KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

Subject : Mathematical Statistics **Semester III**
Subject Code : 18MMP304 **Class :II M.Sc Mathematics**

Instruction Hours / week: L: 4 T: 0 P: 0 **Marks:** Internal: **40** External: **60** Total: **100**
End Semester Exam: 3 Hours

Course Objectives

- To make the Students to understand the basic concepts in probability generating functions, sample moments and their functions, sampling, significance tests, statistical measures, commonly used probability distributions, significance of testing hypothesis and its interpretation, estimation, ANOVA and their applications in various disciplines.

Course Outcomes (COs)

After successfully completed this module the students will be able to

1. Explain the concepts of probability, including conditional probability.
2. Explain the concepts of random variable, probability distribution, distribution function, expected value, variance and higher moments, and calculate expected values and probabilities associated with the distributions of random variables.
3. Construct confidence intervals for unknown parameters.
4. Explain the concepts of random sampling, statistical inference and sampling distribution, and state and use basic sampling distributions.
5. Test statistical hypotheses.
6. Define a probability generating function, a moment generating function, and use them to evaluate moments.
7. Summarize the main features of a data set.
8. Define basic discrete and continuous distributions, be able to apply them and simulate them in simple cases.
9. Explain the concepts of analysis of variance and use them to investigate factorial dependence.
10. Describe the main methods of estimation and the main properties of estimators, and apply them.

UNIT I

PROBABILITY: Random Events – Preliminary remarks – random events and operations performed on them – the system of axioms of the theory of probability – conditional probability – Baye's theorem – Independent Events –functions of random variables –Introduction to hypothesis testing, type of errors, standard errors, confidence interval, confidence limits. Significance level.

UNIT II

SAMPLE MOMENTS AND THEIR FUNCTIONS: Notion of a sample and a statistic - Distribution functions of X, S^2 and (X , S^2) -Chi-square distribution -Student t-distribution - Fisher's Z-distribution -Snedecor's F -distribution -Distribution of sample mean from non-normal populations.

UNIT III

SIGNIFICANCE TEST: Concept of a statistical test -Parametric tests for small samples and large samples Chi-square test -Kolmogorov Theorem-Smirnov Theorem-Tests of Kolmogorov and Smirnov type The Wald-Wolfovitz and Wilcoxon-Mann-Whitney tests -Independence Tests by contingency tables.

UNIT IV

ESTIMATION: Preliminary notion -Consistency estimation -Unbiased estimates -Sufficiency -Efficiency -Asymptotically most efficient estimates -methods of finding estimates -confidence Interval.

UNIT V

ANALYSIS OF VARIANCE: One way classification and two-way classification. Hypotheses Testing: Poser functions -OC function-Most Powerful test -Uniformly most powerful test -unbiased test.

SUGGESTED READINGS

1. Marek Fisz, (1980). Probability Theory and Mathematical Statistics, John Wiley and Sons, New York.
2. Meyer, (2006). Introduction to Probability and Statistical applications, Oxford and IBH Publishing Co. Pvt Ltd. New Delhi.
3. Sheldon M. Ross, (2009). Introduction to probability and statistics for engineers and scientists, Third edition, Academic press.
4. Heinz Bauer,(1996), Probability Theory, Narosa Publishing House, London.
5. Parimal Mukhopadhyay, (2012). Theory of Probability, New central book agency, Calcutta.

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

Subject : Mathematical Statistics **Semester III** **L T P C**
Subject Code : 18MMP304 **Class : II M.Sc Mathematics** **4 0 0 4**

	UNIT- I		
S.No	**Lecture Duration (Hr)**	**Topics to be Covered**	**Support Materials**
1	1	Basic Probability Concepts, Random events,	R5:Chap.5; Pg. No. 5.4-5.14
2	1	Random events and Operations performed on them	R6:Chap.5; Pg. No. 58-60
3	1	The system of Axioms of the theory of probability	R5:Chap.5; Pg. No. 5.16-5.17
4	1	Conditional Probability and Bayes theorem	R6:Chap.5; Pg. No. 57-58, 84
5	1	Independent events and functions of random variables	T1:Chap.5; Pg.No.104-108
6	1	Introduction to hypothesis testing	R6:Chap.5; Pg. No. 292-293
7	1	Type of errors and standard error	R6:Chap.5; Pg. No. 294-296
8	1	About Normal Distribution, Normal curve and its properties,	R6:Chap.5; Pg. No. 248-251
9	1	Confidence interval, confidence limits and Significance level	R6:Chap.5; Pg. No. 252-256
10	1	Recapitulation and discussion on important questions	
Total 10 hrs			

UNIT- II			
S.No	**Lecture Duration (Hr)**	**Topics to be Covered**	**Support Materials**
1	1	Sample moments and their functions: Notion of a sample and a statistic	R1: Chap16,Pg.No.210-211
2	1	Distribution functions of X, S^2 and (X ,S^2)	T1:Chap.6; Pg.No.130-132
3	1	Chi-square distribution, its properties and applications	R1:Chap.9; Pg. No. 196-198
4	1	Students t-Distribution and its properties	R6:Chap.5; Pg. No. 292-296
5	1	Fishzers Z-distribution and its properties	R6:Chap.8; Pg. No. 368-372
6	1	Snedecor's F-distribution of Sample mean from non-normal population Its properties and problems	R6:Chap.8; Pg. No. 379-378
7	1	Recapitulation and discussion on important questions	
Total 7 hrs			

UNIT- III			
S.No	**Lecture Duration (Hr)**	**Topics to be Covered**	**Support Materials**
1	1	Significance test: Concept of a statistical test, difference between Parametric and non- Parametric tests	R5:Chap.12; Pg. No. 12.1-12.2
2	1	Parametric tests for small samples and large samples	R5:Chap.10; Pg. No. 10.37-10.38
3	1	Problems on small sample test (t – test)	R5:Chap.10; Pg. No. 10.38-10.40
4	1	Problems on large sample test (Z – test)	R5:Chap.10; Pg. No. 10.41-10.45
5	1	Chi-square test with illustration	R5:Ch.10; Pg.No.10.62-68, R3:Ch.11; Pg. No.70-72
6	1	More problems on Chi-square test	R3:Chap.11; Pg. No. 75 -79
7	1	Kolmogorov Theorem-Smirnov Theorem-	R5:Chap.12; Pg. No. 12.10-12.11
8	1	Tests based on Kolmogorov and Smirnov	R5:Chap.12; Pg. No. 12.11-12.12
9	1	Problems on The Wald-Wolfowitz test	R6:Chap.8; Pg. No. 741-753
10	1	Problems on Wilcoxon-Mann-Whitney test	R6:Chap.8; Pg. No. 744-750
11	1	Independence Tests by contingency tables.	R6:Chap.8; Pg. No. 752-755
12	1	Recapitulation and discussion on important questions	
Total 12 hrs			

UNIT- IV			
S.No	**Lecture Duration (Hr)**	**Topics to be Covered**	**Support Materials**
1	1	Introduction to Estimation theory and its concepts, -Preliminary notations	R5:Chap.10; Pg. No. 10.3-10.5
2	1	Consistency estimation and unbiased estimates	R5:Chap.10; Pg. No. 10.6-10.7
3	1	Sufficiency estimates with examples	R4:Chap.11; Pg. No. 156-158
4	1	Efficiency estimates with examples	R4:Chap.11; Pg. No. 158-162
5	1	Asymptotically most efficient estimates	R5:Chap.10; Pg. No. 10.12-10.15
6	1	Methods of finding Estimates and Confidence Interval	R5:Chap.10; Pg. No. 10.11-10.13
7	1	Recapitulation and discussion on important questions	
Total 7 hrs			

UNIT- V			
S.No	**Lecture Duration (Hr)**	**Topics to be Covered**	**Support Materials**
1	1	Analysis of variance-Introduction and one way classification	R5:Chap.11; Pg. No. 11.1-11.13
2	1	Analysis of variance - Two way classification with illustration	R6:Chap.8; Pg. No. 435-436
3		Problems on ANOVA	R6:Chap.8; Pg. No. 436-437
4	1	Hypotheses testing-Power Functions	T1:Chap.16,P.No.432-438
5		Hypotheses testing- OC Functions	T1:Chap.16,P.No.438-440
6	1	Most powerful test	T1:Chap.16,P.No.440-445
7		Uniformly most powerful test	T1:Chap.16,P.No.445-450
8	1	Unbiased tests	T1:Chap.16,P.No.450-452
9	1	Recapitulation and discussion on important questions	
10	1	Discussion on important questions from previous year ESE question paper.	
11	1	Discussion on important questions from previous year ESE question paper.	
12	1	Discussion on important questions from previous year ESE question paper.	
Total 12 hrs			
Total Lecture Hours Planned : 48			

TEXT BOOK

1) T1. Marek Fisz, 1980. Probability Theory and Mathematical Statistics, John Wiley and Sons, New York.

REFERENCES

1) Meyer, 1969. Introduction to Probability and Statistical applications, Oxford and IBH Publishing Co. Pvt Ltd. New Delhi.
2) Sheldon M. Ross, 1995. Introduction to probability and statistics for engineers and scientists, Third edition, Academic press.
3) Heinz Bauer, 1995.Probality Theory, Narosa Publishing House, London.
4) Parimal Mukhopadhyay, 1991. Theory of Probability, New central book agency, Calcutta.
5) T.N. Srivastava & Shailaja Rego; Statistics for Management, 2nd Edition.; Mc Graw Hill Education Pvt. Ltd.
6) Aczel.A.D and Sounderpandian.J (2012), Complete Business Statistics (7th Edition.) McGraw Hill Education

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

Subject : Mathematical Statistics	**Semester III**	**L T P C**
Subject Code : 18MMP304	**Class : II M.Sc Mathematics**	**4 0 0 4**

UNIT I

Probability: Random Events – Preliminary remarks – random events and operations performed on them – the system of axioms of the theory of probability – conditional probability – Bayes theorem – Independent Events –functions of random variables –Introduction to hypothesis testing, type of errors, standard errors, confidence interval, confidence limits. Significance level.

SUGGESTED READINGS

1. Marek Fisz, (1980). Probability Theory and Mathematical Statistics, John Wiley and Sons, New York.
2. Meyer, (2006). Introduction to Probability and Statistical applications, Oxford and IBH Publishing Co. Pvt Ltd. New Delhi.
3. Sheldon M. Ross, (2009). Introduction to probability and statistics for engineers and scientists, Third edition, Academic press.
4. Heinz Bauer,(1996),Probability Theory, Narosa Publishing House, London.
5. Parimal Mukhopadhyay, (2012). Theory of Probability, New central book agency, Calcutta.

PROBABILITY

Introduction

The theory of probability has its origin in the games of chance related to gambling such as tossing of a coin, throwing of a die, drawing cards from a pack of cards etc. Jerame Cardon, an Italian mathematician wrote ‘ A book on games of chance’ which was published on 1663. Starting with games of chance, probability has become one of the basic tools of statistics. The knowledge of probability theory makes it possible to interpret statistical results, since many statistical procedures involve conclusions based on samples.

Probability theory is being applied in the solution of social, economic, business problems. Today the concept of probability has assumed greater importance and the mathematical theory of probability has become the basis for statistical applications in both social and decision-making research. Probability theory, in fact, is the foundation of statistical inferences.

Definitions and basic concepts

The following definitions and terms are used in studying the theory of probability.

Random experiment

Random experiment is one whose results depend on chance, that is the result cannot be predicted. Tossing of coins, throwing of dice are some examples of random experiments.

Trial

Performing a random experiment is called a trial.

Outcomes

The results of a random experiment are called its outcomes. When two coins are tossed the possible outcomes are HH, HT, TH, TT.

Event

An outcome or a combination of outcomes of a random experiment is called an event. For example tossing of a coin is a random experiment and getting a head or tail is an event.

Sample space

Each conceivable outcome of an experiment is called a sample point. The totality of all sample points is called a sample space and is denoted by **S**. For example, when a coin is tossed, the sample space is S = { H, T }. H and T are the sample points of the sample space S.

Equally likely events

Two or more events are said to be equally likely if each one of them has an equal chance of occurring. For example in tossing of a coin, the event of getting a head and the event of getting a tail are equally likely events.

Mutually exclusive events

Two or more events are said to be mutually exclusive, when the occurrence of any one event excludes the occurrence of the other event. Mutually exclusive events cannot occur simultaneously.

For example when a coin is tossed, either the head or the tail will come up. Therefore the occurrence of the head completely excludes the occurrence of the tail. Thus getting head or tail in tossing of a coin is a mutually exclusive event.

Exhaustive events

Events are said to be exhaustive when their totality includes all the possible outcomes of a random experiment. For example, while throwing a die, the possible outcomes are {1, 2, 3, 4, 5, 6} and hence the number of cases is 6.

Complementary events

The event ' A occurs' and the event ' A does not occur' are called complementary events to each other. The event ' A does not occur' is denoted by A' or $\bar{A}$ or A^c. The event and its complements are mutually exclusive. For example in throwing a die, the event of getting odd numbers is { 1, 3, 5 } and getting even numbers is

{2, 4, 6 }These two events are mutually exclusive and complement to each other.

Independent events

Events are said to be independent if the occurrence of one does not affect the others. In the experiment of tossing a fair coin, the occurrence of the event ‘ head’ in the first toss is independent of the occurrence of the event ‘ head’ in the second toss, third toss and subsequent tosses.

Definitions of Probability

There are two types of probability. They are Mathematical probability and Statistical probability.

Mathematical Probability (or a priori probability)

If the probability of an event can be calculated even before the actual happening of the event, that is, even before conducting the experiment, it is called***Mathematical probability.***

If the random experiments results in µ Q exhaustive, mutually exclusive and equally likely cases, out of which µ P are favorable to the occurrence of an event A, then the ratio *m/n* is called the probability of occurrence of event A, denoted by P(A), is given by

$$P(A) = \frac{m}{n} = \frac{\text{Number of cases favourable to the event A}}{\text{Total number of exhaustive cases}}$$

Mathematical probability is often called ***classical probability*** or a ***priori probability*** because if we keep using the examples of tossing of fair coin, dice etc., we can state the answer in advance (*prior*), without tossing of coins or without rolling the dice etc.,

The above definition of probability is widely used, but it cannot be applied under the following situations:

(1) If it is not possible to enumerate all the possible outcomes for an experiment.
(2) If the sample points(outcomes) are not mutually independent.
(3) If the total number of outcomes is infinite.
(4) If each and every outcome is not equally likely.

Some of the drawbacks of classical probability are removed in another definition given below:

Statistical Probability (or a posteriori probability)

If the probability of an event can be determined only after the actual happening of the event, it is called ***Statistical probability.***

If an event occurs ***m*** times out of ***n***, its relative frequency is ***m/n.***

In the limiting case, when ***n*** becomes sufficiently large it corresponds to a number which is called the probability of that event.

In symbol, $P(A) = \lim_{n \to \infty} (m/n)$

The above definition of probability involves a concept which has a long term consequence. This approach was initiated by the mathematician Von Mises .

If a coin is tossed 10 times we may get 6 heads and 4 tails or 4 heads and 6 tails or any other result. In these cases the probability of getting a head is **not 0.5** as we consider in Mathematical probability.

However, if the experiment is carried out a large number of times we should expect approximately equal number of heads and tails and we can see that the probability of getting head approaches 0.5. The Statistical probability calculated by conducting an actual experiment is also called a ***posteriori probability*** or ***empirical probability.***

Axiomatic approach to probability

The modern approach to probability is purely axiomatic and it is based on the set theory. The axiomatic approach to probability was introduced by the Russian mathematician A.N. Kolmogorov in the year 1933.

Axioms of probability

Let S be a sample space and A be an event in S and P(A) is the probability satisfying the following axioms:

(1) The probability of any event ranges from zero to one.
i.e $0 \leq P(A) \leq 1$

(2) The probability of the entire space is 1.
i.e $P(S) = 1$

(3) If $A_1, A_2, ..$ is a sequence of mutually exclusive events in S, then

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$$

Interpretation of statistical statements in terms of set theory

$S \Rightarrow$ Sample space

$\bar{A} \Rightarrow$ A does not occur

$A \cup \bar{A} = S$

$A \cap B = \phi \Rightarrow$ A and B are mutually exclusive.

$A \cup B \Rightarrow$ Event A occurs or B occurs or both A and B occur. (at least one of the events A or B occurs)

$A \cap B \Rightarrow$ Both the events A and B occur.

$\bar{A} \cap \bar{B} \Rightarrow$ Neither A nor B occurs

$A \cap \bar{B} \Rightarrow$ Event A occurs and B does not occur

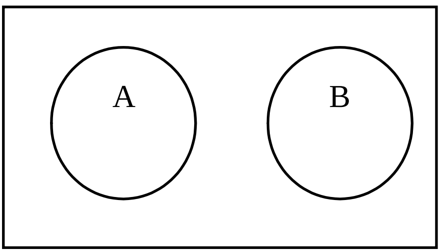
$\bar{A} \cap B \Rightarrow$ Event A does not occur and B occur.

Addition theorem on probabilities

We shall discuss the addition theorem on probabilities for mutually exclusive events and not mutually exclusive events.

Addition theorem on probabilities for mutually exclusive events

If two events A and B are mutually exclusive, the probability of the occurrence of either A or B is the sum of individual probabilities of A and B. ie $P(AUB) = P(A) + P(B)$
This is clearly stated in axioms of probability.



Addition theorem on probabilities for not-mutually exclusive events

If two events A and B are not-mutually exclusive, the probability of the event that either A or B or both occur is given as

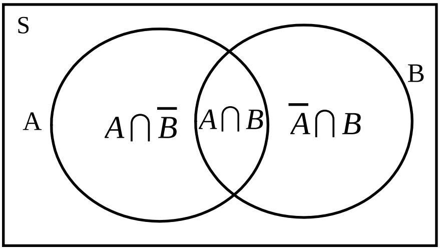
$$P(AUB) = P(A) + P(B) - P(A \cap B)$$

Proof

Let us take a random experiment with a sample space S of N sample points.

Then by the definition of probability ,

$$P(AUB) = \frac{n(AUB)}{n(S)} = \frac{n(AUB)}{N}$$



From the diagram, using the axiom for the mutually exclusive events, we write

$$P(AUB) = \frac{n(A) + n(\overline{A} \cap B)}{N}$$

Adding and subtracting n($A \cap B$) in the numerator,

$$= \frac{n(A) + n(\overline{A} \cap B) + n(A \cap B) - n(A \cap B)}{N}$$

$$= \frac{n(A) + n(B) - n(A \cap B)}{N}$$

$$= \frac{n(A)}{N} + \frac{n(B)}{N} - \frac{n(A \cap B)}{N}$$

$$P(AUB) = P(A) + P(B) - P(A \cap B)$$

Note

In the case of three events A,B,C, P(AUBUC) = P(A) + P(B) + P(C) – P($A \cap B$) – P($A \cap B$) – P($B \cap C$) + P ($A \cap B \cap C$)

Compound events

The joint occurrence of two or more events is called compound events. Thus compound events imply the simultaneous occurrence of two or more simple events.

For example, in tossing of two fair coins simultaneously, the event of getting ‘ atleast one head’ is a compound event as it consists of joint occurrence of two simple events.

Namely,

Event A = one head appears ie A = { HT, TH} and

Event B = two heads appears ie B = {HH}

Similarly, if a bag contains 6 white and 6 red balls and we make a draw of 2 balls at random, then the events that ‘ both are white’ or one is white and one is red’ are compound events.

The compound events may be further classified as

(1) Independent event
(2) Dependent event

Independent events

If two or more events occur in such a way that the occurrence of one does not affect the occurrence of another, they are said to be independent events.

For example, if a coin is tossed twice, the results of the second throw would in no way be affected by the results of the first throw.

Similarly, if a bag contains 5 white and 7 red balls and then two balls are drawn one by one in such a way that the first ball is replaced before the second one is drawn. In this situation, the two events, ‘ the first ball is white’ and ‘ second ball is red’ , will be independent, since the composition of the balls in the bag remains unchanged before a second draw is made.

Dependent events

If the occurrence of one event influences the occurrence of the other, then the second event is said to be dependent on the first.

In the above example, if we do not replace the first ball drawn, this will change the composition of balls in the bag while making the second draw and therefore the event of ‘ drawing a red ball’ in the second will depend on event (first ball is red or white) occurring in first draw.

Similarly, if a person draw a card from a full pack and does not replace it, the result of the draw made afterwards will be dependent on the first draw.

Conditional probability

Let A be any event with p(A) >0. The probability that an event B occurs subject to the condition that A has already occurred is known as the conditional probability of occurrence of the event B on the assumption that the event A has already occurred and is denoted by the symbol P(B/A) or P(B|A) and is read as the probability of B given A.

The same definition can be given as follows also:

Two events A and B are said to be dependent when A can occur only when B is known to have occurred (or vice versa). The probability attached to such an event is called the **conditional probability** and is denoted by P(B/A) or, in other words, probability of B given that A has occurred.

If two events A and B are dependent, then the conditional probability of B given A is

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

Similarly the conditional probability of A given B is given as

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Note

If the events A and B are independent, that is the probability of occurrence of any one of them P(A/B) = P(A) and P(B/A) = P(B)

Multiplication theorem on probabilities

We shall discuss multiplication theorem on probabilities for both independent and dependent events.

Multiplication theorem on probabilities for independent events

If two events A and B are independent, the probability that both of them occur is equal to the product of their individual probabilities. i.e $P(A\cap B) = P(A).P(B)$

Proof:

Out of n_1 possible cases let m_1 cases be favourable for the occurrence of the event A.

$$\therefore P(A) = \frac{m_1}{n_1}$$

Out of n_2 possible cases, let m_2 cases be favourable for the occurrence of the event B

$$\therefore P(B) = \frac{m_2}{n_2}$$

Each of n_1 possible cases can be associated with each of the n_2 possible cases.

Therefore the total number of possible cases for the occurrence of the event 'A' and 'B' is $n_1 \times n_2$. Similarly each of the m_1 favourable cases can be associated with each of the m_2 favourable cases. So the total number of favourable cases for the event 'A' and 'B' is $m_1 \times m_2$

$$\therefore P(A\cap B) = \frac{m_1\, m_2}{n_1\, n_2}$$

$$= \frac{m_1}{n_1} \cdot \frac{m_2}{n_2}$$

$$= P(A).P(B)$$

Note:

The theorem can be extended to three or more independent events. If A,B,C……. be independent events, then

$P(A\cap B\cap C\text{…….}) = P(A).P(B).P(C)\text{…….}$

Note

If A and B are independent then the complements of A and B are also independent. i.e $P(\overline{A} \cap \overline{B}) = P(\overline{A}) . P(\overline{B})$

1.5.2 Multiplication theorem for dependent events:

If A and B be two dependent events, i.e the occurrence of one event is affected by the occurrence of the other event, then the probability that both A and B will occur is

$$P(A \cap B) = P(A)\, P(B/A)$$

Proof:

Suppose an experiment results in n exhaustive, mutually exclusive and equally likely outcomes, m of them being favourable to the occurrence of the event A.

Out of these n outcomes let m_1 be favourable to the occurrence of another event B.

Then the outcomes favourable to the happening of the events ' A and B' are m_1.

$$\therefore P(A \cap B) = \frac{m_1}{n}$$

$$= \frac{m_1}{n} \times \frac{m}{m} = \frac{m\,m_1}{n\,m}$$

$$= \frac{m}{n} \times \frac{m_1}{m}$$

$$\therefore P(A \cap B) = P(A) . P(B/A)$$

Note

In the case of three events A, B, C, $P(A \cap B \cap C) = P(A). P(B/A). P(C/A \cap B)$. ie., the probability of occurrence of A, B and C is equal to the probability of A times the probability of B given that A has occurred, times the probability of C given that both A and B have occurred.

BAYES' Theorem

The concept of conditional probability discussed earlier takes into account information about the occurrence of one event to predict the probability of another event. This concept can be extended to revise probabilities based on new information and to determine the probability that a particular effect was due to specific cause. The procedure for revising these probabilities is known as Bayes theorem.

The Principle was given by Thomas Bayes in 1763. By this principle, assuming certain prior probabilities, the posteriori probabilities are obtained. That is why Bayes' probabilities are also called posteriori probabilities.

Bayes' Theorem or Rule (Statement only):

Let $A_1, A_2, A_3, \ldots, A_i, \ldots, A_n$ be a set of n mutually exclusive and collectively exhaustive events and $P(A_1)$, $P(A_2)$…, $P(A_n)$ are their corresponding probabilities. If B is another event such that P(B) is not zero and the priori probabilities $P(B \mid A_i)$ i =1,2…, n are also known. Then

$$P(A_i \mid B) = \frac{P(B \mid A_i)\, P(A_i)}{\sum_{i=1}^{k} P(B \mid A_i)\, P(A_i)}$$

Basic principles of Permutation and Combination

Factorial

The consecutive product of first *n* natural numbers is known as ***factorial*** **n** and is denoted as **n!** or $\angle n$

That is $n! = 1 \times 2 \times 3 \times 4 \times 5 \ldots \times n$

$3! = 3 \times 2 \times 1$

$4! = 4 \times 3 \times 2 \times 1$

$5! = 5 \times 4 \times 3 \times 2 \times 1$

Also $5! = 5 \times (4 \times 3 \times 2 \times 1) = 5 \times (4!)$

Therefore this can be algebraically written as $n! = n \times (n-1)!$

Note that $1! = 1$ and $0! = 1$.

Permutations

Permutation means arrangement of things in different ways. Out of three things A, B, C taking two at a time, we can arrange them in the following manner.

A B B A

A C C A
B C C B

Here we find 6 arrangements. In these arrangements order of arrangement is considered. The arrangement AB and the other arrangement BA are different.

The number of arrangements of the above is given as the number of permutations of 3 things taken 2 at a time which gives the value 6. This is written symbolically, $_3P_2 = 6$

Thus the number of arrangements that can be made out of *n* things take n*r* at a time is known as the number of permutation of *n* things take n*r* at a time and is denoted as nPr.

The expansion of nPr is given below:

$nPr = n(n-1)(n-2)\ldots\ldots\ldots\ldots\ldots[n-(r-1)]$

The same can be written in factorial notation as follows:

$$nPr = \frac{!n}{(n-r)!}$$

For example, to find $_{10}P_3$ we write this as follows:

$$\begin{aligned} {}_{10}P_3 &= 10(10-1)(10-2) \\ &= 10 \times 9 \times 8 \\ &= 720 \end{aligned}$$

[To find $_{10}P_3$, Start with 10, write the product of 3 consecutive natural numbers in the descending order]

Simplifying $_{10}P_3$ using factorial notation:

$$_{10}P_3 = \frac{!10}{(10-3)!} = \frac{10\times9\times8\times7\times6\times5\times4\times3\times2\times1}{7\times6\times5\times4\times3\times2\times1}$$

$$= 10 \times 9 \times 8$$
$$= 720$$

Note that $nP_o = 1$, $nP_1 = n$, $nP_n = n!$

Combinations

A combination is a selection of objects without considering the order of arrangements.

For example, out of three things A,B,C we have to select two things at a time.

This can be selected in three different ways as follows:

A B A C B C

Here the selection of the object A B and B A are one and the same. Hence the order of arrangement is not considered in combination. Here the number of combinations from 3 different things taken 2 at a time is 3.

This is written symbolically $_3C_2 = 3$

Thus the number of combination of n different things, taken r at a time is given by $nCr = \dfrac{n\,Pr}{!r}$

Or $nCr = \dfrac{!n}{(n-r)!\ !r}$

Note that $nC_0=1$, $nC_1 = n$, $nC_n = 1$

Find $_{10}C_3$. $_{10}C_3 = \dfrac{_{10}P_3}{!3} = \dfrac{10\times 9\times 8}{1\times 2\times 3} = 120$

Find $_8C_4$ $_8C_4 = \dfrac{8\times 7\times 6\times 5}{1\times 2\times 3\times 4} = 70$

[To find $_8C_4$ **:** In the numerator, first write the product of 4 natural numbers starting with 8 in descending order and in the denominator write the factorial 4 and then simplify.]

Compare $_{10}C_8$ and $_{10}C_2$

$$_{10}C_8 = \frac{10\times 9\times \cancel{8}\times \cancel{7}\times \cancel{6}\times \cancel{5}\times \cancel{4}\times \cancel{3}}{1\times 2\times \cancel{3}\times \cancel{4}\times \cancel{5}\times \cancel{6}\times \cancel{7}\times \cancel{8}} = \frac{10\times 9}{1\times 2} = 45$$

$$_{10}C_2 = \frac{10\times 9}{1\times 2} = 45$$

From the above, we find $_{10}C_8 = {_{10}C_2}$

This can be got by the following method also:

$_{10}C_8 = {}_{10}C_{(10-8)} = {}_{10}C_2$

This method is very useful, when the difference between *n* and *r* is very high in *n*C*r*.

This property of the combination is written as $_nC_r = {}_nC_{(n-r)}$.

To find $_{200}C_{198}$ we can use the above formula as follows:

$$_{200}C_{198} = {}_{200}C_{(200-198)} = {}_{200}C_2 = \frac{200 \times 199}{1 \times 2} = 19900.$$

Example:

Out of 13 players, 11 players are to be selected for a cricket team. In how many ways can this be done?

Out of 13 players, 11 players are selected in $_{13}C_{11}$ ways

i.e. $_{13}C_{11} = {}_{13}C_2 = \frac{13 \times 12}{1 \times 2} = 78.$

Example 1:

Three coins are tossed simultaneously Find the probability that

(i) no head (ii) one head (iii) two heads

(iv) atleast two heads. (v) atmost two heads appear.

