MAR GREGORIOS COLLEGE
OF ARTS & SCIENCE

Block No.8, College Road, Mogappair West, Chennai – 37

Affiliated to the University of Madras
Approved by the Government of Tamil Nadu
An ISO 9001:2015 Certified Institution

DEPARTMENT OF MATHEMATICS

SUBJECT NAME: MATHEMATICAL STATISTICS II

SUBJECT CODE: BMA-CSA06

SEMESTER: VI

PREPARED BY: PROF.S.C.PREMILA

UNIVERSITY OF MADRAS

B.Sc. DEGREE COURSE IN MATHEMATICS

SYLLABUS WITH EFFECT FROM 2020-2021

BMA-CSA06

ALLIED: MATHEMATICAL STATISTICS-II

(Common to B.Sc. Maths with Computer Applications)

Learning outcomes:

Students will acquire knowledge

- To provide the foundation of statistical analysis used in varied applications.
- Of Sampling methods, Tests of significance and testing of hypothesis.

UNIT I

Sampling theory – Sampling Distributions – Concept of Standard error – Sampling distribution based on normal distribution- t, Chi Square and F distributions.

UNIT II

Point estimation – Concepts of unbiasedness – consistency – efficiency and sufficiency- Cramer Rao inequality – Methods of estimation- Maximum likelihood- moments - minimum square and their properties (Statement only).

UNIT III

Test of significance – Standard error- Large sample test, Exact test based on normal, t, chi-square and F distribution with respect to population mean/means, proportion/proportions, variance and correlation coefficient. Test of independence of attributes based on contingency tables- Goodness of fit based on chi-square.

UNIT IV

Analysis of Variance: One way, two way classification concepts &Problems.Interval estimation – Confidence intervals for population mean/means- Proportion/proportions and variances based on t, Chi-Square and F.

UNIT V

Test of hypothesis- Type I and II errors- Power of test – Neymann Pearson lemma- Likelihood ratio test- concepts of most powerful test- statements and results only-simple problems.

Reference:

- S.C.Gupta&V.K.Kapoor: Elements of Mathematical Statistics, Sultan Chand & Sons, NewDelhi.
- Hogg R.V. & Craig A.T. (1988): Introduction to Mathematical Statistics, McMillan.
- Mood A.M. &Graybill F.A. &Boes D.G. (1974): Introduction to theory of Statistics, McGraw Hill.
- Snedecor G.W. & Cochran W.G(1967) : Statistical Methods, Oxford and IBH.
- Hoel P.G. (1971) : Introduction to Mathematical Statistics, Wiley.
- Wilks S.S. Elementary Statistical Analysis, Oxford and IBH.

UNIT - 1

In Statistics, the **sampling method** or **sampling technique** is the process of studying the population by gathering information and analyzing those data. It is the basis of the data where the sample space is enormous. The statistical research is of two forms:

- In the first form, each domain is studied, and the result can be obtained by computing the sum of all units.
- In the second form, only a unit in the field of the survey is taken. It represents the domain. The result of these samples extends to the domain. This type of study is known as the sample survey.

In this article, let us discuss the different sampling methods in research such as probability sampling and non-probability sampling methods and various methods involved in those two approaches in detail.

What are the sampling methods?

There are several different sampling techniques available, and they can be subdivided into two groups. All these methods of sampling may involve specifically targeting hard or approach to reach groups.

Types of Sampling Method

In Statistics, there are different sampling techniques available to get relevant results from the population. The two different types of sampling methods are::

- Probability Sampling

Non-probability Sampling

What Is a Sampling Distribution?

A sampling distribution is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population. The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.

In statistics, a population is the entire pool from which a statistical sample is drawn. A population may refer to an entire group of people, objects, events, hospital visits, or measurements. A population can thus be said to be an aggregate observation of subjects grouped together by a common feature.

- A sampling distribution is a statistic that is arrived out through repeated sampling from a larger population.
- It describes a range of possible outcomes that of a statistic, such as the mean or mode of some variable, as it truly exists a population.
- The majority of data analyzed by researchers are actually drawn from samples, and not populations.

What Is the Standard Error?

The standard error (SE) of a statistic is the approximate standard deviation of a statistical sample population. The standard error is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation. In statistics, a sample mean deviates from the actual mean of a population; this deviation is the standard error of the mean.

PROPERTIES:

- The standard error is the approximate standard deviation of a statistical sample population.
- The standard error can include the variation between the calculated mean of the population and one which is considered known, or accepted as accurate.
- The more data points involved in the calculations of the mean, the smaller the standard error tends to be.

The **T Distribution** also called the student's t-distribution and is used while making assumptions about a mean when we don't know the standard deviation. In probability and statistics, the normal distribution is a bell-shaped distribution whose **mean is μ and the standard deviation is σ**. The t-distribution is similar to normal distribution but **flatter and shorter** than a normal distribution. Here, we are going to discuss what is t-distribution, formula, table, properties, and applications.

T- Distribution Definition

The t-distribution is a hypothetical probability distribution. It is also known as the student's t-distribution and used to make presumptions about a mean when the standard deviation is not known to us. It is symmetrical, bell-shaped distribution, similar to the standard normal curve. As high as the degrees of freedom (df), the closer this distribution will approximate a standard normal distribution with a mean of 0 and a standard deviation of 1.

T Distribution Formula

A t-distribution is the whole set of t values measured for every possible random sample for specific sample size or a particular degree of freedom. It approximates the shape of normal distribution.

Let x have a normal distribution with mean 'μ' for the sample of size 'n' with sample mean $\bar{x}$ and the sample standard deviation 's', then the t variable has student's t-distribution with a degree of freedom, d.f = n – 1. The formula for t-distribution is given by;

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

Where, $\bar{x}$ is the mean of first sample.

μ is the mean of second sample.

$\frac{s}{\sqrt{N}}$ = the estimate of the standard error of difference between the means.

T-Test Assumptions

1. The first assumption made regarding t-tests concerns the scale of measurement. The assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale, such as the scores for an IQ test.
2. The second assumption made is that of a simple random sample, that the data is collected from a representative, randomly selected portion of the total population.
3. The third assumption is the data, when plotted, results in a normal distribution, bell-shaped distribution curve. When a normal distribution is assumed, one can specify a level of probability (alpha level, level of significance, *p*) as a criterion for acceptance. In most cases, a 5% value can be assumed.
4. The fourh assumption is a reasonably large sample size is used. A larger sample size means the distribution of results should approach a normal bell-shaped curve.
5. The final assumption is homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.

CHI SQUARED TEST

A **chi-squared test** (symbolically represented as χ^2) is basically a data analysis on the basis of observations of a random set of variables. Usually, it is a comparison of two statistical data sets. This test was introduced by **Karl Pearson** in 1900 for categorical data analysis and distribution. So it was mentioned as **Pearson's chi-squared test**.

The chi-square test is used to estimate how likely the observations that are made would be, by considering the assumption of the null hypothesis as true.

A hypothesis is a consideration that a given condition or statement might be true, which we can test afterwards. Chi-squared tests are usually created from a sum of squared falsities or errors over the sample variance.

Chi-Square Distribution

When we consider, the null speculation is true, the sampling distribution of the test statistic is called as **chi-squared distribution**. The chi-squared test helps to determine whether there is a notable difference between the normal frequencies and the observed frequencies in one or more classes or categories. It gives the probability of independent variables.

Probability Density Function

The chi-square distribution results when *v* independent variables with standard normal distributions are squared and summed. The formula for the probability density function of the chi-square distribution is

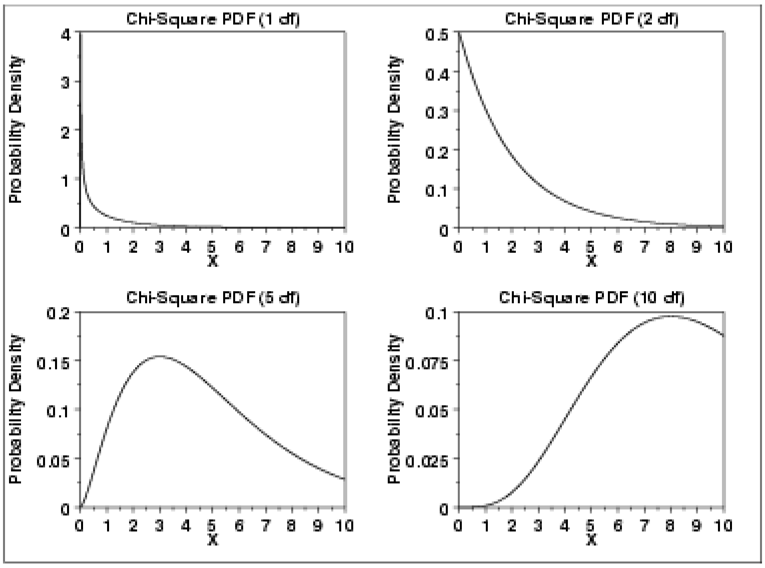
$$f(x)=e-x2xv2-12v2\Gamma(v2)\mathrm{forx}\geq 0$$

where v is the shape parameter and Γ is the gamma function. The formula for the gamma function is

$$\Gamma(a)=\int\infty 0ta-1e-tdt$$

In a testing context, the chi-square distribution is treated as a "standardized distribution" (i.e., no location or scale parameters). However, in a distributional modeling context (as with other probability distributions), the chi-square distribution itself can be transformed with a location parameter, μ, and a scale parameter, σ.

