52	28	30	**110**
Male	50	12	50	**90**
Total	**80**	**40**	**80**	**200**

Is there any convincing evidence that there is a relationship between gender and opinion in the population of interest?

Solution

1. Hypotheses:

   H$_0$: Opinion of adults is independent of gender (or no relationship between the opinion of adults and gender)

   H$_1$: Opinion of adults is dependent on gender (or there is a relationship between the opinion of adults and gender)

2. Calculation of expected frequencies, E$_{ij}$

   E$_{11}$ = 44 E$_{12}$ = 22 E$_{13}$ = 44

   E$_{21}$ = 36 E$_{22}$ = 18 E$_{33}$ = 36

3. Compute the Chi-Square test statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

= 11.129

1. Compute the degrees of freedom, i.e., *df* = (2 – 1) (3 – 1) = 2.

2. Check the critical or table value, i.e., $\chi^2_{0.05,2}$ = 5.99

3. Decision: Since the calculated value is greater than the critical value, the null hypothesis is rejected. Hence, we conclude that the opinion of adults is dependent on gender.

12.3 Goodness-of-fit Test

On several occasions, a decision-maker needs to understand whether an actual sample distribution matches or coincides with a known theoretical probability distribution such as Binomial, Poisson, Normal, and so on. The χ^2 test for goodness-of-fit enables us to determine the extent to which theoretical probability distributions coincide with empirical sample distribution.

To apply the test, a particular theoretical distribution is first hypothesized for a given population and then the test is carried out to determine whether or not the sample data could have come from the population of interest with hypothesized theoretical distribution. The observed frequencies or values come from the sample and the expected frequencies or values come from the theoretical hypothesized probability distribution.

The goodness-of-fit now focuses on the differences between the observed values and the expected values. Large differences between the two distributions throw doubt on the assumption that the hypothesized theoretical distribution is correct. On the other hand, small differences between the two distributions may be assumed to be resulting from sampling error.

Example

A book has 700 pages. The number of pages with various numbers of misprints is recorded below:

No. of misprints	:	0	1	2	3	4	5
No. of pages with misprints	:	616	70	10	2	1	1

Can a Poisson distribution be fitted to this data?

Solution

1. Hypotheses:

   H_0: Poisson distribution fits the data.

   H_1: Poisson distribution does not fit the data.

2. Calculation of expected frequencies (E):

   The number of pages in the book is 700, whereas the maximum possible number of misprints is only 5.Thus we apply Poisson probability distribution to calculate the expected number of misprints in each page of the book as follows:

Misprints (x)	Number of pages (O)		(O)x	$P(X=x)=\frac{\lambda^x e^{-\lambda}}{x!}$	E = n P(X=x)			
0	616		0	0.8607	602.5		1	70
70	0.1291		90.38					
2	10		20	0.0097	6.78			
3	2		6	0.00048	0.34			
4	1	=4	4	0.000019	0.013	=0.353		
5	1		5	0.000000	0			
	n = 700		105					

$$\lambda = \frac{\sum fx}{n} = \frac{105}{700} = 0.15$$

1. Compute the value of the chi-square test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \quad 44.159$$

Note that observed frequencies less than 5 have been grouped together to represent a single category before calculating the test statistics.

1. Compute the degrees of freedom, *df* = n – 1 = 3 – 1 = 2.
2. Check the critical or table value, i.e., $\chi^2_{0.05,2}$ = 5.99
3. Decision: Since the calculated value is greater than the table value, the null hypothesis is rejected.