Solution:

The sample space for the 3 coins is

S = { HHH, HHT, HTH, HTT, THH, THT, TTH, TTT} ; n(S) = 8

(i) No head appear A = {TTT}; n(A) = 1

$\therefore P(A) = \frac{1}{8}$

(ii) One head appear B = {HTT, THT, TTH}; n (B) = 3

$\therefore P(B) = \frac{3}{8}$

(iii) Two heads appear C = {HHT, HTH, THH}; n(c)=3

$\therefore P(C) = \frac{3}{8}$

(iv) Atleast two heads appear

D = { HHT, HTH, THH, HHH}; n(D) = 4

$$\therefore\ P(D) = \frac{4}{8} = 1/2$$

(v) Atmost two heads appear E = { TTT, HTT, THT, TTH,HHT, HTH,THH}

n(E)= 7

$$\therefore P(E) = \frac{7}{8}$$

Example 2:

When two dice are thrown, find the probability of getting doublets (Same number on both dice)

Solution:

When two dice are thrown, the number of points in the sample space is n(S) = 36

Getting doublets: A = {(1,1) , (2,2) , (3,3) , (4,4) , (5,5) , (6,6)}

$$\therefore P(A) = \frac{6}{36} = \frac{1}{6}$$

Example 3:

A card is drawn at random from a well shuffled pack of 52 cards. What is the probability that it is (i) an ace (ii) a diamond card

Solution:

We know that the Pack contains 52 cards $\therefore$n(S)= 52

(i) There are 4 aces in a pack. n(A) = 4

$$\therefore P(A) = \frac{4}{52} = \frac{1}{13}$$

(ii) There are 13 diamonds in a pack $\therefore$n(B) = 13

$$\therefore\ P(B) = \frac{13}{52} = \frac{1}{4}$$

Example 4:

A ball is drawn at random from a box containing 5 green, 6 red, and 4 yellow balls. Determine the probability that the ball drawn is (i) green (ii) Red (iii) yellow (iv) Green or Red (v) not yellow.

Solution:

Total number of balls in the box = 5+6+4 = 15 balls

(i) Probability of drawing a green ball $= \frac{5}{15} = \frac{1}{3}$

(ii) Probability of drawing a red ball $= \frac{6}{15} = \frac{2}{5}$

(iii) Probability of drawing a yellow ball $= \frac{4}{15}$

(iv) Probability of drawing a Green or a Red ball

$$= \frac{5}{15} + \frac{6}{15} = \frac{11}{15}$$

(v) Probability of getting not yellow $= 1 - P\text{ (yellow)}$

$$= 1 - \frac{4}{15}$$

$$= \frac{11}{15}$$

Example 5:

Two dice are thrown, what is the probability of getting the sum being 8 or the sum being 10?

Solution:

Number of sample points in throwing two dice at a time is n(S)=36

Let A= {the sum being 8}

$\therefore$ A= {(6,2), (5,3) , (4,4), (3,5) , (2,6)}; $P(A) = \frac{5}{36}$

B = { the sum being 10}

$\therefore$ B = {(6,4), (5,5) (4,6)} ; $P(B) = \frac{3}{36}$

$A \cap B = \{ 0 \}$; $n(A \cap B) = 0$

$\therefore$ The two events are mutually exclusive

$$\therefore P(AUB) = P(A) + P(B)$$

$$= \frac{5}{36} + \frac{3}{36}$$

$$= \frac{8}{36} = \frac{2}{9}$$

Example 6 :

Two dice are thrown simultaneously. Find the probability that the sum being 6 or same number on both dice.

Solution:

n(S) = 36

The total is 6:

$\therefore$ A = {(5,1) , (4,2), (3,3) , (2,4) , (1,5)}; P(A) = $\frac{5}{36}$

Same number on both dice:

$\therefore$B = {(1,1) (2,2), (3,3), (4,4), (5,5), (6,6)}; P(B) = $\frac{6}{36}$

A$\cap$B = {(3,3)} ; P(A B) =$\frac{1}{36}$

Here the events are not mutually exclusive.

$\therefore$ P(AUB) = P(A) + P(B) – P(A$\cap$B)

$$= \frac{5}{36} + \frac{6}{36} - \frac{1}{36}$$

$$= \frac{5+6-1}{36}$$

$$= \frac{11-1}{36}$$

$$= \frac{10}{36}$$

$$= \frac{5}{18}$$

Example 7:

Two persons A and B appeared for an interview for a job. The probability of selection of A is 1/3 and that of B is 1/2. Find the probability that

(i) both of them will be selected
(ii) only one of them will be selected
(iii) none of them will be selected

Solution:

P(A)= $\frac{1}{3}$, P(B) = $\frac{1}{2}$

P($\overline{A}$) = $\frac{2}{3}$ and P($\overline{B}$) = $\frac{1}{2}$

Selection or non-selection of any one of the candidate is not affecting the selection of the other. Therefore A and B are independent events.

(i) Probability of selecting both A and B

$$P(A\cap B) = P(A).P(B)$$

$$= \frac{1}{3} \times \frac{1}{2}$$

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

(ii) Probability of selecting any one of them

= P (selecting A and not selecting B) + P(not selecting A and selecting B)

i.e $P(A\cap \overline{B}) + P(\overline{A}\cap B) = P(A).\ P(\overline{B}) + P(\overline{A}).\ P(B)$

$$= \frac{1}{3} \times \frac{1}{2} + \frac{2}{3} \times \frac{1}{2}$$

$$= \frac{1}{6} + \frac{2}{6}$$

$$= \frac{3}{6} = \frac{1}{2}$$

(iii) Probability of not selecting both A and B

i.e $P(\overline{A}\cap \overline{B})$

$$= P(\overline{A}).\ P(\overline{B})$$

$$= \frac{2}{3} \cdot \frac{1}{2}$$

$$= \frac{1}{3}$$

TEST OF SIGNIFICANCE

Introduction

It is not easy to collect all the information about population and also it is not possible to study the characteristics of the entire population (finite or infinite) due to time factor, cost factor and other constraints. Thus we need sample. Sample is a finite subset of statistical individuals in a population and the number of individuals in a sample is called the sample size.

Sampling is quite often used in our day-to-day practical life. For example in a shop we assess the quality of rice, wheat or any other commodity by taking a handful of it from the bag and then to decide to purchase it or not.

Parameter and Statistic

The statistical constants of the population such as mean, (μ), variance (σ^2), correlation coefficient (ρ) and proportion (P) are called ‘ Parameters’ .

Statistical constants computed from the samples corresponding to the parameters namely mean ($\bar{x}$), variance (S^2), sample correlation coefficient (r) and proportion (p) etc, are called statistic.

Parameters are functions of the population values while statistic are functions of the sample observations. In general, population parameters are unknown and sample statistics are used as their estimates.

Sampling Distribution

The distribution of all possible values which can be assumed by some statistic measured from samples of same size ‘ n’ randomly drawn from the same population of size N, is called as sampling distribution of the statistic (DANIEL and FERREL).

Consider a population with N values .Let us take a random sample of size n from this population, then there are

$$NC_n = \frac{N!}{n!(N-n)!} = k \text{ (say), possible samples. From each of}$$

these k samples if we compute a statistic (e.g mean, variance, correlation coefficient, skewness etc) and then we form a frequency distribution for these k values of a statistic. Such a distribution is called sampling distribution of that statistic.

For example, we can compute some statistic $t = t(x_1, x_2, \ldots x_n)$ for each of these k samples. Then $t_1, t_2, \ldots, t_k$ determine the sampling distribution of the statistic t. In other words statistic *t* may be regarded as a random variable which can take the values $t_1, t_2, \ldots, t_k$ and we can compute various statistical constants like mean, variance, skewness, kurtosis etc., for this sampling distribution.

The mean of the sampling distribution t is

$$\bar{t} = \frac{1}{K}[t_1 + t_2 + \ldots\ldots + t_k] = \frac{1}{K}\sum_{i=1}^{k} t_i$$

$$\text{and var } (t) = \frac{1}{K}\left[(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + \ldots\ldots\ldots + (t_k - \bar{t})^2\right]$$

$$= \frac{1}{K}\sum (t_i - \bar{t})^2$$

Standard Error

The standard deviation of the sampling distribution of a statistic is known as its standard error. It is abbreviated as S.E. For example, the standard deviation of the sampling distribution of the mean $\bar{x}$ known as the standard error of the mean,

$$\text{Where } v(\bar{x}) = v\left(\frac{x_1 + x_2 + \ldots\ldots\ldots\ldots x_n}{n}\right)$$

$$= \frac{v(x_1)}{n^2} + \frac{v(x_2)}{n^2} + \ldots\ldots\ldots + \frac{v(x_n)}{n^2}$$

$$= \frac{\sigma^2}{n^2} + \frac{\sigma^2}{n^2} + \ldots\ldots + \frac{\sigma^2}{n^2} = \frac{n\sigma^2}{n^2}$$

$\therefore$ The S.E. of the mean is $\frac{\sigma}{\sqrt{n}}$

The standard errors of the some of the well known statistic for large samples are given below, where n is the sample size, σ^2 is the population variance and P is the population proportion and $Q = 1 - P$. n_1 and n_2 represent the sizes of two independent random samples respectively.

Sl.No	Statistic	Standard Error
1.	Sample mean $\bar{x}$	$\dfrac{\sigma}{\sqrt{n}}$
2.	Observed sample proportion p	$\sqrt{\dfrac{PQ}{n}}$
3.	Difference between of two samples means $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
4.	Difference of two sample proportions $p_1 - p_2$	$\sqrt{\dfrac{P_1Q_1}{n_1} + \dfrac{P_2Q_2}{n_2}}$

Uses of standard error

i) Standard error plays a very important role in the large sample theory and forms the basis of the testing of hypothesis.

ii) The magnitude of the S.E gives an index of the precision of the estimate of the parameter.

iii) The reciprocal of the S.E is taken as the measure of reliability or precision of the sample.

iv) S.E enables us to determine the probable limits within which the population parameter may be expected to lie.

Remark:

S.E of a statistic may be reduced by increasing the sample size but this results in corresponding increase in cost, labour and time etc.,

Null Hypothesis and Alternative Hypothesis

Hypothesis testing begins with an assumption called a Hypothesis, that we make about a population parameter. A hypothesis is a supposition made as a basis for reasoning. The conventional approach to hypothesis testing is not to construct a single hypothesis about the population parameter but rather to set up two different hypothesis. So that of one hypothesis is accepted, the other is rejected and vice versa.

Null Hypothesis

A hypothesis of no difference is called null hypothesis and is usually denoted by H_0 " Null hypothesis is the hypothesis" which is tested for possible rejection under the assumption that it is true " by Prof. R.A. Fisher. It is very useful tool in test of significance. For example: If we want to find out whether the special classes (for Hr. Sec. Students) after school hours has benefited the students or not. We shall set up a null hypothesis that "H_0: special classes after school hours has not benefited the students".

Alternative Hypothesis

Any hypothesis, which is complementary to the null hypothesis, is called an alternative hypothesis, usually denoted by H_1, For example, if we want to test the null hypothesis that the population has a specified mean μ_0 (say),

i.e., : Step 1: null hypothesis $H_0: \mu = \mu_0$

then 2. Alternative hypothesis may be

i) $H_1: \mu \neq \mu_0$ (ie $\mu > \mu_0$ or $\mu < \mu_0$)

ii) $H_1: \mu > \mu_0$

iii) $H_1: \mu < \mu_0$

the alternative hypothesis in (i) is known as a two – tailed alternative and the alternative in (ii) is known as right-tailed (iii) is known as left –tailed alternative respectively. The settings of alternative hypothesis is very important since it enables us to decide whether we have to use a single – tailed (right or left) or two tailed test.

Level of significance and Critical value

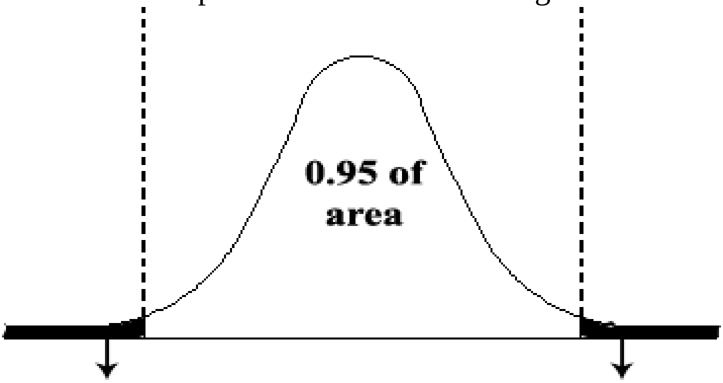
Level of significance

In testing a given hypothesis, the maximum probability with which we would be willing to take risk is called level of significance of the test. This probability often denoted by "α" is generally specified before samples are drawn.

The level of significance usually employed in testing of significance are 0.05(or 5 %) and 0.01 (or 1 %). If for example a 0.05 or 5 % level of significance is chosen in deriving a test of hypothesis, then there are about 5 chances in 100 that we would reject the hypothesis when it should be accepted. (i.e.,) we are about 95 % confident that we made the right decision. In such a case we say that the hypothesis has been rejected at 5 % level of significance which means that we could be wrong with probability 0.05.

The following diagram illustrates the region in which we could accept or reject the null hypothesis when it is being tested at 5 % level of significance and a two-tailed test is employed.

Accept the null hypothesis if the sample statistics falls in this region



Reject the null hypothesis if the sample Statistics falls in these two region

Note: Critical Region: A region in the sample space S which amounts to rejection of H_0 is termed as critical region or region of rejection.

Critical Value

The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the critical value or significant value. It depends upon i) the level of significance used and ii) the alternative hypothesis, whether it is two-tailed or single-tailed

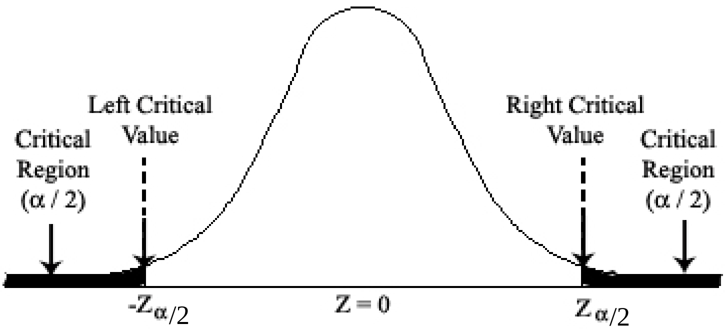
For large samples the standard normal variate corresponding to the statistic t, $Z = \left|\frac{t - E(t)}{S.E.(t)}\right| \sim N(0,1)$

asymptotically as $n \to \infty$

The value of z under the null hypothesis is known as test statistic. The critical value of the test statistic at the level of significance α for a two - tailed test is given by $Z_{\alpha/2}$ and for a one tailed test by Z_α. where Z_α is determined by equation $P(|Z| > Z_\alpha) = \alpha$

$Z\alpha$ is the value so that the total area of the critical region on both tails is α $\therefore$ $P(Z > Z_\alpha) = \frac{\alpha}{2}$. Area of each tail is $\frac{\alpha}{2}$.

$Z\alpha$ is the value such that area to the right of $Z\alpha$ and to the left of $-Z\alpha$ is $\frac{\alpha}{2}$ as shown in the following diagram.



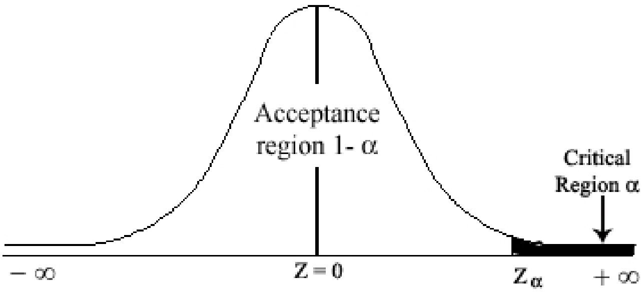
One tailed and Two Tailed tests

In any test, the critical region is represented by a portion of the area under the probability curve of the sampling distribution of the test statistic.

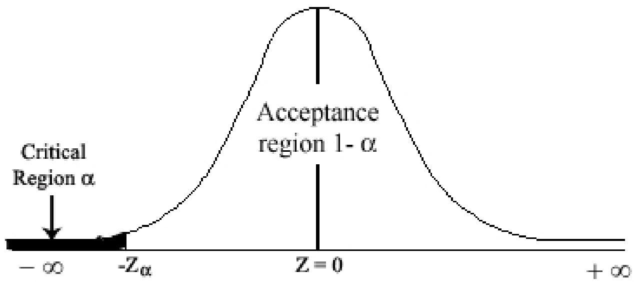
One tailed test: A test of any statistical hypothesis where the alternative hypothesis is one tailed (right tailed or left tailed) is called a one tailed test.

For example, for testing the mean of a population $H_0: \mu = \mu_0$, against the alternative hypothesis $H_1: \mu > \mu_0$ (right – tailed) or $H_1: \mu < \mu_0$ (left –tailed)is a single tailed test. In the right – tailed test $H_1: \mu > \mu_0$ the critical region lies entirely in right tail of the sampling distribution of $\bar{x}$, while for the left tailed test $H_1: \mu < \mu_0$ the critical region is entirely in the left of the distribution of $\bar{x}$.

Right tailed test:



Left tailed test:



Two tailed test:

A test of statistical hypothesis where the alternative hypothesis is two tailed such as, $H_0: \mu = \mu_0$ against the alternative hypothesis $H_1: \mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$) is known as two tailed test and in such a case the critical region is given by the portion of the area lying in both the tails of the probability curve of test of statistic.

For example, suppose that there are two population brands of washing machines, are manufactured by standard process(with mean warranty period μ_1) and the other manufactured by some new technique (with mean warranty period μ_2): If we want to test if the washing machines differ significantly then our null hypothesis is $H_0: \mu_1 = \mu_2$ and alternative will be $H_1: \mu_1 \neq \mu_2$ thus giving us a two tailed test. However if we want to test whether the average warranty period produced by some new technique is more than those produced by standard process, then we have $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 < \mu_2$ thus giving us a left-tailed test.

Similarly, for testing if the product of new process is inferior to that of standard process then we have, $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 > \mu_2$ thus giving us a right-tailed test. Thus the decision about applying a two – tailed test or a single –tailed (right or left) test will depend on the problem under study.

Critical values ($Z\alpha$) of Z

Level of significance α	0.05 or 5%		0.01 or 1%	
	Left	Right	Left	Right
Critical values of Z_α for one tailed Tests	–1.645	1.645	–2.33	2.33
Critical values of $Z_{\alpha/2}$ for two tailed tests	–1.96	1.96	–2.58	2.58

Type I and Type II Errors

When a statistical hypothesis is tested there are four possibilities.

1. The hypothesis is true but our test rejects it (Type I error)
2. The hypothesis is false but our test accepts it (Type II error)
3. The hypothesis is true and our test accepts it (correct decision)
4. The hypothesis is false and our test rejects it (correct decision)

Obviously, the first two possibilities lead to errors.

In a statistical hypothesis testing experiment, a Type I error is committed by rejecting the null hypothesis when it is true. On the other hand, a Type II error is committed by not rejecting (accepting) the null hypothesis when it is false.

If we write ,

α = P (Type I error) = P (rejecting H_0 | H_0 is true)

β = P (Type II error) = P (Not rejecting H_0 | H_0 is false)

In practice, type I error amounts to rejecting a lot when it is good and type II error may be regarded as accepting the lot when it is bad. Thus we find ourselves in the situation which is described in the following table.

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

Test Procedure

Steps for testing hypothesis is given below. (for both large sample and small sample tests)

1. Null hypothesis : set up null hypothesis H_0 .
2. Alternative Hypothesis: Set up alternative hypothesis H_1 , which is complementry to H_0 which will indicate whether one tailed (right or left tailed) or two tailed test is to be applied.
3. Level of significance : Choose an appropriate level of significance (α), α is fixed in advance.
4. Test statistic (or test of criterian):

   Calculate the value of the test statistic, Z = $\dfrac{t - E(t)}{S.E.(t)}$ under the null hypothesis, where t is the sample statistic
5. Inference: We compare the computed value of Z (in absolute value) with the significant value (critical value) $Z_{\alpha/2}$ (or Z_α). If $|Z| > Z_\alpha$, we reject the null hypothesis H_0 at α % level of significance and if $|Z| \leq Z_\alpha$, we accept H_0 at α % level of significance.

Note:

1. Large Sample: A sample is large when it consists of more than 30 items.
2. Small Sample: A sample is small when it consists of 30 or less than 30 items.

Exercise

I. Choose the best answers

1. A measure characterizing a sample such as $\bar{x}$ or s is called
   (a). Population (b). Statistic (c).Universe (d).Mean
2. The standard error of the mean is
   (a). σ^2 (b). $\frac{\sigma}{n}$ (c). $\frac{\sigma}{\sqrt{n}}$ (d). $\frac{\sqrt{n}}{\sigma}$
3. The standard error of observed sample proportion "P" is
   (a). $\sqrt{\frac{P(1-Q)}{n}}$ (b). $\sqrt{\frac{PQ}{n}}$ (c). $\sqrt{\frac{(1-P)Q}{n}}$ (d). $\frac{PQ}{n}$
4. Alternative hypothesis is
   (a). Always Left Tailed (b). Always Right tailed
   (c). Always One Tailed (d). One Tailed or Two Tailed
5. Critical region is
   (a). Rejection Area (b). Acceptance Area
   (c). Probability (d). Test Statistic Value
6. The critical value of the test statistic at level of significance α for a two tailed test is denoted by
   (a). $Z_{\alpha/2}$ (b). Z_α (c). $Z_{2\alpha}$ (d). $Z_{\alpha/4}$
7. In the right tailed test, the critical region is
   (a). 0 (b). 1
   (c). Lies entirely in right tail (d). Lies in the left tail
8. Critical value of $|Z|_\alpha$ at 5% level of significance for two tailed test is
   (a). 1.645 (b). 2.33 (c). 2.58 (d). 1.96

9. Under null hypothesis the value of the test statistic Z is
(a). $\dfrac{t- S.E.(t)}{E(t)}$ (b). $\dfrac{t+E(t)}{S.E.(t)}$ (c). $\dfrac{t - E(t)}{S.E.(t)}$ (d). $\sqrt{\dfrac{PQ}{n}}$
10. The alternative hypothesis $H_1: \mu \neq \mu_0 (\mu > \mu_0$ or $\mu < \mu_0)$ takes the critical region as
(a). Right tail only (b). Both right and left tail
(c). Left tail only (d). Acceptance region
11. A hypothesis may be classified as
(a). Simple (b). Composite
(c). Null (d). All the above
12. Whether a test is one sided or two sided depends on
(a). Alternative hypothesis (b). Composite hypothesis
(c). Null hypothesis (d). Simple hypothesis
13. A wrong decision about H_0 leads to:
(a). One kind of error (b). Two kinds of error
(c). Three kinds of error (d). Four kinds of error
14. Area of the critical region depends on
(a). Size of type I error (b). Size of type II error
(c). Value of the statistics (d). Number of observations
15. Test of hypothesis H_0: $\mu = 70$ vs H_1= $\mu > 70$ leads to
(a). One sided left tailed test (b). One sided right tailed test
(c). Two tailed test (d). None of the above
16. Testing H_0: $\mu = 1500$ against $\mu < 1500$ leads to
(a). One sided left tailed test (b). One sided right tailed test
(c). Two tailed test (d). All the above
17. Testing H_0: $\mu = 100$ vs H_1: $\mu \neq 100$ lead to
(a). One sided right tailed test (b). One sided left tailed test
(c). Two tailed test (d). None of the above

II. Fill in the Blanks

18. n_1 and n_2 represent the __________ of the two independent random samples respectively.
19. Standard error of the observed sample proportion p is ________
20. When the hypothesis is true and the test rejects it, this is called ________

21. When the hypothesis is false and the test accepts it this is called _______
22. Formula to calculate the value of the statistic is __________

III. Answer the following

23. Define sampling distribution.
24. Define Parameter and Statistic.
25. Define standard error.
26. Give the standard error of the difference of two sample proportions.
27. Define Null hypothesis and alternative hypothesis.
28. Explain: Critical Value.
29. What do you mean by level of significance?
30. Explain clearly type I and type II errors.
31. What are the procedure generally followed in testing of a hypothesis ?
32. What do you mean by testing of hypothesis?
33. Write a detailed note on one- tailed and two-tailed tests.

Answers:

I.

1. (b)	2. (c)	3. (b)	4.(d)	5. (a)
6. (a)	7. (c)	8. (d)	9. (c)	10 (b)
11.(d)	12.(a)	13.(b)	14.(a)	15.(b)
16.(a)	17.(c)			

II.

18. size 19. $\sqrt{\frac{PQ}{n}}$ 20. Type I error

21. Type II error

22. $Z = \dfrac{t - E(t)}{S.E.(t)}$

Possible Questions

Unit I

PART-B

1. What is Null Hypothesis? Discuss the steps in testing a Hypothesis.
2. Explain the functions of Random variable by an example.
3. Write a short note about the following terms:
   i. Conditional Probability
   ii. Null and Alternate Hypothesis
   iii. Confidence Limits
4. Explain the Characteristics of Random variable with an example.
5. Write a short note about the following terms:
   i) Confidence Interval
   ii) Type I and Type II Errors
   iii) One tail and Two tail tests
6. (i) Explain the of axioms of the theory probability.
   (ii) State and prove Bayes theorem
7. Write a short note about the following terms:
   i) Type I and Type II Errors
   ii) Standard Error
   iii) One tail and Two tail tests
8. Write a short note about the following terms:
   i) Random Event and Independent Event
   ii) Null and Alternate Hypothesis
   iii) Standard Error

PART-C

9. Write a short note about the following terms:
   a) Conditional Probability
   b) Bayes' Theorem
   c) Type I and Type II Errors
   d) Confidence Interval

e) Level of Significance

DEPARTMENT OF MATHEMATICS						
MATHEMATICAL STATISTICS (18MMP304)						
UNIT- I						
Sl. No.	**Question**	**Option 1**	**Option 2**	**Option 3**	**Option 4**	**Answer**
1	The probability of drawing a card of King from a pack of cards is…………	1/4	1/11	1/12	1/13	1/13
2	In tossing a coin, the probability of getting head is …………...	1/2	1/3	2	0	1/2
3	The probability that a leap year selected at random contain 53 Sundays is…………….	1/7	2/7	3/7	1/53	2/7
4	A bag contains 7 red and 8 black balls. The probability of drawing a red ball is …….	7/15	8/15	1/15	14/15	7/15
5	The probability of drawing a card of clubs from a pack of 52 cards is ……..	0	(1/3)	2/4	1/4	1/4
6	The probability of drawing an ace or queen card from a pack of 52 cards is …	1/13	1/4	2/13	1/52	1/13
7	The total probability is is always equal to…………..	0.5	2	1	0	1
8	A variable whose value is a number determined by the outcome of a random experiment is called a………………………	Sample	Random variable	Outcome	Event	Random variable
9	If a random variable takes only a finite or a countable number of values, it is called ………………..	Finite random space	Continous random variable	Discrete random variable	Infinite random variable	Discrete random variable
10	A continuous random variable is a random variable X which can take any value between….	Interval	Limits	Finite values	Infinite values	Interval

11	Suppose that X be a discrete or continuous random variable, then distribution function is a…………function of x.	Non-decreasing	Decreasing	Neither increasing nor decreasing	Can be increasing and decreasing	Non-decreasing
12	The function f(x) = 5x^4 , 0 < x < 1 can be a…….. of a random variable X.	Probability mass function	Probability density function	Distribution function	Exponential function	Probability density function
13	If F(x) is the cumulative distribution function of a continuous random variable X with p.d.f f(x) then…………	F'(x) = f(x)	F'(x) not equal to f(x)	F'(x) < f(x)	F'(x) >f(x)	F'(x) = f(x)
14	If X is a continuous random variable with p.d.f f(x), then F(b)-F(a)=	P(a>X>b)	P(a<X>b)	P(b<X<a)	P(a<X<b)	P(a<X<b)
15	Which one of the following represents the best estimate of the population mean?	The sample mean	The mean of several sample means	The mode of several sample means	The median of several sample means	The mean of several sample means
16	Which of the following statements are true?	Parameters describe samples and statistics describe populations	Statistics describe samples and populations	Parameters describe populations and statistics describe samples	Both (a) and (b) above	Parameters describe populations and statistics describe samples
17	The narrower the confidence intervals:	The more confidence you can place in your results	The less you can rely on your results	The greater the chance that your results were due to sampling error	Correlation between the two scores	The more confidence you can place in your results

18	Statistical significance:	Is directly equivalent to psychological importance	Does not necessarily mean that results are psychologically important	Depends on sample size	Both (b) and (c) above	Both (b) and (c) above
19	All other things being equal:	The more sample size increases, the more power decreases	The more sample size increases, the more power increases	Sample size has no relationship to power	The more sample size increases, the more indeterminate the power	The more sample size increases, the more power increases
20	Find probability of drawing diamond and a heart card from a pack of 52 cards?	13/102	1/4	2/13	7/16	13/102
21	The probability of drawing king and queen card from a pack of 52 cards is	13/102	1/4	2/13	8/663	8/663
22	Two coins are tossed five times, find the probability of getting an even number of heads ?	0.25	1	0.4	0.25	0.25
23	All other things being equal, the more powerful the statistical test:	The wider the confidence intervals	The more likely the confidence interval will include zero	The narrower the confidence interval	The smaller the sample size	The narrower the confidence interval

24	Power can be calculated by a knowledge of:	The statistical test, the type of design and the effect size	The statistical test, the criterion significance level and the effect size	The criterion significance level, the effect size and the type of design	The criterion significance level, the effect size and the sample size	The criterion significance level, the effect size and the sample size
25	Which of the following constitute continuous variables?	Anxiety rated on a scale of 1 to 5 where 1 equals not anxious, 3 equals moderately anxious and 5 equals highly anxious	Gender	Temperature	Intelligence	Temperature
26	A continuous variable can be described as:	Able to take only certain discrete values within a range of scores	Able to take any value within a range of scores	Being made up of categories	Being made up of variables	Able to take any value within a range of scores
27	Which one of the following represents the best estimate of the population mean?	The sample mean	The mean of several sample means	The mode of several sample means	The median of several sample means	The mean of several sample means

28	Which one of the following represents the best estimate of the population mean?	The sample mean	The mean of several sample means	The mode of several sample means	The median of several sample means	The mean of several sample means
29	The narrower the confidence intervals:	The more confidence you can place in your results	The less you can rely on your results	The greater the chance that your results were due to sampling error	Correlation between the two scores	The more confidence you can place in your results
30	Which of the following could be considered as categorical variables?	Gender	Brand of baked beans	Hair colour	All of the above	All of the above
31	One card is drawn at random from a well-shuffled pack of 52 cards. What is the probability that it will be a diamond ?	1/13	1/4	1/52	1/15	1/4
32	Which of the following is a continous probability distribution?	Normal	Poisson	Binomial	Uniform	Normal
33	For which distribution, mean,meadian and mode coincides?	Poisson	F	Chi square	Normal	Normal
34	The range of standard normal variate is …………….	$-\infty$ to $+\infty$	0 to 1	0 to ∞	1 to ∞	$-\infty$ to $+\infty$

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

Subject : Mathematical Statistics **Semester III** **L T P C**
Subject Code : 18MMP304 **Class : II M.Sc Mathematics** **4 0 0 4**

UNIT II
Sample moments and their functions: Notion of a sample and a statistic - Distribution functions of X, S^2 and (X , S^2) -Chi-square distribution -Student t-distribution -Fisher's Z-distribution -Snedecor's F -distribution -Distribution of sample mean from non-normal populations.

