
# F.Y.B.Sc. COMPUTER SCIENCE

## SEMESTER - II (CBCS)

## PAPER VI

# STATISTICAL METHODS AND TESTING OF HYPOTHESIS

## SUBJECT CODE : USCS206

**Prof. Suhas Pednekar**
Vice Chancellor
University of Mumbai, Mumbai.

**Prof. Ravindra D. Kulkarni**
Pro Vice-Chancellor,
University of Mumbai.

**Prof. Prakash Mahanwar**
Director
IDOL, University of Mumbai.

**Programme Co-ordinator : Prof. Mandar Bhanushe**
Head, Faculty of Science & Technology,
IDOL, University of Mumbai - 400 098.

**Course Co-ordinator : Mr. Sumedh Shejole**
Asst. Professor,
IDOL, University of Mumbai, Mumbai.

**Editor : Mr. Gurav Manav**
Assistant Professor,
Smt. Janakibai Rama Salvi College.

**Course Writers : Sreekala Nair**
The S.I.A College of Higher Education,
Gymkhana road,Gograswadi,
Dombivali East.

**: Geeta Sahu**
Vidyalankar School of Information Technology,
(VSIT), Vidyalankar College Marg,
Wadala (E)-400-037.

**May 2022, Print I**

**Published by**
**Director**
Institute of Distance and Open learning ,
University of Mumbai,Vidyanagari, Mumbai - 400 098.

**DTP COMPOSED AND PRINTED BY**
Mumbai University Press
Vidyanagari, Santacruz (E), Mumbai - 400098.

# CONTENT



*****

# Syllabus

| Course: USCS206 | Statistical Methods and Testing of Hypothesis<br>(Credits : 2 Lectures/Week: 3) | |
|---|---|---|
| **Objectives**:<br>The purpose of this course is to familiarize students with basics of Statistics. This will be essential for prospective researchers and professionals to know these basics.<br>**Expected Learning Outcomes:**<br>Enable learners to know descriptive statistical concepts<br>Enable study of probability concept required for Computer learners | | |
| **Unit I** | **Standard distributions:** random variable; discrete, continuous, expectation and variance of a random variable, pmf, pdf, cdf, reliability,<br>Introduction and properties without proof for following distributions; binomial, normal, chi-square, t, F. Examples | 15L |
| **Unit II** | **Hypothesis testing:** one sided, two sided hypothesis, critical region, p-value, tests based on t, Normal and F, confidence intervals. Analysis of variance : one-way, two-way analysis of variance | 15L |
| Unit III | **Non-parametric tests:** need of non-parametric tests, sign test, Wilicoxon's signed rank test, run test, Kruskal-Walis tests.<br>Post-hoc analysis of one-way analysis of variance : Duncan's test Chi-square test of association | 15L |

**Text book:**


1. Trivedi, K.S.(2009) : Probability, Statistics, Design of Experiments and Queuing theory, with applications of Computer Science, Prentice Hall of India, New Delhi.


**Additional References:**


1. Ross, S.M. (2006): A First course in probability. 6th Edn Pearson
2. Kulkarni, M.B., Ghatpande, S.B. and Gore, S.D. (1999): Common statistical tests. Satyajeet Prakashan, Pune
3. Gupta, S.C. and Kapoor, V.K. (2002) : Fundamentals of Mathematical Statistics, S. Chand and Sons, New Delhi
4. Gupta, S.C. and Kapoor, V.K. (4th Edition) : Applied Statistics, S. Chand and Son's, New Delhi
5. Montgomery, D.C. (2001): Planning and Analysis of Experiments, Wiley.


*****

# UNIT I

# 1

# STANDARD DISTRIBUTIONS
# CONTENTS OF MODULE

**Unit Structure**



## 1.0 OBJECTIVES

Students will be able to:

- Identify the types of random variables.
- Understand the concept of Probability distribution.
- Enable students to understand various types of distributions.

## 1.1 INTRODUCTION

The science of statistics deals with assessing the uncertainty of inferences drawn from random samples of data. This chapter focuses on random variables its types and their probability distribution. To assess the outcome

of an experiment it is desirable to associate a real number X with the possible outcome of an event. The concept of "randomness" is fundamental to the field of statistics. Probability is not only used for calculating the outcome of one event but also can summarize the likelihood of all possible outcomes. The relationship between each possible outcome for a random variable and its probabilities is called a probability distribution. Probability distributions are an important foundational concept in probability and the names and shapes of common probability distributions will be familiar. The structure and type of the probability distribution vary based on the properties of the random variable, such as continuous or discrete, and this, in turn, impacts how the distribution might be summarized or how to calculate the most likely outcome and its probability.

## 1.2 STUDY GUIDANCE

- Understand the basic and concepts
- Practice the questions given in module
- Refer to further reading

## 1.3 STANDARD DISTRIBUTION

### 1.3.1 Random Variable:

A random variable is a real-valued variable or a function that assigns values to each of the outcomes of an experiment. It is used to determine statistical relationships among one another For. Eg. If random variable X is the birth of a male child, then the result of the variable X could be 1 if a male child is born and 0 if a female is born. Eg:2)If the random experiment consists of tossing two coins then the random variable which is the number of heads can be denoted as 0,1 or 2.

There are two types of random variables" discrete and continuous.

**Discrete Random variable**

A random variable that may assume only a finite number or a countably infinite number of values is said to be discrete. For instance, a random variable representing the number of misprints in a book would be a discrete random variable.

**Continuous Random Variable:**

A continuous random variable can assume any value in an interval on the real number line or in a collection of intervals. Since there is an infinite number of values in any interval, it is not meaningful to talk about the probability that the random variable will take on a specific value; instead, the probability that a continuous random variable will lie within a given interval is considered.

$$P(a \le X \le b) = \int_a^b f(x)\,dx$$

### 1.3.2 Probability Mass Function:

The probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable. For a discrete random variable, *x*, the probability distribution is defined by a probability mass function, denoted by *p*(*x*). This function provides the probability for each value of the random variable. Probability distributions always follow the following properties :

(1) *p*(*x*) must be nonnegative for each value of the random variable

i. e, $p(x_i) \ge 0$ for all values of i

(2) The sum of the probabilities for each value of the random variable must be equal to one.

i.e, $\sum_i^n p(x_i) = 1$

The set of values of xi with the corresponding probabilities $p(x_i)$ constitute probability distribution function of discrete random variable X . If X is a discrete random variable then P(X) is called probability mass function (PMF).

The following table shows the discrete distribution random variable

| X | X1 | X2 | X3 | X4 | ….. | ….. | Xn |
|---|---|---|---|---|---|---|---|
| P(X=x) | P1 | P2 | P3 | P4 | ….. | …. | pn |

Eg: The probability distribution of the discrete random variable X is getting head while two coins are tossed

| X | 0 | 1 | 2 |
|---|---|---|---|
| P[X=x] | 1/4 | 2/4 | 1/4 |

### 1.3.3 Probability Density Function:

For a continuous random variable, *x*, the probability distribution is defined by a probability density function (PDF), denoted by *f*(*x*) and the probability density function should satisfy the following conditions:

- For a continuous random variable that takes some value between certain limits, say a and b, The pdf is given by $P[a < X < b] = \int_a^b f(x)\,dx$
- The probability density function is non-negative for all the possible values,

i.e. $f(x) \geq 0$, for all x.

- The area between the density curve and horizontal X-axis is equal to 1,

  i.e. $\int_{-\infty}^{\infty} f(x)dx = 1$

Note: Please note that the probability mass function is different from the probability density function. f(x) does not give any value of probability directly hence the rules of probability do not apply to it.

Eg.: Let X be a continuous random variable with the PDF is given by

$F(X) = \begin{cases} x, 0 < x < 1 \\ 2 - x, 1 < x < 2 \end{cases}$ find p [0.2<x<1.2]

**Solution:**

$$P[0.2 < X < 1.2] = \int_{0.2}^{1.2} f(x)\, dx$$

We can split the integrals by taking the intervals as given below

$$\int_{0.2}^{1} x dx + \int_{1}^{1.2} (2 - x) dx$$

$$= \left(\frac{x^2}{2}\right)_{0.2}^{1} + \left(2x - \frac{x^2}{2}\right)_{1}^{1.2}$$

$$= \frac{1}{2} - 0.02 + (2.4 - 0.72) - \left(2 - \frac{1}{2}\right)$$

$$= \frac{1}{2} - 0.02 + 1.68 - 2 + \frac{1}{2}$$

$$= 0.66$$

### 1.3.4 Expectation of Random Variable (Mean):

**Case 1 Discrete Random variable:**

The The expected value of a random variable is the average value of the random variable over a large number of experiments. In the case of discrete random variables expected value can be found by using the formula

$E(X) = \sum_{i=1}^{n} x_i\, p(x_i)$ where $p(x_i)$ is the probability of each outcome

E.g. Find the expected value of the following probability distribution from the given probability distribution table

| x | -1 | -2 | -3 | 0 | 1 | 2 |
|---|---|---|---|---|---|---|
| P(x) | 0.25 | 0.35 | 0.01 | 0.01 | 0.2 | 0.18 |

**Solution:**

Expected value,

$$E(X)=\sum_{i=1}^{n} x_i P(x_i)$$

$$=(-1\times 0.25)+(-2\times 0.35)+(-3\times 0.01)+(0\times 0.01)(1\times 0.2)+(2\times 0.18)$$

$$=0.25+(-0.7)+(-0.03)+(0)(0.2)+(0.36)$$

= 0.42

**Case 2 Continuous Random variable:**

Let X be a random variable with pdf f(x) then the mathematical expectation of continuous random variable denoted by E(X) and given by

$$E(X)=\int_{-\infty}^{\infty} xf(x)dx$$

For Eg: Let X be a continuous random variable with $f(X)=\begin{cases}3x^2, 0\le x\le 1\\ 0 \quad otherwise\end{cases}$ Find the expected value?

E(X) = $\int_{-\infty}^{\infty} x\, f(x)dx$

For Eg: Let X be a continuous random variable with pdf f(x) = $\begin{cases}3x^2, & 0\le x\le 1\\ 0 & otherwise\end{cases}$, Find the expected value?