12.4 Yate's Correction for Continuity

The distribution of χ^2 –test statistic is continuous but the data under test is categorical, which is discrete. It obviously causes errors, but it is not serious unless we have one degree of freedom, as in a 2 x 2 contingency table. To remove the probability of such errors occurring due to the effect of discrete data, we apply Yate's correction for continuity. The correction factor suggested by Yate in case of a 2 x 2 contingency table is as follows:

a. Decrease by half those cell frequencies, which are greater than expected frequencies and increase by half those which are less than expected. This correction does not affect the row and column totals. For example, in a 2 x 2 contingency table where frequencies are arranged as follows:

Attributes	A	Not A	*Total*
B	*a*	*b*	*a* + *b*
Not B	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d*

The value of χ^2 calculated from independent frequencies is given by

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)}{(a+b)(c+d)(a+c)(b+d)}$$

This value of χ^2 is corrected as:

$$\chi^2_{corrected} = \frac{n(ad-bc-\frac{1}{2}n)^2}{(a+b)(c+d)(a+c)(b+d)}; \quad ad-bc>0$$

and

$$\chi^2_{corrected} = \frac{n(bc-ad-\frac{1}{2}n)^2}{(a+b)(c+d)(a+c)(b+d)}; \quad ad-bc<0$$

b. An alternative formula for calculating the χ^2 test statistic is as follows:

$$\chi^2 = \sum \frac{\{|O-E|-0.5\}^2}{E}$$

12.5 Test for Population Variance

The assumption underlying the χ^2-test is that the population from which the samples are drawn is normally distributed. Let the variance of population be σ^2. The null hypothesis is set up as:

H_0: $\sigma^2 = \sigma_0^2$, where σ_0^2 is hypothesized value of σ^2.

If a sample of size n is drawn from this normal population, then variance of sampling distribution of mean $\bar{X}$ is given by $s^2 = \dfrac{\sum\left(X-\bar{X}\right)^2}{n-1}$. Consequently, the value of χ^2-test statistic is determined as

$$\chi^2 = \frac{\sum\left(X-\bar{X}\right)^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$$

with *df* = n – 1 degrees of freedom.

Decision rule

- Accept H0 if χ^2_{cal} is less than its table value $\chi^2_{\alpha/2}$
- Otherwise reject H0.

12.5.1 Confidence Interval for Variance

We can define 95 per cent, 99 per cent, and so on, confidence limits and intervals for χ^2 test statistic, using table values of χ^2. Such limits of confidence helps to estimate population standard deviation in form of sample standard deviation *s* with $(1 - \alpha)$ per cent confidence as follows:

$$\frac{(n-1)s^2}{\chi^2_{df,U}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{df,L}}$$

or

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{df,U}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{df,L}}}$$

The subscripts U and L stands for upper and lower tails proportion of area under χ^2 –curve. For example, for a 95 per cent confidence interval the lowest 2.5 per cent and the highest 25 per cent of χ^2 distribution curve is excluded, leaving the middle 95 per cent area.

The χ^2 for upper 2.5 per cent (=0.025) area is obtained directly from standard χ^2 table. To obtain value of χ^2 for lower 2.5 per cent (=0.025) area, look under the 0.975 column of χ^2 table for given *df*, because 1 – 0.975 = 0.025.

Example

A random sample of size 25 from a population gives the sample standard deviation of 8.5. Test the hypothesis that the population standard deviation is 10.

Solution

Hypotheses:

H_0: σ = 10 vs H_1: σ ≠ 10

Given that n = 25; *s* = 8.5.

Test statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(8.5)^2}{(10)^2} = 17.34$$

Critical Value:

$$\chi^2_{\alpha,df} = \chi^2_{0.05,24} = 36.415$$

Decision:

Since the calculated χ^2 is less than its critical value, the null hypothesis is accepted. Thus, we conclude that the population variance is 10.

12.5.2 Test of Homogeneity

The test for homogeneity is useful in a case when we intend to verify whether several populations are homogeneous with respect to some characteristic of interest. The null hypothesis specifies that several populations are homogeneous with respect to a characteristic.

This test is different from the test of independence on account of the following reasons:

1. Instead of knowing whether two attributes are independent or not, we may like to know whether different samples come from the same population.
2. Instead of taking only one sample, for this test two or more independent samples are drawn from each population.
3. When the characteristic to be compared consist of two categories, this test is similar to the test of hypothesis of difference between two populations' means or proportions.