SUGGESTED READINGS

1. Marek Fisz, (1980). Probability Theory and Mathematical Statistics, John Wiley and Sons, New York.
2. Meyer, (2006). Introduction to Probability and Statistical applications, Oxford and IBH Publishing Co. Pvt. Ltd. New Delhi.
3. Sheldon M. Ross, (2009). Introduction to probability and statistics for engineers and scientists, Third edition, Academic press.
4. Heinz Bauer,(1996),Probability Theory, Narosa Publishing House, London.
5. Parimal Mukhopadhyay, (2012). Theory of Probability, New central book agency, Calcutta.

Sampling, Sampling Distributions

Samples vs. Populations

Population: A complete set of observations or measurements about which conclusions are to be drawn.

Sample: A subset or part of a population.

Not necessarily random

Statistics vs. Parameters

Parameter: A summary characteristic of a population.

Summary of Central tendency, variability, shape, correlation

E.g., Population mean, Population Standard Deviation, Population Median, Proportion of population of registered voters voting for Bush, Population correlation between Systolic & Diastolic BP

Statistic: A summary characteristic of a sample. Any of the above computed from a sample taken from the population.

E.g., Sample mean, Sample Standard Deviation, median, correlation coefficient

Inferential Statistics

We take a sample and compute a description of a characteristic of the sample – central tendency (usually), variability or shape. That is, we compute the value of a sample statistic.

We use the sample statistic to make an educated guess about the corresponding population parameter.

The basic concept is easy. The devil is in the details.

DEFINITIONS

- A random variable X represents a numerical value associated with each outcome of a probability experiment.
- A random variable is discrete if it has a finite or countable number of possible outcomes that can be listed.
- A random variable is continuous if it has an uncountable number of possible outcomes, represented by an interval on the number line.

The number of calls a salesperson makes in one day is an example of a discrete random variable, while the time in hours he spends making calls in one day is an example of a continuous random variable.

A discrete probability distribution lists each possible value the random variable can assume, together with its probability. A probability distribution must satisfy the following conditions:

The probability of each value of the discrete random variable is between 0 and 1: $0 \le P(x) \le 1$
The sum of all the probabilities is 1: $\sum P(x) = 1$

Guidelines for constructing a discrete probability distribution: (p164)

1. Make a frequency distribution for the possible outcomes
2. Find the sum of the frequencies
3. Find the probability of each possible outcome by dividing its frequency by the sum of the frequencies.
4. Check that each probability is between 0 and 1 and that the sum is 1.

Example (p 164) Individuals are rated on a score of 1 to 5 for passive-aggressive traits, where 1 is extremely passive and 5 is extremely aggressive.

Score, *X*	**Frequency, *f***	***P(X)***
1	24	0.16
2	33	0.22
3	42	0.28
4	30	0.2
5	21	0.14
Total	150	1.00

The mean (also called the expected value) of a discrete random variable is given by:
$Expected\ Value = E(x) = \mu = \sum xP(x)$
Note that each value of x is multiplied by its corresponding probability and the products are added.

Example: Find the mean for passive-aggressive traits above:

X	**P(X)**	**XP(X)**
1	0.16	1*0.16 = 0.16
2	0.22	2*0.22 = 0.44
3	0.28	3*0.28 = 0.84
4	0.2	4*0.20 = 0.80
5	0.14	5*0.14 = 0.70
	$\Sigma P(X) = 1$	$\Sigma XP(X) = 2.94$

The variance of a discrete random variable is the expected value of $(x-\mu)^2$:
$\sigma^2 = E(x-\mu)^2 = \sum (x-\mu)^2 P(x)$

The standard deviation is
$\sigma = \sqrt{\sigma^2}$

Example (p 167) Find the Variance and Standard Deviation of the passive-aggressive measure in the above example

X	**P(X)**	$x-\mu$	$(x-\mu)^2$	$P(x)(x-\mu)^2$
1	0.16	-1.94	3.764	0.602
2	0.22	-0.94	0.884	0.194
3	0.28	0.06	0.004	0.001
4	0.20	1.06	1.124	0.225
5	0.14	2.06	4.244	0.594
X	$\Sigma P(X) = 1$			$\Sigma P(x)(x-\mu)^2 = 1.616$

So,
$Var(x) = \sigma^2 = 1.616$
$\sigma = \sqrt{1.616} \approx 1.27.$

Binomial Distributions

A binomial experiment is a probability experiment that satisfies the following conditions:

1. The experiment is repeated for a fixed number of trials where each trial is independent of the other trials.
2. There are only two possible outcomes of interest for each trial. The outcomes can be classified as a success (S) or as a failure (F).
3. The probability of a success P(S) is the same for each trial.
4. The random variable x counts the number of successful trials.

Notation for Binomial Experiments

Symbol	Description
n	The number of times the trial is repeated
$p = P(S)$	The probability of success in a single trial
$q = P(F)$	The probability of failure in a single trial $(q = 1 - p)$
x	The random variable represents a count of the number of successes in n trials: $x = 0,1,2,3,\ldots,n$

Suppose we have 9 trials. If we let 0 mean failure and 1 mean success, the probability of getting the results: 0 0 1 0 1 1 0 0 0 is p^3q^6. (See Mood and Graybill p 66.) This is a specific way of getting 3 successes: on the third fifth and sixth tries. Each try can be viewed as a box, and the number of ways we can place 3 1's in 9 boxes is the same as the number of ways we can choose the first 3 players from 9 on a baseball team. This is $\binom{9}{3}$. In general the probability of a specific arrangement of x 1's and n-x 0's is $p^x q^{n-x}$ and there are $\binom{n}{x}$ arrangements. This leads to the following formula for the binomial distribution.

In a binomial experiment, the probability of exactly x successes in n trials is:

$$P(x) = {}_nC_x\, p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x} = \binom{n}{x} p^x q^{n-x}, x = 0,1,2,\ldots,n$$

This is often referred to as $b(x; n, p)$.

We can also see how this formula is derived from a simple example: Suppose we perform have 3 trials. The possible results are:

Sample Points	***Probability of sample point***	***Value of x***
SSS	p^3	*3*
SSF	p^2q	*2*
SFS	p^2q	*2*
SFF	pq^2	*1*
FSS	p^2q	*2*
FSF	pq^2	*1*
FFS	pq^2	*1*
FFF	q^3	*0*

So, $P(0) = \binom{3}{0} q^3, P(1) = \binom{3}{1} pq^2, P(2) = \binom{3}{2} p^2 q, P(3) = \binom{3}{3} p^3$

Appendix B, Table 2 gives, for the binomial distribution, the probabilities of x successes in n trials, for values of n = 2-16,20 for x = 0 to n, for various probabilities of success.

Population Parameters of a Binomial Distribution
$\mu = np$
$\sigma^2 = npq$
$\sigma = \sqrt{npq}$

The following are derivations of the mean for n = 1 and 2. (Mendenhall p 123)

$n = 1$	$E(x) = \sum_{x=0}^{1} xp(x) = 0q + 1p = p$
$n = 2$	$E(x) = \sum_{x=0}^{2} xp(x) = 0q^2 + 1 * 2pq + 2p^2 = 2p(q+p) = 2p$

The following is a derivation of the variance for n = 1. (Mendenhall p 123)

$$\sigma^2 = E(x-\mu)^2 = \sum_{x=0}^{1} (x-\mu)^2 p(x) = (0-p)^2 q + (1-p)^2 p = p^2 q + q^2 p = pq(q+p) = pq.$$

Performing the Test

To perform a hypothesis test, you must find a z-score based on the value of the parameter specified in the null hypothesis.

$$z = \left| \frac{\bar{x} - \mu_{H_o}}{\sigma_{\bar{x}}} \right|$$

Note that in forming this z-score, we are using the standard error of the mean in the denominator. That's because your sample mean is distributed normally with that standard deviation, not the standard deviation of the population as a whole. We can rewrite the above like so:

$$z = \left| \frac{x - \mu_{H_o}}{\sigma / \sqrt{n}} \right|$$

We will then compare this to a critical value of z from the standard normal table. If it's greater than the z-critical, we reject the null and accept the alternative hypothesis. Otherwise, we do not reject the null, nor do we accept the alternative.

Example: Let's do the two-tail test on CSUN's graduation time. Let's say we know the standard deviation of the population is 2 years, and we sampled 49 CSUN graduates and found a sample mean of 6.9. Then we calculate:

$$z = \frac{\bar{x} - \mu_{H_o}}{\sigma / \sqrt{n}} = \frac{6.9 - 6.5}{2 / \sqrt{49}} = 1.4$$

We need a z-critical value for a significance level of 0.10. Since this is a two-tail test, we want 0.05 in each tail, so find the value of z in Table 3 that gives you an area as close to 0.95 as possible. This turns out to be 1.64. Since 1.4 < 1.64, we do not reject the null. The administration's claim cannot be rejected.

Example: Now let's do the one-tail test on CSUN's graduation time. All the calculations are the same, except now we want the whole 10% in the right tail. That gives us a z-critical of 1.28.

Since 1.4 > 1.28, we reject the null and accept the alternative. We think the administration has underestimated the true mean graduation time.

NOTE: The test we just did is a right-tail test, because the null hypothesis is rejected only for a sufficiently high sample mean. But what if the null hypothesis had been that CSUN's average graduation time was greater than or equal to 6.5? In that case we would have done a left-tail test. In addition to the z-value calculated above being greater than z-critical, you also need to make sure the sample mean is less than the hypothesized mean (6.5 in this case). Alternatively, just calculate the z-value above without absolute value signs, and then put a negative sign on your z-critical.

Why did we reject in the two-tail case and accept in the one-tail case? Because in the two-tail case, some of the weight of α had to go in the left tail, which turned out to be irrelevant in this case. That meant there was less weight to go in the right tail, and thus less chance of rejecting the null as a result of a high sample mean.

V. Getting Rid of the Bogus Assumptions

We assumed above that true standard deviation was known. Just as with CI's, this is a weird assumption. Why would we know the true standard deviation but not the true mean?
When we have a large sample, we can get away with substituting the sample standard deviation for the true one and continuing to use the z-distribution. This gives us the following z-score formula:

$$z = \left| \frac{\bar{x} - \mu_{H_o}}{s / \sqrt{n}} \right|$$

But what if you don't know the true standard deviation and the sample size is small? Then we have to use the t-distribution. We calculate a t-score instead of a z-score:

$$t = \left| \frac{\bar{x} - \mu_{H_o}}{s / \sqrt{n}} \right|$$

And then we find a t-critical value instead of a z-critical value.

Example: Same example as above, doing a one-tail test. But this time, we don't know the standard deviation is 2, and our sample size was only 17. Our sample standard deviation turns out to be 1.9, and we use this to find our t-score:

$$t = \frac{x - \mu_{H_o}}{\sigma / \sqrt{n}} = \frac{6.9 - 6.5}{1.9 / \sqrt{16}} = 0.84$$

In the t-table, with df = 16 -1 = 15 and 90% confidence level, t-critical is 1.75. We do not reject the null.

If we had wanted a one-tail test, we'd have looked in the column of the table headed by 0.1000 (ignore the 0.8000 confidence level below, because that assumes a two-tail test). We get 1.341. Since 0.84 < 1.341, we do not reject the null.

Example: In the above example, for the two-tail test, the p-value is 2(0.0808) = 0.1616. That's the lowest alpha that will lead to rejection of the null in the two-tail test.

It is possible to find p-values when we're using the t-statistics as well. But the t-table in a book doesn't give us enough information to find the p-value with much precision. A statistical software program can do it for us, though.

Three theoretical facts and one practical fact about the distribution of sample means . . .

The theoretical facts are about 1) central tendency, 2) variability, and 3) shape . . .

1. The mean of the population of sample means will be the same as the mean of the population from which the samples were taken. The mean of the means is the mean. $\mu_M = \mu$.

Implication: The sample mean is an unbiased estimate of the population mean. If you take a random sample from a population, it is just as likely to be smaller than the population mean as it is to be larger than the population mean.

2. The standard deviation of the population of sample means – called the standard error of the mean - will be equal to d original population's standard deviation divided by the square root of N, the size of each sample. (Corty, Eq. 5.1, p 142)

In Corty's notation,

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

The standard deviation (σ_M) is called the standard error of the mean.

Implication: Means are less variable than individual scores. Means are likely to be closer to the population mean than individual scores. You can make a sample mean as close as you want to the population mean if you can afford a large sample.

3. The shape of the distribution of the population of sample means will be the normal distribution if the original distribution is normal or approach the normal as N gets larger in all other cases. This fact is called the Central Limit Theorem. It is the foundation upon which most of modern day inferential statistics rests. See Corty, p. 141.

Why do we care about #3: Because we'll need to compute probabilities associated with sample means when doing inferential statistics. To compute those probabilities, we need a probability distribution.

Practical fact

4. The distribution of Z's computed from each sample, using the formula

$$Z = \frac{\text{X-bar} - \mu_M}{\dfrac{\sigma}{\sqrt{N}}}$$

will be or approach (as sample size gets large) the Standard Normal Distribution with mean = 0 and SD = 1.

Another test question: What are three facts about the distribution of sample means – a fact about central, a fact about variability, and a fact about shape of the distribution of sample means?

Continuous (Normal) Probability Distributions

[NOTE: The following notes were compiled from previous notes used when I taught other statistics courses.]

CHAPTER OBJECTIVES

- Understand the difference between discrete and continuous distributions.
- Compute the mean and the standard deviation for a uniform distribution.
- Compute probabilities using the uniform distribution.
- List the characteristics of the normal probability distribution.
- Define and calculate *z* values.
- Determine the probability an observation will lie between two points using the standard normal distribution.
- Determine the probability an observation will be above or below a given value using the standard normal distribution.
- Use the normal distribution to approximate the binomial probability distribution.

CHARACTERISTICS OF A NORMAL PROBABILITY DISTRIBUTION

- The normal curve is bell-shaped and has a single peak at the exact center of the distribution.
- The arithmetic mean, median, and mode of the distribution are equal and located at the peak.
- Half the area under the curve is above and half is below this center point (peak).
- The normal probability distribution is symmetrical about its mean.
- It is asymptotic - the curve gets closer and closer to the x-axis but never actually touches it.

NOTE

You can also have normal distributions with the same means but different standard deviations.

You can also have normal distributions with the same standard deviation but with different means.

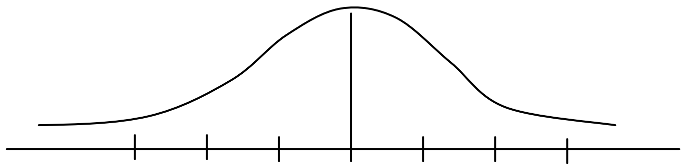
You can also have normal distributions with different means and different standard deviations.

THE STANDARD NORMAL PROBABILITY DISTRIBUTION

A normal distribution with a mean of 0 and a standard deviation of 1 is called the standard normal distribution.

z value: The distance between a selected value, designated X, and the population mean μ, divided by the population standard deviation, σ.

$$\mathbf{z} = \frac{(\mathbf{X} - \mu)}{\sigma} \qquad (5\text{-}2)$$



EXAMPLE 1: Monthly incomes of MBA Graduates.

The monthly incomes of recent MBA graduates in a large corporation are normally distributed with a mean of \$2,000 and a standard deviation of \$200. What is the z value for an income X of \$2,200? \$1,700?

For X = \$2,200 and since $z = (X - \mu)/\sigma$, then $z = (2{,}200 - 2{,}000)/200 = 1$.

For X = \$1,700 and since $z = (X - \mu)/\sigma$, then $z = (1{,}700 - 2{,}000)/200 = -1.5$.

A z value of 1 indicates that the value of \$2,200 is 1 standard deviation above the mean of \$2,000.

A z value of -1.5 indicates that the value of \$1,700 is 1.5 standard deviations below the mean of \$2,000.

<u>AREAS UNDER THE NORMAL CURVE</u> (See Section 5-3 “Applications of Normal Distributions.”)

From the Empirical Rule, we should remember the following:

About 68 percent of the area under the normal curve is within plus one and minus one standard deviation of the mean, written as $\mu \pm 1\sigma$.

About 95 percent of the area under the normal curve is within plus and minus two standard deviations of the mean, written $\mu \pm 2\sigma$.

Practically all (99.74 percent) of the area under the normal curve is within three standard deviations of the mean, written $\mu \pm 3\sigma$.

<u>Form of H_0 and H_1 for one-sample mean:</u>

H_0: $\mu = 115$ H_1: $\mu \neq 115$

- Hypotheses are always about <u>population parameters</u>, not sample statistics

H_0: μ = population value

H_1: $\mu \neq$ population value

- This hypothesis is a *non-directional* (two-tailed) hypothesis
- Null hypothesis: No effect
- Alternative hypothesis: Some effect (doesn't specify an increase or decrease)

Criterion for rejecting H_0: Creating a Decision Rule:

Will compute a test statistic (types vary based on data, design & question)

Then decide if the value of the test statistic is "improbable" under H_0

Traditionally, a test statistic is considered "unlikely" if it is expected to occur:

$\leq$ 5 in a 100: has a probability of .05 or less ($p \leq 0.05$)

- Look in *tails* of *sampling distribution for the unlikely outcomes*
- Divide distribution into two parts:

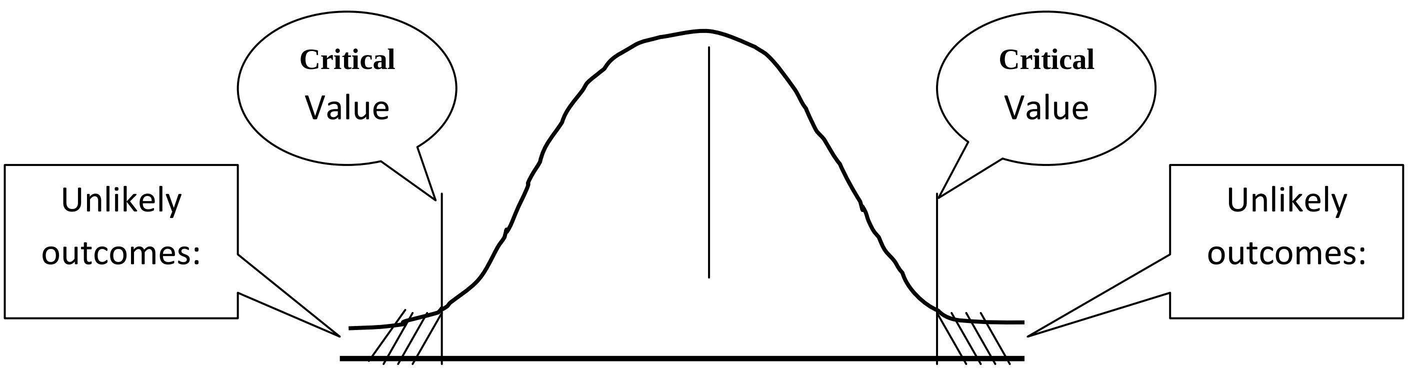
Values that are likely if H_0 is true	Values that are very unlikely if H_0 is true
Values close to H_0	Values far from H_0
Values in the *middle*	Values in the *tails*

Selecting a "significance level": α

Probability chosen as criteria for "unlikely"

Common convention: $\alpha = .05$ (5%)

May set a smaller α to be more conservative ($p \leq 0.01, 0.001$)



Critical value(s) = boundary(ies) b/n likely & unlikely outcomes

Rejection region = area(s) beyond critical value(s); outcomes that lead to a rejection of H_0

Decision rule:

Reject H_0 when *observed* test-statistic equals or exceeds *Critical value*

…when statistic falls in the *rejection region*

Otherwise, Fail to Reject (Retain) H_0

Collect data and Calculate "observed" test statistic:

z-test for one sample mean:

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

z = <u>sample mean – hypothesized population μ</u>

standard error

z = <u>observed difference</u>

difference due to chance

Don't forget to compute standard error first!

$$\sigma_{\overline{\mathbf{X}}} = \frac{\sigma}{\sqrt{n}}$$

Compare Test Statistic to Critical Values:

- Does observed z equal or exceed CV?

  (Does it fall in the *rejection region*?)

- If YES,

  Reject H_0 = "statistically significant" finding

- If NO,

  Fail to Reject H_0 = "non-significant" finding

Interpret results:

- Return to research question
- *statistical significance* = not likely to be due to chance
- Never "prove" H_0 or H_1

Summary of Statistical Hypothesis Testing:

1. Formulate a research question
2. Formulate a research/alternative hypothesis
3. Formulate the null hypothesis
4. Collect data
5. Reference a sampling distribution of the particular statistic assuming that H_0

is true (in the cases so far, a sampling distribution of the mean)

6. Decide on a significance level (α), typically .05
7. Identify the critical value(s)
8. Compute the appropriate test statistic
9. Compare the test statistic to the critical value(s)
10. Reject H_0 if the test statistic is equal to or exceeds the critical value, retain otherwise

Possible Questions

PART-B

1. State properties and applications of *t*- distribution?
2. A group of 5 patients treated with medicine A weigh 39, 48, 60 and 41 kgs; second group of 7 patients treated with medicine B weigh 38, 42, 56, 64, 68, 69, 62 kgs. Do you agree with claim that medicine B increases weight significantly?
   (Use α = 5% and $t_{0.05}$ =1.812)
3. Write the properties normal distribution?
4. Certain pesticide is packed into bags by a machine. Random samples of 10 bags are drawn and their contents are found to weigh (in kg) as follows.
   50 49 52 44 45 48 46 45 49 45
   Test if the average packing can be taken to be 50 kg.
5. Certain brand of rice is packed into bags by a machine. Random samples of 15 bags are drawn and their contents are found to weigh (in kg) as follows.
   50 49 52 44 45 48 46 45 49 45 50 52 54 53 51
   Test if the average packing can be taken to be 50 kg.
6. Mention Snedecor's F-distribution, its properties and applications?

PART-C

7. i) Define Normal Distribution and write its important characteristics.
   ii) Describe the characteristics of χ^2 - distribution.

DEPARTMENT OF MATHEMATICS						
MATHEMATICAL STATISTICS (18MMP304)						
UNIT- II						
Sl. No.	**Question**	**Option 1**	**Option 2**	**Option 3**	**Option 4**	**Answer**
1	The word ---------------- is used to indicate various statistical measures like mean, standard deviation, correlation etc, in the universe.	Statistic	Parameter	Hypothesis	Sample	Parameter
2	The term STATISTIC refers to the statistical measures relating to the…………….	Population	Hypothesis	Sample	Parameter	Sample
3	Degrees of freedom are related to ………………	No. of observations in a set	Hypothesis under test	No. of independent observations in a set	No. of rows of observations	No. of independent observations in a set
4	Student's t-test is applicable in case of …………	Small samples	For sample of size between 25 and 35	Large samples	For sample size of more than 100	Small samples
5	The distribution used to test goodness of fit is………..	F distribution	χ^2 distribution	t distribution	Z distribution	χ^2 distribution
6	The formula for χ^2 is…………	$\sum(O - E)^2 / E$	$(E+O)^2 / E$	$(O-E) / E$	$\sum(O - E)^2 / O$	$\sum(O - E)^2 / E$
7	In sampling distribution the standard error is a………………	Standard mean	Sampling error	Difference error	Type-I error	Sampling error
8	The characteristic of the chi–square test is ………………..	Degree of Freedom	Level of significance	ANOVA	Independence of attributes	Independence of attributes
9	If $S_1^2 > S_2^2$, then the F – statistic is…………	S_1 / S_2	S_2 / S_1	S_1^2 / S_2^2	S_1^3 / S_2^3	S_1^2 / S_2^2

10	Which of the following is the standard deviation of a sampling distribution.	Standard error	Sample standard deviation	Replication error	Meta error	Standard error
11	A good way to get a small standard error is to use a _________.	Repeated sampling	Small sample	Large sample	Large population	Large sample
12	Numerical characteristic of a sample is called……………..	Statistic	Parameter	Hypothesis	Sample	Statistic
13	Which of the following symbols represents a population parameter?	S.D	σ	r	0	σ
14	The distribution of means of all possible samples taken from a population is…………	A sampling distribution	A sample	Population distribution	Parameter distribution	A sampling distribution
15	The mean of the sample means is exactly equal to the ……………….	Sample mean	Population mean	Weighted mean	Combined mean	Population mean
16	The mean of Chi - distribution with n degrees of freedom is…………	n	n-1	2n	2n-1	n
17	The Chi- distribution is…………	Continous	Multimodal	Bimodal	Symmetrical	Continous
18	In sampling without replacement, expectation of samlpe variance is not equal to:	Population variance	Sample mean	Sample S.D	Population mean	Population variance
19	For larger degrees of freedom , t-distribution tends to ------------------ distribution	Standard normal	Binomial	Exponential	Poisson	Standard normal
20	Mode of F-distribution is always………………..	< 1	> 1	1	0	< 1
21	Sampling distribution of F-distribution only depends on …………	Degrees of freedom	Population size	Sample size	Parameters	Degrees of freedom
22	Variance of Chi- distribution with n degrees of freedom is given by	n	2n	n-2	n-3	2n

23	Chi square variate with 1 degree of freedom is the square of ---------------------- variate	Standard normal	Binomial	Normal	Poisson	Standard normal
24	Student's t-distribution was discovered by…….	Karl Pearson	Laplace	Fisher	Gosset	Gosset
25	The most commonly used assumption about the distribution of a variable is…………..	Continuity	Symmetry	Discontinuity	Non-symmetry	Symmetry
26	Which distribution is lower at mean and higher at tail than a normal distribution	t	F	Z	Chi-Square	t
27	F-distribution was devised by--------------------	R.A.Fischer	Snedecor	Gosset	Karl Pearson	Snedecor
28	To test whether or not two population variances are equal, the appropriate distribution is………….	Z distribution	Chi-square distribution	F distribution	t-distribution	F distribution
29	If a statistic t follows student's t distribution,then t^2 follows……….	F distribution	t distribution	Chi-square distribution	Normal distribution	F distribution
30	If F follows $F(n_1,n_2)$,then $\chi 2$ = follows chi square distribution with ------d.f	n_1	n_2	n_1 -1	n_2- 2	n_1
31	The relation between the mean and variance of chi square distribution with n d.f is …….	Mean=2 Variance	Mean=Variance	2 Mean= Variance	Mean<Variance	2 Mean= Variance
32	The range of F- variate is …………	$-\infty$ to $+\infty$	0 to 1	0 to ∞	$-\infty$ to 0	0 to ∞
33	The larger variance in the variance ratio for F-statistic is taken in……………	Denominator	Numerator	As constant	As zero	Numerator
34	From Snedocor's F-distribution, we can devise………………….statistic	Fischer's Z	Students t	Chi square	Normal	Fischer's Z

35	For Fischer's Z-distribution, Z-statistic is ………………..	Z=1/2 log F	Z=2 log F	Z=1/3log F	Z=log F	Z=1/2 log F
36	Z distribution is formulated from…………………..distribution.	F	Chi square	t	Standard Normal	F
37	Which of the following property is not a desirable property of a point estimator?	Consistency	Efficiency	Sufficiency	Bias	Bias
38	Which of the following is most relevant for deriving a point estimator?	Sample size	Confidence desired	Variability in the population	Population size	Sample size
39	Which of the following factor does not usually affect the width of a Confidence interval?	Sample size	Confidence desired	Variability in the population	Population size	Population size
40	Which of the following is not a property of the sample mean?	Unbiased	Efficient	Sufficient	Standardized	Standardized
41	Given the level of confidence as 95% and margin of error as 2%, the minimum sample size required to estimate the population is	1256	2009	2401	2815	2401
42	Which of the following statement about confidence limit for population mean is not true?	50% confidence limits are wider than 95%	90% confidence limits are wider than 95%	95% confidence limits are wider than 99%	99% confidence limits are widest	99% confidence limits are widest
43	Which of the following statement is normally true?	Acceptance region is more than the critical region	Acceptance region is less than the critical region	Acceptance region is equal to the critical region	There is no relationship between acceptance region and critical region	Acceptance region is more than the critical region
44	Which one of the following is not a step in conducting a test of significance?	Set up the Null hypothesis	Decide the level of significance	Decide the power of the Test	Decide on the appropriate statistics	Decide the power of the Test

45	The p-values indicates the :3	Minimum level of significance at which the null hypothesis would be rejected	Maximum level of significance at which the null hypothesis would be accepted	Maximum level of significance at which the null hypothesis would be rejected	Minimum level of significance at which the null hypothesis would be accepted	Maximum level of significance at which the null hypothesis would be accepted
46	If the value of proportion p in the population is not known, the most conservative sample size required for the given margin can be calculated by assuming	p = 1/2	p = 1/3	p = 1/4	p = 3/4	p = 1/2
47	In which distribution the ratio of two variances under the null hypothesis of equal variance is taken?	The t-distribution	The uniform distribution	The Normal distribution	The F-distribution	The F-distribution

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu
Department of Mathematics

Subject : Mathematical Statistics	**Semester III**	**L T P C**
Subject Code : 18MMP304	**Class : II M.Sc Mathematics**	**4 0 0 4**

UNIT III

Significance test: Concept of a statistical test -Parametric tests for small samples and large samples Chi-square test -Kolmogorov Theorem-Smirnov Theorem-Tests of Kolmogorov and Smirnov type The Wald-Wolfovitz and Wilcoxon-Mann-Whitney tests -Independence Tests by contingency tables.