The following is the plot of the chi-square probability density function for 4 different values of the shape parameter.



Cumulative Distribution Function

The formula for the cumulative distribution function of the chi-square distribution is

$$F(x)=\gamma(v2,x2)\Gamma(v2)\mathrm{forx}\geq 0$$

where Γ is the gamma function defined above and *γ* is the incomplete gamma function. The formula for the incomplete gamma function is

$\Gamma_x(a)=\int_0^x t_{a-1}e_{-t}dt$

F Test Formula

A test statistic which has an F-distribution under the null hypothesis is called an F test. It is used to compare statistical models as per the data set provided or available. George W. Snedecor, in honour of Sir Ronald A. Fisher, termed this formula as F-test Formula.

$$FValue=Varianceofset1Varianceofset2=\sigma21\sigma22$$

To compare the variance of two different sets of values, the F test formula is used. To be applied to F distribution under the null hypothesis, we first need to find out the mean of two given observations and then calculate their variance.

$$\sigma_2=\sum(x-\overline{x})_2n-1$$

Where,
σ^2 = Variance
x = Values given in a set of data
$\overline{x}$ = Mean of the data
n = Total number of values

ESTIMATION

Estimation, in statistics, any of numerous procedures used to calculate the value of some property of a population from observations of a sample drawn from the population. A point estimate, for example, is the single number most likely to express the value of the property. An interval estimate defines a range within which the value of the property can be expected (with a specified degree of confidence) to fall.

Estimation, in statistics, any of numerous procedures used to calculate the value of some property of a population from observations of a sample drawn from the population.

A point estimate, for example, is the single number most likely to express the value of the property.

An interval estimate defines a range within which the value of the property can be expected (with a specified degree of confidence) to fall.

The 18th-century English theologian and mathematician Thomas Bayes was instrumental in the development of Bayesian estimation to facilitate revision of estimates on the basis of further information. (*See* Bayes's theorem.)

In sequential estimation the experimenter evaluates the precision of the estimate during the sampling process, which is terminated as soon as the desired degree of precision has been achieved.

The statistical estimation of the population parameter is further divided into two types, (i) Point Estimation and (ii) Interval Estimation

Point Estimation

The objective of point estimation is to obtain a single number from the sample which will represent the unknown value of the population parameter.
Population parameters (population mean, variance, etc) are estimated from the corresponding sample statistics (sample mean, variance, etc).
A statistic used to estimate a parameter is called a point estimator or simply an estimator, the actual numerical value obtained by estimator is called an estimate..

Interval Estimation

A point estimator (such as sample mean) calculated from the sample data provides a single number as an estimate of the population parameter, which can not be expected to be exactly equal to the population parameter because the mean of a sample taken from a population may assume different values for different samples. Therefore we estimate an interval/ range of values (set of values) within which the population parameter is expected to lie with a certain degree of confidence. This range of values used to estimate a population parameter is known as interval estimate or estimate by a confidence interval, and is defined by two numbers, between which a population parameter is expected to lie.

What is an estimator?

In machine learning, an estimator is an equation for picking the "best," or most likely accurate, data model based upon observations in realty. Not to be confused with estimation in general, the estimator is the formula that evaluates a given quantity (the estimand) and generates an estimate. This estimate is then inserted into the deep learning classifier system to determine what action to take.

Uses of Estimators

By quantifying guesses, estimators are how machine learning in theory is implemented in practice. Without the ability to estimate the parameters of a dataset (such as the layers in a neural network or the bandwidth in a kernel), there would be no way for an AI system to "learn."

A simple example of estimators and estimation in practice is the so-called "German Tank Problem" from World War Two. The Allies had no way to know for sure how many tanks the Germans were building every month. By counting the serial numbers of captured or destroyed tanks (the estimand), Allied statisticians created an estimator rule. This equation calculated the maximum possible number of tanks based upon the sequential serial numbers, and apply minimum variance analysis to generate the most likely estimate for how many new tanks German was building.

Types of Estimators

Estimators come in two broad categories—point and interval. Point equations generate single value results, such as standard deviation, that can be plugged into a deep learning algorithm's classifier functions. Interval equations generate a range of likely values, such as a confidence interval, for analysis.

In addition, each estimator rule can be tailored to generate different types of estimates:

- Biased - Either an overestimate or an underestimate.
- Efficient - Smallest variance analysis. The smallest possible variance is referred to as the "best" estimate.
- Invariant: Less flexible estimates that aren't easily changed by data transformations.
- Shrinkage: An unprocessed estimate that's combined with other variables to create complex estimates.
- Sufficient: Estimating the total population's parameter from a limited dataset.

Unbiased: An exact-match estimate value that neither underestimates nor overestimates.

Unbiased and Biased Estimators

We now define unbiased and biased estimators. We want our estimator to match our parameter, in the long run. In more precise language we want the expected value of our statistic to equal the parameter. If this is the case, then we say that our statistic is an unbiased estimator of the parameter.

If an estimator is not an unbiased estimator, then it is a biased estimator. Although a biased estimator does not have a good alignment of its expected value with its parameter, there are many practical instances when a biased estimator can be useful. One such case is when a plus four confidence interval is used to construct a confidence interval for a population proportion.

- **Example for Means**
- To see how this idea works, we will examine an example that pertains to the mean. The statistic
- $(X_1 + X_2 + . . . + X_n)/n$
- is known as the sample mean. We suppose that the random variables are a random sample from the same distribution with mean μ. This means that the expected value of each random variable is μ.
- When we calculate the expected value of our statistic, we see the following:
- $E[(X_1 + X_2 + . . . + X_n)/n] = (E[X_1] + E[X_2] + . . . + E[X_n])/n = (nE[X_1])/n = E[X_1] = μ$.
- Since the expected value of the statistic matches the parameter that it estimated, this means that the sample mean is an unbiased estimator for the population mean.

-).

- **What is Sufficient Estimator?**

- An estimator of a parameter θ which gives as much information about θ as is possible from the sample at hand is called a sufficient estimator. Sufficient estimators exist when one can reduce the dimensionality of the observed data without loss of information. In A/B testing the most commonly used sufficient estimator (of the population mean) is the sample mean (proportion in the case of a binomial metric). A conversion rate of any kind is an example of a sufficient estimator.
- In a more formal expression it can be said that a statistic is sufficient with respect to an unknown parameter and a given family of probability distributions if the sample from which it is calculated gives no additional information as to which of those probability distributions produced it than does the statistic itself. Thus sufficiency refers to how well an estimator utilizes the information in the sample relative to the postulated statistical model.
- Sufficiency is an important quality in hypothesis testing where we are effectively comparing the distribution under the null hypothesis with the actually observed distribution. Having a sufficient estimator makes this process significantly more manageable, especially for large sample sizes.
- be calculated from a sample drawn from a larger population. A consistent estimator is an estimator with the property that the probability of the estimated value and the true value of the population parameter not lying within c units (c is any arbitrary positive constant) of each other approaches zero as the sample size tends to infinity.

- For example, consider a population mean of 10 and an interval of 1 unit either side -- the interval from 9 to 11. As samples get larger, the probability that the sample mean will fall outside that interval diminishes, and approaches zero when the sample gets large enough. For a smaller interval, it takes longer for the probability to approach zero.

Consistent Estimator

Consistent Estimator:
An estimator is a measure or metric intended to be calculated from a sample drawn from a larger population. A consistent estimator is an estimator with the property that the probability of the estimated value and the true value of the population parameter not lying within c units (c is any arbitrary positive constant) of each other approaches zero as the sample size tends to infinity.

For example, consider a population mean of 10 and an interval of 1 unit either side -- the interval from 9 to 11. As samples get larger, the probability that the sample mean will fall outside that interval diminishes, and approaches zero when the sample gets large enough. For a smaller interval, it takes longer for the probability to approach zero.

- **Efficient Estimator**

- An efficient estimator is the "best possible" or "optimal" estimator of a parameter of interest. The definition of "best possible" depends on one's choice of a loss function which quantifies the relative degree of undesirability of estimation errors of different magnitudes.
- When one compares between a given procedure and a notional "best possible" procedure the efficiency can be expressed as relative finite-sample or asymptotic efficiency (a ratio). The relevance to A/B testing is that the more efficient the estimator, the smaller sample size one requires for an A/B test.
- CRAMER RAO INEQUALITY

In statistics, the **Cramér-Rao inequality**, named in honor of Harald Cramér and Calyampudi Radhakrishna Rao, expresses a lower bound on the variance of an unbiased statistical estimator, based on Fisher information.