Solution: E(X) $= \int_{-\infty}^{\infty} x\, f(x)dx$

$= \int_0^1 x\, 3x^2 dx$

$= \int_0^1 3x^3 dx$

$= 3\left[\frac{x^4}{4}\right]_0^1$

$= \frac{3}{4}$

**1.3.5 Variance of A Random Variable:**

**Case 1: Discrete Random variable:**

The variance for a discrete random variable is denoted by V(X) and is defined as where E(X) is the expected value

$V(X)=E(X^2)-\left(E(X)^2\right)$ where E(X) is the expected value

$$E\left(X^2\right)=\sum x^2 p(x)$$

**Case 2: Continuous Random variable:**

The variance for a continuous random variable is denoted by V(X) and is defined as

$V(X)=E\left(X^2\right)-\left(E(X)^2\right)$ where E(X) is the expected value

$$E\left(X^2\right)=\int_{-\infty}^{\infty} x^2 f(x)\,dx$$

Eg: Find the Mean and Variance of the given data

| X | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| P(X) | 0.2 | 0.15 | 0.1 | 0.2 | 0.15 | 0.2 |

$E(X)=\sum_{i=}^{n} x_i p(x_i)$

$=(1\times 0.2)+(2\times 0.15)+(3\times 0.1)+(4\times 0.2)+(5\times 0.15)+(6\times 0.2)$

$=0.2+0.3+0.3+0.8+0.75+1.2$

$=3.55$

$E\left(X^2\right)=\sum x^2 P(x)$

$=\left(1^2\times 0.2\right)+\left(2^2\times 0.15\right)+\left(3^2\times 0.1\right)+\left(4^2\times 0.2\right)+\left(5^2\times 0.15\right)+\left(6^2\times 0.2\right)$

$=(0.2)+(0.6)+(0.9)+(3.2)+(3.75)+(7.2)$

$=15.85$

$V(X)=E\left(X^2\right)-\left(E(X)\right)^2$

$=15.85-(3.55)^2$

$=15.85-12.6025$

$=3.2475$

Mean = 3.55

Variance = 3.2475

Eg: The p.d.f of random variable X is $f(X)=6\left(x-x^2\right), 0\le x\le 1$ Find Mean and variance?

Mean $E(X)=\int_{-\infty}^{\infty} xf(x)dx$

$$=\int_0^1 6x\left(x-x^2\right)dx$$

$$=6\left[\frac{x^3}{3}-\frac{x^4}{4}\right]_0^1$$

$$=6\left(\frac{1}{3}-\frac{1}{4}\right)$$

$$=6\times\frac{1}{12}$$

$$=\frac{1}{2}$$

$$E\left(X^2\right)=\int_{-\infty}^{\infty} x^2 f(x)dx$$

$$=\int_0^1 x^2 6\left(x-x^2\right)dx$$

$$=6\int_0^1 6\left(x^3-x^4\right)dx$$

$$=6\left[\frac{x^4}{4}-\frac{x^5}{5}\right]_0^1$$

$$=6\times\frac{1}{20}$$

$$=\frac{3}{10}$$

$$V(X)=E\left(X^2\right)-\left(E(X)\right)^2$$

$$=\frac{3}{10}-\frac{1}{4}$$

$$=\frac{1}{20}$$

Mean = $=\frac{1}{2}$

Variance $= \frac{1}{20}$

**1.3.6 Cumulative Distribution Function**:

When we are dealing with inequalities, for instance, X < a, the resulting set of the outcome of elements will contain all the elements lesser than a that is " $-\infty$ to a.

When probability function is applied over such inequality, it leads to a cumulative probability value giving the estimate of the value being less than or equal to a particular value. The cumulative distribution function of a random variable is another method to describe the distribution of a random variable.

If X is a continuous random variable with pdf f(x) then the function $\mathrm{F}(x) = \mathrm{P}[X \le x] = f(x)dx, -\infty < x < \infty$ Is called cumulative distribution function (cdf)

**1.3.7 Reliability:**

Reliability is dependent on probability for measuring and describing its characteristics.The probability that the component survives until some time t is called reliability R(t) of the component where X be the lifetime or the time to failure of a component.

Thus, $\mathrm{R}(t) = \mathrm{P}[X > t] = 1 - F(t)$, where F is the distribution function of the component lifetime X. The component is normally (but not always) assumed to be working properly at time t = 0 [i.e., R(0) = 1], and no component can work forever without failure $[i.e. \lim \to +\infty R(t) = 0]$. Also, R(t) is a monotone decreasing function of t. For t less than zero, reliability has no meaning, but we let R(t) = 1 for t < 0. F(t) will often be called the unreliability.

## 1.4 INTRODUCTION-DISTRIBUTION FUNCTIONS

In the previous section we discussed about various types of distribution and its mean and variance. This section focuses on some standard distribution and it properties.

**Bernoulli's Trial:**

Bernoulli's trials are events or experiments which results in two mutually exhaustive outcome one of them is termed as success and the other is failure. For example , when an unbiased coin is tossed we can define success as getting tail and hence getting head is failure

## 1.5 BINOMIAL DISTRIBUTION

Consider 'n' independent Bernoulli's trial which results into either success or failure with probability of success "p" and probability of failure "q".

Let 'X' be a discrete random variable denoting the success in 'n' independent trials the variate X is called random variate and the probability distribution of X is called Binomial distribution and is defined as $P(X=x)=\binom{n}{x}p^x q^{n-z}$

= 0 elsewhere Where x =0, 1 ,2….n,0<p<1 and q =1- p

For example, let's assume an unbiased coin is tossed 10 times and probability of getting a head on one filp is ½.Flip 10 times ,the probability of getting head on any throw is ½ and have a binomial distribution of n=10 and p = ½.''Success" would be "flipping a head" and failure will be" flipping tail'

Properties of Binomial distribution are as follows:

1. Mean of binomial distribution is np

2. Variance of Binomial distribution is npq

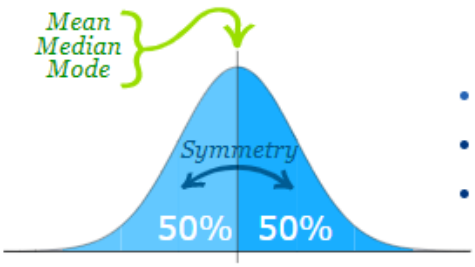
## 1.6 NORMAL DISTRIBUTION

The normal distribution is the most important and most widely used distribution in statistics. In statistics most of the symmetrical distributions are similar to normal distribution.

The equation of the normal curve is $\gamma=\frac{N}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2}/2\sigma^2$ where $\sigma=$ standard $deviation$ $\mu=Arithmetic\ mean$

We can transform the variable x to $z=\frac{x-\mu}{\sigma}$ here z is called normal variate.

The parameters $\mu$ and $\sigma$ are the mean and standard deviation, respectively, and define the normal distribution.




The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

The following are the properties of Normal distribution:

1. It is bell shaped and symmetrical in nature.

2. The mean, median, and mode of a normal distribution are identical.

3. The total area under the normal curve is unity.

4. Normal distributions are denser in the center and less dense in the tails.

5. Normal distributions are defined by two parameters, the mean ($\mu$) and the standard deviation ($\sigma$).

6. 68% of the area of a normal distribution is within one standard deviation of the mean.

7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.

8. The position and shape of the normal curve depend upon $\mu$, $\sigma$ *and* $N$

## 1.7 CHI-SQUARE DISTRIBUTION

The chi-square distribution is a continuous probability distribution that is widely used in statistical inference it is related to the standard normal distribution in that if a random variable Z has the standard normal distribution then that random variable squared has a chi-square distribution with one degree of freedom.

The Chi-Square Distribution, denoted as $X^2$ is related to the standard normal distribution such as, if the independent normal variable, let's say Z assumes the standard normal distribution, then the square of this normal variable $Z^2$ has the chi-square distribution with 'K' degrees of freedom. Here, K is the sum of the independent squared normal variables.

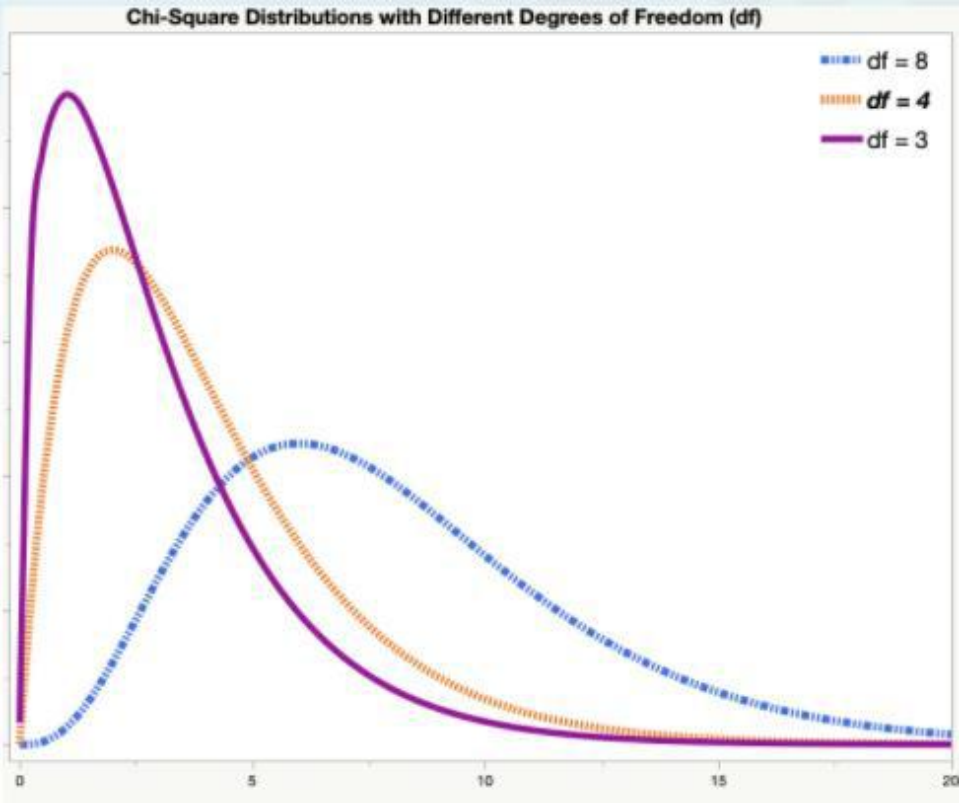
**The properties are as follows:**

1. The chi-square distribution is a continuous probability distribution with the values ranging from 0 to $\infty$ (infinity) in the positive direction. The $X^2$ can never assume negative values.