SUGGESTED READINGS

1. Marek Fisz, (1980). Probability Theory and Mathematical Statistics, John Wiley and Sons, New York.
2. Meyer, (2006). Introduction to Probability and Statistical applications, Oxford and IBH Publishing Co. Pvt. Ltd. New Delhi.
3. Sheldon M. Ross, (2009). Introduction to probability and statistics for engineers and scientists, Third edition, Academic press.
4. Heinz Bauer, (1996), Probability Theory, Narosa Publishing House, London.
5. Parimal Mukhopadhyay, (2012). Theory of Probability, New central book agency, Calcutta.

Hypothesis:

A statistical hypothesis is an assumption that we make about a population parameter, which may or may not be true concerning one or more variables.

According to Prof. Morris Hamburg "A hypothesis in statistics is simply a quantitative statement about a population".

Hypothesis testing:

Hypothesis testing is to test some hypothesis about parent population from which the sample is drawn.

Example:

A coin may be tossed 200 times and we may get heads 80 times and tails 120 times, we may now be interested in testing the hypothesis that the coin is unbiased.

To take another example we may study the average weight of the 100 students of a particular college and may get the result as 110lb. We may now be interested in testing the hypothesis that the sample has been drawn from a population with average weight 115lb.

Hypotheses are two types

1. Null Hypothesis
2. Alternative hypothesis

Null hypothesis:

The hypothesis under verification is known as *null hypothesis* and is denoted by H_0 and is always set up for possible rejection under the assumption that it is true.

For example, if we want to find out whether extra coaching has benefited the students or not, we shall set up a null hypothesis that *"extra coaching has not benefited the students"*. Similarly, if we want to find out whether a particular drug is effective in curing malaria we will take the null hypothesis that *"the drug is not effective in curing malaria*".

Alternative hypothesis:

The rival hypothesis or hypothesis which is likely to be accepted in the event of rejection of the null hypothesis H_0 is called alternative hypothesis and is denoted by H_1 or H_a.

For example, if a psychologist who wishes to test whether or not a certain class of people have a mean I.Q. 100, then the following null and alternative hypothesis can be established.

The null hypothesis would be

$$H_0 : \mu = 100$$

Then the alternative hypothesis could be any one of the statements.

$$H_1 : \mu \neq 100$$
$$(or)\, H_1 : \mu > 100$$
$$(or)\, H_1 : \mu < 100$$

Errors in testing of hypothesis:

After applying a test, a decision is taken about the acceptance or rejection of null hypothesis against an alternative hypothesis. The decisions may be four types.

1) The hypothesis is true but our test rejects it.(type-I error)
2) The hypothesis is false but our test accepts it. .(type-II error)

3) The hypothesis is true and our test accepts it.(correct)
4) The hypothesis is false and our test rejects it.(correct)

The first two decisions are called errors in testing of hypothesis.

i.e.1) Type-I error 2) Type-II error

1) Type-I error: The type-I error is said to be committed if the null hypothesis (H_0) is true but our test rejects it.

2) Type-II error: The type-II error is said to be committed if the null hypothesis (H_0) is false but our test accepts it.

Level of significance:

The maximum probability of committing type-I error is called level of significance and is denoted by α .

α = P (Committing Type-I error)
= P (H_0 is rejected when it is true)

This can be measured in terms of percentage i.e. 5%, 1%, 10% etc…….

Power of the test:

The probability of rejecting a false hypothesis is called power of the test and is denoted by $1-\beta$.

Power of the test =P (H_0 is rejected when it is false)
= 1- P (H_0 is accepted when it is false)
= 1- P (Committing Type-II error)
= 1- β

- A test for which both α and β are small and kept at minimum level is considered desirable.
- The only way to reduce both α and β simultaneously is by increasing sample size.
- The type-II error is more dangerous than type-I error.

Critical region:

A statistic is used to test the hypothesis H_0. The test statistic follows a known distribution. In a test, the area under the probability density curve is divided into two regions i.e. the region of acceptance and the region of rejection. The region of rejection is the region in which H_0 is rejected. It indicates that if the value of test statistic lies in this region, H_0 will be rejected. This region is called critical region. The area of the critical region is equal to the level of significance α . The critical region is always on the tail of the distribution curve. It may be on both sides or on one side depending upon the alternative hypothesis.

One tailed and two tailed tests:

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$, it is called a two tailed test. In this case the critical region is located on both the tails of the distribution.

A test with the null hypothesis $H_0: \theta = \theta_0$ against the alternative hypothesis $H_1: \theta > \theta_0$ (right tailed alternative) or $H_1: \theta < \theta_0$ (left tailed alternative) is called one tailed test. In this case the critical region is located on one tail of the distribution.

$H_0: \theta = \theta_0$ against $H_1: \theta > \theta_0$ ------- right tailed test

$H_0: \theta = \theta_0$ against $H_1: \theta < \theta_0$ ------- left tailed test

Sampling distribution:

Suppose we have a population of size 'N' and we are interested to draw a sample of size 'n' from the population. In different time if we draw the sample of size n, we get different samples of different observations i.e. we can get ${}^N c_n$ possible samples. If we calculate some particular statistic from each of the ${}^N c_n$ samples, the distribution of sample statistic is called sampling distribution of the statistic. For example if we consider the mean as the statistic, then the distribution of all possible means of the samples is a distribution of the sample mean and it is called sampling distribution of the mean.

Standard error:

Standard deviation of the sampling distribution of the statistic t is called standard error of t.

i.e. S.E (t)= $\sqrt{Var(t)}$

Utility of standard error:

1. It is a useful instrument in the testing of hypothesis. If we are testing a hypothesis at 5% l.o.s and if the test statistic i.e. $|Z| = \left|\frac{t - E(t)}{S.E(t)}\right| > 1.96$ then the null hypothesis is rejected at 5% l.o.s otherwise it is accepted.
2. With the help of the S.E we can determine the limits with in which the parameter value expected to lie.
3. S.E provides an idea about the precision of the sample. If S.E increases the precision decreases and vice-versa. The reciprocal of the S.E i.e. $\frac{1}{S.E}$ is a measure of precision of a sample.
4. It is used to determine the size of the sample.

Test statistic:

The test statistic is defined as the difference between the sample statistic value and the hypothetical value, divided by the standard error of the statistic.

i.e. test statistic $Z = \frac{t - E(t)}{S.E(t)}$

Procedure for testing of hypothesis:

1. Set up a null hypothesis i.e. $H_0: \theta = \theta_0$.
2. Set up a alternative hypothesis i.e. $H_1: \theta \neq \theta_0$ or $H_1: \theta > \theta_0$ or $H_1: \theta < \theta_0$
3. Choose the level of significance i.e. α.

4. Select appropriate test statistic Z.
5. Select a random sample and compute the test statistic.
6. Calculate the tabulated value of Z at α % l.o.s i.e. Z_α .
7. Compare the test statistic value with the tabulated value at α % l.o.s. and make a decision whether to accept or to reject the null hypothesis.

Large sample tests:

The sample size which is greater than or equal to 30 is called as large sample and the test depending on large sample is called large sample test.

The assumption made while dealing with the problems relating to large samples are

Assumption-1: The random sampling distribution of the statistic is approximately normal.

Assumption-2: Values given by the sample are sufficiently closed to the population value and can be used on its place for calculating the standard error of the statistic.

Large sample test for single mean (or) test for significance of single mean:

For this test

The null hypothesis is $H_0 : \mu = \mu_0$

against the two sided alternative $H_1 : \mu \neq \mu_0$

where μ is population mean

μ_0 is the value of μ

Let $x_1, x_2, x_3, ..., x_n$ be a random sample from a normal population with mean μ and variance σ^2

i.e. if $X \sim N(\mu, \sigma^2)$ then $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$, Where $\bar{x}$ be the sample mean

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\bar{x} - E(\bar{x})}{S.E(\bar{x})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Note: if the population standard deviation is unknown then we can use its estimate s, which will be calculated from the sample. $s = \sqrt{\dfrac{1}{n-1}\sum(x - \bar{x})^2}$.

Large sample test for difference between two means:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances $\sigma_1^{\ 2}$ and $\sigma_2^{\ 2}$ respectively

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means for the first and second populations respectively

Then $\bar{x}_1 \sim N\left(\mu_1, {\sigma_1^{\ 2}}/{n_1}\right)$ and $\bar{x}_2 \sim N\left(\mu_2, {\sigma_2^{\ 2}}/{n_2}\right)$

Therefore $\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \dfrac{\sigma_1^{\ 2}}{n_1} + \dfrac{\sigma_2^{\ 2}}{n_2}\right)$

For this test

The null hypothesis is $H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$

against the two sided alternative $H_1 : \mu_1 \neq \mu_2$

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{\sigma_1^{\ 2}}{n_1} + \dfrac{\sigma_2^{\ 2}}{n_2}}} \sim N(0,1) \text{[since } \mu_1 - \mu_2 = 0 \text{ from } H_0]$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Note: If $\sigma_1^{\ 2}$ and $\sigma_2^{\ 2}$ are unknown then we can consider $S_1^{\ 2}$ and $S_2^{\ 2}$ as the estimate value of $\sigma_1^{\ 2}$ and $\sigma_2^{\ 2}$ respectively..

Large sample test for single standard deviation (or) test for significance of standard deviation:

Let $x_1, x_2, x_3, \ldots\ldots\ldots\ldots\ldots, x_n$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2,

for large sample, sample standard deviation s follows a normal distribution with mean σ and variance ${\sigma^2}/{2n}$ i.e. $s \sim N\left(\sigma, {\sigma^2}/{2n}\right)$

For this test

The null hypothesis is $H_0 : \sigma = \sigma_0$

against the two sided alternative $H_1 : \sigma \neq \sigma_0$

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{s - E(s)}{S.E(s)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{s - \sigma}{\sigma / \sqrt{2n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for difference between two standard deviations:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances $\sigma_1{}^2$ and $\sigma_2{}^2$ respectively

Let s_1 and s_2 be the sample standard deviations for the first and second populations respectively

Then $s_1 \sim N\left(\sigma_1, \sigma_1{}^2 / 2n_1\right)$ and $\bar{x}_2 \sim N\left(\sigma_2, \sigma_2{}^2 / 2n_2\right)$

Therefore s_1 - $s_2 \sim N\left(\sigma_1 - \sigma_2, \dfrac{\sigma_1{}^2}{2n_1} + \dfrac{\sigma_2{}^2}{2n_2}\right)$

For this test

The null hypothesis is $H_0 : \sigma_1 = \sigma_2 \Rightarrow \sigma_1 - \sigma_2 = 0$

against the two sided alternative $H_1 : \sigma_1 \neq \sigma_2$

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(s_1 - s_2) - E(s_1 - s_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2) - (\sigma_1 - \sigma_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2)}{\sqrt{\frac{\sigma_1{}^2}{2n_1} + \frac{\sigma_2{}^2}{2n_2}}} \sim N(0,1) \text{[since } \sigma_1 - \sigma_2 = 0 \text{ from } H_0\text{]}$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for single proportion (or) test for significance of proportion:

Let x is number of success in n independent trails with constant probability p, then x follows a binomial distribution with mean np and variance npq.

In a sample of size n let x be the number of persons processing a given attribute then the sample proportion is given by $\hat{p} = \frac{x}{n}$

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

And $V(\hat{p}) = V\left(\frac{x}{n}\right) = \frac{1}{n^2}V(x) = \frac{1}{n^2}npq = \frac{pq}{n}$

$S.E(\hat{p}) = \sqrt{\frac{pq}{n}}$

For this test

The null hypothesis is $H_0 : p = p_0$

against the two sided alternative $H_1 : p \neq p_0$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p} - E(\hat{p})}{S.E(\hat{p})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for single proportion (or) test for significance of proportion:

let x_1 and x_2 be the number of persons processing a given attribute in a random sample of size n_1 and n_2 then the sample proportions are given by $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$

Then $E(\hat{p}_1) = p_1$ and $E(\hat{p}_2) = p_2 \Rightarrow E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

And $V(\hat{p}_1) = \frac{p_1 q_1}{n_1}$ and $V(\hat{p}_2) = \frac{p_2 q_2}{n_2} \Rightarrow V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

$$S.E(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}} \text{ and } S.E(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}} \Rightarrow S.E(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

For this test

The null hypothesis is $H_0 : p_1 = p_2$

against the two sided alternative $H_1 : p_1 \neq p_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p}_1 - \hat{p}_2 - E(\hat{p}_1 - \hat{p}_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \sim N(0,1) \quad \text{Since } p_1 = p_2 \text{ from } H_0$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

When p is not known p can be calculated by $p = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ and $q = 1 - p$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

- As σ is unknown,

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = \left[\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right]$$

Step 2: If μ_0 falls into the above confidence intervals, then

do *not* reject H_0 . Otherwise, reject H_0 .

Example 1:

The average starting salary of a college graduate is \$19000 according to government's report. The average salary of a random sample of 100 graduates is \$18800. The standard error is 800.

(a) Is the government's report reliable as the level of significance is 0.05.

(b) Find the p-value and test the hypothesis in (a) with the level of significance $\alpha = 0.01$.

(c) The other report by some institute indicates that the average salary is \$18900. Construct a 95% confidence interval and test if this report is reliable.

[solutions:]

(a)

$$H_0 : \mu = \mu_0 = 19000 \text{ vs. } H_a : \mu \neq \mu_0 = 19000,$$
$$n = 100, \bar{x} = 18800, s = 800, \alpha = 0.05$$

Then,

$$|z| = \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| = \left| \frac{18800 - 19000}{800/\sqrt{100}} \right| = |-2.5| = 2.5 > z_{\alpha/2} = z_{0.025} = 1.96.$$

Therefore, reject H_0.

(b)

$$\text{p-value} = P(|Z| > |z|) = P(|Z| > 2.5) = 2 \cdot P(Z > 2.5) = 0.0124 > 0.01$$

Therefore, ***not*** reject H_0.

(c)

$$H_0 : \mu = \mu_0 = 18900 \text{ vs } H_a : \mu \neq \mu_0 = 18900,$$

A 95% confidence interval is

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 18800 \pm z_{0.025} \frac{800}{\sqrt{100}} = 18800 \pm 1.96 \cdot 80 = [18643.2, 18956.8].$$

Since $\mu_0 = 18900 \in [18643.2, 18956.8]$, Therefore, ***not*** reject H_0.

Example 2:

A sample of 49 provides a sample mean of 38 and a sample standard deviation of 7. Let $\alpha = 0.05$. Please test the hypothesis

$$H_0 : u = 40 \ \ vs. \ \ H_a : u \neq 40.$$

based on

(a) classical hypothesis test

(b) p-value

(c) confidence interval.

[solution:]

$$\bar{x} = 38, \ s = 7, \ u_0 = 40, \ n = 49, \ z = \frac{\bar{x} - u_0}{s/\sqrt{n}} = \frac{38 - 40}{7/\sqrt{49}} = -2.$$

(a)

$|z| = 2 > 1.96 = z_{0.025}$ we reject H_0.

(b)

$p - value = P(|Z| > |z|) = P(|Z| > 2) = 2 * (1 - 0.9772) = 0.0456 < 0.05 = \alpha$ we reject H_0.

(c)

$100 \times (1 - \alpha)\% = 95\%$ confidence interval is

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 38 \pm z_{0.025} \frac{7}{\sqrt{49}} = 38 \pm 1.96 = [36.04, 39.96].$$

Since $40 \notin [36.04, 39.96]$, we reject H_0.

Hypothesis Testing for the Mean (Small Samples)

For samples of size less than 30 and when σ is unknown, if the population has a normal, or nearly normal, distribution, the *t*-distribution is used to test for the mean μ.

Using the t-Test for a Mean μ when the sample is small		
Procedure	**Equations**	**Example**
State the claim mathematically and verbally. Identify the null and alternative hypotheses	State H_0 and H_a	$H_0 : \mu \geq 16500$ $H_a : \mu < 16500$ $n = 14, \bar{x} = 15700, s = 1250$

Specify the level of significance	Specify α	$\alpha = 0.05$
Identify the degrees of freedom and sketch the sampling distribution	$d.f = n-1$	$d.f. = 13$
Determine any critical values. If test is left tailed, use One tail, α column with a negative sign. If test is right tailed, use One tail, α column with a positive sign. If test is two tailed, use Two tails, α column with a negative and positive sign.	Table 5 (t-distribution) in appendix B	The test is left-tailed. Since test is left tailed and $d.f = 13$, the critical value is $t_0 = -1.771$
Determine the rejection regions.	The rejection region is $t < t_0$	The rejection region is $t < -1.771$
Find the standardized test statistic	$t = \dfrac{\bar{x}-\mu}{\sigma_{\bar{x}}} \approx \dfrac{\bar{x}-\mu}{s/\sqrt{n}}$	$t = \dfrac{15700-16500}{1250/\sqrt{14}} \approx -2.39$
Make a decision to reject or fail to reject the null hypothesis	If t is in the rejection region, reject H_0, Otherwise do not reject H_0	Since $-2.39 < -1.771$, reject H_0
Interpret the decision in the context of the original claim.		Reject claim that mean is at least 16500.

Chi-Square Tests and the F-Distribution

Goodness of Fit

DEFINITION A **chi-square goodness-of-fit test is** used to test whether a frequency distribution fits an expected distribution.

The test is used in a multinomial experiment to determine whether the number of results in each category fits the null hypothesis:

H_0: The distribution fits the proposed proportions

H_1: The distribution differs from the claimed distribution.

To calculate the test statistic for the chi-square goodness-of-fit test, you can use observed frequencies and expected frequencies.

DEFINITION The **observed frequency O** of a category is the frequency for the category observed in the sample data.

The **expected frequency E** of a category is the calculated frequency for the category. Expected frequencies are obtained assuming the specified (or hypothesized) distribution. The expected frequency for the *i*th category is

$E_i = np_i$

where n is the number of trials (the sample size) and p_i is the assumed probability of the ith category.

The Chi-square Goodness of Fit Test: The sampling distribution for the goodness-of-fit test is a chi-square distribution with $k-1$ degrees of freedom where k is the number of categories. The test statistic is

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

where O represents the observed frequency of each category and E represents the expected frequency of each category. To use the chi-square goodness of fit test, *the following must be true* .

1. The observed frequencies must be obtained using a random sample.
2. The expected frequencies must be ≥ 5.

Performing the Chi-Square Goodness-of-Fit Test (p 496)		
Procedure	Equations	Example **(p 497)**
Identify the claim. State the null and alternative hypothesis.	State H_0 and H_1	H_0 : Classical 4% Country 36% Gospel 11% Oldies 2% Pop 18% Rock 29%
Specify the significance level	Specify α	$\alpha = 0.01$
Determine the degrees of freedom	d.f. = #categories - 1	$d.f. = 6-1 = 5$
Find the critical value	χ^2_α: Obtain from Table 6 Appendix B	$\varphi^2_{0.01}(d.f = 5) = 15.086$
Identify the rejection region	$\chi^2 \geq \chi^2_\alpha$	$\chi^2 \geq 15.086$
Calculate the test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	Survey results, n = 500 Classical O= 8 E = .04*500 = 20 Country O = 210 E = .36*500 = 180 Gospel O = 7 E = .11*500 = 55

		Oldies O = 10 E = .02*500 = 10 Pop O = 75 E = .18*500 = 90 Rock O= 125 E = .29*500 = 145 Substituting $\chi^2 = 22.713$
Make the decision to reject or fail to reject the null hypothesis	Reject if χ^2 is in the rejection region Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\le \alpha$	Since 22.713 > 15.086 we reject the null hypothesis Equivalently $P(X \ge 22.713) < 0.01$ so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.)
Interpret the decision in the context of the original claim		Music preferences differ from the radio station's claim.

Using Minitab to perform the Chi-Square Goodness-of-Fit Test (Manual p 237)
The data from the example above (Example 2 p 497) will be used.
Enter Three columns: Music Type: Classical, etc, Observed: 8 etc, Distribution 0.04, etc. (Note the names of the columns 'Music Types', 'Observed' and 'Distribution' are entered in the gray row at the top.)

Music Type	Observed	Distribution	Expected
Classical	8	0.04	20
Country	210	0.36	180
Gospel	72	0.11	55
Oldies	10	0.02	10
Pop	75	0.18	90
Rock	125	0.29	145

Next calculate the chi-square statistic, $(O-E)^2/E$ as follows: and should contain the calculated values.

7.2000
5.0000
41.8909
0.0000
2.5000
2.7586

The chi-square statistic is displayed in the session window as follows:

= 22.7132

Next calculate the P-value: The following is displayed on the Session Window.

Cumulative Distribution Function
Chi-Square with 5 DF
x P(X <= x)
22.7132 0.999617

$P(X \leq 22.7132) = 0.999617$ So the P-value = 1 – 0.999617 = 0.000383. This is less that α = 0.01 so we reject the null hypothesis.

Instead of calculating the P-value, we could have found the critical value from the Chi-Square table (Table 6 Appendix B) for 5 degrees of freedom as we did above. The value is 15.086, and since our test statistic is 22.7132, we reject the null hypothesis.

Chi-Square with M&M's

H_0: Brown: 13%, Yellow: 14%, Red: 13%, Orange: 20%, Green 16%, Blue 24%
Significance level: α = 0.05
Degrees of freedom: number of categories – 1 = 5
Critical Value: $\chi^2{}_{0.05}(d.f.=5) = 11.071$
Rejection Region: $\chi^2 \geq 11.071$
Test Statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$, where O is the actual number of M&M's of each color in the bag and E is the proportions specified under H_0 times the total number.
Reject H_0 if the test statistic is greater than the critical value (1.145)

Section 10.2 Independence

This section describes the chi-square test for independence which tests whether two random variables are independent of each other.

DEFINTION An ***r x c* contingency table** shows the observed frequencies for the two variables. The observed frequencies are arranged in r rows and c columns. The intersection of a row and a column is called a **cell.**

The following is a contingency table for two variables A and B where f_{ij} is the frequency that A equals A_i and B equals $B_{j.}$

	A₁	**A₂**	**A₃**	**A₄**	**A**
B₁	f_{11}	f_{12}	f_{13}	f_{14}	$f_{1.}$
B₂	f_{21}	f_{22}	f_{23}	f_{24}	$f_{2.}$
B₃	f_{31}	f_{32}	f_{33}	f_{34}	$f_{3.}$
B	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	f

If A and B are independent, we'd expect

$$f_{ij} = prob(A = A_i) * prob(B = B_j) * f = \left(\frac{f_{i.}}{f}\right)\left(\frac{f_{.j}}{f}\right) f = \frac{(f_{i.})(f_{.j})}{f}$$

$$\frac{(sum of\ row\ i) * (sum of\ column\ j)}{sample\ size}$$ (

Example 1 Determining the expected frequencies of CEO's ages as a function of company size under the assumption that age is independent of company size.

	<= 39	40 - 49	50 - 59	60 - 69	>= 70	Total
Small/midsize	42	69	108	60	21	300
Large	5	18	85	120	22	250
Total	47	87	193	180	43	550

	<= 39	40 - 49	50 - 59	60 - 69	>= 70	Total
Small/midsize	$\frac{300*47}{550} \approx 25.64$	$\frac{300*87}{550} \approx 47.45$	$\frac{300*193}{550} \approx 105.27$	$\frac{300*180}{550} \approx 98.18$	$\frac{300*43}{550} \approx 23.45$	300
Large	$\frac{250*47}{550} \approx 21.36$	$\frac{250*87}{550} \approx 39.55$	$\frac{250*193}{550} \approx 87.73$	$\frac{250*180}{550} \approx 81.82$	$\frac{250*43}{550} \approx 19.55$	250
Total	47	87	193	180	43	550

After finding the expected frequencies under the assumption that the variables are independent, you can test whether they are independent using the chi-square independence test.

DEFINITION A **chi-square independence test** is used to test the independence of two random variables. Using a chi-square test, you can determine whether the occurrence of one variable affects the probability of occurrence of the other variable.

To use the test,

1. The observed frequencies must be obtained from a random sample
2. Each expected frequency must be ≥ 5

The sampling distribution for the test is a chi-square distribution with

$(r-1)(c-1)$

degrees of freedom, where r and c are the number of rows and columns, respectively, of the contingency table. The test statistic for the chi-square independence test is

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

where O represents the observed frequencies and E represents the expected frequencies.

To begin the test we state the null hypothesis that the variables are independent and the alternative hypothesis that they are dependent.

Performing a Chi-Square Test for Independence (p 507)		
Procedure	Equations	Example2 **(p 507)**
Identify the claim. State the null and alternative hypotheses.	State H_0 and H_1	H_0 : CEO's ages are independent of company size H_1 : CEO's ages are dependent on company size.
Specify the level of significance	Specify α	$\alpha = 0.01$
Determine the degrees of freedom	$d.f. = (r-1)(c-1)$	$d.f. = (2-1)(5-1) = 4$
Find the critical value.	χ^2_α : Obtain from Table 6, Appendix B	$\chi^2_\alpha \geq 13.277$
Identify the rejection region	$\chi^2 \geq \chi^2_\alpha$	$\chi^2 \geq 13.277$
Calculate the test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	$\sum \frac{(O-E)^2}{E} \approx 77.9$ Note that O is in the table of actual CEO's ages above, and E is in the table of Expected CEO's ages (if independent of size) above
Make a decision to reject or fail to reject the null hypothesis	Reject if χ^2 is in the rejection region. Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\leq \alpha$	Since 77.9 > 13.277 we reject the null hypothesis Equivalently $P(X \geq 77.0) < \alpha$ so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.)
Interpret the decision in the context of the original claim		CEO's ages and company size are dependent.

The test statistic (77.887) is greater than the critical value obtained from Table 6, Appendix B (13.277) so the null hypothesis is rejected. (Alternatively the P-Value (0.000) is less than the level of significance, α (0.01) so the null hypothesis is rejected.)

3. An urban geographer randomly samples 20 new residents of a neighborhood to determine their ratings of local bus service. The scale used is as follows: 0–very dissatisfied, 1–dissatisfied, 2– neutral, 3–satisfied, 4–very satisfied. The 20 responses are 0,4,3, 2,2,1,1,2,1,0,01,2,1,3,4,2,0,4,1. Use the sign test to see whether the population median is 2.

   Solution:

There are 5 observations above the hypothesized median. Because the sample size is larger than 10, we test using the sample proportion p = 5/20 = 0.25. Using the PROB-VALUE method the steps in this test are:

1) H_0: $\pi = 0.5$ and H_A: $\pi \neq 0.5$
2) We will use the *Z*-distribution
3) We will use the 5%-level, thus $\alpha = 0.05$
4) The test statistic is $z = (0.25 - 0.5)/\sqrt{0.25/20} = -2.24$
5) Table A-4 shows that $P(|Z| > 2.24) \approx 0.025$.
6) Because PROB-VALUE < α, we reject H_0. We conclude π is different than 0.5, and thus the median is different than 2.

4. A course in statistical methods was team-taught by two instructors, Professor Jovita Fontanez and Professor Clarence Old. Professor Fontanzez used many active learning techniques, whereas Old employed traditional methods. As part of the course evaluation, students were asked to indicate their instructor preference. There was reason to think students would prefer Fontanez, and the sample obtained was consistent with that idea: of the 12 students surveyed, 8 preferred Professor Fontanez and 2 preferred Professor Old. The remaining students were unable to express a preference. Test the hypothesis that the students prefer Fontanez. (*Hint:* Use the sign test.)