To apply this test, random sample is drawn from each population, and then in each sample, the proportion falling into each category or stratum is determined. The sample data so obtained is arranged in a contingency table. The procedure for testing the hypothesis is the same as discussed for test of independence.

Summary

In this study session, you have learnt the following:

1. Contingency Table

   A (*r* x *c*) contingency table shows the observed frequencies for two categorical variables arranged in *r* rows and *c* columns.

2. Expected value E_{ij}

$$E_{ij} = \frac{R_i}{n} \times \frac{C_j}{n} \times n = \frac{R_i \times C_j}{n}$$

3. Test of Independence of attributes & Goodness-of-fit Test

   test-statistic $$\chi^2 = \sum \frac{(O-E)^2}{E}$$

4. Yate's Correction factor for a 2 x 2 Contingency Table

$$\chi^2_{corrected} = \frac{n(ad-bc-\frac{1}{2}n)^2}{(a+b)(c+d)(a+c)(b+d)}; \quad ad-bc>0$$

   and

$$\chi^2_{corrected} = \frac{n(bc-ad-\frac{1}{2}n)^2}{(a+b)(c+d)(a+c)(b+d)}; \quad ad-bc<0$$

5. Test of Population Variance

   test statistic $$\chi^2 = \frac{\sum\left(X-\bar{X}\right)^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$$

6. Confidence Interval for variance

$$\frac{(n-1)s^2}{\chi^2_{df,U}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{df,L}}$$

Self-Assessment Questions (SAQs) for Study Session 12

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 12.1 (Testing Learning Outcomes 12.1)

Test for independence of attributes, that is, analyze contingency tables;

SAQ 12.2 (Testing Learning Outcomes 12.2)

Test for goodness-of-fit

SAQ 12.3 (Testing Learning Outcomes 12.3)

Apply Yate's correction for continuity

SAQ 12.4 (Testing Learning Outcomes 12.4)

Test for population variance and Test for homogeneity.

Study Session 13: Introduction to Analysis of Variance (ANOVA)

Introduction

In Study Session 9 and 10, we introduced hypothesis testing procedures to test the significance of differences between two means to understand whether the means of two populations are equally based upon two independent random samples.

However, there may be situations where more than two populations are involved, and we need to test the significance of differences between three or more sample means.

We also need to test the null hypothesis that three or more populations from which independent samples are drawn have equal (or homogeneous) means against the alternative hypothesis that the population means are not all equal.

Under certain assumptions, a method known as *analysis of variance* or *ANOVA,* developed by R. A. Fisher, is used to test the significance of the difference between several population means.

Learning Outcomes for Study Session

When you have studied this session, you should be able to:

13.1 Explain the concept of ANOVA;

13.2 List some areas where ANOVA can be applied;

13.3 List the assumptions required for analysis of variance; and

13.4 Understand some basic terms.

13.1 The Concept of ANOVA

ANOVA is a widely used technique for comparing the means of several populations, given samples of observations from those populations. It is based on an analysis of the total variation displayed by the data, splitting this into variation between the samples and variation within the samples, and then comparing these components.

TOTAL VARIATION = BETWEEN SAMPLES VARIATION + WITHIN SAMPLES VARIATION

Let $\mu_1, \mu_2, \ldots, \mu_k$ be the mean value for population 1, 2, … , k respectively. Then from sample data, we intend testing the following hypotheses:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ against H_1: Not all μ_j are equal (j = 1, 2, …, k)

In other words, the null and alternative hypotheses of population means imply that the null hypothesis should be rejected if any of the k sample means is different from the others.

13.2 Areas of Applications

The following are a few examples, involving more than two populations where it is necessary to conduct a comparative study to arrive at a statistical inference.

- Effectiveness of different promotional devices in terms of sales.
- Quality of a product produced by different manufacturers in terms of an attribute.
- Production volumes in different shifts in a factory.
- Yield from plots of land due to varieties of seeds, fertilizers, and cultivation methods.
- Effectiveness of different treatment methods or drugs for a particular disease.