2. The shape of the chi-square distribution depends on the number of degrees of freedom 'v'. When 'v' is small, the shape of the curve tends to be skewed to the right, and as the 'v' gets larger, the shape becomes more symmetrical and can be approximated by the normal distribution.

3. The mean of the chi-square distribution is equal to the degrees of freedom, i.e. E($X^2$) = 'v'. While the variance is twice the degrees of freedom, Viz. n($X^2$) = 2v.

4. The $X^2$ distribution approaches the normal distribution as v gets larger with mean v and standard deviation as$\sqrt{2}$ $X^2$.




**Chi-square distribution with different degree of freedom.**

## 1.8 T - DISTRIBUTION

The t-Distribution, also known as Student's t-distribution is the probability distribution that estimates the population parameters when the sample size is small and the population standard deviation is unknown.

It resembles the normal distribution and as the sample size increases the t-distribution looks more normally distributed with the values of means and standard deviation of 0 and 1 respectively.

**Properties of t-Distribution:**

1. The graph of the t distribution is also bell-shaped and symmetrical with a mean zero.
2. The *t*-distribution is most useful for small sample sizes, when the population standard deviation is not known, or both.
3. The student distribution ranges from $-\infty$ to $\infty$ (infinity).
4. The shape of the t-distribution changes with the change in the degrees of freedom.
5. The variance is always greater than one and can be defined only when the degrees of freedom $v \geq 3$
6. It is less peaked at the center and higher in tails, thus it assumes a platykurtic shape.

7. The t-distribution has a greater dispersion than the standard normal distribution. And as the sample size 'n' increases, it assumes the normal distribution. Here the sample size is said to be large when $n \geq 30$.

## 1.9 F-TEST DISTRIBUTION

The distribution which is used to compute the behavior of two variances, taken from two independent populations is called F-distribution. The distribution of all possible values of f- statistic is called f-distribution with degrees of freedom $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$. There are several properties of F-distribution:

1. The curve is not symmetrical but skewed to the right.
2. The F-distribution is positively skewed with an increase in the degree of freedom $v_1$ and $v_2$, its skewness increases.
3. The *F* statistic is greater than or equal to zero.
4. As the degrees of freedom for the numerator and the denominator gets larger, the curve approximates the normal.
5. The statistic used to calculate the value of mean and variance is:

$$\text{Mean} = v_1 = \frac{v_2}{-v_2 - 2}, \text{for } v_2 > 2$$

$$\text{Variance} = = \frac{2v_2^{\,2}\left(v_1 + v_2 - 2\right)}{v_1\left(v_2 - 2\right)^2\left(v_2 - 4\right)},$$

6. The shape of the F-distribution depends on its parameters í1 and í2 degrees of freedom.
7. The values of the area lying on the left-hand side of the distribution can be found out by taking the reciprocal of F values corresponding to the right-hand side and the degrees of freedom in the numerator and the denominator are interchanged.

## 1.10 SUMMARY

We discussed about random variable and its different types. There are two types of probability distribution, discrete and continuos.A random variable assumes only a finite or countably infinite number of values are called a discrete random variable. A continuous random variable can assumes values uncountable number of values. Discrete random variable is associated with probability mass function and that of continuous related with probability density function. Expected value and variance of the discrete and continuous distribution were defined. We learnt some standard distributions and its properties and these distributions will be applicable in testing of hypothesis. The application methods of probability

can be seen in modeling of text and Web data, network traffic modeling, probabilistic analysis of algorithms and graphs, reliability modeling, simulation algorithms, data mining, and speech recognition.

## 1.11 UNIT END QUESTIONS

1, Let X be a continuous random variable with the following PDF

$f_x(x) = \begin{cases} ae^{-x} & x \geq 0 \\ o & elsewhere \end{cases}$

Where a is a positive constant

(i) Find a

(ii) Find CDF of X ,$F_X(x)$

(iii) Find P(1<$X$<3)

2. Let X be a random variable with PDF given by

$f_x(x) = \begin{cases} ae^{-x} & |x| \leq 1 \\ o & elsewhere \end{cases}$

Where a is a positive constant

(i) Find a

(ii) Find E(X) and V(X)

3. Check whether the following can define probability distribution

(i) $f(x) = \frac{x}{15}$ where x =0 , 1 , 2 , 3 , 4 ,5

(ii) $f(x) = \frac{2-x}{3}$ where x = 3, 4, 5

(iii) $f(x) = \frac{1}{2}$ where x = 1, 2

4. Consider tossing of a fair coin 3 times Define X = number of times tails occurred

| Value | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Probability | 1/8 | 3/8 | 3/8 | 1/8 |

Find E(X) and V(X)

5. Find mean and variance x given the following probability distribution

| x | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| P(x) | 0.3 | 0.2 | 0.2 | 0.2 | 0.1 |

6. A random variable x has following probability distribution

| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| P(x) | k | 2k | 3k | 5k | 4k | 2k | K |

Find k. Hence find E(x).

7. A bag contains 4 Red and 6 White balls. Two balls are drawn at random and gets Rs.10 for each red and Rs.5 for each white ball.. Find his mathematical expectation.

8. A continuous distribution of a variable X in the range (-3, 3) is defined by

   (i) Verify that the area under the curve is unity.

   (ii) Find the mean and variance of the above distribution.

$$F(x) = \frac{1}{16}(3+x)^2 \qquad -3 \le x \le -1$$
$$= \frac{1}{16}(6-2x)^2 \qquad -1 \le x \le 1$$
$$= \frac{1}{16}(3-x)^2 \qquad 1 \le x \le 3$$

(i) Verify that the area under the curve is unity.

(ii) Find the mean and variance of the above distribution.

## 1.12 REFERENCES

1. Probability and Statistics with Reliability, Queuing and Computer Science Applications, Kishor S. Trivedi, 2016 by John Wiley &Sons, Inc., 1946.

2. Fundamentals of Mathematical Statistics by S.C. Gupta , 10th Edition, 2002.

## 1.13 FURTHER READING

1. Introductory Business Statistics, Alexander Holmes et.al., 2018.

*****

# UNIT II

# 2

# HYPOTHESIS TESTING

**Unit Structure**



## 2.0 OBJECTIVE

- Statistics is referred to as a process of collecting, organizing and analyzing data and drawing conclusions.
- The statistical analysis gives significance to insignificant data or numbers.
- Statistics is "a branch of mathematics that deals with the collection, analysis, interpretation, and presentation of masses of numerical data.

## 2.1 INTRODUCTION

- The science of collecting, organizing, analyzing and interpreting data in order to make decisions.
- Statistics is used to describe the data set and to draw conclusion about the population from the data set.

**Statistical methods are of two types:**

**Descriptive Method:** This method uses graphs and numerical summaries.

**Inferential Method:** This method uses confidence interval and significance test which are part of applied statistics.

## 2.2 HYPOTHESIS TESTING

**Definition**

- Hypothesis is a claim or idea about a group or population.
- Hypothesis refers to an educated guess or assumption that can be tested.
- Hypothesis is formulated based on previous studies.

## 2.3 NULL HYPOTHESIS (*H*O)

A statistical hypothesis which is formulated for the purpose of rejecting or nullifying it.

## 2.4 ALTERNATE HYPOTHESIS (*H*1)

**Any hypothesis which differs from the given null hypothesis:**

The alternative hypothesis is what you might believe to be true or hope to prove true.

Null Hypothesis

Null Hypothesis can be a statement of equality.

There is no difference in mean scores of VSIT and SIES.

H0: $\mu 1=\mu 2$

Null Hypothesis can be a statement of no relationship.

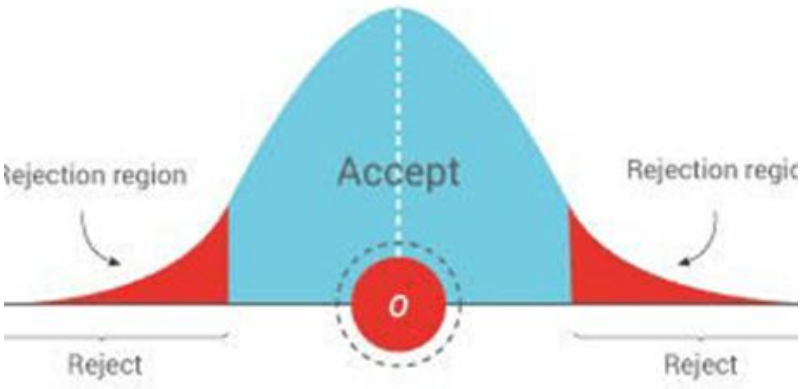
Example - There is no relation between personality and job success.

Plant growth is not affected by light intensity.

Hypothesis Testing -Two tailed and One Tailed Test

Hypothesis Testing

- **Decision-making process for evaluating claims about a population.**
- **Whether to accept or reject Ho**




**Type I and Type II Errors:**

**Type I Error:**

When we reject a hypothesis when it should be accepted

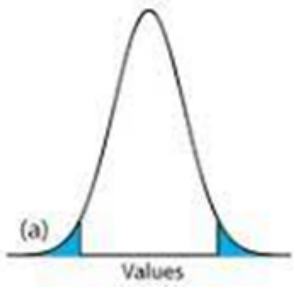
**Type II Error:**

**When we accept a hypothesis when it should be rejected**

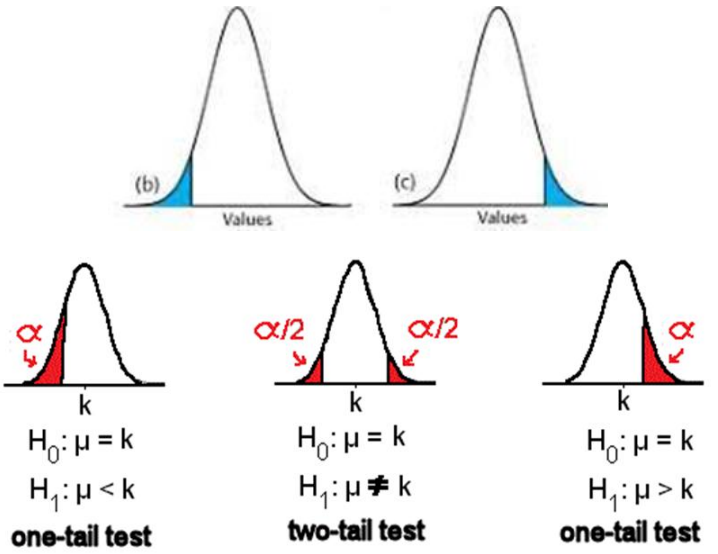
| | Null hypothesis is TRUE | Null hypothesis is FALSE |
|---|---|---|
| Reject null hypothesis | Type I Error (False positive) | Correct outcome! (True positive) |
| Fail to reject null hypothesis | Correct outcome! (True negative) | Type II Error (False negative) |

**Two Tailed and One Tailed Test:**

- Two Tail Test: critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values.
- If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis.