It states that the reciprocal of the Fisher information, $\mathcal{I}(\theta)$, of a parameter θ, is a lower bound on the variance of an unbiased estimator of the parameter (denoted $\hat{\theta}$).

$$\operatorname{var}\left(\hat{\theta}\right) \geq \frac{1}{\mathcal{I}(\theta)} = \frac{1}{\mathrm{E}\left[\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right]}$$

In some cases, no unbiased estimator exists that realizes the lower bound.

The Cramér-Rao inequality is also known as the **Cramér-Rao bounds** (CRB) or **Cramér-Rao lower bounds** (CRLB) because it puts a lower bound on the variance of an estimator $\hat{\theta}$.

Estimation method

In the method of Estimation

- a sample is used to make statements about the probability distribution that generated the sample;
- the sample is regarded as the realization of a random vector , whose unknown joint distribution function, denoted by , is assumed to belong to a set of distribution functions , called statistical model;
- MAXIMUM LIKELIHOOD ESTIMATOR

- It seems reasonable that a good estimate of the unknown parameter θ would be the value of θ that **maximizes** the probability, errrr... that is, the **likelihood**... of getting the data we observed. (So, do you see from where the name "maximum likelihood" comes?) So, that is, in a nutshell, the idea behind the method of maximum likelihood estimation. But how would we implement the method in practice? Well, suppose we have a random sample X1,X2,⋯,Xn for which the probability density (or mass) function of each Xi is f(xi;θ). Then, the joint probability mass (or density) function of X1,X2,⋯,Xn, which we'll (not so arbitrarily) call L(θ) is:

- L(θ)=P(X1=x1,X2=x2,…,Xn=xn)=f(x1;θ)·f(x2;θ)⋯f(xn;θ)=∏i=1nf(xi;θ)
- The first equality is of course just the definition of the joint probability mass function. The second equality comes from that fact that we have a random sample, which implies by definition that the Xi are independent. And, the last equality just uses the shorthand mathematical notation of a product of indexed terms. Now, in light of the basic idea of maximum likelihood estimation, one reasonable way to proceed is to treat the "**likelihood function**" L(θ) as a function of θ, and find the value of θ that maximizes it.
-

Minimum chi-square estimation

- In statistics, **minimum variance to be** chi-square estimation **is a method of estimation of unobserved quantities based on observed data.**[1]
- In certain chi-square tests, one rejects a null hypothesis about a population distribution if a specified test statistic is too large, when that statistic would have approximately a chi-square distribution if the null hypothesis is true. In minimum chi-square estimation, one finds the values of parameters that make that test statistic as small as possible.
- Among the consequences of its use is that the test statistic actually does have approximately a chi-square distribution when the sample size is large. Generally, one reduces by 1 the number of degrees of freedom for each parameter estimated by this method.

UNIT III

TESTS OF SIGNIFICANCE (Large sample)

INTRODUCTION:

Any statistical investigation usually deals with the study of some characteristics of a collection of objects

SAMPLING:

Definition:

A finite subset of population is called a sample and the number of objects in a sample is called the sample size.

Some of the important types of sampling are (i) purposive sampling (ii) Random sampling (iii) Simple sampling (iv) stratified sampling.

(i)Purposive sampling:

If the sample elements are selected with a definite purpose in mind then the sample selected is called purposive sample.

(ii) Random sampling:

A random sample is one in which each element of the population has an equal chance of inclusion in the sample.

(iii) Simple sampling:

Simple sampling is a special type of random sampling in which each element of the population has an equal and independent chance of being included in the sample.

(iv) Stratified sampling:

The sample which is the aggregate of the sampled individuals of each stratum is called stratified sample and the technique of selecting such sample is called stratified sampling.

TESTS OF SIGNIFICANCE FOR LARGE SAMPLES

I. Tests for proportion or percentage

(A) Single proportion (B) Difference of proportions.

II. Tests for means.

(A) (i) Test for single mean if standard deviation of the population σ is known. (ie) $H_{0:}\mu = \mu_0$,ς is known.

(B) (ii) Tests for single mean if σ is not known H_0:$\mu = \mu_0$,σ is unknown.

(C) (i) Test for equality of means of 2 normal populations with Known standard deviations (ie) $H_0:\mu_1 = \mu_2; \varsigma_1, \varsigma_2$ is known.

(ii) Test for equality of means of 2 normal populations with same standard deviation though unknown $H_0:\mu_1 = \mu_2, \ \varsigma_1 = \varsigma = \varsigma_2$.

III. Test for standard deviations.

(A) : Test for single standard deviation $H_0:\varsigma = \varsigma_0$

(B) : Test for equality for 2 standard deviation (ie)$H_0:\varsigma_1 = \varsigma_2$

1. Test of significance for proportions and percentages.

I(A) **Single proportions.**

If X is the number success in independent trials with constant probability of success **P** for each trial we have E(X)=nP and V(X)=variance(X)=nPQ where Q=1-P. It has been proved that for large n, the binomial distribution tends to a normal distribution. Hence for large n,

$$X \ \sim N(np, nPQ)$$

$$\therefore Z = \frac{X - E(X)}{S.E\ of\ (X)} = \frac{X - nP}{\sqrt{nPQ}} \sim N(0,1)$$

I (B)Difference of proportions.

Suppose we want to compare 2 distinct populations with regard to possession of an attributes. Let a sample of size $n_1$be chosen from the first population and sample of size $n_2$be chosen from the second population.

Let X_1be number of persons possessing the attribute A in the first sample andX_2 be the number of persons possessing the same attribute in the second sample

$$p_1 = \frac{X_1}{n_1}; p_2 = \frac{X_2}{n_2}$$

As before $E(p_1) = P_1$and$E(p_2) = P_2$where $P_1\ and \quad P_2$ are the proportions in the populations. $V(p_1) = \frac{P_1Q_1}{n_1}\ and\ V(p_2) = \frac{P_2Q_2}{n_2}$.

$$\therefore Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

An unbiased estimate of population proportion P based on both the samples is given by $P = \frac{(n_1p_1 + n_2p_2)}{n_1 + n_2}$ .Suppose the population proportion$s\ P_1 and P_2$are given to be different (ie)$P_1 \neq P_2$.

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} \sim N(0,1)$$

Problem:

A coin is tossed 144 times and a person gets 80 heads. Can we say that the coin is unbiased one?

Solution:

Set the null hypothesis H_0: the coin is unbiased. Given n=144.

Probability of getting a head in a toss P=1/2.HenceQ=1/2.Let X=number of successes=number of getting heads=80.

$$Z = \frac{80 - 144\left(\frac{1}{2}\right)}{\sqrt{144\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = \frac{80 - 72}{\sqrt{36}} = \frac{8}{6} = 1.33 < 1.96$$

.

Since |Z|<1.96. We accept the hypothesis at 5% level of significance.

Hence the coin is unbiased.

Problem:

A die is thrown 10000 times and a throw of 1 or 2 was obtained 4200 times. On the assumption of random throwing do the data indicate an unbiased die?

Solution:

P=Probability of getting 1 or 2=1/3 .Hence Q=2/3

Given n=10000,X=4200.The null hypothesisH_0:the die is unbiased.

$$\therefore Z = \frac{X - nP}{\sqrt{nPQ}} = \frac{4200 - 10000\left(\frac{1}{3}\right)}{\sqrt{10000(\frac{2}{9})}} = \frac{4200 - 3333.3}{47.14} = \frac{866.7}{47.14} = 18.4$$

Since |z|>3, $H_0$is rejected and hence the die is biased one.

Problem:

A manufacturer claimed that at least 95% of the equipment which he supplied to a factory conformed to specification. An examination of a sample of 200 pieces of

equipment revealed that 18 were faulty. Test his claim at a significant level of (i)5% (ii)1%.

Solution:

Out of a sample of 200 equipments 18 were faulty.

X=200-80=182

$$p = \frac{182}{200} = 0.91$$.

Set the null hypothesis H_0:P=0.95,Q=0.05.$H_1: P < 0.95$.

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{.91 - .95}{\sqrt{\frac{.95 \times .05}{200}}} = -\frac{.04}{.0154} = -2.6$$

(i) Since the alternative hypothesis is left tailed and the significant value of Z at 5% level of significant for left tail is -1.645.

Z=-2.6<-1.645.

Hence we accept the null hypothesis at 5% level of significance.

(ii) The critical value of Z at 1% value of significance for left tailed test is 2.33 and Z=-2.6<-2.33.Hence$H_0$is accepted at 1% level.

Problem:

A sample of 1000 products from a factory are examined and found to be 2.5% defective. Another sample of 1500 similar products from another factory are found to have only 2% defective. Can we conclude that the products of the first factory are inferior to those of the second?