13.3 Assumptions for Analysis of Variance

The following assumptions are required for analysis of variance:

- Each population has a normal distribution.
- The populations from which the samples are drawn have equal variances, that is, $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ for k populations.
- Each sample is drawn randomly and is independent of other samples.

Some Basic Terms

In analysis of variance, we often use terms such as “response variable”, “a factor or criterion”, and “a treatment”. Let us illustrate with an example. Suppose the production levels in three shifts in a factory are to be compared to determine if production level is higher/lower on any day of the week.

There are two variables

- Days of the week and
- Volume of production in each shift. If one of the objectives is to determine whether mean production volume is the same during days of the week, then the variable of interest is, that is, the **response**, which is the mean production volume. The variables, qualitative or quantitative, that are related to a response variable are called **factors,** that is, a day of the week is the **factor** or **independent variable** and the value assumed by a factor in an experiment is called a level**.**

  The combinations of levels of the factors for which the response will be observed are called **treatments.** In this example, days of the week are treatments. These treatments define the populations or samples, which are differentiated in terms of production volume and we have interest to compare them with each other.

  Other examples of treatments or populations can be different drug dosages given for the treatment of a disease, assembly line techniques for manufacturing a product, and so on.

Summary

In this study session, you have learnt the following:

1. ANOVA is a technique used for comparing the means of several populations.
2. The following assumptions are required for analysis of variance:

- Each population has a normal distribution.
- The populations from which the samples are drawn have equal variances, that is, $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ for k populations.

Each sample is drawn randomly and is independent of other samples

Self-Assessment Questions (SAQs) for Study Session 13

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 13.1 (Testing Learning Outcomes 13.1)

Explain the concept of ANOVA

SAQ 13.2 (Testing Learning Outcomes 13.2)

List some areas where ANOVA can be applied

SAQ 13.3 (Testing Learning Outcomes 13.3)

List the assumptions required for analysis of variance

SAQ 13.4 (Testing Learning Outcomes 13.4)

Understand some basic terms.

Study Session 14: Anova One-Way Classification

Introduction

The basic idea of analysis of variance is to express a measure of the total variability of a set of data as a sum of terms, each of which can be attributed to a specific source, or cause, of variation. The observations in the sample data may be classified according to one factor (criterion) or two factors (criteria).

The classifications, according to one factor and two factors, are called one-way classification and two-way classification, respectively.

We shall consider the former in this study session

Learning Outcomes for Study Session 14

When you have studied this session you should be able to:

14.1 Understand the layout of one-way classification;

14.2 Construct ANOVA for one-way classification; and

14.3 Test hypothesis for the equality of three or more population means.

14.1 The Layout of One-way Classification

Suppose that there are k treatments (j = 1, . . . , k) and that treatment j has n_j observations. Xij is the ith observation on the jth treatment. The observations (measurements) obtained for k independent samples based on one-criterion classification can be arranged as shown in the following table:

Observations (Measurements)	Treatments (Number of Samples)			
	1	2	…	K
1	x_{11}	x_{12}	…	x_{1k}
2	x_{21}	x_{22}	…	x_{2k}
.	.	.		.
.	.	.		.
.	.	.		.
r	x_{r1}	x_{r2}	…	x_{rk}
Sum	T_1	T_2	…	T_k
Mean	$\bar{X}_1$	$\bar{X}_2$		$\bar{X}_k$

Where

$T_j = \sum_{i=1}^{r} x_{i1}$ is called the treatment or sample totals

$T = \sum_{j=1}^{k} T_j$ is called the grand total of all observations (or measurements)

$\bar{X}_j = \frac{1}{r} \sum_{i=1}^{r} x_{i1}$ is called the treatment or sample means

$\bar{\bar{X}} = \frac{1}{rk} \sum_{j=1}^{k} \bar{X}_j = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{k} x_{ij}$ is called the grand mean of all observations (or measurements)

Since there are r rows and k columns, then total number of observations is rk = n, provided each row has equal number of observations. But, if the number of observations in each column varies, then the total number of observations is $n_1 + n_2 + \ldots + n_k = n$.