- One tail test: A one-tailed test is a statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both.
- If the sample being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis.




One-tailed tests are applied to answer for the questions: Is our finding significantly greater than our assumed value? Or: Is our finding significantly less than our assumed value?

Two-tailed tests are applied to answer the questions: Are the findings different from the assumed mean?

**Level of Significance:**

- Maximum allowable probability of making type I error.
- This probability is denoted by $\alpha$
- A significance level of 0.05 (5%) *or* 0.01 (1%) is common.

## 2.5 CRITICAL REGION

| LOS Test | α=0.05 (5 %) | α=0.01 (1 %) |
|---|---|---|
| Two-tailed Test | Zc=1.96 | Zc= 2.58 |
| One-tailed Test | Zc=1.645 | Zc= 2.33 |

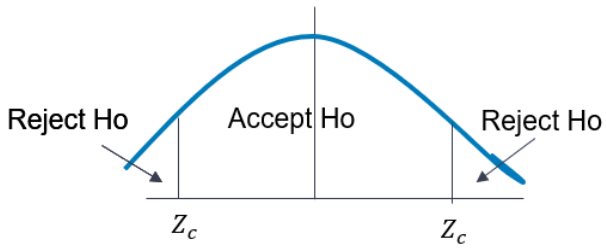
## 2.6 P -VALUE

**Z Score:**

- **Mean $Z= (\bar{X}-\mu)/(\sigma/\sqrt{N})$**
- **Proportion $Z= (P -p)/\sqrt{(pq/N)}$**

Steps for hypothesis testing

1. Propose Ho and H1.
2. Identify test-
   - one tailed (if <, >)
   - two tailed (if ≠)
3. Get table value Zc according to LOS mentioned in the problem.
4. Find Z score using the formula.
5. Inference-
   - If Z < Zc , accept Ho.
   - If Z > Zc , reject Ho




**Question:**

**Qn)** The breaking strength of cables produced by a manufacturer have a mean of 1800 $lb$ and a standard deviation of $100lb$. By a new technique in the manufacturing process, it is claimed that the breaking strength can be increased. To test this claim, a sample of 50 cables is tested and it is found that the mean breaking strength is $1850lb$. Can we support the claim at the 0.01 significance level?

**Solution:**

**Step 1- Write given values**

$\mu = 1800\ lb$

$\sigma = 100\ lb$

Population Parameter

**N = 50**

- $\bar{X} = 18500\ lb$

- **LOS =** $\alpha$ **= 0.01= 1 %**

**Step 2- Propose H0**

$H_0$: $\mu = 1800lb$ and there is really no change in breaking strength.

$H_1$: $\mu > 1800lb$ and there is a change in breaking strength.

**Step 3- Identify Test**

As > sign is there, use One tailed Test

**Step 4-** Get table value of Zc for LOS α=0.01 (1 %)

Zc= 2.33

**Step 5-** Find Z score using formula:

$Z= (\bar{X}-\mu)/(\sigma/\sqrt{N})$

Z=3.5355

| | α=0.05 (5 %) | α=0.01 (1 %) |
|---|---|---|
| Two-tailed Test | Zc=1.96 | Zc= 2.58 |
| One-tailed Test | Zc=1.645 | Zc= 2.33 |

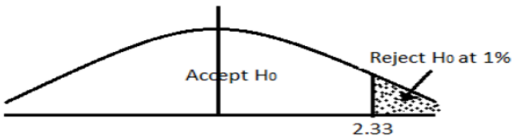
| | |
|---|---|
| M | 1800 |
| Σ | 100 |
| N | 50 |
| $\bar{X}$ | 1850 |

**Step 6 – Inference**

$Z = 3.5355\ ,\ Z_c = 2.33$

**Z > Zc, reject Ho**

- **Therefore, we can support the claim at 0.01 LOS. i.e., the cable strength is increased.**




Reject $H_0$

By new technique breaking strength has increased.

**Qn)** On an examination given to students at a large number of different schools, the mean grade was 74.5 and standard deviation was 8.0. At one particular school where 200 student took the examination, the mean grade was 75.9. Discuss the significance of this result at he 0.05 level from the view point of

a. One tailed test

b. Two tailed test

**Step 1- Write given values**

$\mu=74.5$

$\sigma=8$

N = 200

$\bar{X}=75.9$

LOS = $\alpha = 0.05 = 5\ \%$

**One tailed Test**

**Step 2- Propose H0**

H0: $\mu=74.5$; performance of school is same as population

H1: $\mu>74.5$; performance of school is better than population

**Two tailed Test**

**Step 3- Propose H0**

H0: $\mu=74.5$; performance of school is same as population

H1: $\mu\neq74.5$; performance of school is different than population

| | $\alpha=0.05$ (5 %) | $\alpha=0.01$ (1 %) |
|---|---|---|
| Two-tailed Test | Zc=1.96 | Zc= 2.58 |
| One-tailed Test | Zc=1.645 | Zc= 2.33 |

**One tailed Test**

**Step 4- Get table value of Zc for LOS α=0.05 (5 %)**

Zc= 1.645

**Step 5- Find Z score using formula;**

$Z= (\bar{X}-\mu)/(\sigma/\sqrt{N})$

Z=2.4748

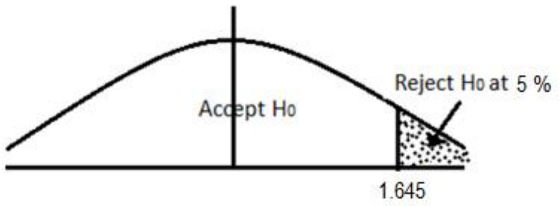
| | |
|---|---|
| μ | 74.5 |
| σ | 8 |
| N | 200 |
| $\bar{X}$ | 75.9 |

**Step 6 – Inference**

Z =2.4748, Zc=1.645

As Z > Zc, reject Ho.

Therefore, we can support the claim at 0.05 LOS. i.e., the performance of the school is better than population




Reject $H_0$

**Two tailed Test:**

**Step 4:** Get table value of Zc for LOS α=0.05 (5 %)

Zc= 1.96

**Step 5:**

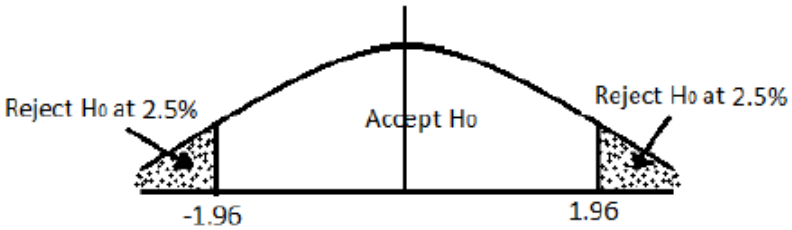
Z =2.4748

**Step 6 : Inference**

As Z > Zc, reject Ho.

Therefore, we can support the claim at 0.05 LOS. i.e., the performance of the school is different than the population

| | **α=0.05 (5 %)** | **α=0.01 (1 %)** |
|---|---|---|
| **Two-tailed Test** | Zc=1.96 | Zc= 2.58 |
| **One-tailed Test** | Zc=1.645 | Zc= 2.33 |




**Z Score**

**Mean**

$Z= (\bar{X}-\mu)/(\sigma/\sqrt{N})$

**Proportion**

$Z= (P -p)/\sqrt{(pq/N)}$

**Steps for hypothesis testing**

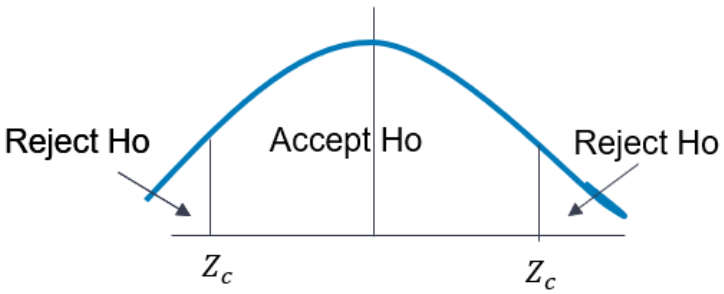
1. Write given values.
2. Propose Ho and H1.
3. Identify test-

   one tailed (if <, >)

   two tailed (if ≠)
4. Get table value Zc according to LOS mentioned in the problem.
5. Find Z score using the formula.
6. Inference-

   If Z < Zc , accept Ho.

   If Z > Zc , reject Ho.



## Question

**Qn)** The manufacturer of a patent medicine claims that it is 90% effective in relieving an allergy for a period of 8 hours. In a sample of 200 people who had the allergy, the medicine provided relief for 160 people. Determine whether the manufacturer's claim is legitimate at 5% level of significance.

**Step 1- Write given values**

p=90/100=0.9 (Population Parameter)

q=1-0.9=0.1 (Population Parameter)

N = 200

P= 160/200=0.8 (Sample Data)

LOS = α = 0.05= 5 % (Sample Data)

**Step 2- Propose H0**

$H_0: p = 90\% = 0.9$(Manufacturer's claim is valid)

$H_1: p < 0.9$ (Manufacturer's claim is not valid)

**Step 3-** Identify Test

As < sign is there, use One tailed Test

**Step 4-** Get table value of Zc for LOS α=0.05 (5 %)

Zc= 1.645

**Step 5-** Find Z score using formula-

Z= (P -p)/√(pq/N)

Z=- 4.714

| | **α=0.05 (5 %)** | **α=0.01 (1 %)** |
|---|---|---|
| **Two-tailed Test** | Zc=1.96 | Zc= 2.58 |
| **One-tailed Test** | Zc=1.645 | Zc= 2.33 |

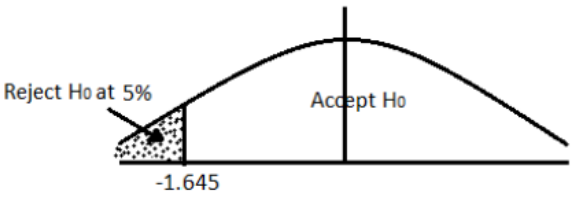
| | |
|---|---|
| p | 0.9 |
| q | 0.1 |
| P | 0.8 |
| N | 200 |

**Step 6 – Inference**

Z = -4.714, Zc=-1.645

As Z falls in critical region, reject Ho.