Solution:

Given $n_1 = 1000, n_2 = 1500.$ Proportion of defectives in the first factory $p_1 = {}^{25}/_{1000} = .025 \quad p_2 = {}^{30}/_{1500} = .020$

Proportion of defective in the second factory $p_2 = {}^{30}/_{1500} = .020$

$$\therefore P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{25 + 30}{1000 + 1500} = .022$$

Hence $.022 = .978.$ Q=1-

Null hypothesis $H_0: P_1 = P_2$

Test hypothesis $Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$

$$Z = \frac{.025 - .020}{\sqrt{.022(.978)\left(\frac{1}{1000} + \frac{1}{1500}\right)}} = \frac{.005}{\sqrt{{.022(.978)}/{600}}} = .83 < 1.9$$

The difference of proportion is not significant on 5% level. Hence this hypothesis is accepted and the two factories are producing similar products. Hence one is not inferior to the other.

Problem:

A machine puts out 16 imperfect articles in a sample of 500 articles. After the machine overhauled it. Puts out 3 defective articles in sample of 100.Has the machine improved?

Solution:

Given $n_1 = 500, n_2 = 100. p_1 =$Proportion defectatives in the first sample=16/500=.032

$$p_2 = \frac{3}{100} = .03$$

Set the null hypothesis $H_0: P_1 = P_2$
Alternative hypothesis $H_1: P_1 > P_2$

$$\therefore P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{16+3}{500+100} = \frac{19}{600} = .032\,; Q = 1 - .032 = .968$$

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{.032 - .030}{\sqrt{.032 \times .968\left(\frac{1}{500} + \frac{1}{100}\right)}} = \frac{.002}{\sqrt{.032 \times \left({}^{6}/_{500}\right)}} = \frac{.002}{.019}$$
$$= .105.$$

Sine Z<1.645 it is not significant at 5% level of significance. Hence we can accept the null hypothesis and conclude that the machine has not been improved.

II (B) Test of significance for difference of sample means.

Consider two different normal populations with μ_1 and μ_2 and s.d σ_1 and σ_2 respectively. Let a sample of size n_1 be drawn from the first population and an independent sample of size n_2 be drawn from the second population. Let $\overline{x_1}$ be the mean of the first sample from the first population and $\overline{x_2}$ be the mean of second sample from the second population. If the sample sizes are large we know $\overline{x_1}$ is a normal variate with mean μ_1 and variance $\frac{\sigma_1^2}{n_1}$ and $\overline{x_2}$ is an independent normal variate and normal variate with mean μ_2 and variance $\frac{\sigma_2^2}{n_2}$.

The test statistic becomes $\frac{2}{2}$ $Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma^2}{n}}}$ which can be tested at any level of significance.

Problem:

The number of accidents per day were studied for 144 days in Madras city and for 100 days in Delhi city. The mean numbers of accidents and the s.ds were respectively 4.5 and 1.2 for Madras city and 5.4 and 1.5 for Delhi city. Is Madras city more prone to accidents than Delhi city?

Solution:

Given $n_1 = 144\,; \bar{x}_1 = 4.5; \overline{x_2} = 5.4$.
$n_2 = 100; \sigma_1 = 1.2; \sigma_2 = 1.5$.

Set the null hypothesis $H_0: \mu_1 = \mu_2$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}} = \frac{4.5 - 5.4}{\left({1.2^2}/_{144}\right) + \left({1.5^2}/_{100}\right)} = -4.99$$

.

$\therefore |Z| = 4.99 > 3$ we reject the hypothesis that the two cities have the same accident rates. However since Delhi city has higher rate of accident than Madras city. Therefore Delhi more prone to accidents.

Problem:

The mean yields of rice from two places in a district were 210 kgs and 220 kgs per acre from 100 acres and 150 acres respectively. Can it be regarded that the sample were drawn from the same district which has the s.d of 11kgs per acre?

Solution:

$$n_1 = 100; \bar{x}_1 = 210;\ \sigma = 11$$

$$n_2 = 150; \bar{x}_2 = 220$$

Set the null hypothesis $H_0: \mu_1 = \mu_2$

$$\therefore Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\left({1}/_{n_1}\right) + \left({1}/_{n_2}\right)}}$$

$$Z = \frac{210 - 220}{11\sqrt{\left({1}/_{100}\right) + \left({1}/_{150}\right)}} = \frac{-10}{11\sqrt{{250}/_{15000}}} = -7.04$$

$|Z| = 7.04 > 3$.The value is highly significant and hence we reject the null hypothesis. Hence the samples are certainly not from the same district with the s.d 11.

Test of significance for equality of standard deviations of a normal population.

If we want to test whether the two independent samples with known standard deviations s_1 and s_2 have come from the same population with standard deviation σ. Under the hypothesis $H_0: \sigma_1 = \sigma_2$ the test statistics is $Z = \frac{s_1 - s_2}{\sigma\sqrt{\left({1}/_{2n_1}\right) + \left({1}/_{2n_2}\right)}}$.

Problem:

The s.d of weight of all students in a first grade college was found to be 4 kgs. Two samples are drawn. The s.ds of the weight of 100 undergraduate students is 3.5kgs and 50 post graduate students are 3 kgs. Test the significance of the difference of standard deviations of the samples at 5% level.

Solution:

Given $n_1 = 100; s_1 = 3.5; \sigma = 4; n_2 = 50; s_2 = 3$.

Set the null hypothesis $H_0: \sigma_1 = \sigma_2$.Then$H_1: \sigma_1 \neq \sigma_2$

$$\therefore Z = \frac{s_1 - s_2}{\sigma\sqrt{\left(1/_{2n_1}\right) + \left(1/_{2n_2}\right)}} = \frac{3.5 - 3}{4\sqrt{(1/_{200}) + (1/_{100})}} = 1.02$$

$|Z|$=1.02<1.96. It is not significant at 5% level of significance.

Problem:

The mean production of wheat of a sample of 100 plots is 200kgs per acre with s.d of 10 kgs. Another sample of 150 plots gives the mean production of wheat as 220kgs. With s.d of 12kgs. Assuming the s.d of the 11kgs for the universe find at 1% level of significance ,whether two results are consistent.

Solution:

Given σ=11 and

	size	mean	S.D
Sample 1	$n_1 = 100$	$\bar{x}_1 = 200$	$s_1 = 10$
Sample 2	$n_2 = 150$	$\bar{x}_2 = 220$	$s_2 = 12$

Set the null the hypothesis $H_0: \mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$ .For $H_0: \mu_1 = \mu_2$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{(1/_{n_1}) + (1/_{n_2})}} = \frac{200 - 220}{11\sqrt{(1/_{100}) + \left(1/_{150}\right)}} = \frac{-20}{11\sqrt{10/_{600}}} = \frac{-155}{11} = -14.1$$

$\therefore |Z|$=14.1>3.Hence the two means differ significantly at 5% level even a 1% level.

For $H_0: \sigma_1 = \sigma_2$.

$$Z = \frac{s_1 - s_2}{\sigma\sqrt{\left(1/_{2n_1}\right) + \left(1/_{2n_2}\right)}} = \frac{10-12}{11\sqrt{(1/_{200}) + (1/_{300})}} = -1.99$$

$\therefore |Z| = 1.99 > 1.96$ and $|Z| = 1.99 < 2.58$.

Hence the difference of s.d is significant at 5% level and not significant and 1% level.

∴At 1% level the difference between s.d is not significant but between means it is significant. Hence we can conclude that at 1% level the two results are not consistent.

TEST OF SIGNIFICANCE (SMALL SAMPLES)

TEST OF SIGNIFICANCE BASED ON t-DISTRIBUTION (t-test)

Consider a normal population with mean μ and s.d σ . Let $x_1, x_2, \dots x_n$. be a random sample of size n with mean$\bar{x}$ and standard deviation s. We know that $Z = \frac{x-\mu}{\varsigma \quad \overline{n}}$is the standard normal variate N(0,1).

Hence the test statistics is in small sample becomes

$Z = \frac{\bar{x}-\mu}{\left(s\sqrt{n/_{n-1}}\right)/\sqrt{n}} = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}}$. Now let us define $t = \frac{\bar{x}-\mu}{s/_{\sqrt{n-1}}}$.

This follows students‟s t-distribution with n-1 degrees of freedom

1.Test for the difference between the mean of a sample and that of a population

Under the null hypothesis $H_0 \colon \mu = \bar{x}$.

The test statistic

$t = \frac{\bar{x}-\mu}{s/_{\sqrt{n-1}}} \sim t_{n-1}$. Which can be tested at any level of significance with n-1 degrees of freedom.

II. Test for the difference between the means of two samples

II.A. If$\bar{x}_1$and $\bar{x}_2$are the means of two independent samples of sizes $n_1$and $n_2$from a normal population with mean μ and standard deviation σ. It found that $\sigma \quad \frac{\bar{x}_1-\bar{x}_2}{\sqrt{(1/_{n_1})+(1/_{n_2})}} \sim N(0,1)$.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Which follows a t-distribution with d.f $v = (n_1 + n_2 - 2)$.