Solution:

Although the sample is large enough for a normal approximation, we will use the binomial distribution to illustrate its application. Of the 12 observations, 8 preferred Prof. Fontanez, thus we need the probability of observing 8 or more successes in 12 trials of a Bernoulli process with the probability of success equal to 0.5. From Table A-1, we get

$$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12)$$

$$P(X \geq 8) = 0.1208 \quad + 0.0537 \quad + 0.0161 \quad + 0.0029 \quad + 0.002 \quad = 0.1937$$

Adopting the 5% uncertainty level, we see that PROB-VALUE > α. Thus we fail to reject H_0. We cannot conclude students prefer Fontanez.

5. Use the data in Table 10-8 to perform two Mann–Whitney tests: (a) compare uncontrolled intersections and intersections with yield signs, and (b) compare uncontrolled intersections and intersections with stop signs.

   Solution:

   (a) The rank sums are 119.5 and 90.5 for the yield-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.272. Adopting a 5% level of uncertainty, we fail to reject the hypothesis of no difference. We cannot conclude the samples were drawn from different populations.

(b) The rank sums are 130.5 and 59.5 for the stop-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.013. Adopting a 5% level of uncertainty, we reject the hypothesis of no difference. We conclude the samples were drawn from different populations.

6. Solid-waste generation rates measured in metric tons per household per year are collected randomly in selected areas of a township. The areas are classified as high-density, low density, or sparsely settled. It is thought that generation rates probably differ because of differences in waste collection and opportunities for on-site storage. Do the following data support this hypothesis?

High Density	Low Density	Sparsely Settled
1.84	2.04	1.07
3.06	2.28	2.31
3.62	4.01	0.91
4.91	1.86	3.28
3.49	1.42	1.31

Solution:

We will use the multi-sample Kruskal-Wallis test with an uncertainly level □ = 0.1. The null hypothesis is that all samples have been drawn from the same population. The rank sums are 55, 39 and 26 for the high density, low density, and sparsely settled samples respectively. The Kruskal-Wallis statistic is

$$H = \frac{12}{15(15+1)}\left(\frac{55^2}{5}+\frac{39^2}{5}+\frac{26^2}{5}\right) - 3(15+1) = 4.22$$

Using the □2 distribution with $3 - 1 = 2$ degrees of freedom, the associated PROB-VALUE is 0.121. We fail to reject the null hypothesis. The sample does not support the hypothesis of differing waste generation rates.

7. The distances travelled to work by a random sample of 12 people to their places of work in 1996 and again in 2006 are shown in the following table.

	Distance (km)			Distance (km)	
Person	1996	2006	Person	1996	2006
1	8.6	8.8	7	7.7	6.5

2	7.7	7.1	8	9.1	9
3	7.7	7.6	9	8	7.1
4	6.8	6.4	10	8.1	8.8
5	9.6	9.1	11	8.7	7.2
6	7.2	7.2	12	7.3	6.4

Has the length of the journey to work changed over the decade?

Solution:

The sample can be considered as twelve paired observations. By taking differences between paired values, we get measures of the change for each individual. If the median change for the population is zero, we expect a sample to have a median difference near zero. Thus we will do a sign test for the median difference with a hypothesized value of zero. In other words, the hypotheses are $H_0: \eta = 0$ and $H_A: \eta \neq 0$ We denote samples values whose distance decreased with a minus sign. Sample values with a positive difference get a plus sign. The sample becomes

S = {-,+,+,+,+,0,+,-,+,-,+,+}

Ignoring the tie, this is a sample of size 11 with 8 values above the hypothesized median. We are using Format (C) of Table 10-1, thus the PROB-VALUE is $2P(X \geq 8)$ where X is a binomial variable with = 0.5. From the equation for the binomial, the PROB-VALUE is found to be 0.113. At the = 10% level, we fail the reject the null hypothesis. We cannot conclude there has been a change in distance.

Nonparametric Hypothesis Testing

EARNING OBJECTIVES

By the end of this chapter, you will be able to:

1. Identify and cite examples of situations in which nonparametric tests of hypothesis are appropriate.
2. Explain the logic of nonparametric hypothesis testing for ordinal variables as applied to the Mann-Whitney *U* and runs tests.
3. Perform Mann-Whitney *U* and runs tests using the five-step model as a guide, and correctly interpret the results.
4. Select an appropriate nonparametric test.

1 INTRODUCTION

The chi square test, a **nonparametric test** of hypothesis. In this section, we will consider two other nonparametric tests, the **Mann-Whitney *U* test** and the **runs test**. Both tests are appropriate in the two-sample case when the variable of interest is measured at the ordinal level and, under conditions to be specified, may be considered as alternatives to the differences-in-means tests presented in Chapter 8 of the text. Before introducing the tests themselves, let us consider the situations where they can or should be used.

The Mann-Whitney *U* and the runs test represent a large class of tests of significance called "nonparametric" or "distribution-free" tests. These tests differ from the tests introduced in Chapters 7, 8, and 9 of the text in that they require no particular assumption about the shape of the population distribution. Recall that we needed to assume a normal sampling distribution in order to test proportions and means for significance. This assumption is satisfied only if sample size is large or if the additional assumption of a normally distributed population can be made (see the theorems presented in Chapter 5 of the text). Thus, means and proportions based on small samples can be tested for their significance only when we can make assumptions about the particular shape of the population distribution. Needless to say, researchers often find themselves in situations where they are uncomfortable with such precise assumptions about an unknown distribution.

Nonparametric tests, on the other hand, do not require the assumption of a normally shaped population distribution. In fact, with nonparametric tests, we do not need to assume any particular shape for the population distribution. The tests to be presented here, therefore, have wide applicability in situations where the researcher is unsure of the form of the population distribution. Since the requirement of a normal population can be relaxed with large samples, these tests are particularly useful when working with small samples.

Besides the assumptions of a normal sampling distribution, the test for means also requires the assumption of interval-ratio level of measurement. The Mann-Whitney *U* and the runs test may also be used when the researcher is uncertain of the appropriateness of this assumption.

To summarize, you should seriously consider ordinal-level, nonparametric tests of significance when either the assumption of interval-ratio measurement or the assumption of normal sampling distribution is in doubt for a test of sample means. Both of these assumptions are included in the model assumptions (step 1) for tests of means. The model can be thought of as the mathematical foundation for the rest of the test, and the researcher should be very certain of the appropriateness of these assumptions before placing confidence in the test results. If the model assumptions can be satisfied, tests of the differences in means are preferable to the nonparametric alternatives because the assumption of interval-ratio level permits more sophisticated mathematical procedures to be performed on the data. Thus, a greater volume of more precise information can be amassed with respect to the variable of interest. If the model assumptions for means cannot be satisfied, the Mann-Whitney *U* and the runs test are very useful alternative tests for investigating the significance of the difference between two samples.

2 THE MANN-WHITNEY *U* TEST

In many ways, this test is similar to the test of significance for the difference in sample means. In both cases we compare random samples as a way of making inferences about the possible differences between two populations. In both cases, the test statistics are computed from the samples and compared with the sampling distribution of all possible sample outcomes. Instead of computing means as the sample statistic, however, the Mann-Whitney test is based on the ranking of the sample scores. This is appropriate since ranking is the most sophisticated mathematical operation that can be performed on ordinal-level data.

The computation of Mann-Whitney *U* test is straightforward. First, the scores from both samples are pooled and ranked from highest to lowest. Second, the ranks for the two samples are totaled and compared. If the two samples represent very different populations, the cases from each sample should be grouped together. The greater the extent to which the two samples differ, the greater the difference in the sum of the total ranks. If the populations are not significantly different from each other, then the cases from the two samples should be intermixed, and the sum of the total ranks would have similar values.

An example should clarify the underlying logic as well as the computational routines for this test. Assume that you are concerned with sex differences in the level of satisfaction with the social life available on your campus. You have administered survey instruments to randomly selected samples of male and female students and have devised a scale on which a high score indicates great satisfaction. The scores, by sex, are presented in Table 1.

TABLE 1 SCORES ON SATISFACTION SCALE FOR MALE AND FEMALE STUDENTS

Sample 1 (Male)		**Sample 2 (Female)**	
Case	**Score**	**Case**	**Score**
1	42	**13**	45
2	35	**14**	40
3	30	**15**	32
4	25	**16**	30
5	19	**17**	28
6	17	**18**	27
7	15	**19**	26
8	14	**20**	24
9	9	**21**	20
10	5	**22**	10
11	4	**23**	8
12	2	**24**	7

You may be tempted to compute a mean for each of these samples and test the difference for its significance. If you keep in mind that these scores are only ordinal in level of measurement, you should be able to restrain yourself long enough to compute the Mann-Whitney *U* test. First, pool the scores and rank them from high to low. If you encounter any tied scores, assign all of them the average of the ranks they would have used up if they had not been tied. For example, cases 3 and 16 are tied with scores of 30. If they had not been tied, they would have used up ranks 6 and 7. Therefore, assign both cases the rank 6.5 [(6 + 7)/2]. Ranked scores are presented in Table 2.

TABLE 2 RANKED SCORES ON SATISFACTION SCALE FOR MALE AND FEMALE STUDENTS

Sample 1 (Male)			**Sample 2 (Female)**		
Case	Score	Rank	Case	Score	Rank
1	42	2	**13**	45	1
2	35	4	**14**	40	3
3	30	6.5	**15**	32	5
4	25	11	**16**	30	6.5
5	19	14	**17**	28	8
6	17	15	**18**	27	9
7	15	16	**19**	26	10
8	14	17	**20**	24	12
9	9	19	**21**	20	13
10	5	22	**22**	10	18
11	4	23	**23**	8	20
12	2	24	**24**	7	21
$\sum R$		173.5			126.5

Next, the ranks are summed and the U statistic is computed. The formula for U is presented in Formula 1.

FORMULA 1 $$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1$$

where: n_1 = number of cases in sample 1

n_2 = number of cases in sample 2

$\sum R_1$ = the sum of the ranks for sample 1

For our sample problem above

$$U = (12)(12) + \frac{12(13)}{2} - 173.5$$

$$U = 144 + \frac{156}{2} - 173.5$$

$$U = 48.5$$

Note that we could have computed the *U* by using data from sample 2. This alternative solution, which we will label *U´* (*U* prime), would have resulted in a larger value for *U*. The smaller of the two values, *U* or *U´*, is always taken as the value of *U*. Once *U* has been calculated, *U´* can be quickly determined by means of Formula 2:

FORMULA 2 $U' = n_1 n_2 - U$

Thus, for the sample problem, *U´* would be (12)(12) - 48.5, or 95.5. Remember that the lower of these values is always taken as *U*.

Once the value of *U* has been determined, we must still conduct the test of significance. In step 1 of the five-step model, we assume ordinal level of measurement and make no assumption about the shape of the population distribution. The null hypothesis in step 2 is, as usual, a statement of "no difference:" the two populations represented by these samples are identical. Note that if this assumption is true, then the differences in total ranks should be small.

The alternative or research hypothesis is usually a statement to the effect that the two populations are different. This form for H_1 would direct the use of a two-tailed test. It is perfectly possible to use one-tailed tests with Mann-Whitney *U* when a direction for the difference can be predicted, but, to conserve space and time, we will consider only the two-tailed case.

In step 3, we will take advantage of the fact that, when total sample size (the combined number of cases in the two samples) is greater than or equal to 20, the sampling distribution of *U* approximates normality. This will allow us to use the *Z*-score table (see Appendix A of the textbook) to find the critical region as marked by *Z* (critical).

To compute the Mann-Whitney *U* test statistic (step 4), the necessary formulas are

FORMULA 3 $$Z \text{ (obtained)} = \frac{U - \mu_U}{\sigma_U}$$

where *U* = the sample statistic

$_U$ = the mean of the sampling distribution of sample *U*'s

σ_U = the standard deviation of the sampling distribution of sample *U*'s

FORMULA 4 $$_U = \frac{n_1 n_2}{2}$$

FORMULA 5 $$\sigma_U = \sqrt{\frac{n_1\, n_2\, (n_1 + n_2 + 1)}{12}}$$

For the sample problem above, □$_U$ will be $\frac{(12)(12)}{2} = 72$. The standard deviation of the sampling distribution is

$$\sigma_U = \sqrt{\frac{(12)(12)(12+12+1)}{12}}$$

$$\sigma_U = \sqrt{300}$$

$$\sigma_U = 17.32$$

We now have all the information we need to conduct a test of significance for *U*.

Step 1. Making Assumptions.

Model: Independent random samples

Level of measurement is ordinal

Step 2. Stating the Null Hypothesis

H_0: The populations from which the samples are drawn are identical on the variable of interest.

(H_1: The populations from which the samples are drawn are different on the variable of interest.)

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

Sampling distribution = *Z* distribution

Alpha = 0.05

Z (critical) = ± 1.96

Step 4. Calculating the Test Statistic. With *U* equal to 48.5, μ_U equal to 72, and σ_U of 17.32,

$$Z \text{ (obtained)} = \frac{U - \mu_U}{\sigma_U}$$

$$Z \text{ (obtained)} = \frac{48.5 - 72}{17.32}$$

$$Z \text{ (obtained)} = -1.36$$

Step 5. Making a Decision. The test statistic, a *Z* (obtained) of -1.36, does not fall in the critical region as marked by the *Z* (critical) of ± 1.96. Therefore, we fail to reject the null of no difference. Male students are not significantly different from female students in terms of their level of satisfaction with the social life available on campus. Note that if we had used the U' value of 95.5 instead of *U* in computing the test statistic, the value of *Z* (obtained) would have been +1.36, and our decision to fail to reject the null would have been exactly the same.

3 THE RUNS TEST

The runs test, also called Wald–Wolfowitz runs test, is very similar in logic and form to the Mann-Whitney *U* test. The null hypothesis is, again, the assumption that there is no significant difference between the populations from which the samples come. To conduct the test, the scores from both samples are pooled and then ranked from high to low as if they were a single sample. If the null is true, then the scores should be intermixed and there should be many runs. A **run** is defined as *any sequence of one or more scores from the same sample*. If the null is false and the populations are different on the variable being measured, then there should be very few runs.

To illustrate with the data on social satisfaction presented above, if we pooled the two samples and designated a female student with an **F** and a male student with an **M**, we would get the following sequences:

F	**M**	**F**	**M**	**F**	**M**	**FFFF**	**M**	**FF**	**MMMM**	**F**	**M**	**FF**	**MMM**
1	2	3	4	5	6	7	8	9	10	11	12	13	14

where the **F** at the far left represents the individual with the highest score (i.e., case 13 with a score of 45, who is a **F**emale), the next letter, **M**, represents the individual with the second highest score (case 1 with a score of 42, who is a **M**ale), and so on. The underlinings represent runs, and by counting we find that there are 14 runs in these data. So, for example, the 7th run is denoted as "FFFF" since there is a sequence of four females grouped side-by-side in this run, with the scores: 30, 28, 27, and 26; the 8th run reflects the cluster of the single male score: 25; the 9th run reflects the cluster of the two females scores: 24 and 20; the 10th run reflects the cluster of the four males scores: 19, 17, 15, and 14.

Although no prediction of the direction of the difference need be made with the runs test, the smaller the number of runs, the greater the likelihood that the difference between the two samples is significant. The lowest number of runs possible is, of course, two. If there had been sex differences in social satisfaction in the data above, with all female students expressing greater satisfaction than any male student, the data would have looked like this:

FFFFFFFFFFFF	**MMMMMMMMMMMM**
1	2

The fewer the differences between the two populations, the greater the intermixing between the two samples. More intermixing means a higher number of runs and, consequently, a lower probability of being able to reject the null hypothesis.

For situations in which total sample size is greater than or equal to 20, the sampling distribution of all possible sample runs approximates normality. Thus, we can use the Z-score table (Appendix A of the textbook) to find the critical region as marked by Z (critical). The runs test statistic, *Z* (obtained), is computed with Formula 6:

FORMULA 6 $Z \text{ (obtained)} = \dfrac{R - \mu_R}{\sigma_R}$

where R = the number of runs

μ_R = the mean of the sampling distribution of sample R's

σ_R = the standard deviation of the sampling distribution of sample R's

FORMULA 7 $\mu_R = \dfrac{2n_1\, n_2}{n_1 + n_2 + 1}$

FORMULA 8 $\sigma_R = \sqrt{\dfrac{2n_1\, n_2\left(2n_1\, n_2 - n_1 - n_2\right)}{\left(n_1 + n_2\right)^2\left(n_1 + n_2 - 1\right)}}$

To illustrate with an example, assume that randomly selected samples of male and female patients at a hospital have been asked to rate the level of pain they are feeling. Scores on the pain scale range from 1–35 with higher scores indicating greater pain. Is there a significant difference in pain levels reported between the sexes? The scores are reported in Table 3.

TABLE 3 SCORES ON PAIN SCALE FOR MALE AND FEMALE PATIENTS

Male		**Female**	
Case	**Score**	**Case**	**Score**
1	29	**23**	35
2	29	**24**	33
3	25	**25**	30
4	25	**26**	30
5	25	**27**	27
6	24	**28**	26
7	23	**29**	26
8	23	**30**	22
9	21	**31**	22
10	18	**32**	20
11	18	**33**	20
12	17	**34**	19
13	15	**35**	19
14	13	**36**	16
15	12	**37**	14
16	8	**38**	14
17	8	**39**	10
18	7	**40**	9
19	6	**41**	4
20	5	**42**	4
21	3	**43**	2
22	3	**44**	2

First, the scores must be pooled and then ranked; then we can count the number of runs. Note that tied scores can be a serious problem with this particular test because the number of runs can be affected by how the tied cases are arranged. If you encounter tied scores, probably the best (safest) thing to do is to compute all possible numbers of runs by rearranging the tied cases. If all possible solutions lead to the same decision (reject H_0 or fail to reject H_0), then we are on safe ground in making that decision. If the different arrangements lead to different decisions, then we clearly must opt for another test of significance. In our example, there are no tied cases across the samples (ties within samples won't make any difference in the number of runs).

Designating males with **M** and females with **F**, we can array the scores as follows:

<u>**FFFF**</u>	<u>**MM**</u>	<u>**FFF**</u>	<u>**MMMMMM**</u>	<u>**FF**</u>	<u>**M**</u>	<u>**FFFF**</u>	<u>**MMM**</u>	<u>**F**</u>	<u>**M**</u>	<u>**FF**</u>
1	2	3	4	5	6	7	8	9	10	11

<u>**MM**</u>	<u>**FF**</u>	<u>**MMMMM**</u>	<u>**FF**</u>	<u>**MM**</u>	<u>**FF**</u>
12	13	14	15	16	17

There are 17 runs in these data, and we can now proceed with the formal test of significance.

Step 1. Making Assumptions.

Model: Independent random sampling

Level of measurement is ordinal

Step 2. Stating the Null Hypothesis.

H_0: The two populations are identical on level of pain.

(H_1: The two populations are different on level of pain.)

Step 3. Selecting the Sampling Distribution and Establishing the Critical Region.

Sampling distribution = Z distribution

Alpha = 0.05

Z (critical) = ±1.96

Step 4. Calculating the Test Statistic. Before solving for Z (obtained), both μ_R (the mean of the sampling distribution) and σ_R (the standard deviation of the sampling distribution) must be calculated.

$$\mu_R = \frac{2n_1 n_2}{n_1 + n_2 + 1} \qquad \mu_R = \frac{2(22)(22)}{22 + 22 + 1} \qquad \mu_R = 23$$

$$\sigma_R = \sqrt{\frac{2n_1 n_2(2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} \qquad \sigma_R = \sqrt{\frac{[2(22)(22)][(2(22)(22) - 22 - 22)]}{(22 + 22)^2 \ (22 + 22 - 1)}} \qquad \sigma_R = 3.28$$

$$Z\text{ (obtained)} = \frac{R - \mu_R}{\sigma_R} = \frac{17 - 23}{3.28} = -1.83$$

Step 5. Making a Decision. The test statistic, a Z (obtained) of -1.83, does not fall in the critical region as marked by the Z (critical) of ±1.96. Therefore, we must fail to reject the null hypothesis. Male and female hospital patients do not differ significantly on levels of pain reported.

4 CHOOSING AN ORDINAL TEST OF SIGNIFICANCE

Several other ordinal tests of significance have been developed by statisticians. Most are based on logic similar to that presented in Sections 2 and 3 above, although some are appropriate only in specific circumstances. For example, when we have more than two samples, we could use a nonparametric equivalent of the analysis of variance called the Kruskal–Wallis one-way analysis of variance by ranks test.

Ordinal data independent groups.
Kruskal-Wallis test

k- groups comparison, $k \geq 2$

Null hypothesis : k sampled populations are equivalent in location

The observations from all groups are combined and ranked, with the average rank assigned in the case of ties.

If the populations are identical in location, the ranks should be randomly mixed between the k samples

The test statistic is $K = \frac{12}{n(n+1)}\left(\sum \frac{R_j^2}{n_j}\right) - 3(n+1)$,

where n_j is the number of observations in the j^{th} sample,
n is the total number of observations, and
R_j is the sum of ranks for the j^{th} sample.

If each $n_j \geq 5$ and the null hypothesisis true, then the distribution of K is χ^2 with dof $= c - 1$, wherec is the number of samplegroups.

In the case of ties, a corrected statistic should be computed:

$$K_c = \frac{K}{1 - \left[\frac{\sum (t_j^3 - t_j)}{n^3 - n}\right]}$$

where t_j is the number of ties in the j^{th} sample.

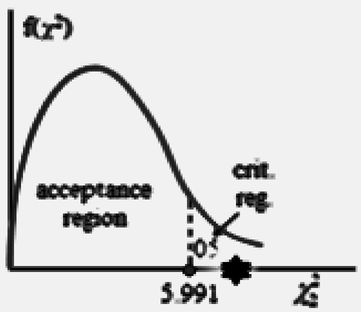
Kruskal-Wallis Test Example: Test at the 5% level whether average employee performance is the same at 3 firms, using the following standardized test scores for 20 employees.

Firm 1		Firm 2		Firm 3	
score	rank	score	rank	score	rank
78		68		82	
95		77		65	
85		84		50	
87		61		93	
75		62		70	
90		72		60	
80				73	
$n_1 = 7$		$n_2 = 6$		$n_3 = 7$	

We rank all the scores. Then we sum the ranks for each firm. Then we calculate the K statistic.

Firm 1		Firm 2		Firm 3	
score	rank	score	rank	score	rank
78	12	68	6	82	14
95	20	77	11	65	5
85	16	84	15	50	1
87	17	61	3	93	19
75	10	62	4	70	7
90	18	72	8	60	2
80	13			73	9
$n_1 = 7$	$R_1 = 106$	$n_2 = 6$	$R_2 = 47$	$n_3 = 7$	$R_3 = 57$

$$K = \frac{12}{n(n+1)}\left(\sum \frac{R_i^2}{n_i}\right) - 3(n+1) = \frac{12}{20(21)}\left(\frac{106^2}{7} + \frac{47^2}{6} + \frac{57^2}{7}\right) - 3(21) = 6.641$$



From the χ^2 table, we see that the 5% critical value for a χ^2 with 2 dof is 5.991.

Since our value for K was 6.641, we reject H_0 that the means are the same and accept H_1 that the means are different.

The two tests presented here are among the more popular general tests of significance for ordinal data, and choosing between them may be something of a problem. In most cases, which test is selected makes little difference, because the decision made in step 5 will be the same. One criterion you can use in making the choice is the number of ties across the samples. As was pointed out above, a large number of such ties makes the runs test rather troublesome and, in such a case, you should opt for the Mann-Whitney *U* test. Also, remember that, when the more restrictive model assumptions can be satisfied, the test for the significance of the difference in sample means is preferred over either of the tests presented here.

NONPARAMETRIC TESTS FOR COMPARING TWO POPULATIONS

In situations where the normality of the population(s) is suspect or the sample sizes are so small

that checking normality is not really feasible, it is sometimes preferable to use nonparametric tests to make inferences about "average" value.
You have data for a random sample of n_1 subjects from Population 1 and a random sample of n_2 subjects from Population 2. You'd like to test the significance of the difference between those two samples. What should you do? Carry out the traditional t test? Perhaps. But you probably should use the Kolmogorov-Smirnov test.

You have randomly assigned a random sample of n subjects to two treatment conditions (experimental and control), with n_1 subjects in the experimental group and n_2 subjects in the control group, where $n_1 + n_2 = n$, and you'd like to test the statistical significance of the effect on the principal outcome variable. What should you use there? The t test? No. Again, the better choice is the Kolmogorov-Smirnov test.

What is the Kolmogorov-Smirnov test? In what follows I will try to explain what it is, why it has not been used very often, and why it is an "all-purpose" significance test as well as an "all-purpose" procedure for constructing confidence intervals for the difference between two independent samples.

What it is

The Kolmogorov-Smirnov test, hereinafter referred to as the K-S test, for two independent samples was developed by Smirnov (1939), based upon previous work by Kolmogorov (1933). [See Hodges, 1957.] It compares the differences between two cumulative relative frequency distributions.

Consider the following example, taken from Goodman (1954):

Sample 1: 1, 2, 2, 2, 2, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5 ($n_1 = 15$)

Sample 2: 0, 0, 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 5, 5, 5 ($n_2 = 15$)

The frequency distributions for Sample 1 are:

Value	Freq.	Rel. Freq.	Cum. Freq.	Cum. Rel. Freq.
0	0	0/15 = 0	0	0/15 = 0
1	1	1/15 = .067	1	1/15 = .067
2	4	4/15 = .267	5	5/15 = .333
3	0	0/15 = 0	5	5/15 = .333
4	4	4/15 = .267	9	9/15 = .600
5	6	6/15 = .400	15	15/15 = 1.000

The corresponding frequency distributions for Sample 2 are:

Value	Freq.	Rel. Freq.	Cum. Freq.	Cum. Rel. Freq.
0	4	4/15 = .267	4	4/15 = .267
1	2	2/15 = .133	6	6/15 = .400
2	4	4/15 = .267	10	10/15 = .667
3	2	2/15 = .133	12	12/15 = .800
4	0	0/15 = 0	12	12/15 = .800
5	3	3/15 = .200	15	15/15 = 1.000

The test statistic for the K-S test is the <u>largest difference</u>, D, between corresponding cumulative relative frequencies for the two samples. For this example the largest difference is for scale value 3, for which D = .800 - .333 = .467. How likely is such a difference to be attributable to chance? Using the appropriate formula and/or table and/or computerized routine (more about those later) the corresponding p-value is .051 (two-tailed). If the pre-specified level of significance, α, is .05 and the alternative hypothesis is non-directional the null hypothesis of no difference between the two population distributions cannot be rejected.

Nice. But what's wrong with using the t test? And why hasn't the K-S test been used more often? Let me take the first question first.

<u>Why not t?</u>

There are at least three things wrong with using the t test for such data:

1. The t test tests only the significance of the difference between the two sample means...nothing else. The K-S test is sensitive to differences throughout the entire scale.

2. The data might have come from a six-point Likert-type ordinal scale for which means are not appropriate; e.g., if the 0 is the leftmost scale value and might have nothing at all to do with "none". (As you undoubtedly know, there has been a never-ending controversy regarding treating ordinal scales as interval scales, to which I have contributed [Knapp, 1990, 1993], and to little or no avail. But see Marcus-Roberts & Roberts [1987] for the best resolution of the controversy that I've ever read.)

3. Even if means are appropriate the t test assumes that the two populations from which the samples have been drawn have normal distributions and equal variances. Those two asumptions are often very difficult to justify, robustness considerations to the contrary notwithstanding.

But isn't there a problem with loss of power by using this test? Not really. Wilcox (1997) has shown that the power of the K-S test can be quite high compared to that of various methods for testing differences between means.

Now for the second question.

Why has the K-S test for independent samples not been used more often?

At the beginning of his article Goodman (1954) says: "Recent results and tables on this topic have been prepared which contribute toward establishing the Kolmogorov-Smirnov statistic as a standard nonparametric tool of statistical analysis." (p. 160). That was in 1954. It's now more than fifty years later, and the K-S test is definitely not a "standard nonparametric tool", as Wilcox (1997) has documented. There are several reasons:

1. It's not even mentioned in some nonparametric statistics textbooks, chapters within general statistics textbooks, and methodological articles. Gibbons (1993), for example, treats the Sign, Wilcoxon (Mann-Whitney), Kruskal-Wallis, and Friedman tests, but there is nary a word about the K-S test.

2. Some people might be under the impression that the K-S test is strictly a goodness-of-fit test. There is indeed a K-S test of goodness-of-fit, but researchers seem to have been able to distinguish the chi-square test for independent samples from the chi-square test of goodness-of-fit, so if they can handle two chi-square tests they should be able to handle two K-S tests. (And the K-S test for two samples has even been extended to the case of three or more samples, just like the two-sample chi-square test has. See Conover [1980] and Schroer & Trenkler [1995] for details.)

3. There might be further concerns that it's too complicated and the necessary computer software is not readily available. Both concerns would be unfounded. It's simple to carry out, even by hand, as the above example in Goodman (1954) and comparable examples in Siegel and Castellan (1988) attest. Tables for testing the significance of D have been around for a long time (as have formulas for the case of two large samples) and there are at least two excellent internet sites where all the user need do is enter the data for the two samples and the software does the rest (see below).