Therefore, we cannot support the claim at 0.05 LOS. i.e., the medicine is not 90% effective.




**Qn)** A pair of dice is tossed 100 times and it is observed that 23 times sum of numbers appearing on uppermost faces is 7. Test the hypothesis that the dice are fair by using a two-tailed test at 5% significance level.

**Step 1- Write given values:**

p=probability of getting sum 7

E= {(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)}

n(E) = 6

n(S) = 36

p=6/36= 1/6 = 0.167

*q = 1-p = 0.833*

*N = 100*

P=23/100=0.23

*LOS = α = 0.05= 5 %*

**Step 2- Propose H0:**

$$H_0: p = \frac{1}{6} \text{ (i.e. the dice are fair)}$$

$$H_0: p \neq \frac{1}{6} \text{(dice are not fair)}$$

**Step 3 Identify Test:**

As ≠ sign is there, use Two tailed Test

**Step 4:** Get table value of Zc for LOS α=0.05 (5 %)

Zc= 1.96

**Step 5**: Find Z score using formula

| | α=0.05 (5 %) | α=0.01 (1 %) |
|---|---|---|
| **Two-tailed Test** | Z_c=1.96 | Z_c= 2.58 |
| **One-tailed Test** | Z_c=1.645 | Z_c= 2.33 |

$Z= (P -p)/\sqrt{(pq/N)}$

Z=0.1689

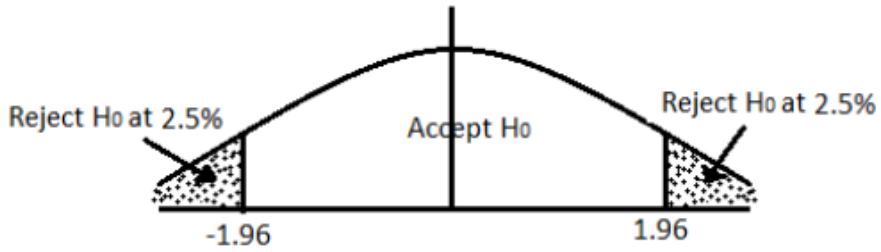
| | |
|---|---|
| p | 0.167 |
| q | 0.833 |
| P | 0.23 |
| N | 100 |

**Step 6:** Inference

Z =0.1689, Zc=1.96

As Z < Zc, Accept Ho.

Therefore, we can support the claim at 0.05 LOS. i.e., the dice are fair.




## 2.7 TEST BASED ON T

**Student's t distribution:**

**Degrees of freedom:**

The number of independent pieces of information that went into calculating the estimate.

Degrees of freedom = N-1

z score, or z statistic is replaced by a suitable t score, or t statistic.

$$t = \frac{\bar{X}-\mu}{\sigma_{\bar{x}}}\sqrt{N-1}$$

**Q.10** individuals are chosen at random from a population and their height (in inches) is found to be – 63, 63, 64, 65, 66, 69, 69, 70, 70, 71. Find students t by considering population mean to be 65.

**Solution:**

Formula-

$$t = \frac{\bar{X}-\mu}{\sigma_{\bar{x}}}\sqrt{N-1}$$

Given-

N = 10

μ = 65

$$\bar{X} = \frac{63+63+64+65+66+69+69+70+70,+71}{10} =$$

$$\sigma_{\bar{x}} = \sqrt{\frac{(X-\bar{X})^2}{N}}$$

$$t = \frac{\bar{X}-\mu}{\sigma_{\bar{x}}}\sqrt{N-1}$$

$$t = 2.0226$$

| X | $X - \bar{X}$ | $(X - \bar{X})^2$ |
|---|---|---|
| 63 | -4 | 16 |
| 63 | -4 | 16 |
| 64 | -3 | 9 |
| 65 | -2 | 4 |
| 66 | -1 | 1 |
| 69 | 2 | 4 |
| 69 | 2 | 4 |
| 70 | 3 | 9 |
| 70 | 3 | 9 |
| 71 | 4 | 16 |
| **670** | | **88** |

Q. In the past, a machine has produced washers having a thickness of 0.050 *in*. To determine whether the machine is in proper working order, a sample of 10 washers is chosen, for which the mean thickness is 0.053 *in* and the standard deviation is 0.003 *in*. Test the hypothesis that the machine is in proper working order at 5% and 1% LOS. (tc at 5% LOS =2.26, t_c at 1% LOS = 3.25)

**Given:**

μ = 0.050 in

N = 10

$\bar{X}$ = 0.053 in

**Propose Hypothesis:**

Let $H_0: \mu = 0.050$ and the machine is in proper working order.

$H_1: \mu \neq 0.050$ and the machine is not in proper working order.

$$t = \frac{\bar{X}-\mu}{\sigma_{\bar{x}}}\sqrt{N-1}$$

t= 3

At 5% LOS

tc= 2.26

t = 3

As t > tc à Reject Ho at 5% LOS

At 1% LOS

tc=3.25

t = 3

As t < tc à Accept Ho at 1% LOS

## 1.8 NORMAL AND F DISTRIBUTION

- Called as Fisher's F Distribution.
- z score, or z statistic is replaced by a suitable F score, or F statistic.

$$F = \frac{\frac{N_1 S_1^2}{(N_1-1)\sigma_1^2}}{\frac{N_2 S_2^2}{(N_2-1)\sigma_2^2}}$$

**Where,**

N1= Sample 1 size

N2= Sample 2 size

σ1 = Population 1 SD

σ2= Population 2 SD

S1= Sample 1 SD

S2= Sample 2 SD

**Q.** Two samples of sizes 9 and 12 are drawn from two normally distributed populations having variances 16 and 25 respectively. If the sample variances are 20 and 8, determine whether the first sample has a significantly larger variance than the second sample at significance levels of (a)0.05 (b) 0.01

(F0.95=2.95, F0.99=4.74)

**Solution:**

**Given:**

N1 = 9

N2 = 12

σ1^2 = Population 1 variance =16

σ2^2 = Population 2 variance = 25

S1^2 = Sample 1 variance = 20

S2^2 = Sample 2 variance = 8

- Formula-

$$F = \frac{\frac{N_1 S_1^2}{(N_1-1)\sigma_1^2}}{\frac{N_2 S_2^2}{(N_2-1)\sigma_2^2}}$$

$$F = \frac{(9)(20)/(9-1)(16)}{(12)(8)/(12-1)(25)}$$

F = 4.03

At 5% LOS

Fc= 2.95

F = 4.03

As F > Fc à We can conclude that the variance of sample 1 is significantly larger than that for sample 2.

At 1% LOS

Fc =4.74

F = 4.03

As F < Fc à Variance of sample 1 is not larger than that for sample 2.

## 2.9 ANALYSIS OF VARIANCE (ANOVA)

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

The Formula for ANOVA is: = MST/MSE

where:

F=ANOVA coefficient

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error
The ANOVA test is the initial step in analysing 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.

## 2.10 ONE WAY ANALYSIS OF VARIANCE

The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. This guide will provide a brief introduction to the one-way ANOVA, including the assumptions of the test and when you should use this test. If you are familiar with the one-way ANOVA,

The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other. Specifically, it tests the null hypothesis:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

where μ = group mean and k = number of groups. If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypothesis ($H_A$), which is that there are at least two group means that are statistically significantly different from each other.

## 2.11 TWO-WAY ANALYSIS OF VARIANCE

A two-way ANOVA is used to estimate how the mean of a quantitative variable changes according to the levels of two categorical variables. Use a two-way ANOVA when you want to know how two independent variables, in combination, affect a dependent variable.

**Example:** You are researching which type of fertilizer and planting density produces the greatest crop yield in a field experiment. You assign different plots in a field to a combination of fertilizer type (1, 2, or 3) and planting density (1=low density, 2=high density), and measure the final crop yield in bushels per acre at harvest time.

You can use a two-way ANOVA to find out if fertilizer type and planting density influence average crop yield.

A two-way ANOVA with interaction tests three null hypotheses at the same time:

- There is no difference in group means at any level of the first independent variable.
- There is no difference in group means at any level of the second independent variable.
- The effect of one independent variable does not depend on the effect of the other independent variable (a.k.a. no interaction effect).

A two-way ANOVA without interaction (a.k.a. an additive two-way ANOVA) only tests the first two of these hypotheses.

**The following columns provide all of the information needed to interpret the model:**

- Df shows the degrees of freedom for each variable (number of levels in the variable minus 1).
- Sum sq is the sum of squares (a.k.a. the variation between the group means created by the levels of the independent variable and the overall mean).
- Mean sq shows the mean sum of squares (the sum of squares divided by the degrees of freedom).
- F value is the test statistic from the F-test (the mean square of the variable divided by the mean square of each parameter).
- Pr(>F) is the p-value of the F statistic, and shows how likely it is that the F-value calculated from the F-test would have occurred if the null hypothesis of no difference was true.

## 2.12 SUMMARY

At the end of this chapter one can draw conclusion based on the data available. Data will be processed, summarized and results can be generated and in graphs it will be displayed.

## 2.13 UNIT END QUESTIONS

**Q1.** Compute student's t for data below:

-4 -2 -2 0 2 2 3 3

**Take mean of universe to be zero.**

**Q2.** In a city, it is claimed that average IQ of students is 102. The intelligence quotients (IQs) of 16 students from one area of a city showed a mean of 107 and a standard deviation of 10. Test the claim at 5% LOS. (tc at 5% LOS =2.144)

**Q3.** Two samples of sizes 10 and 15 are drawn from two normally distributed populations having variances 40 and 60, respectively. If the sample variances are 90 and 50, determine whether the sample 1 variance is significantly greater than the sample 2 variance at significance levels of (a) 0.05 and (b) 0.01. (F0.95=2.645, F0.99=4.029)

**Q4.** Two samples of sizes 8 and 12 are drawn from two normally distributed populations having variances 25 and 49, respectively. If the sample variances are 36 and 60, determine whether Summary.