II.B. suppose the sample sizes are equal (ie)$n_1 = n_2 = n$.Then we have n pairs of values. Further we assume that the $n$pair are independent .Then the test statistic t in (1) becomes

$t = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{n\left(s_1^2+s_2^2\right)}{2n-2}\left(\frac{2}{n}\right)}}$.

$$\therefore t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{{(s_1^2+s_2^2)}/{(n-1)}}}$$ is a students t variate with

$v = n + n - 2 = 2n - 2.$

II. (C) suppose the sample size are equal and if then n pairs of values in this case are not independent.

The test statistic $t = \frac{x-\mu}{\frac{s}{\overline{n-1}}}$ to test whether the means of differences is significantly different from zero. In this case the d.f is n-1.

Confidence limits (Fiducial limits). Ifσ is not known and $n$is small then

1. 95% confidence limits for μ is $\left(\bar{x} - \frac{st_{.05}}{\sqrt{n-1}}, \bar{x} + \frac{st_{.05}}{\sqrt{n-1}}\right)$
2. 99% confidence limits for μ is$\left(\bar{x} - \frac{st_{.01}}{\sqrt{n-1}}, \bar{x} + \frac{st_{.01}}{\sqrt{n-1}}\right)$

Problem:

A random sample of 10 boys has the following I.Q (intelligent quotients). 70, 120, 110, 101, 88, 95, 98, 107, 100. Do these data support the assumption of a population mean of a population mean I.Q of 100?

Solution:

Given n=10; μ=100 . Set H_0 : μ=100

Under H_0 test statistics $\frac{\bar{x}-\mu}{{s}/{\sqrt{n-1}}} \sim t_{(n-1)}$ where$\bar{x}$ and s can be calculated from the sample data as $\bar{x} = {972}/{10} = 97.2$ and

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{1833.60}{10} = 183.36.$$

Hence s =13.54.

$$\therefore t = \frac{97.2 - 100}{{13.54}/{9}} = \frac{-2.8 \times 3}{13.54} = -.6204$$

$$\therefore |t| = .62\text{(nearly)}.$$

The table value for 9 d.f at 5% level of significance is $t_{.05}$=2.26

$\therefore$ $|t|= .62 < t_{.05}$.Hence the difference is not significant at 5% level. Hence H_0 may be accepted at 5% level hence the data support the assumption of population mean 100.

Problem:

It was found that a machine has produced pipes having a thickness .05 mm. to determine whether the machine is in proper working order a sample of 10 pipe is chosen for which the mean thickness is .53mm and s.d is 0.3mm .test the hypothesis that the machine is in proper working order using a level of significance of (1) .05 (2) .01

Solution :

Given μ= .50,$\bar{x}$=.53;$s = .03; n = 10$.

Set the null hypothesis H_0 :μ=50

Under the null hypothesis the test statistic is $t = \frac{\bar{x}-\mu}{s/\sqrt{n-1}} = \frac{.53-.50}{.03} \times \sqrt{9}$

$$= \frac{.03 \times 3}{.03} = 3.$$

(i)The table value for $v = 9$d.f at 5% level of significance is $t_{.05}$=2.26 (ie)|t|=3> $t_{.05}$.

∴The difference is significant at 5% level of significance.

∴The null hypothesis is rejected at 5%level of significance .

(ii) The table value for $v = 9$ d.f at 1% level of significance is$t_{.01} = 3.25$.

Hence |t|=3<$t_{.01}$.

∴The difference is not significant at 1% level of significant .

Hence the null hypothesis is accepted at 1% level of significance.

Problem:

A group of 10 rats fed on a diet A and another group of 8 rats fed on a different diet B recorded the following increase in weight in gms.

Diet A	5	6	8	1	12	4	3	9	6	10
Diet B	2	3	6	8	1	10	2	8	-	-

Test whether diet A is superior to diet B .

Solution :

Given $n_1 = 10; n_2 = 8$.

Mean of the first sample $\bar{x}_1 = \frac{5+6+\cdots+10}{10} = \frac{64}{10} = 6.4$.

Mean of the second sample $\overline{x}_2 = \frac{2+3+\cdots+8}{8} = \frac{40}{8} = 5.0$.

Standard deviation$s_1$ands_2 of the first and second sample can be found as

$$s_1^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = 10.24 \text{ and } \ s_2^2 = 10.25$$

Set the null hypothesis $H_0 : \mu_1 = \mu_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{6.4 - 5}{\sqrt{\frac{10 \times 10.24 + 8 \times 10.25}{10 + 8 - 2}\left(\frac{1}{10} + \frac{1}{8}\right)}}$$

$$= \frac{1.4}{\sqrt{11.525(.1 + .125)}} = .875.$$

Table value for t at 5% level of significance for $(n_1 + n_2 - 2) = 16\ d.f\ is\, t_{.05}$ =2.12.

Since t=.875 <$t_{.05}$ the difference is not significant at 5% level of significance .

Hence the null hypothesis may be accepted.

Problem:

The table gives the biological values of protein from 6 cows milk and 6 buffalo‟s milk . Examine whether the differences are significant .

Cow‟s milk	Buffalo‟s milk
1.8	2.0
2.0	1.8
1.9	1.8
1.6	2.0
1.8	2.1
1.5	1.9

Solution:

Mean value of protein of cow‟s milk =1.6

Mean value of protein of buffalo‟s milk =1.9

Variance of protein of cow‟s milk =.03

Variance of protein in buffalo‟s milk=0.1

We notice that the two sets of observations are independent .

Given $n_1 = n_2 = 6; \bar{x}_1 = 1.8; \bar{x}_2 = 1.9; s_1^2 = .03;$

$$s_2^2 = .01.$$

Set null hypothesis $H_0: \bar{x}_1 = \bar{x}_2$. Under this null hypothesis the test statistic is $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n-1}}}$ and the d.f $v = 2n - 2 = 10.$

$$= \frac{-.1}{\sqrt{(.03 + .01)/_5}} = \frac{-.1}{\sqrt{.04/_5}} = -1.11$$

.

The table value for $v = 10$d.f at 5% level of significance

is 2.23.

|t|=1.11<2.23.Hence the difference is not significant .

Hence the hypothesis is accepted .

Problem:

Ten soldiers participated in a shooting competition in the first week. After intensive training they participated in the competition in the second week. Their scores before and after coaching were given as follows.

Soldiers	1	2	3	4	5	6	7	8	9	10
Score before(x)	67	24	57	55	63	54	56	68	33	43
Score after(y)	70	38	58	58	56	67	68	75	42	38

Do the data indicate that the soldier have been identified by the training ?

Solution:

Here we are connected with the same set of the soldiers in the 2 competitions and their scores which are related to each other because of the intensive training .we compute the difference in their scores $z = y - x$and calculate the mean $\bar{z}$ and the s.d z as follow

x	y	$z = y - x$	$z - \bar{z}$	$(z - \bar{z})^2$

67	70	3	-2	4
24	38	14	9	81
57	58	1	-4	16
55	58	3	-2	4
63	56	-7	-12	144
54	67	13	8	64
56	68	12	7	49
68	75	7	2	4
33	42	9	4	16
43	38	-5	-10	100
-	-	50	-	482

$z = \frac{50}{10} = 5; s2 = \frac{\qquad}{10} \quad \frac{\quad}{10} z - z \; 2 = 482 = 48.2$

Set the null hypothesis $H_0 \colon z$=0._

Under the null hypothesis the test statistic is $t = \frac{\bar{z}-0}{s/\sqrt{n-1}} = \frac{5}{\sqrt{48.2}} \times \sqrt{9} = \frac{15}{6.94} = 2.16$

The table value for $v = 9$d.f at 55 level of significance is $t_{.05} = 2.26$.

$\therefore |t| = 2.16 < t_{.05}$.

The difference is not significant on 5% level of significance .

Hence the null hypothesis is accepted .We can conclude that there is no significant improvement in the training .

TEST BASED ON χ^2- DISTRIBUTION

INTRODUCTION:

The χ^2 distribution has number of application in statistics. It has three important applications based on χ^2 distribution. I. χ^2- test for population variance.

II. χ^2-test to test the goodness of fit.

III.χ^2-test to test the independence of attributes.

χ^2-TEST.

$I.\chi^2$-test for population variance

Let $x_1, x_2, \dots x_n$ be a random sample from a normal population with variance σ^2.Set the null hypothesis $H_0: \sigma^2 = \sigma_0^2$. Then the test statistic is $\chi^2 = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_0}\right)^2 = \frac{ns^2}{\sigma_0^2}$ where s^2 is the variance of the sample. Then $\chi^2 = \frac{ns^2}{\sigma_0^2}$ defined above follows aχ^2 distribution with $n-1$ degrees of freedom.

Problem:

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

Solution:

Given σ=10,n=25,s=8.5 $H_0: \sigma = 10$

$$\chi^2 = \frac{ns^2}{\sigma_0^2} = \frac{25 \times 8.5^2}{100} = 18.06$$

.

The table value ofχ^2 for 24 d.f =36.415 at 5% level of significance.

It is not a significant. Hence the null hypothesis is accepted.

Problem:

Test the hypothesis that σ=8 given that s=10 for a random sample of size 51.