14.2. Steps for Constructing the ANOVA

- **Step 1 Calculate the Total Variation**

  The total variation is represented by the "sum of squares total" (SST) and is equal to the sum of the squared differences between each sample value from the grand mean.

  $$\text{SST} = \sum_{i=1}^{r}\sum_{j=1}^{k}(X_{ij} - \bar{\bar{X}})^2$$

- **Step 2 Calculate the Variation between Sample Means**

  This is usually called the "sum of squares between" (SSB). It measures the variation between samples due to treatments. In statistical terms, variation between samples means is also called the *between-column variance*.

  $$\text{SSB} = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{\bar{X}})^2$$

- **Step 3 Calculate the Variation Within Samples**

  This is usually called the "sum of squares within" (SSW). It measures the difference within samples due to chance error. Such variation is also called *within sample variance* or *error sum of squares*.

  $$\text{SSW} = \sum_{i=1}^{r}\sum_{j=1}^{k}(X_{ij} - \bar{X}_j)^2$$

  Alternatively, SSW = SST – SSB

- **Step 4 Calculate the Average Variation Between and Within Samples –Mean Squares**

  Since k independent samples are being compared, therefore k – 1 degrees of freedom are associated with SSB. Also, as each of the k samples contributes n_j – 1 degrees of freedom for each independent sample within itself, therefore there are n – k degrees of freedom associated with SSW. Thus, total degrees of freedom equal to the degrees of freedom associated with SSB and SSW. That is

$$\text{Total } df = \text{SSB } df + \text{SSW } df$$

$$n - 1 = k - 1 + n - k$$

When these "sum of squares" are divided by their associated degrees of freedom, we get the following variances or *mean square* terms:

$$MSB = \frac{SSB}{k-1}; \qquad MSW = \frac{SSW}{n-k}$$

- **Step 5 Compute the F-statistic**

  The F-statistic is given by

$$F = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{SSB/(k-1)}{SSW/(n-k)} = \frac{MSB}{MSW}$$

The following table shows the general arrangement of the ANOVA table for one-factor analysis of variance:

ANOVA Summary Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Test-Statistic or F-Value
Between Samples	SSB	k – 1	MSB	F = MSB/MSW
Within Samples	SSW	n – k	MSW	
Total	SST	n - 1		

14.3 Testing for the Equality of Population Means

- **Hypotheses**

  H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ against H_1: Not all ar μ_j ual (j = 1, 2, …, k)

- **Decision Rule**

  ➢ Reject H_0 if the calculated value of F > its critical value $F_{\alpha(k-1, n-k)}$

  ➢ Otherwise accept H_0.

Note:

The F distribution is a family of distributions, each identified by a pair of degrees of freedom. The first number refers to the number of degrees of freedom in the numerator of the F ratio, and the second refers to the number of degrees of freedom in the denominator. In the F table, columns represent the degrees of freedom for the numerator and the row represents the degrees of freedom for the denominator.

Example

There are five treatments (A, B, C, D, and E) for lowering blood pressure. An initial test is needed to ascertain whether there is any real difference between them. Each treatment is given to a different randomly chosen sample of people with high blood pressure. The results, using suitable units, are as follows. Test at 5% level of significance, whatever the different treatments, has the same mean effect.