K-S test confidence intervals

Many researchers prefer confidence intervals to significance tests, and they argue that you get significance tests "for free" when you establish confidence intervals around the test statistics. The Kolmogorov-Smirnov test for two independent samples includes a procedure for constructing confidence intervals for D.

K-S test: There are two stand-alone routines that can carry out the two-sample K-S test. One of requires the entry of the raw data for each subject in each sample, with n_1 and n_2 each between 10 and 1024. The other one requires the entry of the frequency (actual, not relative) of each of the observations for each of the samples. Here are the data:

Sample 1:

Value	Freq.	Rel. Freq	Cum. Freq.	Cum. Rel. Freq.
1	0	0/70 = 0	0	0/70 = 0
2	52	52/70 = .743	52	52/70 = .743
3	11	11/70 = .157	63	63/70 = .900
4	0	0/70 = 0	63	63/70 = .900
5	7	7/70 = .100	70	70/70 = 1.000

Sample 2:

Value	Freq.	Rel. Freq	Cum. Freq.	Cum. Rel. Freq.
1	37	37/70 = .529	37	37/70 = .529
2	10	10/70 = .143	47	47/70 = .671
3	0	0/70 = 0	47	47/70 = .671
4	0	0/70 = 0	47	47/70 = .671
5	23	23/70 = .329	70	70/70 = 1.000

This set of data is a natural for the K-S test but they do not discuss it among their suggested nonparametric alternatives to the t test. Although their article is concerned primarily with dependent samples, they introduce alternatives to t via this example involving independent samples. The two sample means are identical (2.457) but they argue, appropriately, that there are some very large differences at other scale points and the t test should not be used. (The p-value for t is 1.000.) They apply the Wilcoxon test and get a p-value of .003. Using the K-S test for the same data, I get D = 0.5285714 (for the scale value of 0) and the p-value is 0.000000006419. [So many decimal places, D is huge and that p is tiny.]

Possible Questions

PART-B

1. How Z-test is used for testing significance of proportions?
   In a referendum submitted to the student body and 850 men and 550 women voted. Out of these, 530 of men and 310 of women voted 'yes'. Does this indicate a significant difference in opinion on matter between men and women students? (Use α = 5% and $Z_{(0.05)}$= 1.96)
2. Test Median class size for Math is larger than the median class size for English for the following data using Mann – Whitney U test.

Class size (Math, M)	23	45	34	78	34	66	62	95	81
Class size (English, E)	30	47	18	34	44	61	54	28	40

3. According to the IQ level and the economic conditions of their homes 1000 students at a college were graded. Use χ^2 test to find out whether there is any association between economic condition at home and IQ.

Economic Conditions	IQ		Total
	High	Low	
Rich	**460**	**140**	600
Poor	**240**	**160**	400
Total	700	300	1000

(Note: The level of significance is 0.05 and table value is 3.84).

4. Test Median class size for Math is larger than the median class size for English for the following data using Mann – Whitney U test.

Class size (Math, M)	23	45	34	78	34	66	62	95	81
Class size (English, E)	30	47	18	34	44	61	54	28	40

5. Explain Kolomogrov-Smirnov test for 2 populations.
6. In order to increase the efficiency, one group of operators class room training, and the other group was provided on the job training. After the training, the times to complete a certain job, in minutes, was recorded for both the groups, the data recorded is given in the below table. Use Mann – Whitney U test to test whether both the methods of imparting training are equality effective.

Class room training	**Operator No.**	1	2	3	4	5	6	7	8	9
	Time	35	39	51	63	48	31	29	41	55
Class room training	**Operator No.**	1	2	3	4	5	6	7	8	
	Time	85	28	42	37	61	54	36	57	

PART-C

7. Mr. Gowtham, Personal Manager is concerned about absenteeism. He decides to sample the records to determine if absenteeism is distributed evenly throughout the six-day work-

week. The null hypothesis to be tested is: absenteeism is distributed evenly throughout the week. The sample results are as follows:

Day	**Number of Absentees**
Monday	12
Tuesday	9
Wednesday	11
Thursday	10
Friday	9
Saturday	9

i. Using χ^2 test of significance, compute χ^2 value.

ii. Is the null hypothesis rejected?

iii. Specifically, what does this indicate to the Personal Manager?

(Note: The level of significance is 0.01 and table value is 15.086).

DEPARTMENT OF MATHEMATICS

MATHEMATICAL STATISTICS (18MMP304)

UNIT- III

Sl. No.	Question	Option 1	Option 2	Option 3	Option 4	Answer
1	Level of significance is the probability of…………………	Type I error	Type II error	No error	Standard error	Type I error
2	If the calculated value is less than the table value, then we accept the --------------- hypothesis.	Alternative	Null	Sample	Statistics	Null
3	Small sample test is also known as ………………	Exact test	t – test	normal test	F - test	t – test
4	An example in a two-sided alternative hypothesis is:	$H1: \mu < 0$	$H1: \mu > 0$	$H1: \mu \geq 0$	$H1: \mu \neq 0$	$H1: \mu \neq 0$
5	Larger group from which the sample is drawn is called …………..	Statistic	Sampling	Universe	Parameter	Universe
6	Any hypothesis concerning a population is called a…………………	Sample	Population	Statistical measure	Statistical hypothesis	Statistical hypothesis
7	Rejecting null hypothesis when it is true leads to ……………..	Type I error	Type II error	Correct decision	Type III error	Type I error
8	Accepting null hypothesis when it is true leads to	Type I error	Type II error	Correct decision	Type III error	Correct decision
9	Type II error occurs only if………..	Reject Ho when it is true	Accept Ho when it is false	Accept Ho when it is true	Reject Ho when it is false	Accept Ho when it is false
10	The correct decision is ……………	Reject Ho when it is true	Accept Ho when it is false	Reject Ho when it is false	Level of significance	Reject Ho when it is false
11	The maximum probability of committing type I error, which we specified in a test is known as	Null hypothesis	Alternative hypothesis	Degrees of freedom	Level of significance	Level of significance

12	If the computed value is less than the critical value, then …………….	Null hypothesis is accepted	Null hypothesis is rejected	Alternative hypothesis is rejected	Level of significance	Null hypothesis is accepted
13	If the computed value is greater than the critical value, then……………..	Null hypothesis is accepted	Null hypothesis is rejected	Alternative hypothesis is rejected	Level of significance	Null hypothesis is rejected
14	A critical function provides the basis for …………..	Accepting H_0	Rejecting H_0	No decision about H_0	No decision about H1	No decision about H1
15	Degree of freedom for statistic chi-square incase of contingency table of order 2 x 2 is….	4	3	2	1	1
16	If the sample size is greater than 30, then the sample is called	Large sample	Small sample	Population	Normal	Large sample
17	If the sample size is less than 30, then the sample is called………..	Large sample	Small sample	Population	Normal	Small sample
18	Z – test is applicable only when the sample size is ………….	Zero	2	Small	Large	Large
19	The degrees of freedom for two samples in t – test is ………..	$n_1 + n_2 + 1$	$n_1 + n_2 – 2$	$n_1 + n_2 + 2$	$n_1 + n_2 + 1$	$n_1 + n_2 – 2$
20	The test-statistic t has d.f = _________:	n	n-1	n-2	n-3	n-1
21	An assumption of t – test is population of the sample is …….	Binomial	Poisson	Normal	Exponential	Normal
22	The degrees of freedom of chi – square test in contigency tables is …..	(r – 1)(c – 1)	(r + 1)(c + 1)	(r + 1)(c – 1)	(r – 1)(c + 1)	(r – 1)(c – 1)
23	In chi – square test, if the values of expected frequency are less than 5, then they are combined together with the neighbouring frequencies. This is known as………	Goodness of fit	Degree of Freedom	Level of significance	Pooling	Pooling
24	In F – test, the variance of population from which samples are drawn are…..	Equal	Unequal	Small	Large	Equal

25	If the data is given in the form of a series of variables, then the DOF is…….	n	n – 1	n +1	(r – 1)(c – 1)	n – 1
26	The value of Z test at 5% level of significance is …………..	3.96	2.96	1.96	0.96	1.96
27	From the following which one of the following is taken as null hypothesis?	$P_1 = P_2$	$P_1 > P_2$	$P_1 < P_2$	$P_1 \neq P_2$	$P_1 = P_2$
28	Most of the non-parametric methods utilise measurements on………..	Interval scale	Ratio scale	Ordinal scale	Nominal scale	Ordinal scale
29	For a non-parametric test,the distibution ……………	Should be normal	Should be binomial	Need not be normal	Should be Poisson	Need not be normal
30	Which of the following is a non-parametric test?	Chi square test	F-test	t-test	Z-test	Chi square test
31	To test goodness of fit for a non-normal distribution, we use………..	Kolomogrov-Smirnov test	Chi square test	F-test	t-test	Kolomogrov-Smirnov test
32	In tests using rank methods, the null hypothesis is rejected if calculated value is	> tabulated value	< tabulated value	>= tabulated value	<=tabulated value	<tabulated value
33	The chi sqaure test statistic is defined as………….	Chi square = Sum [{(O - E) * (O - E)} / E]	Chi square = Sum [(O - E)	Chi square = Sum [{(O + E) * (O + E)} / E]	Chi square = Sum [{(O + E) * (O + E)}	Chi square = Sum [{(O - E) * (O - E)} / E]
34	The chi square variate is always……….	≥ 0	<0	>1	1	≥ 0
35	Kolomogrov Smirnov test is for testing………………..	Equality of several means	Comparing two populations	Equality of 2 means	Equality of several variances	Comparing two populations
36	The null hypothesis for Kolomogrov Smirnov test is that two populations are from…….	Same population	Normal population	Different populations	Non-normal populations	Same population
37	Mann-Whitney U-test is used for testing……………	Equality of two means	Equality of three means	Equality of several means	Equality of 2 sets of several rankings	Equality of two means

38	The non-parametric test analogous to ANOVA is…………	Kruskal-Wallis test	Mann-Whitney U-test	Kolomogrov Smirnov test	Chi square test	Kruskal-Wallis test
39	Mann-Whitney U-test is analogous to ……………..	t-test	Chi-Square test	F-test	Z-test	t-test
40	Wilcoxon-Wilcox test is used for testing ……………	Equality of several means	Equality of 2 variances	Equality of two means	Equality of several variances	Equality of several means
41	Wilcoxon-Wilcox test will compare equality of…………………..	All pairs of means	All means	2 means	More than 3 means	All pairs of means
42	The independence of attributes can be tested by using…………	Contigency tables	Normal table	Pooling	Z-test	Contigency tables
43	The degrees of freedom in a 3x3 contigency table is…………….	8	4	3	0	4
44	The degrees of freedom in a r x s contigency table is……………..	r-1	s-1	r+1	(r-1)(s-1)	(r-1)(s-1)
45	Most of the Non-Parametric methods utilize measurements on :	Interval Scale	Ratio Scale	Ordinal Scale	Nominal Scale	Ordinal Scale
46	Kolmogorow - Smirnov have evolved tests for	Goodness of fit of a distribution	Comparing two Populations	Both (a) and (b)	Neither (a) nor (b)	Both (a) and (b)
47	The most commonly used assumption about the distribution of a variable is:	Continuity of the distribution	Symmetry of the distribution	Both (a) and (b)	Neither (a) nor (b)	Symmetry of the distribution
48	Which one of the following statement is false:	α is called type I error	1 - α is called power of the test	β is called type II error	1 - β is called power of the test	1 - β is called power of the test
49	Which one of the following is **not** an alternative hypothesis?	$H_1 : m \neq m_0$	$H_1 : m > m_0$	$H_1 : m < m_0$	$H_1 : m = m_0$	$H_1 : m \neq m_0$

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

Subject : Mathematical Statistics	**Semester III**	**L T P C**
Subject Code : 18MMP304	**Class : II M.Sc Mathematics**	**4 0 0 4**

UNIT IV

Estimation: Preliminary notion -Consistency estimation -Unbiased estimates -Sufficiency - Efficiency -Asymptotically most efficient estimates -methods of finding estimates -confidence Interval.

SUGGESTED READINGS

1. Marek Fisz, (1980). Probability Theory and Mathematical Statistics, John Wiley and Sons, New York.
2. Meyer, (2006). Introduction to Probability and Statistical applications, Oxford and IBH Publishing Co. Pvt. Ltd. New Delhi.
3. Sheldon M. Ross, (2009). Introduction to probability and statistics for engineers and scientists, Third edition, Academic press.
4. Heinz Bauer,(1996), Probability Theory, Narosa Publishing House, London.
5. Parimal Mukhopadhyay, (2012). Theory of Probability, New central book agency, Calcutta.

Basic (inferential) Statistics: Estimation, Confidence Intervals and Testing

Based on a (relatively small) random sample, taken from a population, we will try to draw conclusions about this population.

We built a model with statistical assumptions for the actual **observations $x_1, ..., x_n$**: the experiment can be repeated under the same conditions and repetition will lead to new, different observations. The model stays the same, but the realizations are different.

A **random sample $X_1, ..., X_n$** is a sequence of independent random variables that all have the same distribution with expectation μ and variance σ^2.

Estimation

Example: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an estimator of the (unknown) population parameter μ: it is a random variable.

If the experiment is really executed, the observed value $\bar{x}$ is its estimate.

An **estimator** is a statistic (= function of the sample variables), used to get an idea about the real value of a population parameter.

An **estimate** is the observed value of an estimator after actually executing the sample.

Frequently used estimators:

Pop. par.	Estimator	Estimate
μ	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
$\boldsymbol{\sigma^2}$	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$
p	$\hat{p} = \frac{X}{n}$	$\hat{p} = \frac{x}{n}$

An estimator T of a parameter θ is **unbiased** if E(T) = θ

The **bias** of T is E(T) - θ.

Estimators can be compared by using the **Mean Squared Error** as a criterion:

$$\mathbf{E(T-\theta)^2}$$

Property:

$$\mathbf{E(T-\theta)^2 = (ET-\theta)^2 + var(T)}$$

MSE = bias^2 + variance of T

If T is an unbiased estimator of **θ**,

so E(T) = **θ,** then: $E(T-\theta)^2 = var(T)$

Comparing of two estimators T_1 and T_2 of the parameter **θ** (assuming that T_1 and T_2 are based on samples): the estimator having the **smallest mean squared error** is the **best!**

Pop.par.	Estimator	Biased?	Standard error *) of the estimator
μ	$\overline{X}$	$E(\overline{X}) = \mu$	$\frac{\sigma}{\sqrt{n}}$
$\boldsymbol{\sigma^2}$	S^2	$E(S^2) = \sigma^2$	--
p	$\hat{p} = \frac{X}{n}$	$E(\hat{p}) = p$	$\frac{p(1-p)}{\sqrt{n}}$

***) =** the **standard deviation** of an estimator

Confidence Intervals

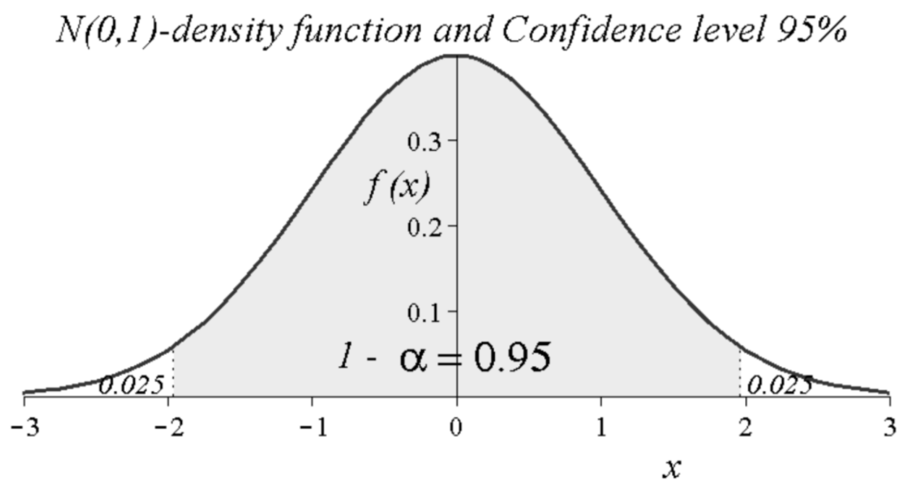
Estimators are also called point estimators. Often we want to quantify the variation by giving an interval as estimation: interval estimation or confidence intervals.

Case 1: a confidence interval for μ, given σ Statistical assumptions (probability model): $X_1, \ldots, X_n$ are independent and all $X_i \sim N(\mu, \sigma^2)$

From probability theory we know: $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ or : $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

For this $Z \sim N(0,1)$ and **confidence level 95%** we can find c, such that $P(-c \le Z \le c) = 0.95$.

Using the N(0, 1)-table we find: **c = 1.96**



$$P\left(-c \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le c\right) = 0.95 \quad \Leftrightarrow$$

$$P\left(\overline{X} - c \times \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + c \times \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- $\left(\overline{X} - c \times \frac{\sigma}{\sqrt{n}}, \overline{X} + c \times \frac{\sigma}{\sqrt{n}}\right)$ is called the **stochastic 95%-confidence interval for μ**. Note that $\overline{X}$ is not a number: $\overline{X}$ is a random variable for which a probability holds
- If we have the result of the sample (the real values $x_1, \ldots, x_n$) we compute a numerical interval, given the values of c=1.96, n and σ:

  $\mathbf{95\% - CI}(\mu) = \left(\overline{\mathbf{X}} - \mathbf{c}\frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}, \overline{\mathbf{X}} + \mathbf{c}\frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}\right)$
- **Interpretation**: "When repeating the n observations often and computing the interval as often, about 95% of these numerical confidence intervals will contain the unknown value of μ".
- In general: $\mathbf{CI}(\mu) = \left(\overline{\mathbf{X}} - \mathbf{c}\frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}, \overline{\mathbf{X}} + \mathbf{c}\frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}\right)$

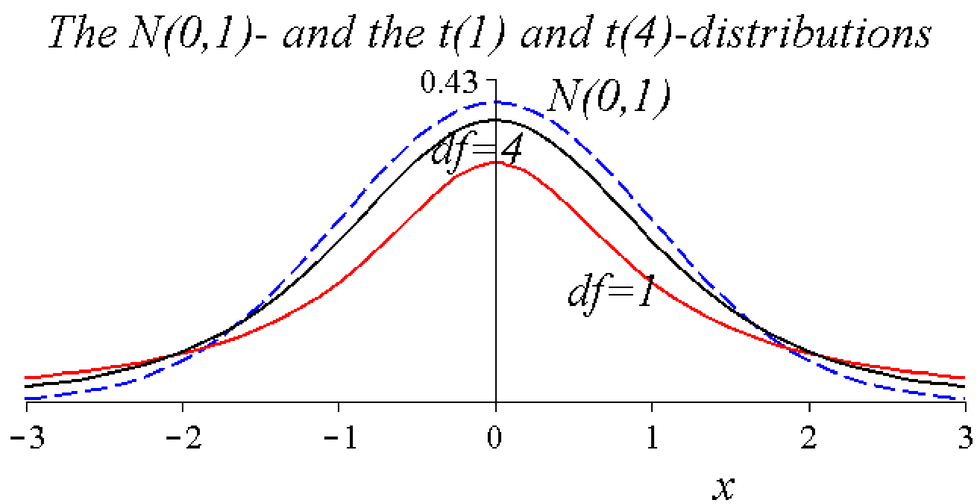
Case 2: N(μ, σ²)-model, unknown σ² and μ

Statistical assumptions: $X_1, \ldots, X_n$ are independent and all $X_i \sim N(\mu, \sigma^2)$, unknown σ^2

- We cannot simply substitute **s** as an estimate for σ in the formula $\left(\bar{X} - c\frac{\sigma}{\sqrt{n}}, \bar{X} + c\frac{\sigma}{\sqrt{n}}\right)$
- Instead we use the **Student's t-distribution:**

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}, \text{where } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- T has a Student's t-distribution with **n – 1 degrees of freedom**. Notation: **T ~ t(n-1)**
- The graph of T is similar as the N(0,1)-graph



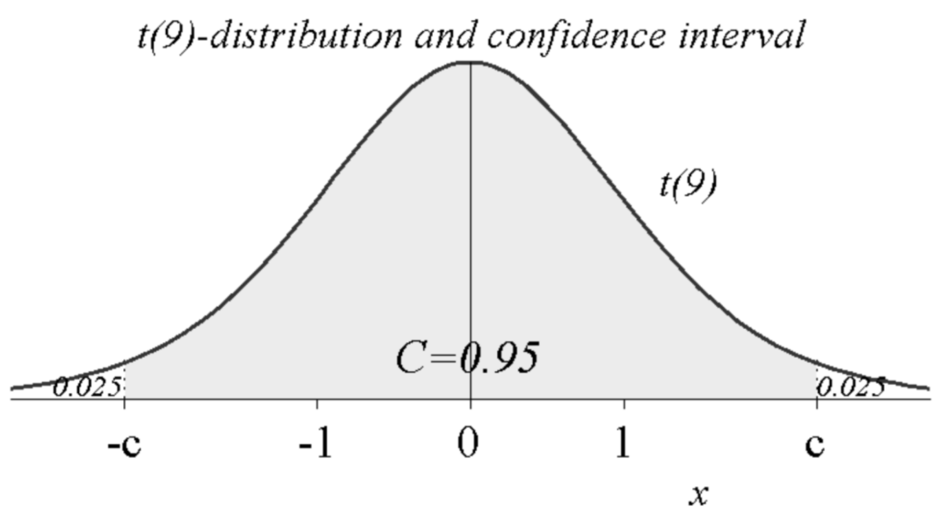
Now we use the table of critical values of the t-distribution to find c such that

$$P\left(-c \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq c\right) = 0.95$$

Solving μ we find the numerical interval:

$$\mathbf{95\% - CI}(\mu) = \left(\bar{\mathbf{x}} - \mathbf{c}\frac{\mathbf{s}}{\sqrt{\mathbf{n}}}, \bar{\mathbf{x}} + \mathbf{c}\frac{\mathbf{s}}{\sqrt{\mathbf{n}}}\right)$$

Note that c depends on n -1 and that for 0.95 you should find c using **upper tail area**: P(T ≥ c) =0.025 of the t(n-1)-table.



For confidence level **1-α**: $P(T \geq c) = \frac{\alpha}{2}$:

$\mathbf{(1-\alpha)100\% - CI}(\mu) = \left(\bar{\mathbf{x}} - \mathbf{c}\frac{\mathbf{s}}{\sqrt{\mathbf{n}}}, \bar{\mathbf{x}} + \mathbf{c}\frac{\mathbf{s}}{\sqrt{\mathbf{n}}}\right)$

a **Conf. Int. for σ^2** in a $N(\mu, \sigma^2)$-model:

Statistical assumptions: $X_1,\ldots,X_n$ constitute a random sample from the $N(\mu, \sigma^2)$, where σ^2 and μ are unknown.

We will use $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$:

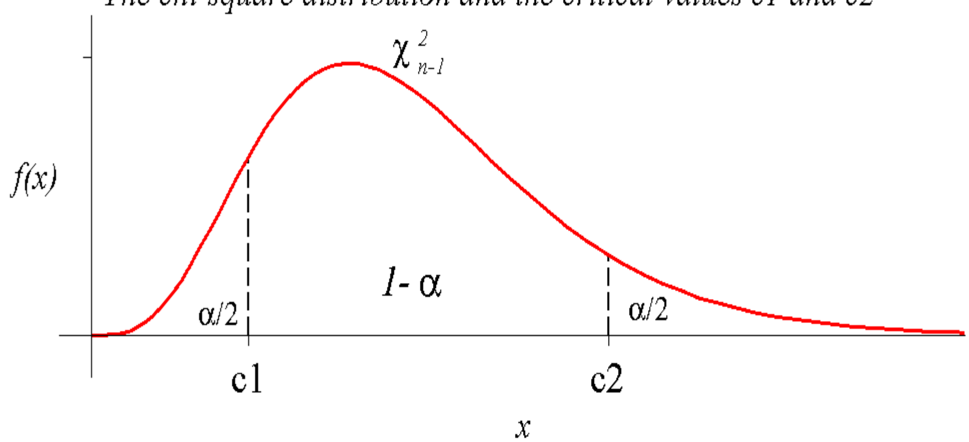
an unbiased estimator of σ^2: $E(S^2) = \sigma^2$

The "standardized" form $\frac{(n-1)S^2}{\sigma^2}$

has a **chi − squared distribution with n − 1 degrees degrees of freedom**.

Notation: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

The chi-square distribution and the critical values c1 and c2



Solving σ^2 from $P\left(c_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq c_2\right) = 1 - \alpha$,

we find $\frac{(n-1)S^2}{c_2} \leq \sigma^2 \leq \frac{(n-1)S^2}{c_1}$ or numerically:

$$\mathbf{(1-\alpha)100\% - CI}(\boldsymbol{\sigma}^2) = \left(\frac{\mathbf{(n-1)S^2}}{\mathbf{c_2}}, \frac{\mathbf{(n-1)S^2}}{\mathbf{c_1}}\right)$$

And:

$$\mathbf{(1-\alpha)100\% - CI}(\boldsymbol{\sigma}) = \left(\sqrt{\frac{\mathbf{(n-1)S^2}}{\mathbf{c_2}}}, \sqrt{\frac{\mathbf{(n-1)S^2}}{\mathbf{c_1}}}\right)$$

A confidence interval for the success proportion p of a population:

Statistical assumptions:

X = “the number of successes in a random sample of length n”: **X ~ B(n, p).**

For large n (np(1-p) > 5) we approximate:

X ~ N(np, np(1-p)) and for $\hat{p} = \frac{X}{n}$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Using the N(0, 1)-table we find c such that:

$$P(-c \le Z \le c) = P\left(-c \le \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \le c\right) = 1 - \alpha$$

Estimating p(1-p) in the denominator by $\hat{p}(1-\hat{p})$, we find:

$$\mathbf{CI(p)} = \left(\hat{\mathbf{p}} - \mathbf{c}\sqrt{\frac{\hat{\mathbf{p}}(\mathbf{1}-\hat{\mathbf{p}})}{\mathbf{n}}}, \hat{\mathbf{p}} + \mathbf{c}\sqrt{\frac{\hat{\mathbf{p}}(\mathbf{1}-\hat{\mathbf{p}})}{\mathbf{n}}}\right)$$

Testing hypotheses

Statistical tests are used when we want to verify a statement (hypothesis) about a population on the basis of a random sample.

- The hypotheses should be expressed in the population parameters like μ, σ^2 and p.
- The **null hypothesis (H_0)** contains the "old/ common situation" or the "prevailing view"
- The **alternative hypothesis (H_1)** contains the denial of H_0: the statement that we want to proof (statistically), by using the sample.

Examples: 1. Test H_0: p = ½ versus H_1: p > ½

2. Test H_0: μ = 28 versus H_1: μ ≠ 28

H_0: p = ½ is a **simple** hypothesis (one value of p) and H_0: p ≤ ½ is a **compound** hypothesis.