Solution:

Given n_1=51,σ=8,s=10.

Let $H_0: \sigma = 8$.

$$\chi^2 = \frac{ns^2}{\sigma_0^2} = \frac{51 \times 10^2}{8^2} = 79.7$$

.

Since$Z = \sqrt{2\chi^2}$-$\sqrt{2n-1}$=$\sqrt{2 \times 79.7}$-$\sqrt{2 \times 51 - 1}$

=2.58

$$|z| = 2.58 > 1.96.$$

Hence the difference is significant at 5% level of significance and hence the hypothesis is rejected at 5% level of significance.

$II.\chi^{2}$ -TEST TO TEST THE GOODNESS OF FIT

The χ^2 -distribution can be used to test the goodness of fit. This test can also be applied to test for compatibility of observed frequencies and theoretical frequencies. Let $o_1, o_2, \dots o_n$ be the observed frequencies and $e_1, e_2, \dots e_n$ be the corresponding expected frequencies such that $\sum_{i=1}^{n} o_i = N = \sum_{i=1}^{n} e_i$ where N is the number of members in the population.

Define $\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$ .It is a χ^2 variable with n-1 degrees of freedom.

Problem:

The theory predicts that the proportion of an object available in four groups A,B,C,D should be 9:3:3:1. In an experiment among 1600 items of this object the members in the four groups were 882,313,287 and188.use χ^2-test to verify whether the experimental result support the theory.

Solution:

The observed frequencies o_i are 882,313,287,118.

$\sum o_i$=882+313+287+118=1600

The expected frequencies are in the ratio 9:3:3:1.

$\therefore$ The expected frequencies e_i are 900,300,300,100.

$\sum e_i$=1600=$\sum o_i$.

$$\therefore \chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}.$$

$$= \frac{(882-900)^2}{900} + \frac{(313-300)^2}{300} + \frac{(287-300)^2}{300} + \frac{(118-100)^2}{100} = 4.7266$$

Degrees of Table value of χ^2 for 3 d.f at 5% level of significance is 7.851.

Calculated χ^2=4.7266<7.852=table value ofχ^2.It is not significant. Hence the null hypothesis may be accepted at 5% level of significance and hence we may conclude that experiment results support the theory.

Problem:

Fit a poisson distribution for the following data and test the goodness of fit.

x	0	1	2	3	4	5	6	Total
f	273	70	30	7	7	2	1	390

Solution:

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{70+60+21+28+10+6}{273+70+30+7+7+2+1} = \frac{195}{390}$$

$\lambda = {}^{1}/_{2}.$

The theoretical frequencies of the poisson distribution are given by

$$f(x) = \frac{Ne^{-\lambda}\lambda^x}{x!} = \frac{390}{\sqrt{e}x!}\left(\frac{1}{2}\right)^x ;x = 0,1,2,\ldots 6$$.

freedom =4-1 =3 .

The expected frequencies are by$f(0) = \frac{390}{\sqrt{e}}\left(\frac{1}{2}\right)^0 = 236.4;$

$f(1) = \frac{390}{\sqrt{e}1!}\left(\frac{1}{2}\right)^1 = 118.2$ ………………….,

$$f(6) = \frac{390}{\sqrt{e}6!}\left(\frac{1}{2}\right)^6 = 0.005$$.

Thus the observed and expected frequencies can be shown below

o_i	273	70	30	7	7	2	1	390
e_i	236.4	118.2	29.5	4.9	.6	.1	0	389.7

Since the sum of the expected frequencies is 389.7.It can be adjusted in the last two frequencies by adding .3.

o_i	273	70	30	17	390

e_i	236.4	118.2	29.5	5.9	390

Set up the null hypothesis H_0:Poisson distribution can be fitted well.

The test statistics is $\chi^2 = \frac{\sum(o_i - e_i)^2}{e_i}$.

$$= \frac{(273-236.4)^2}{236.4} + \frac{(70-118.2)^2}{118.2} + \frac{(30-29.5)^2}{29.5} + \frac{(17-5.9)^2}{5.9} = 46.3.$$

Degrees of freedom=7-1-1-3=2.

The table value of 2 d.f. at 5% level is 5.99.

Since χ^2=46.3>5.99=The table value of $\chi^2_{.05}$ it is much significant at 5% level of significance.

Hence the hypothesis is rejected at 5% level and hence the poisson distribution is not a good fit to the data.

UNIT IV

ANALYSIS OF VARIANCEWhat Does the Analysis of Variance Reveal?

The ANOVA test is the initial step in analyzing factors that affect a given data set. Once the test is finished, an analyst performs additional testing on the methodical factors that measurably contribute to the data set's inconsistency. The analyst utilizes the ANOVA test results in an f-test to generate additional data that aligns with the proposed regression models.

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

If no real difference exists between the tested groups, which is called the null hypothesis, the result of the ANOVA's F-ratio statistic will be close to 1. The distribution of all possible values of the F statistic is the F-distribution. This is actually a group of distribution functions, with two characteristic numbers, called the numerator degrees of freedom and the denominator degrees of freedom.

- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.
- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

Example of How to Use ANOVA

A researcher might, for example, test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.

The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. Analysis of variance is employed if there is no access to statistical software resulting in computing ANOVA by hand. It is simple to use and best suited for small samples. With many experimental designs, the sample sizes have to be the same for the various factor level combinations.

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups and within groups.

One-Way ANOVA Versus Two-Way ANOVA

There are two main types of ANOVA: one-way (or unidirectional) and two-way. There also variations of ANOVA.

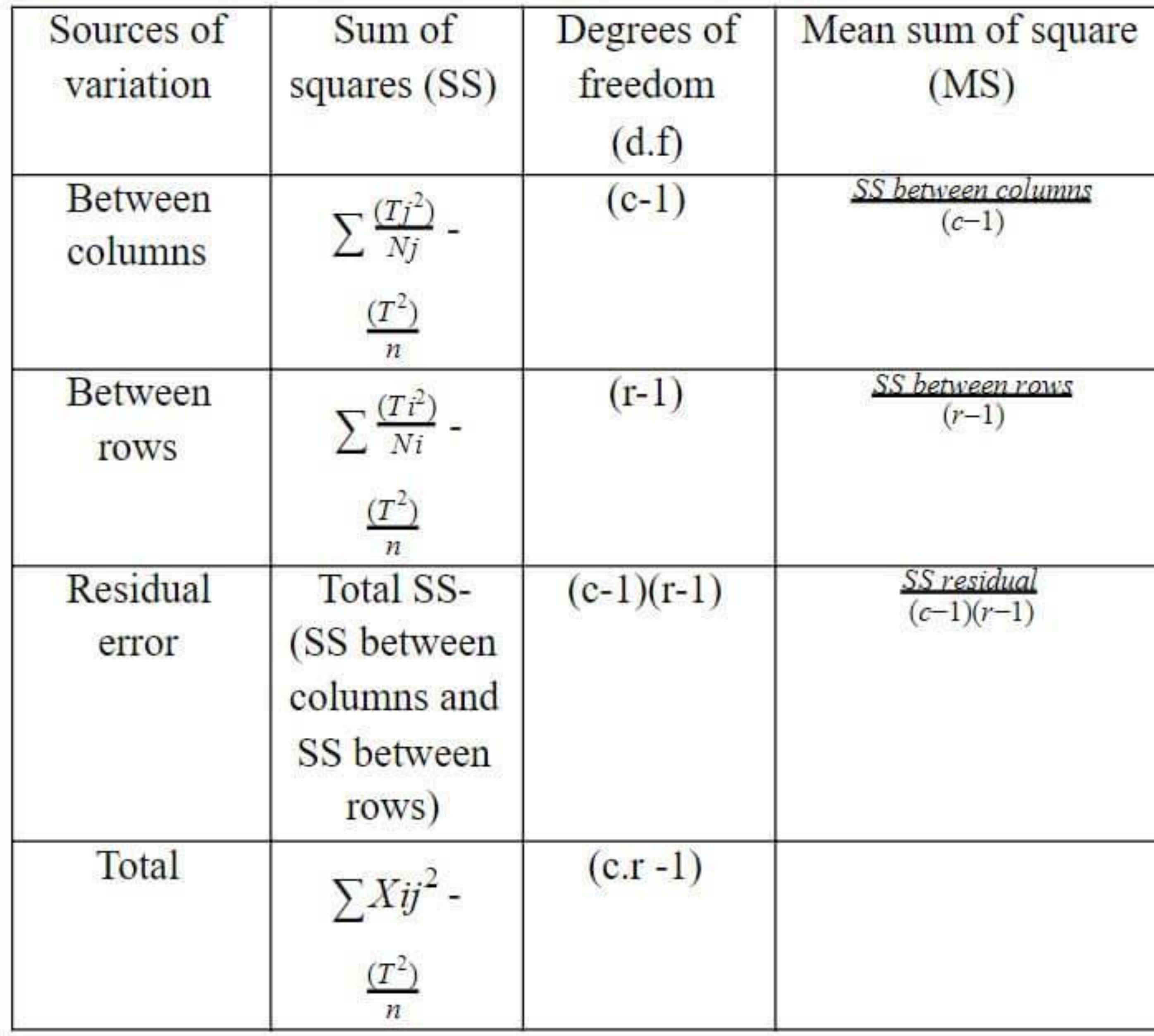
For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time. One-way or two-way refers to the number of independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.