Observations	Treatments				
	A	B	C	D	E
1	12	10	3	7	16
2	6	15	2	8	18
3	5	14	7	7	21
4	7	13	8	10	19
5	10	12	3		21
6		12	1		
7		12			

Solution

- Hypothesis

  H_0: the means of the treatments are equal

  H_1: the means of the treatments are not all equal

- Computations

  For these data, k = 5

  $n_1 = 5$, $n_2 = 7$, $n_3 = 6$, $n_4 = 4$, $n_5 = 5$ and n = 27

  The treatment means are

  $\bar{X}_1 = 8$, $\bar{X}_2 = 13$, $\bar{X}_3 = 4$, $\bar{X}_4 = 8$, $\bar{X}_5 = 19$

  The grand mean is $\bar{\bar{X}} = 10.44$

  Within Samples

  Sum of squares, SSW = 370

  Degrees of freedom, n – k = 22

Mean square, MSW = 16.82

Between Samples

Sum of squares, SSB = 738.6

Degrees of freedom, k – 1 = 4

Mean square, MSB = 184.6

The test statistic F = 184.6/16.82 = 11.0

ANOVA TABLE

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Test-Statistic or F-Value
Between Samples	738.6	4	184.6	F = 11.0
Within Samples	370	22	16.82	
Total	1108.6	26		

- Decision

The critical value for $F_{4,22}$ at the 5% significance level is 2.82.

Since 11.0 > 2.82, the null hypothesis is rejected. The different treatments do not have the same mean effect.

Summary

In this study session, you have learnt the following:

1. Treatment Totals $T_j = \sum_{i=1}^{r} x_{i1}$

2. Grand Total $T = \sum_{j=1}^{k} T_j$

3. Treatment Means $\bar{X}_j = \frac{1}{r} \sum_{i=1}^{r} x_{i1}$

4. Grand Mean $\bar{\bar{X}} = \frac{1}{rk} \sum_{j=1}^{k} \bar{X}_j = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{k} x_{ij}$

5. Total Variation SST = $\sum_{i=1}^{r} \sum_{j=1}^{k} (X_{ij} - \bar{\bar{X}})^2$

6. Variation Between SamplesSSB = $\sum_{j=1}^{k} n_j (\bar{X}_j - \bar{\bar{X}})^2$

7. Variation Within Samples SSW = $\sum_{i=1}^{r} \sum_{j=1}^{k} (X_{ij} - \bar{X}_j)^2$ = SST - SSB

Self-Assessment Questions (SAQs) for Study Session 14

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 14.1 (Testing Learning Outcomes 14.1)

Understand the layout of one-way classification

SAQ 14.2 (Testing Learning Outcomes 14.2

Construct ANOVA for one-way classification

SAQ 14.3 (Testing Learning Outcomes 14.3)

Test hypothesis for the equality of three or more population means.

Study Session 15: ANOVA Two-Way Classification

Introduction

In one-way ANOVA, we divided the total variation in the sample data into two: variation among the samples due to different samples or treatments and variation within samples due to random error.

However, there might be a possibility that some of the variation left in the random error from one-way analysis of variation was not due to random error or chance, but due to some other measurable factor. In two-way analysis of variance, we are introducing another term called "blocking factor" to remove the undesirable accountable variation.

Learning Outcomes for Study Session 15

When you have studied this session you should be able to:

15.1 Explain the concept of blocking;

15.2 Construct ANOVA for two-way classification

15.3 Test equality of several population means with block-restricted samples.

15.1 The Concept of Blocking

The term "blocking" refers to a block of land and, thus, has agricultural origin. For instance, if the aim of an experiment is to compare the effects of several different fertilizers, an experimental field may be divided into fairly small sections (often called plots).

A stream flows alongside one edge of the field, so there is likely to be a natural fertility gradient sideways across the field; plots near the stream might well be naturally more fertile than those further away. So, it is sensible to divide the field into strips running parallel to the stream, each strip having several plots in it, and to make sure that all the fertilizers are used in each strip.

The results within each strip can be compared with each other, but there may be consistent differences from strip to strip. In this experiment, each strip is a block. The variability within a block would be expected to be less than that within the whole population being investigated.

The block-restriction is usually applied when an extraneous (nuisance) factor is suspected to have some influence on the observations, and observations within a block share some commonality based on the extraneous factor. If the block effects are found to be all equal, then the extraneous factor has no influence on the observations. Therefore, the observations within a column can be treated as an unrestricted sample.