- Our aim is usually to reject H_0 in favour of the statement in H_1: the conclusion should be "reject H_0" or "not reject H_0" (accept H_0)
- Our "proof" on the basis of observations is never 100% certain. We will choose a small probability of falsly rejecting H_0: this is called **the significance level α**.
- The test statistic is usually an estimator of the population parameter, e.g. $\hat{\mathbf{p}}$ if H_0: p = ½, or a linked variable, e.g. X or $(\hat{p} - \frac{1}{2}) / \sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{n}}$

8-steps-procedure for testing hypotheses:

(Formulate/state/compute successively:)

1. The research question (in words)
2. The statistical assumptions (model)
3. The hypotheses and level of significance
4. The statistic and its distribution
5. The observed value (of the statistic)

6. The rejection region (for H_0) or p-value
7. The statistical conclusion
8. The conclusion in words (answer question)

Applying the testing procedure in an example:
1. Does the majority of Gambians find Coca Cola (CC) preferable to Pepsi Cola (PC)?
2. **p** = "the proportion of Gambian cola drinkers who prefer CC" and
X = "the number of the 400 test participants who prefer CC": $X \sim B(400, p)$
3. Test H_0: $p = ½$ and H_1: $p > ½$ if $\alpha = 0.01$
4. Statistic $X \sim B(400, ½)$ if H_0: $p = ½$ is true
So X is approximately N(200, 100)-distrib.
5. The observed value: X = 225
6. **TEST**: $\mathbf{X \geq c} \;\boxed{\Rightarrow}\;$ **reject H_0**
$P(X \geq c \mid p = ½) \leq \alpha = 0.01$ (use cont.corr.)
$P\left(\frac{X-200}{\sqrt{100}} \geq \frac{c-\frac{1}{2}-200}{10} |p = \frac{1}{2}\right) \leq 0.01$, or:
$P(Z \leq \frac{c-\frac{1}{2}-200}{10}) \geq 0.99 \Rightarrow \frac{c-\frac{1}{2}-200}{10} \geq 2.33$
$c \geq 23.3 + 200.5 = 223.8 \Rightarrow$ **c = 224**
7. $X = 225 > 224 = c \Rightarrow$ reject H_0
8. At significance level 1% we have proven that Gambian cola drinkers prefer CC

Deciding by computing the **p-value** (or observed significance level): comparing the observed X = 225 and expected value E(X) = 200 if H_0 is true, the p-value is the probability that X deviates this much or more:

6. **TEST: p-value $\leq \alpha$** $\boxed{\Rightarrow}$ **reject H_0**
p-value = $P(X \geq 225 \mid H_0: p = ½)$
$= P(Z \geq$ **2.45) = 0.69**% $> \alpha \rightarrow$ **reject** H_0

The choice of the test statistic: we chose X, but we also could have chosen $\hat{p}$ or the standardized $\hat{p}$. Then the rejection region should be adjusted:

Statistic	observed	Rejection region
The number X	X = 225	{224, 225,..., 400}
proportion $\hat{p} = \frac{x}{n}$	$\hat{p} = 0.5\mathbf{6}25$	$\left[\frac{224}{400}, 1\right]$
$Z = \frac{\hat{p} - 0.5}{\sqrt{\frac{0.5 \times 0.5}{400}}}$	Z = 2.5	$[2.33, \infty)$

Errors in testing are shown in this table:

		The reality is	
		H_0 is true	H_1 is false
Test result	accept H_0	Correct decision	Type II error
	reject H_0	Type I error	Correct decision

P(Type I error) = $P(X \geq 224 \mid p = ½)$
$\approx P(Z > 2.35) = 0.94\% < \alpha$
P(Type II error) depends on the value of p, chosen from H_1: $p > ½$.
E.g. if $p = 0.6$ we find:
P(Type II error) = $P(X < 224 \mid p = 0.6)$
$\approx P(Z \leq -1.68) = 4.65\%$
The **power** of the test = 1- P(Type II error).
So if p = 0.6, the power of the test is 95.35%.
Test for μ in a N(μ, σ²)-model, unknown σ²

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \text{ is the test statistic for } H_0: \mu = \mu_0$$

(Not $\overline{X}$: var($\overline{X}$) contains the unkown σ^2)

$T \sim t(n - 1)$ if $H_0: \mu = \mu_0$ is true

Example

1. Statement: the mean IQ of Gambians is higher than that of Senegalese (mean 101)
2. $X_1,\dots, X_n$ are independent IQ's, $X_i \sim N(\mu, \sigma^2)$
3. Test H_0: $\mu = 101$, H_1: $\mu > 101$ for $\alpha = 0.05$
4. Statistic $T = \frac{\overline{X}-101}{S/\sqrt{20}} \sim t(20\text{-}1)$ if H_0: $\mu = 101$
5. Observations for n= 20: $\bar{x} = 104$ and $s^2 = 81$ So observed value $t = \frac{104-101}{9/\sqrt{20}} \approx 1.491$
6. **TEST: $t \geq c$ => reject H_0.**
   $P(T_{19} \geq c) = 0.05$ => **c** = 1.729
7. $t \approx 1.491 < 1.729$=> accept H_0.
8. There is not enough evidence to maintain the statement that Gambians are smarter than Senegalese at 5%-level.

Using the p-value:

6`. TEST: p-value ≤ α $\boxed{\Rightarrow}$ **reject H_0**
p-value = $P(T \geq 1.491 \mid H_0)$ is between
5% and 10%
7`. p-value > 5% => do not reject H_0.

Note 1: If we know σ^2 we can execute the test procedure using $Z = \frac{\bar{X}-101}{\sigma/\sqrt{20}}$ and Z ~ N(0, 1).

Note 2: If n is large (> 100) we can use the N(0, 1)-distribution as an approximation for T.

Note 3: If the population does not have a normal distribution and n is large**,** we can use $Z = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$ as a test statistic, which is approximately N(0, 1)-distributed as $\mu = \mu_0$

A test for σ^2 (or σ) , normal model

To test whether σ^2 has a specific value we use S^2 : if H_0: $\sigma^2 = \sigma_0^2$ is true then $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$

An example:

The variation of the quantity active substance in a medicine is an important aspect of quality. Suppose the standard deviation of the quantity active substance in a tablet should not exceed 4 mg. For testing the quality, a random sample of 10 quantities of tablets is available: the sample variance turned out to be 25 mg^2.

The chi-square test for σ^2

1. Is the quantity active substance in the medicine greater higher than permittted?
2. The quantities $X_1,..., X_{10}$ are independent and $X_i \sim N(\mu, \sigma^2)$ for i = 1, ..., 10
3. Test H_0: $\sigma^2 = 4^2$ and H_1: $\sigma^2 > 16$ for α =5%
4. Test Statistic S^2: $\frac{9S^2}{16} \sim \chi^2_9$ if H_0 is true
5. Observed value $s^2 = 25$
6. **TEST: $s^2 \geq c$ => reject H_0.**

   $P(S^2 \geq \mathbf{c} \mid \sigma^2 = 16) = P\left(\frac{9S^2}{16} > \frac{9c}{16}\right) = 0.05$

   => $\frac{9c}{16} = 16.9$ => c = 30.0
7. $s^2 = 25 < 30.0 = c$ => do not reject H_0
8. The sample did not proof that the variation of quantity active substance is greater than allowed, at 5% level of significance.

Two sided tests

1. Normal model, test for μ, unknown σ^2

   Hypotheses H_0: $\mu = \mu_0$ and H_1: $\mu \neq \mu_0$ $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ has a two sided (symmetric) rejection region:$T \leq -c$ or $T \geq c$ =>reject H_0

   p-value = $2P(T \geq |t|)$.
2. Normal model, test for σ^2, unknown μ.

   Hypotheses H_0: $\sigma^2 = \sigma_0^2$ and H_1: $\sigma^2 \neq \sigma_0^2$ S^2 has a two sided (asymmetric) rejection region: $S^2 \leq c_1$ or $S^2 \leq c_2$ => reject H_0

   p-value = $2P(S^2 \geq s^2)$ or $2P(S^2 \leq s^2)$
3. Binomial model, test for p:

   Hypotheses H_0: $p = p_0$ and H_1: $p \neq p_0$. $Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$ has a two sided (symmetric)

rejection region:
$Z \leq -c$ or $Z \geq c$ => reject H_0
p-value = $2P(Z \geq |z|)$.

Confidence Intervals for the Mean (Large Samples)

- A **point estimate** is a single value estimate for a population parameter. The most unbiased estimate of the population mean μ is the sample mean $\bar{x}$.
- An **interval estimate** is an interval, or range of values, used to estimate a population parameter.
- The **level of confidence c** is the probability that the interval estimate contains the population parameter.

Since we can hardly expect that point estimates based on samples always hit the parameters they are supposed to estimate exactly, it is often desirable to give an interval rather than a single number. We can then assert with a certain probability (or degree of confidence) that such an interval contains the parameter it is intended to estimate. (Freund p 214)

For large samples, the Central Limit Theorem applies. From the CLT, when $n \geq 30$, the sampling distribution of the sample mean $\bar{x}$ is normal. The level of confidence c is the area under the standard normal curve between **the critical values**, $-z_c$ and z_c. For example, if c = 95%, then 2.5% is less than $-z_c$ and 2.5% is greater than z_c. Looking up the z-score in table A16, we see that $-z_{.95} = -1.96$ and . $z_{.95} = 1.96$.

The distance between the point estimate and the actual parameter value is called the **error of estimate**. When estimating μ the error of estimate is the distance $|\bar{x} - \mu|$.

Given a level of confidence c, the **maximum error of estimate** (sometimes called the margin of error or error tolerance) **E** is the greatest possible distance between the point estimate and the value of the parameter it is estimating.

$$E = z_c \sigma_{\bar{x}} = z_c \frac{\sigma}{\sqrt{n}}$$

When $n \geq 30$, the sample standard deviation s can be used in place of σ.

In example 1 and example 2 on pages 270 – 272, there are 54 samples and the sample mean and sample standard deviation are:

$$\bar{x} = \frac{\Sigma x}{n} = 12.4$$

$$s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}} \approx 5.0$$

Substituting

$$E = z_c \frac{\sigma}{\sqrt{n}} \approx z_c \frac{s}{\sqrt{n}} = 1.96 * \frac{5.0}{\sqrt{54}} = 1.3$$

So we are 95% sure that the maximum error of estimate for the population mean is about 1.3.
The **c-confidence interval** for the population mean is $\bar{x} - E < \mu < \bar{x} + E$

In the above example the 95% left endpoint (often called the lower confidence limit or **LCL**) of the confidence interval is 12.4 – 1.3 = 11.1 and the right endpoint (often called the upper confidence limit or **UCL**) is 12.4 + 1.3 = 13.7. So the 95% confidence interval is
$11.1 < \mu < 13.7$.

The confidence interval is often denoted in the following ways
$\bar{x} \pm E$

$$(\bar{x} - E, \bar{x} + E) = (\bar{x} - z_c \frac{\sigma}{\sqrt{n}}, \bar{x} + z_c \frac{\sigma}{\sqrt{n}})$$

Summary for finding confidence interval for population mean (p 273)

What to do	**Equations**	**Example (from above)**
Find the sample statistics n and $\bar{x}$	$\bar{x} = \frac{\Sigma x}{n}$	$n = 54, \bar{x} = 12.4$
Specify σ if known. Otherwise, if $n \geq 30$, find the sample standard deviation, s	$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$	$s = 5.0$
Find the critical value z_c that corresponds to the given level of confidence	Use the Standard Normal Table to find the value z_c such that the area to the right of $z_c = (1-c)/2$.	$z_{.95} = 1.96$
Find the maximum error of the estimate E. Note this is the critical value times the standard error of the mean.	$E = z_c \sigma_{\bar{x}} = z_c \frac{\sigma}{\sqrt{n}} \approx z_c \frac{s}{\sqrt{n}}$	$E = 1.3$
Find the left and right endpoints and form the confidence interval	Left endpoint (LCL): $\bar{x} - E$ Right endpoint (UCL): $\bar{x} + E$ Interval: $\bar{x} - E < \mu < \bar{x} + E$	(11.1,13.7)

In summary, $CI = (\bar{x} - z_c \frac{\sigma}{\sqrt{n}}, \bar{x} + z_c \frac{\sigma}{\sqrt{n}})$

Example 5, **p 275** Take a sample of size 20 from a Normal distribution with standard deviation = 1.5. The sample mean is 22.9. What is the 90% CI?
Looking in Table 4 p A16 (Standard Normal Distribution) $z_{.05} = 1.645$

$$CI = (\bar{x} - z_c \frac{\sigma}{\sqrt{n}}, \bar{x} + z_c \frac{\sigma}{\sqrt{n}}) =$$

$$(22.9 - 1.645\frac{1.5}{\sqrt{20}},\ 22.9 + 1.645\frac{1.5}{\sqrt{20}}) = (22.9 - 0.55,\ 22.9 + 0.55) = (22.35,\ 23.45)$$

One-Sample Z
The assumed standard deviation = 5

N	Mean	SE Mean	95% CI
100	50.0000	0.5000	(49.0200, 50.9800)

Determining Sample Size (p 276)
How large a sample size (n) is needed to guarantee a certain level of confidence for a given maximum error of estimate (E)? This can be derived from the formula for E above

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

Solving for n gives:

$$n = \left(\frac{z_c \sigma}{E}\right)^2$$

If σ is unknown, s can be used as an estimate if there is a preliminary sample size of at least 30.
Example 6, p 276 We want to estimate the mean number of sentences in a magazine ad. How many ads must be in the sample if you want to be 95% confident that the sample mean is within one sentence of the population mean?

From Example 2, p 272, s = 5.0, so

$$n = \left(\frac{z_c \sigma}{E}\right)^2 = \left(\frac{1.96 * 5.0}{1}\right)^2 \approx 96.04.$$ So you need a sample of size 97.

Confidence Intervals for the Mean (Small Samples)
When the sample size is small (less than 30), the sample standard deviation s is not good enough to assume that the Central Limit Theorem applies. However when the random variable x is drawn from an approximately normal distribution, the distribution of the following random variable t is known and is called the **t-distribution**.

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

- The t-distribution is bell shaped and symmetric about the mean.
- The t-distribution is a family of curves, each determined by a parameter called the degrees of freedom (d.f.) where $d.f. = n - 1$ (n is the sample size)
- The total area under the t-curve is 1.

- The mean, median, and mode of the t-distribution are equal to zero.
- As the degrees of freedom increase, the t-distribution approaches the standard normal distribution

Constructing confidence interval using the t-distribution is similar to constructing it for the normal distribution as the following table indicates

Procedure	**Equations**	**Example 2 p 286**
Identify the sample statistics $n, \bar{x}$, and s	$\bar{x} = \frac{\Sigma x}{n}$, $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$	$n = 16, \bar{x} = 162, s = 10$
Identify the degrees of freedom, the level of confidence c, and the critical value t_c	$d.f. = n-1$ t_c is found in Table 5 Appendix B	$c = .95$ $d.f. = n-1 = 15$ $t_c = 2.131$
Estimate the maximum error of estimate E	$E = t_c \frac{s}{\sqrt{n}}$	$E = 2.131 * \frac{10}{\sqrt{16}} = 5.3275$
Find the confidence interval	$(\bar{x}-E, \bar{x}+E)$	(156.6725,167.3275)

Summary of when the normal distribution or the t-distribution can be used (p 288)

- If $n \geq 30$, the normal distribution can be used, and s can be used to estimate σ.
- If $n < 30$ and the population is normally or approximately normally distributed, use the normal distribution if σ is known, otherwise use the t-distribution.
- If $n < 30$ and the population is not approximately normally distributed, a CI cannot be constructed.

Note that the 95% Confidence Interval is (22.7617, 25.2383).
The results are presented in the session window as follows:

One-Sample T

N	Mean	StDev	SE Mean	95% CI
10	50.0000	5.0000	1.5811	(46.4232, 53.5768)

Confidence Intervals for Population Proportions

The for p, the population proportion of success is given by the proportion of successes in a sample and is denoted by

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

where n is the sample size and x is the number of successes in the sample. The point estimate for the number of failures is $\hat{q} = 1 - \hat{p}$. The symbols $\hat{p}$ and $\hat{q}$ are read as “p hat” and “q hat”. Note this is derived from the equation in section 5.5 where the mean of $\bar{X}$ was obtained.
The mean and standard deviation of the estimate $\hat{p}$ are:

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

This is the **standard error of the mean** (section 5.5) when the random variable X_i only can take on the value 0 or 1.

Relate this to the sample mean of a random variable that has a binomial distribution:

$$\hat{p} = \frac{\bar{x}}{n},\ E(\hat{p}) = E\left(\frac{\bar{x}}{n}\right) = \frac{1}{n}E(\bar{x}) = \frac{np}{n} = p, \quad Var(\hat{p}) = Var\left(\frac{\bar{x}}{n}\right)^2 = \frac{1}{n^2}Var(\bar{x}) = \frac{1}{n^2}npq = \frac{pq}{n}$$

The following table explains how to construct a confidence interval for the population proportion.

Constructing a Confidence Interval for the Population Proportion (p 294)		
Procedure	**Equations**	**Examples 1, 2 (p 293, 295)**
Identify the sample statistics.	n is the number of trials and x is the number of successes	$n = 1024,\ x = 287$
Find the point estimate $\hat{p}$. Also find the estimate of the standard deviation (the standare error of the mean): $s_{\hat{p}}$	$\hat{p} = \frac{x}{n},\ s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$	$\hat{p} = \frac{287}{1027} \approx 0.28$
Verify that the sampling distribution of $\hat{p}$ can be approximated by the normal distribution	$n\hat{p} \geq 5$ $n\hat{q} \geq 5$	$n\hat{p} = 1024 * 0.28 \approx 287$ $n\hat{q} = 1024 * 0.72 \approx 737$
Find the critical value z_c that corresponds to the given level of confidence c.	Use the Standard Normal Table to find the value z_c such that the area to the right of $z_c = (1-c)/2$	$z_{.95} = 1.96$
Find the maximum error of the estimate E. This is the critical value times the standard error of the mean.	$E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}$	$E \approx 1.96\sqrt{\frac{0.28 * 0.72}{1024}} \approx 0.028$
Find the left and right endpoints of the confidence interval	$(\hat{p} - E, \hat{p} + E)$	$(0.28 - 0.028, 0.28 + 0.028) =$ $(0.252, 0.308)$

Finding the minimum sample size is done by substituting in the formula that was derived above:

$n = \left(\frac{z_c \sigma}{E}\right)^2 = \hat{p}\hat{q}\left(\frac{z_c}{E}\right)^2$ Note that $\hat{p}\hat{q}$ is the estimate of the standard deviation of the proportion.

Example 4 **p 297**. We want to estimate the proportion of voters who support our candidate with 95% confidence that we are within 3% of the actual proportion. Since there is no preliminary estimate for $\hat{p}$ we use 0.5. Substituting into the above equation we gives:

$$n = 0.5 * 0.5\left(\frac{1.96}{0.03}\right)^2 \approx 1067.11$$

Rounding up, we need at least 1068 registered voters to be included in the sample.

Hypothesis Testing with One Sample

Introduction to Hypothesis Testing

A **null hypothesis** H_0 is a statistical hypothesis that contains a statement of equality, such as $\leq, =, or \geq$.

The **alternative hypothesis** H_a is the complement of the null hypothesis. It is a statement that must be true if H_0 is false and it contains a statement of inequality such as $>, \neq, or <$.

- A **Type I error** occurs if the null hypothesis is rejected when it is actually true.
- A **Type II error** occurs if the null hypothesis is not rejected when it is actually false.

A university claims that the proportion of students who graduate in four years is 82%	H_0: p = 0.82 (Claim) H_a: p <> 0.82
A water faucet manufacturer claims that the mean flow rate of a faucet is less than 2.5 gpm	H_0: p >= 2.5 gpm H_a: p < 2.5 gpm (Claim)
A cereal company claims that the mean weight of the contents of its 20-ounce size cereal boxes is more than 20 oz	H_0: mean <= 20 oz H_a: mean > 20 oz (Claim)
An automobile battery manufacturer claims that the mean live of a certain battery type is 74 months	H_0: mean = 74 (Claim) H_a: mean <> 74
A television manufacturer claims that the variance of the life of a certain type of TV is <= 3.5	H_0: variance <= 3.5 (Claim) H_a: variance > 3.5
A radio station claims that its proportion of the local listening audience is greater than 39%	H_0: p <= 0.39 H_a: p > 0.39 (Claim)

DEFINITION In a hypothesis test, the **level of significance** is your maximum allowable probability of making **a Type I error**. It is denoted by α. Three commonly used levels of significance are $\alpha = 0.10, 0.05, 0.01$. Note that making α small means that we want a very small chance that we will reject a null hypothesis that is true.

The probability of **a type II** error is denoted by β.

The following table summarizes this:

	H_0 **True**	H_a **True**
Do not reject H_0	Correct decision	Type II Error (Probability = β)
Reject H_0	Type I error (Probability = α)	Correct decision

The statistic that is compared to the parameter in the null hypothesis is called the test statistic. The following table shows the relationships between population parameters and their corresponding test statistics, sampling distributions, and standardized test statistics. **(p 325)**

Population Parameter	**Test statistic**	**Sampling Distribution**	**Standardized test statistic**
μ	$\bar{x}$	If $n \geq 30$, Normal If $n < 30$, Student t	z t
p	$\hat{p}$	Normal	z

DEFINITION: Assuming the null hypothesis is true, a **P-value** (or **probability value**) of a hypothesis is the probability of obtaining a sample statistic with a value as extreme or more extreme than the one determined from the sample data.

- If the alternative hypothesis H_a contains the less-than inequality symbol (<), the hypothesis test is a **left-tailed test,** i.e. P is the area of the standard normal curve to the left of z.
- If the alternative hypothesis H_a contains the greater-than inequality symbol (>), the hypothesis test is a **right-tailed test.,**i.e. P is the area of the standard normal curve to the right of z.
- If the alternative hypothesis H_a contains the not-equal-to symbol ($\neq$), the hypothesis test is **two-tailed test.** In a two-tailed test, each tail has an area of $\frac{1}{2}P$.

Possible Questions

PART-B

1. Describe the properties of Point Estimation in detail.
2. In order to introduce some incentive for higher balance in savings accounts, a random sample of size 64 savings accounts in a branch was studied to estimate the average monthly balance in saving bank accounts. The mean and standard deviation were found to be Rs, 8500 and Rs. 2000 respectively. Find (i) 90% (ii) 95% (iii) 99% confidence interval for the population mean.
   (Use α : $Z_{(0.10)}$ = 1.645, $Z_{(0.05)}$= 1.96 and $Z_{(0.01)}$ = 2.575)
3. Describe the different methods of estimating the parameters in Point Estimation.
4. One of the properties of a good quality paper is its bursting strength. Suppose a sample of 16 specimens' yields mean bursting strength of 25 units, and it is known from the history of such tests that the standard deviation among specimens is 5 units, assuming normality of test results, what are the (i) 95% and (ii) 98% confidence limits for the mean bursting strength from this sample? Use α :$Z_{(0.05)}$= 1.96 and $Z_{(0.08)}$ = 2.327)
5. Write the different factors affecting the width of a confidence interval.
6. One of the properties of a good quality paper is its bursting strength. Suppose a sample of 16 specimens' yields mean bursting strength of 25 units, and it is known from the history of such tests that the standard deviation among specimens is 5 units, assuming normality of test results, what are the (i) 95% and (ii) 98% confidence limits for the mean bursting strength from this sample? Use α :$Z_{(0.05)}$= 1.96 and $Z_{(0.08)}$ = 2.327)

PART-C

7. In order to introduce some incentive for higher balance in savings accounts, a random sample of size 64 savings accounts in a branch was studied to estimate the average monthly balance in saving bank accounts. The mean and standard deviation were found to be Rs, 8500 and Rs. 2000 respectively.

   Find (i) 90%
   (ii) 95%
   (iii) 99% confidence interval for the population mean.

   (Use α : $Z_{(0.10)}$ = 1.645, $Z_{(0.05)}$= 1.96 and $Z_{(0.01)}$ = 2.575)

DEPARTMENT OF MATHEMATICS						
MATHEMATICAL STATISTICS (18MMP304)						
UNIT- IV						
Sl. No.	**Question**	**Option 1**	**Option 2**	**Option 3**	**Option 4**	**Answer**
1	The process of making estimates about the population parameter from a sample is called:	Statistical independence	Statistical inference	Statistical hypothesis	Statistical decision	Statistical inference
2	Statistical inference has namely two branches,they are………	Level of confidence and degrees of freedom	Biased estimator and unbiased estimator	Point estimator and unbiased estimator	Estimation of parameter and testing of hypothesis	Estimation of parameter and testing of hypothesis
3	Estimation is possible only in case of a…………..	Parameter	Universe	Random sample	Population	Random sample
4	The numerical value which we determine from the sample for population parameter is called:	Estimation	Estimate	Estimator	Confidence coefficient	Estimate
5	Estimation is of two types.They are………..	One sided and two sided	Type I and Type II	Point eastimation and interval estimation	Biased and unbiased	Point eastimation and interval estimation
6	A formula or rule used for estimating the parameter is called:	Estimation	Estimate	Estimate	Interval estimate	Estimator
7	A single value used to estimate a population values is called:	Interval estimate	Point estimate	Level of confidence	Degrees of freedom	Point estimate
8	A value of an estimator is called:	Estimation	Estimate	Variable	Constant	Estimate
9	Standard error is the standard deviation of the sampling distribution of an:	Estimate	Estimation	Estimator	Error of estimation	Estimator
10	An estimator is a random variable because it varies from:	Population to sample	Population to population	Sample to sample	Sample to population	Sample to sample

11	If T is the estimator of parameter t, then T is called unbiased if………	E(T)>t	E(T)<t	E(T) not equal to t	E(T)=t	E(T)=t
12	Estimates given in the form of confidence intervals are called ………………	Point estimates	Interval estimates	Confidence limits	Degree of freedom	Interval estimates
13	Interval estimate is associated with:………	Probability	Non-probability	Range of values	Number of Parameters	Range of values
14	Range or set of values which have chances to contain value of population parameter with particular confidence level is considered as……………	Secondary interval estimation	Confidence interval estimate	Population interval estimate	Sample interval estimate	Confidence interval estimate
15	Sample means are:	Point estimates of sample means	Interval estimates of population means	Interval estimates of sample means	Point estimates of population means	Point estimates of population means
16	The process of making estimates about the population parameter from a sample is called:	Statistical independence	Statistical inference	Statistical hypothesis	Statistical decision	Statistical inference
17	Statistical inference has namely two branches,they are………	Level of confidence and degrees of freedom	Biased estimator and unbiased estimator	Point estimator and unbiased estimator	Estimation of parameter and testing of hypothesis	Estimation of parameter and testing of hypothesis
18	Estimation is possible only in case of a…………..	Parameter	Universe	Random sample	Population	Random sample
19	The numerical value which we determine from the sample for population parameter is called:	Estimation	Estimate	Estimator	Confidence coefficient	Estimate
20	Estimation is of two types.They are………..	One sided and two sided	Type I and Type II	Point eastimation and interval estimation	Biased and unbiased	Point eastimation and interval estimation

21	A formula or rule used for estimating the parameter is called:	Estimation	Estimate	Estimate	Interval estimate	Estimator
22	A single value used to estimate a population values is called:	Interval estimate	Point estimate	Level of confidence	Degrees of freedom	Point estimate
23	A value of an estimator is called:	Estimation	Estimate	Variable	Constant	Estimate
24	Standard error is the standard deviation of the sampling distribution of an:	Estimate	Estimation	Estimator	Error of estimation	Estimator
25	An estimator is a random variable because it varies from:	Population to sample	Population to population	Sample to sample	Sample to population	Sample to sample
26	If T is the estimator of parameter t, then T is called unbiased if………	E(T)>t	E(T)<t	E(T) not equal to t	E(T)=t	E(T)=t
27	Estimates given in the form of confidence intervals are called ………………	Point estimates	Interval estimates	Confidence limits	Degree of freedom	Interval estimates
28	Interval estimate is associated with:………	Probability	Non-probability	Range of values	Number of Parameters	Range of values
29	Method in which sample statistic is used to estimate value of parameters of population is classified as	Estimation	Valuation	Probability calculation	Limited theorem estimation	Estimation
30	Range or set of values which have chances to contain value of population parameter with particular confidence level is considered as……………	Secondary interval estimation	Confidence interval estimate	Population interval estimate	Sample interval estimate	Confidence interval estimate
31	Upper and lower boundaries of interval of confidence are classified as…….	Error biased limits	Marginal limits	Estimate limits	Confidence limits	Confidence limits
32	Criteria of selecting point estimator must includes information of……	Consistency	Biasedness	Inefficiency	Population	Consistency
33	Considering sample statistic, if mean of sampling distribution is equal to population mean then sample statistic is classified as……	Unbiased estimator	Biased estimator	Interval estimation	Hypothesis estimator	Unbiased estimator

34	Which of the following is an estimate of the variability of estimates of the mean in different samples?	Standard error of the mean	Average	Variance	Standard deviation	Standard error of the mean
35	If point estimate is 8 and margin of error is 5 then confidence interval is	3 to 13	4 to 14	5 to 15	6 to 16	3 to 13
36	To develop interval estimate of any parameter of population, value which is added or subtracted from point estimate is classified as………	Margin of efficiency	Margin of consistency	Margin of biasedness	Margin of error	Margin of error
37	In confidence interval estimation, confidence efficient is denoted by……	$1 + \beta$	$1 - \beta$	$1 - \alpha$	$1 + \alpha$	$1 - \alpha$
38	In confidence interval estimation, interval estimate is also classified as….	Confidence efficient	Confidence deviation	Confidence mean	Marginal coefficient	Confidence efficient
39	Value of any sample statistic which is used to estimate parameters of population is classified as……..	Point estimate	Population estimate	Sample estimate	Parameter estimate	Point estimate
40	Distance between true value of population parameter and estimated value of population parameter is called………….	Error of central limit	Error of confidence interval	Error of estimation	Error of hypothesis testing	Error of estimation
41	In confidence interval estimation, formula of calculating confidence interval is………….	Point estimate * margin of error	Point estimate ± margin of error	Point estimate - margin of error	Point estimate + margin of error	Point estimate ± margin of error
42	Difference between value of parameter of population and value of unbiased estimator point is classified as……….	Sampling error	Marginal error	Confidence error	Population error	Sampling error
43	Considering sample statistic, if sample statistic mean is not equal to population parameter then sample statistic is considered as……………..	Unbiased estimator	Biased estimator	Interval estimation	Hypothesis estimator	Biased estimator

44	If true value of population parameter is 10 and estimated value of population parameter is 15 then error of estimation is…	5	25	0.667	150	5
45	A confidence interval will be widened if:	The confidence level is increased and the sample size is reduced	The confidence level is increased and the sample size is increased	The confidence level is decreased and the sample size is decreased	The confidence level is decreased and the sample size is increased	The confidence level is increased and the sample size is reduced
46	A 95% confidence interval for the mean of a population is such that:	It contains 95% of the values in the population	There is a 95% chance that it contains all the values in the population.	There is a 95% chance that it contains the standard deviation of the population	There is a 95% chance that it contains the mean of population	There is a 95% chance that it contains the mean of population
47	If a standard error of a statistic is less than that of another then what is the former is said to be…………..	efficient	unbiased	consistent	sufficient	efficient
48	_________ are the values that mark the boundaries of the confidence interval.	Confidence intervals	Confidence limits	Levels of confidence	Margin of error	Confidence limits
49	When S is used to estimate σ, the margin of error is computed by using…….	normal distribution	t distribution	sample mean	population mean	t distribution
50	For the interval estimation of μ when σ is known and the sample is large, the proper distribution to use is……………….	t distribution with n +1 degrees of freedom	t distribution with n-1 degrees of freedom	t distribution with n degrees of freedom	normal distribution	normal distribution

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established under Section 3 of UGC Act, 1956)

Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

Subject : Mathematical Statistics	**Semester III**	**L T P C**
Subject Code : 18MMP304	**Class : II M.Sc Mathematics**	**4 0 0 4**

UNIT V

Analysis of Variance: One way classification and two-way classification. Hypotheses Testing: Poser functions -OC function-Most Powerful test -Uniformly most powerful test -unbiased test.