ANOVA Table

- ANOVA table is a tabular form of all the data and calculations performed during the test.
- This makes it more convenient for the observation and calculation of data.

The following is the ANOVA table for two-way ANOVA:



Sources of variation	Sum of squares (SS)	Degrees of freedom (d.f)	Mean sum of square (MS)
Between columns	$\sum \frac{(Tj^2)}{Nj} - \frac{(T^2)}{n}$	(c-1)	$\frac{SS\ between\ columns}{(c-1)}$
Between rows	$\sum \frac{(Ti^2)}{Ni} - \frac{(T^2)}{n}$	(r-1)	$\frac{SS\ between\ rows}{(r-1)}$
Residual error	Total SS-(SS between columns and SS between rows)	(c-1)(r-1)	$\frac{SS\ residual}{(c-1)(r-1)}$
Total	$\sum Xij^2 - \frac{(T^2)}{n}$	(c.r -1)	

In the table,
c= number of columns
r= number of rows
T= the total of the values of individual items

Tj= the sum of the values in the column
Ti= the sum of the values in the row

- The ANOVA table shows the statistics used to test hypotheses about the population means.
- Here, the F-ratios for rows and columns are compared with their corresponding table values, for the given degree of freedom and given level of significance.
- If the calculated F-ratio is found to be equal or higher than its table value, the differences or variation among the columns are considered significant.
- A similar process is employed for rows to determine the significance of the variation.

Interval Estimation (Confidence Intervals)

Let X_1X1, X_2X2, X_3X3, ..., X_nXn be a random sample from a distribution with a parameter θθ that is to be estimated. Suppose that we have observed $X_1=x_1$X1=x1, $X_2=x_2$X2=x2, $\cdots$⋯, $X_n=x_n$Xn=xn. So far, we have discussed point estimation for θθ. The point estimate $\hat{\theta}$θ^ alone does not give much information about θθ. In particular, without additional information, we do not know how close $\hat{\theta}$θ^ is to the real θθ. Here, we will introduce the concept of **interval estimation**. In this approach, instead of giving just one value $\hat{\theta}$θ^ as the estimate for θθ, we will produce an interval that is likely to include the true value of θθ. Thus, instead of saying

$\hat{\theta}=34.25,$θ^=34.25,

we might report the interval

$[\hat{\theta}_l,\hat{\theta}_h]=[30.69,37.81],$[θ^l,θ^h]=[30.69,37.81],

which we hope includes the real value of θθ. That is, we produce two estimates for θθ, a *high estimate* $\hat{\theta}_h$θ^h and a low estimate $\hat{\theta}_l$θ^l. In interval estimation, there are two important concepts. One is the **length** of the reported interval, $\hat{\theta}_h-\hat{\theta}_l$θ^h−θ^l. The length of the interval shows the precision with which we can estimate θθ. The smaller the interval, the higher the precision with which we can estimate θθ. The second important factor is the **confidence level** that shows how confident we are about the interval. The confidence level is the probability that the interval that we construct includes the real value of θθ. Therefore, high confidence levels are desirable. We will discuss these concepts in this section.

Confidence Interval

In Statistics, a **confidence interval** is a kind of interval calculation, obtained from the observed data that holds the actual value of the unknown parameter. It is associated with the confidence level that quantifies the confidence level in which the interval estimates the deterministic parameter. Also, we can say, it is based on Standard Normal Distribution, where Z value is the z-score. Here, let us look at the definition, formula, table, and the calculation of the confidence level in detail.

Confidence Interval Definition

The confidence level represents the proportion (frequency) of acceptable confidence intervals that contain the true value of the unknown parameter. In other terms, the confidence intervals are evaluated using the given confidence level from an endless number of independent samples. So that the proportion of the range contains the true value of the parameter that will be equal to the confidence level.

Mostly, the confidence level is selected before examining the data. The commonly used confidence level is 95% confidence level. However, other confidence levels are also used, such as 90% and 99% confidence levels.

Confidence Interval Formula

The confidence interval is based on the mean and standard deviation. Thus, the formula to find CI is

$\bar{X} \pm Z\alpha/2 \times [\ \sigma / \sqrt{n}\ ]$

Where

$\bar{X}$ = Mean

Z = Confidence coefficient

α = Confidence level

σ = Standard deviation

N = sample space

The value after the ± symbol is known as the margin of error.

Note: This interval is only accurate when the population distribution is normal. But, in the case of large samples from other population distributions, the interval is almost accurate by the Central Limit Theorem.

Confidence Interval Table

The confidence interval table for Z values are given as follows

Confidence Interval	**Z Value**
80%	1.282
85%	1.440
90%	1.645
95%	1.960

99%	2.576
99.5%	2.807
99.9%	3.291

How to Calculate Confidence Interval?

To calculate the confidence interval, go through the following procedure.

Step 1: Find the number of observations n(sample space), mean $\bar{X}$, and the standard deviation σ.

Step 2: Decide the confidence interval of your choice. It should be either 95% or 99%. Then find the Z value for the corresponding confidence interval given

Step 3: Finally, substitute all the values in the formula.

Confidence Interval Example

Question: In a tree, there are hundreds of apples. You are randomly choosing 46 apples with a mean of 86 and a standard deviation of 6.2. Determine that the apples are big enough.

Solution:

Given: Mean, $\bar{X} = 86$

Standard deviation, $\sigma = 6.2$

Number of observations, $n = 46$

Take the confidence level as 95%. Therefore, the value of $z = 1.960$ (from the table)

The formula to find the confidence interval is

$\bar{X} \pm Z\alpha/2 \times [\sigma / \sqrt{n}]$

Now, substitute the values in the formula, we get

$86 \pm 1.960 \times [6.2 / \sqrt{46}]$

$86 \pm 1.960 \times [6.2 / 6.78]$

$86 \pm 1.960 \times 0.914$

86 ± 1.79

Here, the margin of error is 1.79

Therefore, all the hundreds of apples are likely to be between in the range of 84. 21 and 87.79.

Independent Samples

Depending on the sample types and whether or not the population standard deviation is **known** will depend on whether we employ either a *z-test* or *t-test*.

Suppose the population standard deviation is known, which is highly unlikely.
In that case, we will use a z-test and follow the following formula for constructing a confidence interval for the difference of means.

$$\left(\bar{x}_1-\bar{x}_2\right)\pm z^{*}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\quad,\quad z^{*}=z_{\left(1-\frac{\alpha}{2}\right)}$$

Confidence Interval Formula For Two Sample Mean

But for two independent random samples where the standard deviation is **unknown**, and the sample size is sufficiently large, then we will have to use a t-test, which involves a t-distribution with degrees of freedom, as well as the possibility of pooled variances.

$$\left(\bar{x}_1-\bar{x}_2\right)\pm t^{*}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\quad,\quad t^{*}=t_{\left(\frac{\alpha}{2},df\right)}\quad,\quad df=n_1+n_2-2$$

Formula For Two Sample Mean With Unknown Standard Deviation

$$\left(\bar{x}_1-\bar{x}_2\right)\pm t^*\left(s\right.$$

$$s_p=\sqrt{\frac{\left(n_1-1\right)s_1^2}{n_1+}}$$

$$t^*=t_{\left(\frac{\alpha}{2},df\right)}\ ,\ df$$

UNIT V

TESTING OF HYPOTHESIS

Hypothesis Testing Formula

We run a hypothesis test that helps statisticians determine if the evidence are enough in a sample data to conclude that a research condition is true or false for the entire population. For finding out hypothesis of a given sample, we conduct a Z-test. Usually, in Hypothesis testing, we compare two sets by comparing against a synthetic data set and idealized model.

The Z test formula is given as:

z=x¯¯¯−μσn−−√

Where,
x¯¯¯ is the sample mean
μ is the population mean
σ is the standard deviation and *n* is the sample size.

Solved Examples

Question: What will be the z value when the given parameters are sample mean = 600, population mean = 585, the standard deviation is 100 and the sample size is 150?

Solution:

Given parameters are,
Sample mean, x¯ = 600
Population mean, μ = 585,
Standard deviation, σ = 100
Sample size, n = 150
The formula for hypothesis testing is given as,

z=x¯¯¯−μσn√
z=600−585100150√

=1.837

Hypothesis Definition

In Statistics, the determination of the variation between the group of data due to true variation is done by hypothesis testing. The sample data are taken from the population parameter based on the assumptions. The hypothesis can be classified into various types. In this article, let us discuss the hypothesis definition, various types of hypothesis and the significance of hypothesis testing, which are explained in detail.