Generally, for c samples or treatments (columns), r blocks (rows), and number of observations n = r x c, the separation of total variation in the sample data is shown below:

TOTAL VARIATION (SST) = VARIATION BETWEEN SAMPLES (SSB or SSC) + VARIATION BETWEEN BLOCKS (SSR) + VARIATION DUE TO RANDOM ERROR (SSE)

15.2 Steps for Constructing the ANOVA

- Step 1 Calculate the Total Variation

$$\text{SST} = \sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \bar{\bar{X}})^2$$

- Step 2 Calculate the Variation between Treatment (Column) Means

$$\text{SSC} = r\sum_{j=1}^{c}(\bar{X}_j - \bar{\bar{X}})^2$$

- Step 3 Calculate the Variation between Blocks (Rows) Means

  SSR = $c\sum_{i=1}^{r}(\bar{X}_i - \bar{\bar{X}})^2$

- Step 3 Calculate the Variation Due to Random Error

  SSE = $\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \bar{X}_i - \bar{X}_j + \bar{\bar{X}})^2$

  Alternatively, SSE = SST – SSC – SSR

- Step 4 Calculate the Average Variations

  Total *df* = SSC *df* + SSR *df* + SSE *df*

  cr – 1 = c – 1 + r – 1 + (c-1)(r-1)

  When these "sums of squares" are divided by their associated degrees of freedom, we get the following variances or *mean square* terms.

  $$MSC = \frac{SSC}{c-1}; \qquad MSR = \frac{SSR}{r-1}; \quad MSE = \frac{SSE}{(c-1)(r-1)}$$

- Step 5 Compute the F-statistic

  The test statistic F for analysis of variance is given by

  $$F_1 = \frac{MSC}{MSE}$$

  $$F_2 = \frac{MSR}{MSE}$$

The following shows the general arrangement of the ANOVA table for two-way analysis of variance:

ANOVA Summary Table for Two-way Classification

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Test-Statistic or F-Value
Between columns	SSC	c – 1	MSC	F_1 = MSC/MSE
Between Rows	SSR	r - 1	MSR	F_2 = MSR/MSE
Residual Error	SSE	(c – 1)(r -1)	MSE	
Total	SST	cr - 1		

15.3 Testing for the Equality of Population Means

- **Hypotheses**

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ against H_1: Not all μ_j's re equal (j = 1, 2, …, c)

- **Decision Rule**

  - Reject H_0 if the calculated value of F_1 > its critical value $F_{\alpha,(c-1),(c-1)(r-1)}$
  - Otherwise accept H_0.

15.4 Testing for the Equality of Block Effects

- **Hypotheses**

H_0: $b_1 = b_2 = \ldots = b_r$ against H_1: Not all b_is are equal (i = 1, 2, …, r)

- **Decision Rule**

  - Reject H_0 if the calculated value of F_2 > its critical value $F_{\alpha,(r-1),(c-1)(r-1)}$
  - Otherwise accept H_0.

Example

An experimenter wants to study how weight gain in rats is influenced by the source of protein. There are six different diets (called treatments) as source of protein, and a total of 60 experimental rats are available for the study. Prior to the experiment, the rats are divided into b=10 homogeneous groups (called blocks) of size 6 based on initial body weight (i.e., rats within a group or block have more or less the same initial body weight).

Within each of the 10 blocks, six rats are randomly assigned one of the six diets (treatments). The weight gain (in grams), after a fixed period of time, is measured for each of the experimental rats as given in the following table.