## 2.14 REFERENCES FOR FUTURE READING


- https://www.basic-concept.com/c/basics-of-statistical-analysis.
- https://www.scribbr.com/statistics/two-way-anova/
- Problems are taken from Schaum's Outline, Statistics, Fourth Edition by Murray R, Larry Stephens.


*****

# UNIT III

# 3

# NON-PARAMETRIC TESTS

**Unit Structure**



## 3.0 OBJECTIVE

This type of statistics can be used without the mean, sample size, standard deviation, or the estimation of any other related parameters when none of that information is available. Since nonparametric statistics makes fewer assumptions about the sample data, its application is wider in scope than parametric statistics.

## 3.1 INTRODUCTION

A non-parametric test (sometimes called a distribution free test) does not assume anything about the underlying distribution (for example, that the data comes from a normal distribution). That's compared to parametric test, which makes assumptions about a population's parameters (for example, the mean or standard deviation); When the word "non parametric" is used in stats, it doesn't mean that you know nothing about the population. It usually means that you know the population data does not have a normal distribution.

## 3.2 NON-PARAMETRIC TEST DEFINITION

A non-parametric test does not assume anything about the underlying distribution (for example, that the data comes from a normal distribution). That's compared to parametric test, which makes assumptions about a

population's parameters (For example, the mean or standard deviation); When the word "non-parametric" is used in statistics it means that the population data does not have a normal distribution.

## 3.3 NEED OF NON-PARAMETRIC TEST DEFINITION

- Use nonparametric tests only when assumptions like normality are being violated. Nonparametric tests can perform well with non-normal continuous data with large sample size (generally 15-20 items in each group). Non-parametric tests are used when your data isn't normal. For nominal scales or ordinal scales, use non-parametric statistics.

## 3.4 SIGN TEST

**A few nonparametric tests are:**

- 1-sample sign test: This test is used to estimate the median of a population and compare it to a reference value or target value.
- 1-sample Wilcoxon signed rank test. With this test, estimate the population median and compare it to a reference/target value. However, the test assumes the data comes from a symmetric distribution (eg- Cauchy distribution or uniform distribution).
- Kruskal-Wallis test. Use this test instead of a one-way ANOVA to find out if two or more medians are different. Ranks of the data points are used for the calculations, rather than the data points themselves.
- The Mann-Kendall Trend Test looks for trends in time-series data.
- Mann-Whitney test. Use this test to compare differences between two independent groups when dependent variables are either ordinal or continuous.

**Sign Test:**

The sign test compares the sizes of two groups. It is a non-parametric or "distribution free" test, which means the test doesn't assume the data comes from a particular distribution, like the normal distribution. The sign test is an alternative to a one sample t test or a paired t test. It can also be used for ordered (ranked) categorical data. The null hypothesis for the sign test is that the difference between medians is zero.

**How to Calculate a Paired/Matched Sample Sign Test?**

1. The data should be from two samples.
2. The two dependent samples should be paired or matched. For example, depression scores from before a medical procedure and after.

Set the data in a table. This set of data represents test scores at the end of Spring and the beginning of the Fall semesters. The hypothesis is that summer break means a significant drop in test scores.

- $H_0$: No difference in median of the signed differences.
- $H_1$: Median of the signed differences is less than zero.

**Step1: Subtract set 2 from set 1 and put the result in the third column.**

| | Set 1 | Set 2 | Set 1 – Set 2 | Sign |
|---|---|---|---|---|
| 1 | 443 | 57 | 386 | + |
| 2 | 421 | 352 | 69 | + |
| 3 | 436 | 587 | -151 | - |
| 4 | 376 | 415 | -39 | - |
| 5 | 458 | 458 | 0 | NA |
| 6 | 408 | 424 | -16 | - |
| 7 | 422 | 463 | -41 | - |
| 8 | 431 | 583 | -152 | - |
| 9 | 459 | 432 | 27 | + |
| 10 | 369 | 379 | -10 | - |
| 11 | 360 | 370 | -10 | - |
| 12 | 431 | 584 | -153 | - |
| 13 | 403 | 422 | -19 | - |
| 14 | 436 | 587 | -151 | - |
| 15 | 376 | 415 | -39 | - |
| 16 | 370 | 419 | -49 | - |
| 17 | 443 | 57 | 386 | + |

**Step 2: Add a fourth column indicating the sign of the number in column** 3

**Step 3: Count the number of positives and negatives.**

- **4 positives.**
- **12 negatives.**

**Step 3:** Add up the number of items in the sample and subtract, we get a difference of zero for (in column 3). The sample size in this question was 17, with one zero, so n = 16.

**Step 4:** Find the p-value using a binomial distribution table or use a binomial calculator.

- .5 for the probability. The null hypothesis is that there are an equal number of signs (i.e., 50/50). Therefore, the test is simple binomial experiment with a .5 chance of the sign being negative and .5 of it being positive (assuming the null hypothesis is true).
- 16 for the number of trials.

- 4 for the number of successes. "Successes" here is the smaller of either the positive or negative signs from Step 2.

The p-value is 0.038, which is smaller than the alpha level of 0.05. We can reject the null hypothesis and there is a significant difference.

## 3.5 WILCOXON'S SIGNED RANK TEST

**Definition:**

The Wilcoxon Signed Rank Test is the non-parametric version of the paired t-test. It is used to test whether there is a significant difference between two population means. Use the Wilcoxon Signed Rank test when you would like to use the paired t-test but the distribution of the differences between the pairs is severely non-normally distributed.

Eg: Q. A basketball coach wants to know if a certain training program increases the number of free throws made by his players. To test this, he has 15 players shoot 20 free throws each before and after the training program.

**Solution:** Since each player can be "paired" with themselves, the coach had planned on using a paired t-test to determine if there was a significant difference between the mean number of free throws made before and after the training program.

However, the distribution of the differences turns out to be non-normal, so the coach instead uses a Wilcoxon Signed Rank Test.

The following table shows the number of free throws made (out of 20 attempts) by each of the 15 players, both before and after the training program:

| Player | Before | After |
|---|---|---|
| Player #1 | 14 | 15 |
| Player #2 | 17 | 17 |
| Player #3 | 12 | 15 |
| Player #4 | 15 | 15 |
| Player #5 | 15 | 17 |
| Player #6 | 9 | 14 |
| Player #7 | 12 | 9 |
| Player #8 | 13 | 14 |
| Player #9 | 13 | 11 |
| Player #10 | 15 | 16 |
| Player #11 | 19 | 18 |
| Player #12 | 17 | 20 |
| Player #13 | 14 | 20 |
| Player #14 | 14 | 10 |
| Player #15 | 16 | 17 |

**Step 1: State the null and alternative hypotheses.**

- $H_0$: The median difference between the two groups is zero.
- $H_A$: The median difference is negative. (e.g., the players make less free throws before participating in the training program)

**Step 2: Find the difference and absolute difference for each pair.**

| Player | Before | After | Difference | Abs. Difference |
|---|---|---|---|---|
| Player #1 | 14 | 15 | -1 | 1 |
| Player #2 | 17 | 17 | 0 | 0 |
| Player #3 | 12 | 15 | -3 | 3 |
| Player #4 | 15 | 15 | 0 | 0 |
| Player #5 | 15 | 17 | -2 | 2 |
| Player #6 | 9 | 14 | -5 | 5 |
| Player #7 | 12 | 9 | 3 | 3 |
| Player #8 | 13 | 14 | -1 | 1 |
| Player #9 | 13 | 11 | 2 | 2 |
| Player #10 | 15 | 16 | -1 | 1 |
| Player #11 | 19 | 18 | 1 | 1 |
| Player #12 | 17 | 20 | -3 | 3 |
| Player #13 | 14 | 20 | -6 | 6 |
| Player #14 | 14 | 10 | 4 | 4 |
| Player #15 | 16 | 17 | -1 | 1 |

**Step3:**

| Player | Before | After | Difference | Abs. Difference | Rank |
|---|---|---|---|---|---|
| Player #2 | 17 | 17 | 0 | 0 | - |
| Player #4 | 15 | 15 | 0 | 0 | - |
| Player #1 | 14 | 15 | -1 | 1 | 3 |
| Player #8 | 13 | 14 | -1 | 1 | 3 |
| Player #10 | 15 | 16 | -1 | 1 | 3 |
| Player #11 | 19 | 18 | 1 | 1 | 3 |
| Player #15 | 16 | 17 | -1 | 1 | 3 |
| Player #5 | 15 | 17 | -2 | 2 | 6.5 |
| Player #9 | 13 | 11 | 2 | 2 | 6.5 |
| Player #3 | 12 | 15 | -3 | 3 | 9 |
| Player #7 | 12 | 9 | 3 | 3 | 9 |
| Player #12 | 17 | 20 | -3 | 3 | 9 |
| Player #14 | 14 | 10 | 4 | 4 | 11 |
| Player #6 | 9 | 14 | -5 | 5 | 12 |
| Player #13 | 14 | 20 | -6 | 6 | 13 |

**Step 4: Find the sum of the positive ranks and the negative ranks**.

| Player | Before | After | Difference | Abs. Difference | Rank | Negative Ranks | Positive Ranks |
|---|---|---|---|---|---|---|---|
| Player #2 | 17 | 17 | 0 | 0 | - | | |
| Player #4 | 15 | 15 | 0 | 0 | - | | |
| Player #1 | 14 | 15 | -1 | 1 | 3 | -3 | |
| Player #8 | 13 | 14 | -1 | 1 | 3 | -3 | |
| Player #10 | 15 | 16 | -1 | 1 | 3 | -3 | |
| Player #11 | 19 | 18 | 1 | 1 | 3 | | 3 |
| Player #15 | 16 | 17 | -1 | 1 | 3 | -3 | |
| Player #5 | 15 | 17 | -2 | 2 | 6.5 | -6.5 | |
| Player #9 | 13 | 11 | 2 | 2 | 6.5 | | 6.5 |
| Player #3 | 12 | 15 | -3 | 3 | 9 | -9 | |
| Player #7 | 12 | 9 | 3 | 3 | 9 | | 9 |
| Player #12 | 17 | 20 | -3 | 3 | 9 | -9 | |
| Player #14 | 14 | 10 | 4 | 4 | 11 | | 11 |
| Player #6 | 9 | 14 | -5 | 5 | 12 | -12 | |
| Player #13 | 14 | 20 | -6 | 6 | 13 | -13 | |
| | | | | | **Sum** | **-61.5** | **29.5** |

**Step 5: Reject or fail to reject the null hypothesis.**

The test statistic, W, is the smaller of the absolute values of the positive ranks and negative ranks. In this case, the smaller value is 29.5. Thus, our test statistic is W = 29.5.