Hypothesis Definition in Statistics

In Statistics, a hypothesis is defined as a formal statement, which gives the explanation about the relationship between the two or more variables of the specified population. It helps the researcher to translate the given problem to a clear explanation for the outcome of the study. It clearly explains and predicts the expected outcome. It indicates the types of experimental design and directs the study of the research process.

Types of Hypothesis

The hypothesis can be broadly classified into different types. They are:

Simple Hypothesis

A simple hypothesis is a hypothesis that there exists a relationship between two variables. One is called a dependent variable, and the other is called an independent variable.

Complex Hypothesis

A complex hypothesis is used when there is a relationship between the existing variables. In this hypothesis, the dependent and independent variables are more than two.

Null Hypothesis

In the null hypothesis, there is no significant difference between the populations specified in the experiments, due to any experimental or sampling error. The null hypothesis is denoted by H_0.

Alternative Hypothesis

In an alternative hypothesis, the simple observations are easily influenced by some random cause. It is denoted by the H_a or H_1.

Empirical Hypothesis

An empirical hypothesis is formed by the experiments and based on the evidence.

Statistical Hypothesis

In a statistical hypothesis, the statement should be logical or illogical, and the hypothesis is verified statistically.

Apart from these types of hypothesis, some other hypotheses are directional and non-directional hypothesis, associated hypothesis, casual hypothesis.

Characteristics of Hypothesis

The important characteristics of the hypothesis are:

- The hypothesis should be short and precise
- It should be specific
- A hypothesis must be related to the existing body of knowledge
- It should be capable of verification

Type I and Type II errors are subjected to the result of the null hypothesis. In case of type I or type-1 error, the null hypothesis is rejected though it is true whereas type II or type-2 error, the null hypothesis is not rejected even when the alternative hypothesis is true. Both the error type-i and type-ii are also known as "**false negative**". A lot of statistical theory rotates around the reduction of one or both of these errors, still, the total elimination of both is explained as a statistical impossibility.

Type I Error

A type I error appears when the null hypothesis (H_0) of an experiment is true, but still, it is rejected. It is stating something which is not present or a false hit. A type I error is often called a false positive (an event that shows that a given condition is present when it is absent). In words of community tales, a person may see the bear when there is none (raising a false alarm) where the null hypothesis (H_0) contains the statement: "There is no bear".

The type I error significance level or rate level is the probability of refusing the null hypothesis given that it is true. It is represented by Greek letter α (alpha) and is also known as alpha level. Usually, the significance level or the probability of type i error is set to 0.05 (5%), assuming that it is satisfactory to have a 5% probability of inaccurately rejecting the null hypothesis.

Type II Error

A type II error appears when the null hypothesis is false but mistakenly fails to be refused. It is losing to state what is present and a miss. A type II error is also known as false negative (where a real hit was rejected by the test and is observed as a miss), in an experiment checking for a condition with a final outcome of true or false.

A type II error is assigned when a true alternative hypothesis is not acknowledged. In other words, an examiner may miss discovering the bear when in fact a bear is present (hence fails in raising the alarm). Again, H0, the null hypothesis, consists of the statement that, "There is no bear", wherein, if a wolf is indeed present, is a type II error on the part of the investigator. Here, the bear either exists or does not exist within given circumstances, the question arises here is if it is correctly identified or not, either missing detecting it when it is present, or identifying it when it is not present.

The rate level of the type II error is represented by the Greek letter β (beta) and linked to the power of a test (which equals 1−β).

Table of Type I and Type II Error

The relationship between truth or false of the null hypothesis and outcomes or result of the test is given in the tabular form:

Error Types	**When H_0 is True**	**When H_0 is False**
Don't Reject	Correct Decision (True negative) Probability = 1 – α	Type II Error (False negative) Probability = β
Reject	Type II Error (False Positive) Probability = α	Correct Decision (True Positive) Probability = 1 – β

Type I and Type II Errors Example

Check out some real-life examples to understand the type-i and type-ii error in the null hypothesis.

Example 1: Let us consider a null hypothesis – A man is not guilty of a crime.

Then in this case:

Type I error (False Positive)	Type II error (False Negative)
He is condemned to crime, though he is not guilty or committed the crime.	He is condemned not guilty when the court actually does commit the crime by letting the guilty one go free.

Example 2: Null hypothesis- A patient's signs after treatment A, are the same from a placebo.

Type I error (False Positive)	Type II error (False Negative)

Treatment A is more efficient than the placebo	Treatment A is more powerful than placebo even though it truly is more efficient.

POWER OF THE TEST

The probability of correctly rejecting H_0 when it is false is known as *the power of the test.* The larger it is, the better. Suppose you want to calculate the power of a hypothesis test on a population mean when the standard deviation is known. Before calculating the power of a test, you need the following:

- The previously claimed value of

  μ

  in the null hypothesis,

  $\mathrm{H}_0 : \mu = \mu_0$

- The one-sided inequality of the alternative hypothesis (either < or >), for example,

  $\mathrm{H}_\partial : \mu > \mu_0$

- The mean of the observed values

  (denoted $\overline{X}$)

- The population standard deviation

  (denoted σ)

- The sample size (denoted *n*)
- The level of significance

A Likelihood Ratio Test

The likelihood ratio test is a test of the sufficiency of a smaller model versus a more complex model. The null hypothesis of the test states that the smaller model provides as good a fit for the data as the larger model. If the null hypothesis is rejected, then the alternative, larger model provides a significant improvement over the smaller model.

To use the likelihood ratio test, the null hypothesis model must be a model nested within, that is, a special case of, the alternative hypothesis model.

For example, the scaled identity structure is a special case of the compound symmetry structure, and compound symmetry is a special case of the unstructured matrix. However, the autoregressive and compound symmetry structures are not special cases of each other.

The likelihood ratio test can be used to test repeated effect or random effect covariance structures, or both at the same time.

For example, it is possible to test a model that has an identity structure for a random effect and an autoregressive structure for the repeated effect, versus a model that has a compound symmetry structure for the random effect and an unstructured matrix for the repeated effect. Simply make sure that the covariance structure for each effect in one model is nested within the covariance structures for the effects in the other model.

e Neyman Pearson Lemma is all well and good for deriving the best hypothesis tests for testing a simple null hypothesis against a simple alternative hypothesis, but the reality is that we typically are interested in testing a simple null hypothesis, such as $H0:\mu=10$ against a composite alternative hypothesis, such as $HA:\mu>10$. The good news is that we can extend the Neyman Pearson Lemma to account for composite alternative hypotheses, providing we take into account each simple alternative specified in H_A. Doing so creates what is called a **uniformly most powerful** (or **UMP**) **test**.

Uniformly Most Powerful (UMP) test

> A test defined by a critical region *C* of size α is a **uniformly most powerful (UMP) test** if it is a most powerful test against each simple alternative in the alternative hypothesis HA. The critical region *C* is called a **uniformly most powerful critical region of size α**.

Let's demonstrate by returning to the normal example from the previous page, but this time specifying a composite alternative hypothesis.

Example

Suppose X1,X2,:,Xn is a random sample from a normal population with mean μ and variance 16. Find the test with the best critical region, that is, find the uniformly most powerful test, with a sample size of $n=16$ and a significance level $\alpha = 0.05$ to test the simple null hypothesis $H0:\mu=10$ against the composite alternative hypothesis $HA:\mu>10$.

Answer

For each simple alternative in HA,μ=μa, say, the ratio of the likelihood functions is:

L(10)L(μα)=(32π)−16/2exp[−(1/32)∑i=116(xi−10)2](32π)−16/2exp[−(1/32)∑i=116(xi−μα)2]≤k

Simplifying, we get:

exp[−(132)(∑i=116(xi−10)2−∑i=116(xi−μα)2)]≤k

And, simplifying yet more, we get:

$$\exp\left[-\left(\frac{1}{32}\right)\left(\cancel{\sum x_i^2}-2(10)\sum x_i+16\left(10^2\right)-\cancel{\sum x_i^2}+2\mu_a\sum x_i-16\mu_a^2\right)\right]\le k$$

Taking the natural logarithm of both sides of the inequality, collecting like terms, and multiplying through by 32, we get:

−2(μα−10)∑xi+16(μα2−102)≤32ln(k)

Moving the constant term on the left-side of the inequality to the right-side, and dividing through by −16(2(μα−10)), we get:

116∑xi≥−116(2(μα−10))(32ln(k)−16(μα2−102))=k∗

In summary, we have shown that the ratio of the likelihoods is small, that is:

L(10)L(μα)≤k

if and only if:

x¯≥k∗

Therefore, the best critical region of size α for testing H0:μ=10 against each simple alternative HA:μ=μa, where μa>10, is given by:

C={(x1,x1,...,xn):x¯≥k∗}

where k∗ is selected such that the probability of committing a Type I error is α, that is:

α=P(X¯≥k∗) when μ=10

Because the critical region *C* defines a test that is most powerful against each simple alternative $\mu a>10$, this is a uniformly most powerful test, and *C* is a uniformly most powerful critical region of size α.