Weight gains of rats in 10 blocks with 6 diets

Block	**Diet-1**	**Diet-2**	**Diet-3**	**Diet-4**	**Diet-5**	**Diet-6**
1	*90*	*87*	*83*	*107*	*96*	*111*
2	*94*	*70*	*82*	*102*	*72*	*100*
3	*86*	*95*	*85*	*102*	*76*	*102*
4	*63*	*71*	*63*	*93*	*70*	*93*
5	*81*	*75*	*72*	*111*	*79*	*101*
6	*89*	*84*	*85*	*128*	*89*	*104*
7	*63*	*62*	*64*	*56*	*70*	*72*
8	*82*	*72*	*80*	*97*	*91*	*92*
9	*63*	*81*	*82*	*80*	*63*	*87*
10	*81*	*93*	*83*	*103*	*102*	*112*

Based on the data, test whether all the six diets have the same effect on weight again (on an average).

Solution

Total sample size = r x c = 10 x 6 = 60

Treatment (Column) Means: $\overline{X}_1 = 79.2, \overline{X}_2 = 79.0, \overline{X}_3 = 77.9, \overline{X}_4 = 97.9, \overline{X}_5 = 80.8, \overline{X}_6 = 97.4$

Block (Row) Means:

$$\overline{X}_1 = 95.667, \overline{X}_2 = 86.667, \overline{X}_3 = 91.0, \overline{X}_4 = 75.5, \overline{X}_5 = 86.5,$$
$$\overline{X}_6 = 96.5, \overline{X}_7 = 64.5, \overline{X}_8 = 85.667, \overline{X}_9 = 76.0, \overline{X}_{10} = 95.667$$

Grand Mean

$\overline{\overline{X}} = 85.367$

SST = 13642.5187, SSC = 4570.5340, SSR = 5946.6402, SSE = 3125.3445

MSC = 914.1068, MSR = 660.7378, MSE = 69.4521

F_1 = 13.162 F_2 = 9.5136

Two-way ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Test-Statistic or F-Value
Treatments (Columns)	4570.534	5	914.1068	F_1 = 13.162
Blocks (Rows)	5946.6402	9	660.7378	F_2 = 9.5136
Residual Error	3125.3442	45	69.4521	
Total	13642.5184	59		

$F_{\alpha,(c-1),(c-1)(r-1)}$ = $F_{0.05,\ 5,45}$ = 2.422. Since F_1 > 2.422, we reject H_0

Also, $F_{\alpha,(r-1),(c-1)(r-1)}$ = $F_{0.05,9,45}$ = 2.096. Since F_2 > 2.096, we reject H_0

The data indicates that there is sufficient reason to believe that the six diets have different effects on the weight gain in rats. Also, the weight gain is influenced by the initial body weight of the rats.

1. **Treatment Totals** $T_j = \sum_{i=1}^{r} x_{i1}$

2. **Block Totals** $B_i = \sum_{j=1}^{c} x_{1j}$

3. **Grand Total** $T = \sum_{j=1}^{c} T_j$

4. **Treatment Means** $\bar{X}_j = \frac{1}{r}\sum_{i=1}^{r} x_{i1}$

5. **Block Means** $\bar{X}_i = \frac{1}{c}\sum_{j=1}^{c} x_{1j}$

6. **Grand Mean** $\bar{\bar{X}} = \dfrac{1}{cr}\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}$

7. **Total Variation** SST = $\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \bar{\bar{X}})^2$

8. **Variation Between Treatments** SSC = $r\sum_{j=1}^{c}(\bar{X}_j - \bar{\bar{X}})^2$

9. **Variation Between Blocks** SSR = $c\sum_{i=1}^{r}(\bar{X}_i - \bar{\bar{X}})^2$

10. **Variation Due to Random Errors**

    a. SSE = $\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \bar{X}_i - \bar{X}_j + \bar{\bar{X}})^2$

    b. Alternatively, SSE = SST – SSC – SSR

Summary

In this study session, you have learnt the following:

Self-Assessment Questions (SAQs) for Study Session 15

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ 15.1 (Testing Learning Outcomes 15.1)

Explain the concept of blocking;

SAQ 15.2 (Testing Learning Outcomes) 15.2

Construct ANOVA for two-way classification

SAQ 15.3(Testing Learning Outcomes) 15.3

Test equality of several population means with block-restricted samples.