To determine if we should reject or fail to reject the null hypothesis, we can reference the critical value found in the Wilcoxon Signed Rank Test Critical Values Table that corresponds with n and our chosen alpha level.

If our test statistic, W, is less than or equal to the critical value in the table, we can reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

The critical value that corresponds to an alpha level of 0.05 and n = 13 (the total number of pairs minus the two we didn't calculate ranks for since they had an observed difference of 0) is 17.

Since in test statistic (W = 29.5) is not less than or equal to 17, we fail to reject the null hypothesis

**Refer Wilcoxon Signed Rank Test Critical Values Table**

| | Alpha value | | | | |
|---|---|---|---|---|---|
| n | 0.005 | 0.01 | 0.025 | 0.05 | 0.10 |
| 5 | - | - | - | - | 0 |
| 6 | - | - | - | 0 | 2 |
| 7 | - | - | 0 | 2 | 3 |
| 8 | - | 0 | 2 | 3 | 5 |
| 9 | 0 | 1 | 3 | 5 | 8 |
| 10 | 1 | 3 | 5 | 8 | 10 |
| 11 | 3 | 5 | 8 | 10 | 13 |
| 12 | 5 | 7 | 10 | 13 | 17 |
| 13 | 7 | 9 | 13 | 17 | 21 |
| 14 | 9 | 12 | 17 | 21 | 25 |
| 15 | 12 | 15 | 20 | 25 | 30 |
| 16 | 15 | 19 | 25 | 29 | 35 |
| 17 | 19 | 23 | 29 | 34 | 41 |
| 18 | 23 | 27 | 34 | 40 | 47 |
| 19 | 27 | 32 | 39 | 46 | 53 |
| 20 | 32 | 37 | 45 | 52 | 60 |
| 21 | 37 | 42 | 51 | 58 | 67 |
| 22 | 42 | 48 | 57 | 65 | 75 |
| 23 | 48 | 54 | 64 | 73 | 83 |
| 24 | 54 | 61 | 72 | 81 | 91 |
| 25 | 60 | 68 | 79 | 89 | 100 |
| 26 | 67 | 75 | 87 | 98 | 110 |
| 27 | 74 | 83 | 96 | 107 | 119 |
| 28 | 82 | 91 | 105 | 116 | 130 |
| 29 | 90 | 100 | 114 | 126 | 140 |
| 30 | 98 | 109 | 124 | 137 | 151 |

Source: This Question and Solution is taken from the link: How to Perform the Wilcoxon Signed Rank Test - Statology

## 3.6 RUN TEST

**What Is a Runs Test?**

A runs test is a statistical procedure that examines whether a string of data is occurring randomly from a specific distribution. The runs test analyzes

the occurrence of similar events that are separated by events that are different.

- Wolfowitz runs test, which was developed by mathematicians Abraham Wald and Jacob Wolfowitz.
- A runs test is a statistical analysis that helps determine the randomness of data by revealing any variables that might affect data patterns.
- Technical traders can use a runs test to analyze statistical trends and help spot profitable trading opportunities.
- For example, an investor interested in analyzing the price movement of a particular stock might conduct a runs test to gain insight into possible future price action of that stock.
- A nonparametric test for randomness is provided by the theory of runs. To understand what a run is, consider a sequence made up of two symbols, a and b, such as
- aa bbb a bb aaaaa bbb aaaa
- The problem discussed is from Schaum' Outline series by Murray Spiegel, fouth edition.
- In tossing a coin, for example, a could represent ''heads'' and b could represent ''tails.'' Or in sampling the bolts produced by a machine, a could represent ''defective'' and b could represent ''nondefective.
- A run is defined as a set of identical (or related) symbols contained between two different symbols or no symbol (such as at the beginning or end of the sequence).
- Proceeding from left to right in sequence (10), the first run, indicated by a vertical bar, consists of two a's; similarly, the second run consists of three b's, the third run consists of one a, etc. There are seven runs in all.
- It seems clear that some relationship exists between randomness and the number of runs. Thus, for the sequence
- a b a b a b a b a b a b
- there is a cyclic pattern, in which we go from a to b, back to a again, etc., which we could hardly believe to be random. In such case we have too many runs (in fact, we have the maximum number possible for the given number of a's and b's). On the other hand, for the sequence
- **aaaaaa bbbb aaaaa bbb**

- There seems to be a trend pattern, in which the a's and b's are grouped (or clustered) together. In such case there are too few runs, and we would not consider the sequence to be random. Thus, a sequence would be considered nonrandom if there are either too many or too few runs, and random otherwise.

- To quantify this idea, suppose that we form all possible sequences consisting of N1 a's and N2 b's, for a total of N symbols in all N1 + N2 = N. The collection of all these sequences provides us with a sampling distribution: Each sequence has an associated number of runs, denoted by V. In this way we are led to the sampling distribution of the statistic V. It can be shown that this sampling distribution has a mean and variance given, respectively, by the formulas

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 \qquad \sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)}$$

-

$$z = \frac{V - \mu_V}{\sigma_V}$$

By using formulas, we can test the hypothesis of randomness at appropriate levels of significance. It turns out that if both N1 and N2 are at least equal to 8, then the sampling distribution of V is very nearly a normal distribution. Thus, it is normally distributed with mean 0 and variance 1.

## 3.7 KRUSKAL-WALIS TEST

The Kruskal–Wallis Non-Parametric Hypothesis Test (1952) is a non-parametric. It is generally used when the measurement variable does not meet the normality assumptions of one-way ANOVA. It is also a popular nonparametric test to compare outcomes among three or more independent (unmatched) groups.

**Assumptions of the Kruskal-Wallis Test:**

- All samples are randomly drawn from their respective population.
- Independence within each sample.
- The measurement scale is at least ordinal.
- Mutual independence among the various samples

**Procedure to conduct Kruskal-Wallis Test:**

- First pool all the data across the groups.

- Rank the data from 1 for the smallest value of the dependent variable and next smallest variable rank 2 and so on… (if any value ties, in that case it is advised to use mid-point), N being the highest variable.
- Compute the test statistic
- Determine critical value from Chi-Square distribution table
- Finally, formulate decision and conclusion

**Calculation of the Kruskal-Wallis Non-Parametric Hypothesis Test:**

The Kruskal–Wallis Non-Parametric Hypothesis Test is to compare medians among k groups (k > 2). The null and alternative hypotheses for the Kruskal-Wallis test are as follows:

- Null Hypothesis $H_0$: Population medians are equal
- Alternative Hypothesis $H_1$: Population medians are not all equal

Kruskal-Wallis test pools the observations from the k groups into one combined sample, and then ranks from lowest to highest value (1 to N), where N is the total number of values in all the groups.

**The test statistic for the Kruskal Wallis test denoted as H is given as follows**:

$$H = \frac{12}{N(N+1)} \sum \frac{T_i^2}{N_i} - 3(N+1)$$

**Where $T_i$ = rank sum for the ith sample i = 1, 2…,k**

In Kruskal-Wallis test, the H value will not have any impact for any two groups in which the data values have same ranks. Either increasing the largest value or

decreasing the smallest value will have zero effect on H. Hence, the extreme outliers (higher and lower side) will not impact this test.

**Example of Kruskal-Wallis Non-Parametric Hypothesis Test:**

In a manufacturing unit, four teams of operators were randomly selected and sent to four different facilities for machining techniques training. After the training, the supervisor conducted the exam and recorded the test scores. At 95% confidence level does the scores are same in all four facilities.

| Facility 1 | Facility 2 | Facility 3 | Facility 4 |
|---|---|---|---|
| 88 | 77 | 71 | 52 |
| 82 | 76 | 56 | 65 |
| 86 | 84 | 64 | 68 |
| 87 | 59 | 51 | 81 |

- Null Hypothesis $H_0$: The distribution of operator scores is same
- Alternative Hypothesis $H_1$: The scores may vary in four facilities
- N=16

**Rank the score in all the facilities:**

| | Facility 1 | Facility 2 | Facility 3 | Facility 4 |
|---|---|---|---|---|
| | 88(16) | 77(10) | 71(8) | 52(2) |
| | 82(12) | 76(9) | 56(3) | 65(6) |
| | 86(14) | 84(13) | 64(5) | 68(7) |
| | 87 (15) | 59 (4) | 51 (1) | 81 (11) |
| $T_i$ | 57 | 36 | 17 | 26 |

$$H = \frac{12}{N(N+1)}\sum\frac{T_i^2}{N_i} - 3(N+1)$$

$$H = \frac{12}{16(17)}\left(\frac{57^2 + 36^2 + 17^2 + 26^2}{4}\right) - 3(17)$$

$$H = \frac{12}{16(17)}\left(\frac{5510}{4}\right) - 3(17) = 9.77$$

**Right tailed chi-square test with 95% confidence level, and df =3, critical $\chi^2$ value is 7.815**

| | Area in the Right Tail | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0.999 | 0.995 | 0.990 | 0.975 | 0.950 | 0.900 | 0.100 | 0.050 | 0.025 | 0.010 |
| Degrees of Freedom | | | | | | | | | | |
| 1 | 0.000 | 0.000 | 0.000 | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 |
| 2 | 0.002 | 0.010 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | 5.991 | 7.378 | 9.210 |
| 3 | 0.024 | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | 6.251 | 7.815 | 9.348 | 11.345 |
| 4 | 0.091 | 0.207 | 0.297 | 0.484 | 0.711 | 1.064 | 7.779 | 9.488 | 11.143 | 13.277 |
| 5 | 0.210 | 0.412 | 0.554 | 0.831 | 1.145 | 1.610 | 9.236 | 11.070 | 12.833 | 15.086 |
| 6 | 0.381 | 0.676 | 0.872 | 1.237 | 1.635 | 2.204 | 10.645 | 12.592 | 14.449 | 16.812 |

Calculated $\chi^2$ value is greater than the critical value of $\chi^2$for a 0.05 significance level. $\chi^2_{calculated} > \chi^2_{critical}$ hence reject the null hypotheses

There is a difference in test scores exists for four teaching methods at different facilities.

## 3.8 POST-HOC ANALYSIS OF ONE-WAY ANALYSIS OF VARIANCE

ANOVA test tells the overall difference between the groups, but it does not tell you which specific groups differed – post hoc tests do that. Because post hoc tests are run to confirm where the differences occurred between groups, they should only be run when showed an overall

statistically significant difference in group means (i.e., a statistically significant one-way ANOVA result). Post hoc tests attempt to control the experiment wise error rate (usually alpha = 0.05) in the same manner that the one-way ANOVA is used instead of multiple t-tests.

## 3.9 DUNCAN'S TEST CHI-SQUARE TEST OF ASSOCIATION

**Chi square test:**

A chi-square ($\chi^2$) statistic is a test that measures how expectations compare to actual observed data (or model results).

$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

Where o-observed frequency

e-expected frequency

Degree of freedom for mxn table-

$\vartheta$=(no of rows-1)(no of columns-1)

$\vartheta$=(m-1)(n-1)

Eg: In an experiment to study the dependence of hypertension on smoking habits, the following data is taken from 180 individuals

| | No Smokers | Moderate Smokers | Heavy smokers |
|---|---|---|---|
| Hypertension | 21 | 36 | 30 |
| No hypertension | 48 | 26 | 19 |

Test the hypothesis at 5 % LOS that the presence or absence of hypertension is independent of smoking. (Given- χ_tab^2=5.99)

**Solution:**

Ho: Presence or absence of hypertension is independent of smoking.

H1: Presence or absence of hypertension is dependent of smoking.

| | No Smokers O | Moderate Smokers o | Heavy smokers o | |
|---|---|---|---|---|
| Hypertension | 21 | 36 | 30 | RT1=87 |
| No hypertension | 48 | 26 | 19 | RT2= 93 |
| Total=180 | CT1 =69 | CT2=62 | CT3=49 Total=180 | |

**RT=Row Total and CT=Column Total**

| | No Smokers | Moderate | Heavy |
|---|---|---|---|

| | E | Smokers e | smokers e |
|---|---|---|---|
| Hypertension | (RT1 x CT1)/Total | RT1xCT2/Total | RT1xCT3/Total |
| No hypertension | (RT2 x CT1)/Total | (RT2 x CT2)/Total | (RT2 x CT3)/Total |

| | No Smokers E | Moderate Smokers e | Heavy smokers e |
|---|---|---|---|
| Hypertension | 87*69/180 | 87*62/180 | 87*49/180 |
| No hypertension | 93*69/180 | 93*62/180 | 93*49/180 |

| | No Smokers O | Moderate Smokers O | Heavy smokers O | Total |
|---|---|---|---|---|
| Hypertension | 21 | 36 | 30 | 87 |
| No hypertension | 48 | 26 | 19 | 93 |
| Total | 69 | 62 | 49 | 180 |

| o | e | (0-e)2/e |
|---|---|---|
| 21 | 33.35 | 4.5734 |
| 36 | 29.967 | 1.2177 |
| 30 | 23.683 | 1.6849 |
| 48 | 35.65 | 4.2780 |
| 26 | 32.033 | 1.1363 |
| 19 | 25.316 | 1.5761 |
| | | $\chi^2 = \sum \frac{(o-e)^2}{e}$ = 14.46 |

χ^2=14.46

χ_tab^2=5.99

As χ^2> χtab^2, Reject H0 at 5% LOS.

Therefore, we can conclude that Presence or absence of hypertension is dependent of smoking.

The Chi-square test of independence determines whether there is a statistically significant relationship between categorical variables. It is a hypothesis test that answers the question—do the values of one categorical variable depend on the value of other categorical variables? This test is also known as the chi-square test of association.

- **Null hypothesis:** There are no relationships between the categorical variables. If one variable is known, it does not help you predict the value of another variable.
- **Alternative hypothesis:** There are relationships between the categorical variables. Knowing the value of one variable does help you predict the value of another variable.

The Chi-square test of association works by comparing the distribution that you observe to the distribution that you expect if there is no relationship between the categorical variables.

For a Chi-square test, a p-value that is less than or equal to your significance level indicates there is sufficient evidence to conclude that the observed distribution is not the same as the expected distribution. You can conclude that a relationship exists between the categorical variables.

A Chi-square test of independence to determine whether there is a statistically significant association between shirt color and deaths. We need to use this test because these variables are both categorical variables. Shirt color can be only blue, gold, or red. Fatalities can be only dead or alive.

The problem discussed is from https://statisticsbyjim.com/hypothesis-testing/chi-square-test-independence-example/

Eg -The color of the uniform represents each crewmember's work area. We will statistically assess whether there is a connection between uniform color and the fatality rate.

| Color | Areas | Crew | Fatalities |
|---|---|---|---|
| Blue | Science and Medical | 136 | 7 |
| Gold | Command and Helm | 55 | 9 |
| Red | Operations, Engineering, and Security | 239 | 24 |
| Ships' Total | All | 430 | 40 |

For example, we will determine whether the observed counts of deaths by uniform color are different from the distribution that we'd expect if there is no association between the two variables.

| Color | Status | Frequency |
|---|---|---|
| Blue | Dead | 7 |
| Blue | Alive | 129 |
| Gold | Dead | 9 |
| Gold | Alive | 46 |
| Red | Dead | 24 |
| Red | Alive | 215 |

```
Using frequencies in Frequency

Rows: Uniform Color   Columns: Status

          Alive     Dead      All

Blue        129        7      136
          94.85     5.15   100.00      Count < Expected
         123.35    12.65   136.00
         0.2589   2.5243        *

Gold         46        9       55
          83.64    16.36   100.00      Count > Expected
          49.88     5.12    55.00
         0.3024   2.9481        *

Red         215       24      239
          89.96    10.04   100.00      Count = Expected
         216.77    22.23   239.00
         0.0144   0.1405        *

All         390       40      430
          90.70     9.30   100.00
         390.00    40.00   430.00
              *         *        *

Cell Contents:       Count
                     % of Row
                     Expected count
                     Contribution to Chi-square


Pearson Chi-Square = 6.189, DF = 2, P-Value = 0.045
Likelihood Ratio Chi-Square = 6.132  DF = 2, P-Value = 0.047
```

**Both p-values are less than 0.05. Reject the null hypothesis and there is a relationship between shirt color and deaths.**

## 3.10 SUMMARY

In statistics, nonparametric tests are methods of statistical analysis that do not require a distribution to meet the required assumptions to be analyzed (especially if the data is not normally distributed).

It is also referred to as distribution-free tests. Nonparametric tests serve as an alternative to parametric tests such as T-test or ANOVA that can be employed only if the underlying data satisfies certain criteria and assumptions.

## 3.11 UNIT END QUESTIONS

**Q1.**

Using the data given in below table to decide whether we can conclude that standard of a salesman has significant effect on hD performance in field selling at 5% level of significance.

| | Performance in field | | | Total |
|---|---|---|---|---|
| | Disappointing | Satisfactory | Excellent | |
| Poor dressed | 21 | 15 | 6 | 42 |
| Well dressed | 24 | 35 | 26 | 85 |
| Very well dressed | 35 | 80 | 58 | 173 |
| Total | 80 | 130 | 90 | 300 |

**Given- χ_tab^2=9.49**

**Q2.** The PQR Company claims that the lifetime of a type of battery that it manufactures is more than 250 hours (h). A consumer advocate wishing to determine whether the claim is justified measures the lifetimes of 24 of the company's batteries; the results are listed below. Assuming the sample to be random, determine whether the company's claim is justified at the 0.05 significance level. Work the problem first by hand, supplying all the details for the sign test

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 271 | 230 | 198 | 275 | 282 | 225 | 284 | 219 |
| 253 | 216 | 262 | 288 | 236 | 291 | 253 | 224 |
| 264 | 295 | 211 | 252 | 294 | 243 | 272 | 268 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| + | − | − | + | + | − | + | − |
| + | − | + | + | − | + | + | − |
| + | + | − | + | + | − | + | + |

**Q3.**A sample of 40 grades from a statewide examination is shown below. Test the hypothesis at the 0.05 significance level that the median grade for all participants is (a) 66 and (b) 75. Work the problem first by hand, supplying all the details for the sign test.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 71 | 67 | 55 | 64 | 82 | 66 | 74 | 58 | 79 | 61 |
| 78 | 46 | 84 | 93 | 72 | 54 | 78 | 86 | 48 | 52 |
| 67 | 95 | 70 | 43 | 70 | 73 | 57 | 64 | 60 | 83 |
| 73 | 40 | 78 | 70 | 64 | 86 | 76 | 62 | 95 | 66 |

**Q4.**A company wishes to purchase one of five different machines: A, B, C, D, or E. In an experiment designed to determine whether there is a performance difference between the machines, five experienced operators each work on the machines for equal times. The table below shows the number of units produced by each machine. Test the hypothesis that there is no difference between the machines at the (a) 0.05 and (b) 0.01 significance levels. Work the problem first by hand, supplying all the details for the Kruskal–Wallis H test.

| | | | | | |
|---|---|---|---|---|---|
| *A* | 68 | 72 | 77 | 42 | 53 |
| *B* | 72 | 53 | 63 | 53 | 48 |
| *C* | 60 | 82 | 64 | 75 | 72 |
| *D* | 48 | 61 | 57 | 64 | 50 |
| *E* | 64 | 65 | 70 | 68 | 53 |

**Q5.** In 30 tosses of a coin the following sequence of heads (H) and tails (T) is obtained:

H T T H T H H H T H H T T H T

H T H H T H T T H T H H T H T

(a) Determine the number of runs, V.

(b) Test at the 0.05 significance level whether the sequence is random. Work the problem first by hand, supplying all the details of the runs test for randomness.

## 3.12 REFERENCES FOR FUTURE READING


- Schaum' Outline series by Murray Spiegel, fouth edition.
- https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/parametric-and-non-parametric-data/
- https://www.statisticshowto.com/sign-test/
- How to Perform the Wilcoxon Signed Rank Test - Statology
- https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/bs704_nonparametric_print.html
- https://www.statisticshowto.com/kruskal-wallis/
- https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide-4.php


*****