SUGGESTED READINGS

1. Marek Fisz, (1980). Probability Theory and Mathematical Statistics, John Wiley and Sons, New York.
2. Meyer, (2006). Introduction to Probability and Statistical applications, Oxford and IBH Publishing Co. Pvt. Ltd. New Delhi.
3. Sheldon M. Ross, (2009). Introduction to probability and statistics for engineers and scientists, Third edition, Academic press.
4. Heinz Bauer,(1996), Probability Theory, Narosa Publishing House, London.
5. Parimal Mukhopadhyay, (2012). Theory of Probability, New central book agency, Calcutta.

Analysis of Variance (ANOVA)

I. Introduction

In Regression, the decomposition of the total sum of squares (SST) into the "explained" sum of squares (SSR) and the "unexplained" sum of squares (SSE) took place in the Analysis of Variance or ANOVA table. However, ANOVA also refers to a statistical technique used to test for diffferences between the means for several populations. While the procedure is related to regression, in ANOVA the independent variable(s) are qualitative rather than quantitative. In both regression and ANOVA the dependent variable is quantitative.

Example 1: As city manager, one of your responsibilities is purchasing. The city is looking to buy lightbulbs for the city's streetlights. Aware that some brands' lightbulbs might outlive other brands' lightbulbs, you decide to conduct an experiment. Seven lightbulbs each are purchased from four brands (GE, Dot, West, and a generic) and placed in streetlights. The lifetime of each of the 28 lightbulbs is then recorded in the file "**Lightbulbs**."

In this example, the lifetime of a lightbulb, in thousands of hours, is the quantitative dependent variable of interest. The company marketing the lightbulb, i.e., the brand-name, is the qualitative independent variable. The variable "brand name" has four possible values (or four "levels" in the terminology of ANOVA). The letter "*k*" will be used for the number of "levels" of the independent variable or "factor". Here, $k = 4$ for the four brands being tested. We say, "the factor *brand-name* has four levels: GE, Dot, West, and generic."

The "populations" referred to in these notes are simply the different levels of the factor. So that, in this example, we are interested in whether the mean lifetimes for the four populations of lightbulbs differ. Since we cannot know with certainty, however, the true mean lifetime of all lightbulbs carrying a certain brand-name, we rely upon statistics to determine if the differences observed *between* samples drawn from the four brands are statistically significant. (Non-significant differences are those that can plausibly be attributed to chance, i.e., sample-to-sample, variation alone.)

II. The (one-way) ANOVA Model

In order to perform tests of statistical significance, a model is assumed. The model used in ANOVA is similar in many respects to the model employed in regression. In fact, You may find it useful in these notes to make analogies between the model and formulas in ANOVA and the corresponding model and formulas in regression. In the model below, recall that the dependent (or **response**) variable is quantitative as in regression, but the independent (or **factor**) variable is now qualitative. We begin with a model in which a single independent variable is used to describe the dependent variable. This **One-Way** ANOVA is analogous to simple regression. The one-way analysis of variance model is

Example 1 (continued): For the lightbulb problem,

- the response is the lifetime of a particular lightbulb (in thousands of hours)
- the factor is the brand-name
- there are four levels or treatments: GE, Dot, West, and generic
- □$_{\text{□GE}}$ is the mean lifetime of all GE bulbs, □$_{\text{□Dot}}$ is the mean lifetime of all Dot bulbs, etc.

III Hypothesis Test

As usual, we rely on a hypothesis test to determine if the sample means for the k samples drawn (one from each population) differ enough for the difference to be statistically significant (more than would likely occur due to random chance alone).

Example 1 (continued): It is important that the student understand why probability is important here. It is not unusual for one manufacturer to source a product marketed under many brand-names. For example, there are only a handfull of companies manufacturing denim jeans, but there are dozens of brand-name jeans available to the consumer. Similarly, not all lightbulbs are manufactured by the companies marketing them. It is not inconceivable, therefore, that all four brands of lightbulbs being tested by the city come off of the same assembly line. Yet, when tested, they would still yield four *different* sample means simply because of sample-to-sample variation. As city manager, you might be more than a little embarassed to discover that the brand that you've touted as superior to all others is actually different in name only! Lawsuits have been lost for far less.

Hypotheses:

- **H_0:**, i.e., all population means are equal. This is equivalent to saying that the *k* treatments have no differential effect upon the value of the response.
- **H_A:** At least two of the means differ. This says that different treatments produce different values of the response variable, on average.

Test Statistic:

F = $\frac{MSR}{MSE}$ **,** where ***MSR*** = the **<u>M</u>ean <u>S</u>quare** for **<u>T</u>reatments**,
and ***MSE*** = the **<u>M</u>ean <u>S</u>quare** for **<u>E</u>rror**

Note: What I'm calling ***MSR*** is often called *MST* in the literature. I've chosen to continue the use of *MSR* to highlight the similarity between regression and analysis of variance. MSE remains the same for both regression and analysis of variance. Formulas for the mean squares are given later in the notes.

Logic:

The analysis of variance uses the ratio of two ***variances***, *MSR* and *MSE*, to determine whether population ***means*** differ; hence the name "analysis of variance." Recall that one of the assumptions of the model is that the variance □□ is the same for all populations. ***MSE*** provides an unbiased estimate of □□□in ANOVA just as it does in regression (see regression notes). If the population means are all equal, which is the null hypothesis, it can be shown that *MSR also* provides an unbiased estimate of □□ . If all of the population means are equal, therefore, we would expect **F** to be nearly equal to **1** since *MSR* and *MSE* should yield similar estimates of the variance □□ .

If some population means differ from others, however, *MSR* will tend to be bigger than *MSE* resulting in an **F - Ratio** substantially larger than **1**. Thus we reject **H_0** for large values of **F**, just as in regression.

IV. The ANOVA Table: Sums of Squares and Degrees of Freedom

A. Introduction

At the heart of any analysis of variance is the ANOVA Table. The formulas for the sums of squares in ANOVA are simplified if the *k* samples are all of the same size n_S. In the interests of simplicity, therefore, the following discussion assumes that all *k* samples contain the same number of observations n_S.

B. Notation

- The index i represents the i^th population or treatment, where i ranges from 1 to *k*
- The index j represents the j^th obsevation within a sample, where j ranges from 1 to n_S
- *n* is the total number of observations from all samples
- y_{ij} is the value of the j^th observation in the i^th sample
- $\bar{y}_i$ is the mean of the i^th sample
- $\bar{\bar{y}}$ (read "y double-bar") is the mean of all *n* observations, $\bar{\bar{y}} = \frac{\mathbf{1}}{n}\sum_{i=\mathbf{1}}^{k}\sum_{j=\mathbf{1}}^{n_S} y_{ij}$, or the mean of the sample means (hence the "double-bar" in the name), $\bar{\bar{y}} = \frac{\bar{y}_{\mathbf{1}} + \bar{y}_{\mathbf{2}} + \cdots + \bar{y}_k}{k}$

C. Sums of Squares

<u>S</u>um of <u>S</u>quares for <u>T</u>reatments, $\boldsymbol{SSR} = n_S\sum_{i=1}^{k}\left(\bar{y}_i - \bar{\bar{y}}\right)^2$ is the "Between Group" variation, where the *k* "groups" or populations are represented by their sample means. If the sample means differ substantially then SST will be large.

<u>S</u>um of <u>S</u>quares for <u>E</u>rror, $\boldsymbol{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{n_S}\left(y_{ij} - \bar{y}_i\right)^2$ is the "Within Group" variation and represents the random or sample-to-sample variation

<u>Total</u> <u>S</u>um of <u>S</u>quares, $\boldsymbol{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_S}\left(y_{ij} - \bar{\bar{y}}\right)^2$ is the total variation in the values of the response variables over all *k* samples. (Note: SST is the same as in regression)

D. Degrees of Freedom

Degrees of freedom for treatments, $\mathbf{df_{SSR}} = \boldsymbol{k}\mathbf{-1}$. Rather than memorizing this formula, just imagine the number of dummy variables that you would have to create to conduct the equivalent analysis in regression. Since you always leave one possibility out in regression, you would need to create *k* – 1 dummy variables. Since the resulting regression model would have *k* – 1 independent variables, SSR (SST here) would have *k* – 1 degrees of freedom.

Degrees of freedom for error, $\mathbf{df_{SSE}} = \boldsymbol{n} - \boldsymbol{k}$.

Total degrees of freedom, $\mathbf{df_{SST}} = \boldsymbol{n} - \mathbf{1}$. This is the same result obtained in regression.

Note: The two component degrees of freedom sum to the total degrees of freedom, just as in regression.

E. Mean Squares

<u>M</u>ean <u>S</u>quare for <u>T</u>reatments, $MSR = \dfrac{SSR}{k - 1}$ is equivalent to MSR in regression

<u>M</u>ean <u>S</u>quare for <u>E</u>rror, $MSE = \dfrac{SSE}{n - k}$ is the same as MSE in regression. As in regression, MSE is an unbiased estimator of the common population variance □□.

F. F – Ratio

The statistic used to test the null hypothesis is $F = \dfrac{MSR}{MSE}$. As mentioned earlier, if the null hypothesis is correct then this ratio should be close to one. If some of the sample means differ substantially, however, the ratio will be much larger. Large values of F therefore correspond to strong evidence for rejecting H_0. Statgraphics reports a *P*-value for the test.

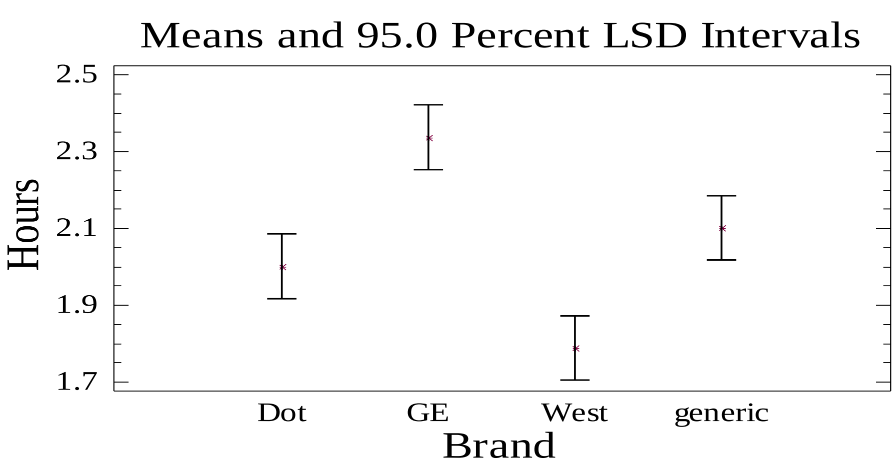
G. Summary

The ANOVA Table below summarizes some of the information in this section

ANOVA Table for One-Way Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Between groups	***SSR***	k - 1	***MSR* = *SSR*/(*k*-1)**	***F* = *MSR/MSE***	
Within groups	***SSE***	n - k	***MSE* = *SSE*/(n-k)**		
Total (Corr.)	***SST***	n - 1			

Below is the *Means Plot*. There is clearly evidence, at the 5% level of significance that the GE bulbs last longer, on average, than bulbs from the other brands. Similarly, there is evidence, at the 5% level, that the West bulbs fail sooner, on average, than bulbs from the other brands. The sample differences between the Dot and generic bulbs, however, may be due to chance alone. (We don't actually *know* that Dot and generic bulbs are interchangeable, but the sample doesn't provide strong enough evidence to discount the possibility.)



V. Two-Way ANOVA

When the effects of two qualitative factors upon a quantitative response variable are investigated, the procedure is called two-way ANOVA. Although a model exists for two-way analysis of variance, similar to the multiple regression model, it will not be covered in this class. Neither will we cover the details of the ANOVA Table. Nevertheless, there are some new considerations in two-way ANOVA stemming from the presence of the second factor in the model.

Example 2: The EPA (Environmental Protection Agency) tests public bodies of water for the presence of *coliform* bacteria. Aside from being potentially harmful to people in its own right, this bacteria tend to proliferate in polluted water, making the presence of *coliform* bacteria a surrogate for polution. Water samples are collected off public beaches, and the number of *coliform* bacterial per cc is determined. (See the file "**Bacteria**."

The EPA is interested in determining the factors that affect *coliform* bacterial formation in a particular county. The county has beaches adjacent to the ocean, a bay, and a sound. The EPA beleives that the amount of "flushing" a beach gets may affect the ability of polution to accumulate in the waters off the beach. The EPA also believes that the geographical location of the beach may be significant. (There could be several reasons for this: the climate may be different in different parts of the county, or the land-use may vary across the county, etc.)

As luck would have it, there is at least one beach for each combination of type (ocean, bay, sound) and location (west, central, east) within the county. Because of this, the EPA decides to sample a beach at each of the 9 possible combinations of type and location and conduct a two-way analysis of variance for *coliform* bacterial count. Two independent samples are taken at each beach to allow for an estimation of the natural variation in *coliform* bacterial count (this "repetition" is needed for the computation of MSE, which estimates the sample-to-sample variance in bacterial counts).

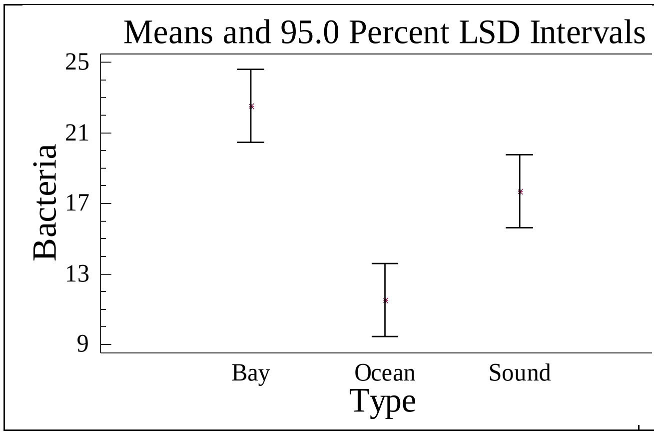
Before interpreting the results in the ANOVA table above, we should consider the role that interaction plays. If the effect of beach type on bacteria formation depends on the location of the beach then it is better to investigate the *combinations* of the levels of the factors type and location for their affect on bacteria. It will come as no surprise to you that there is a hypothesis test for interactions.

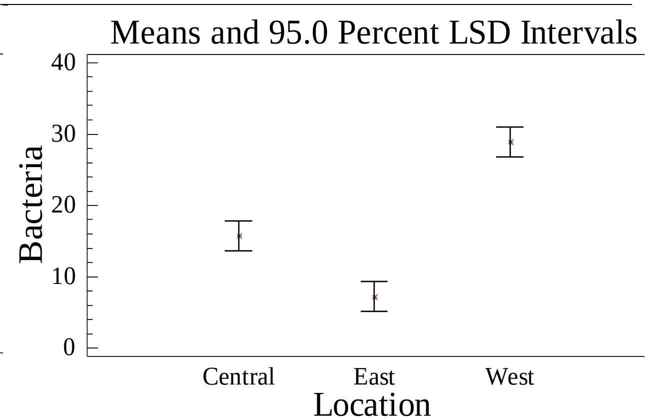
H_0: The factors Type and Location do *not* interact.
H_A: The factors Typ and Location *do* interact

To check for interaction, use the right mouse button and *Analysis Options* and enter "2" for the *Maximum Order Interaction*. The resulting output for our example below shows a *P*-value of 0.3047 for the test for interactions. Thus the evidence for interaction is not particularly strong. The practical effect of discounting interaction is that we are able to return to the previous output (the one without interactions) and interpret the *P*-values for the factors Type and Location separately. Since the *P*-values for both factors are significant, we conclude that factors affect bacteria growth.

Source	Sum of Squares	Df	Mean Square	F-Rat
MAIN EFFECTS				
A:Type	364.778	2	182.389	18.
B:Location	1430.11	2	715.056	70.
INTERACTIONS				
AB	57.2222	4	14.3056	1.
RESIDUAL	91.0	9	10.1111	
TOTAL (CORRECTED)	1943.11	17		

Having determined that the type of beach and the beach's location are both significant, we next investigate the nature of the relationship between these factors and bacteria count. Once again we turn to the means plots under *Graphical Options*. Statgraphics dfaults to a means plot for the factor Type because this was the first factor entered in the *Input Dialog Box*. To get a means plot for the factor Location, use *Pane Options* to select it. The two means plots appear below.





Individually, these means plots are interpreted as in one-way ANOVA. There is evidence, at the 5% level of significance, that the mean bacteria count at ocean beaches is less than for other types, and that the mean count is highest at bay beaches. Similarly, the mean count is lowest in the east and greatest in the west, with all differences being statistically significant at the 5% level of significance. Furthermore, ***because interactions were judged not-significant***, we can add the main effects together and say that the least polluted beaches tend to be located in the east on the ocean, while the most polluted tend to be in the west on bays. We could not have added the separate (or main) effects in this way if there had been significant interact, for in that case the effect upon bacteria count at a particular type of beach (ocean, for example) may be very different for different locations.

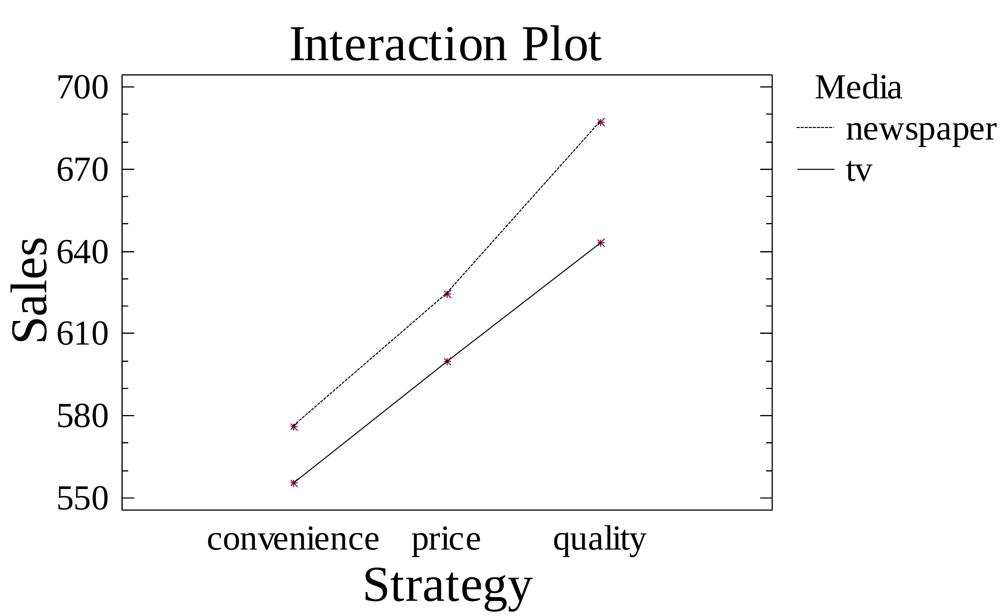
Example 3: The last two examples are based on a marketing study. A new apple juice product was entering the marketplace. It had three distinct advantages relative to existing apple juices. First, it was not a concentrate and was therefore considered to be of higher "quality" than many similar products. Second, as one of the first juices packaged in cartons, it was cheaper than competing products. Third, partly because of the packaging, it was more convenient. The director of marketing for the company would like to know which advantage should be emphasized in advertisements. The director would also like to know whether local television or newspapers are better for sales.

Consequently, six cities with similar demographics are chosen, and a different combination of "Marketing Strategy " and "Media" is tried in each. The unit sales of apple juice for the ten weeks immediately following the start of the ad campaigns are recorded for each city in the file **Apple Juice (two-way)**. The two-way table below describes the city assignments for the six possible combinations of levels for the two factors. Below the assignment table is the ANOVA Table for interactions.

	Convenience	Quality	Price
Local Television	City 1	City 3	City 5
Newspaper	City 2	City 4	City 6

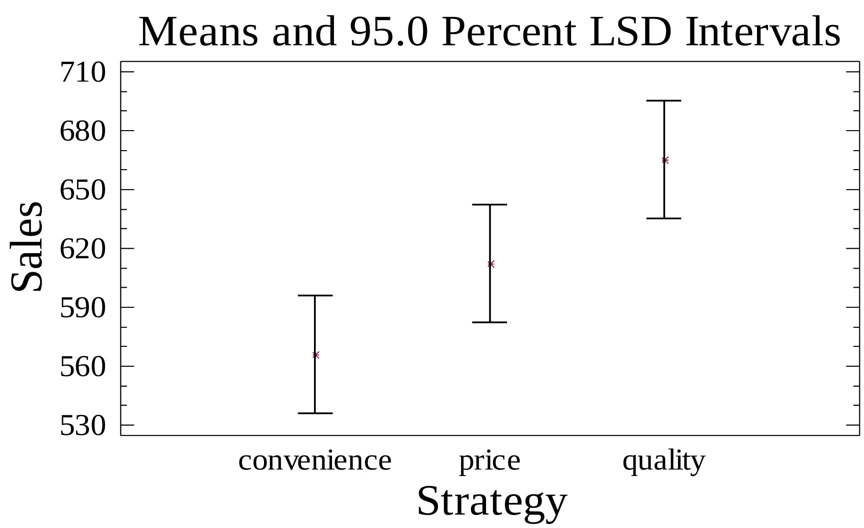
```
Source                     Sum of Squares      Df     Mean Square     F-Rat
--------------------------------------------------------------------------
MAIN EFFECTS
 A:Strategy                       98838.6       2         49419.3        5.
 B:Media                          13172.0       1         13172.0        1.

INTERACTIONS
 AB                               1609.63       2         804.817        0.

RESIDUAL                         501137.0      54         9280.31
--------------------------------------------------------------------------
TOTAL (CORRECTED)                614757.0      59
```

Interactions are not significant to the model (p-value equals 0.9171), a fact which is reinforced by looking at the *Interaction Plot* under *Graphical Options*. Note that the two curves are almost parallel, a sign that interactions are not significant.



Removing interactions, we obtain the ANOVA Table below, from which we conclude that the marketing strategy is significant, but the media used probably isn't. Since only marketing strategy apppears to affect sales, we'll restrict ourselves to the means plot for the factor Strategy below. Only the difference in mean sales when emphasizing quality versus emphasizing convenience is statistically significant at the 5% level of significance.

Source	Sum of Squares	Df	Mean Square	F-Rat
MAIN EFFECTS				
A:Strategy	98838.6	2	49419.3	5.
B:Media	13172.0	1	13172.0	1.
RESIDUAL	502746.0	56	8977.61	
TOTAL (CORRECTED)	614757.0	59		

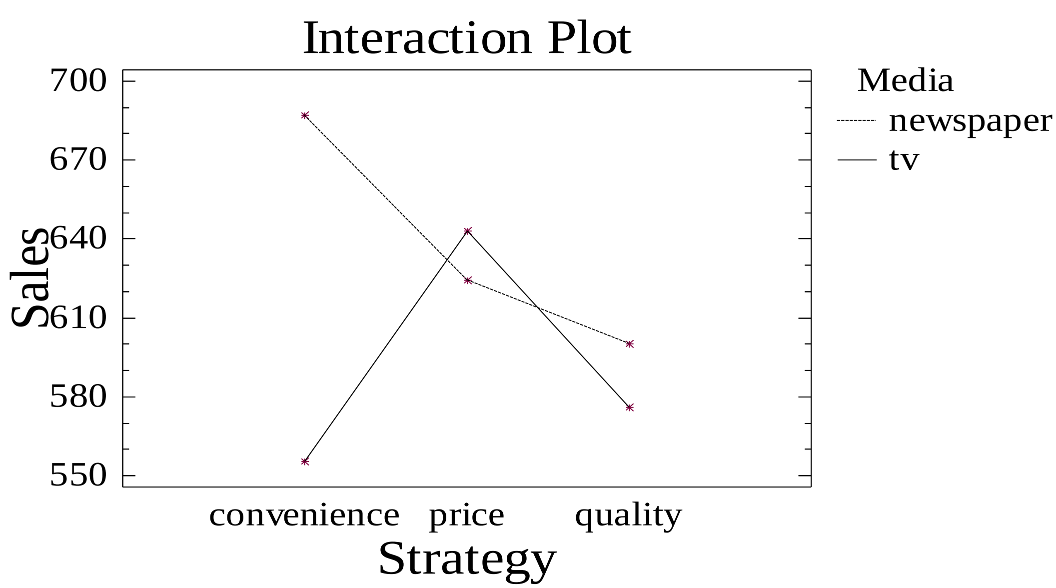


Example 4: This is just the apple juice problem revisited (see file "**Apple Juice – Remix**"). By a judicious rearrangement of sales figures, I've created a marketing study in which interactions are significant. (See the two-way table below for the new assignments.) The comparison of the interaction plots for this example and example 3 should help to clarify the role of interactions in the interpretation of ANOVA output. The small *P*-value of 0.0474 for the hypothesis test of interactions implies that certain combinations of marketing strategy and media are important to sales.

	Convenience	Quality	Price
Local Television	City 1	City 2	City 3
Newspaper	City 4	City 5	City 6

Source	Sum of Squares	Df	Mean Square	F-Rat
MAIN EFFECTS				
A:Strategy	22393.6	2	11196.8	1.
B:Media	31327.4	1	31327.4	3.
INTERACTIONS				
AB	59899.3	2	29949.6	3.
RESIDUAL	501137.0	54	9280.31	
TOTAL (CORRECTED)	614757.0	59		

Looking at he interaction plot, notice that emphasizing convenience lead to both the lowest and highest mean sales, depending upon whether local television or newspapers were used. Thus, it wouldn't make sense to talk about the effect of emphasizing convenience without consideration of the media used, i.e., we should only interpret levels of the two factors taken together (the combinations). Therefore, we will not investigate the means plots for Strategy and Media. From the interaction plot, it appears that the most effective campaign would emphasize convenience in newspapers. The least effective combination is to emphasize convenince on local television. (Note: Since the interaction plot doesn't display confidence intervals for the six possible combinations, we cannot attach a particular significance level to our conclusions as we could with the means plots.)



Power Functions

DEFINITION: *A **power function** is a function of the form* $f(x) = kx^p$ *where* k *and* p *are constants.*

EXAMPLE 1: Which of the following functions are power functions? For each power function, state the value of the constants k and p in the formula $y = kx^p$.

a. $b(x) = 5(x-3)^4$ ***b.*** $m(x) = 7\sqrt[4]{x}$ ***c.*** $l(x) = 3 \cdot 2^x$

SOLUTIONS:

a. *The function* $b(x) = 5(x-3)^4$ *is not a power function because we cannot write it in the form* $y = kx^p$.

b.*The function* $m(x) = 7\sqrt[4]{x}$ *is a power function because we can rewrite its formula as* $m(x) = 7 \cdot x^{1/4}$. *So* $k = 7$ *and* $p = \frac{1}{4}$.

c. *The function* $l(x) = 3 \cdot 2^x$ *is not a power function because the power is not constant. In fact,* $l(x) = 3 \cdot 2^x$ *is an exponential function.*

As is the case with linear functions and exponential functions, given two points on the graph of a power function, we can find the function's formula.

EXAMPLE 2: Suppose that the points $(1, 81)$ and $(3, 729)$ are on the graph of a function f. Find an algebraic rule for f assuming that it is a power function.

SOLUTION:

Since f is a power function we know that its rule has form $f(x) = kx^p$. We can use the two given points to find two equations involving k and p:

$$(1, 81) \Rightarrow f(1) = 81 = k(1)^p$$
$$(3, 729) \Rightarrow f(3) = 729 = k(3)^p.$$

We can use the first equation to immediately find k.

$$81 = k(1)^p$$
$$\Rightarrow k = 81$$

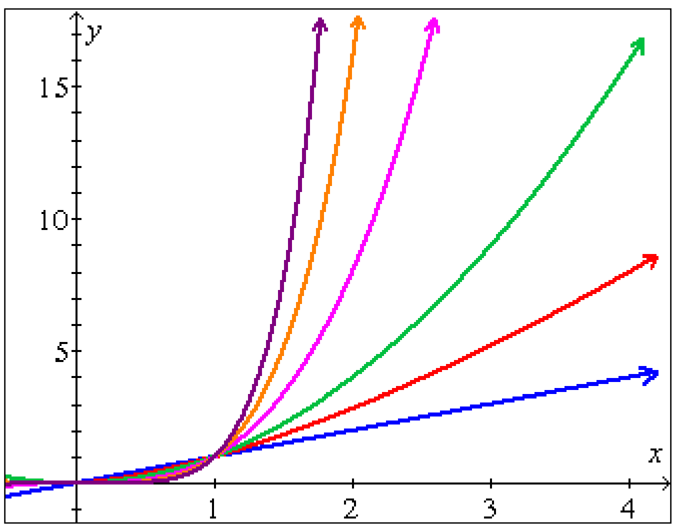
Now we can find p by substituting $k = 81$ into the second equation:

$$729 = 81(3)^p$$
$$\Rightarrow \frac{729}{81} = 3^p$$
$$\Rightarrow \quad 9 = 3^p$$
(note that this could be solved with logarithms if the solution weren't so obvious)
$$\Rightarrow \quad p = 2$$

Thus, if f is a power function, its rule is $f(x) = 81x^2$.

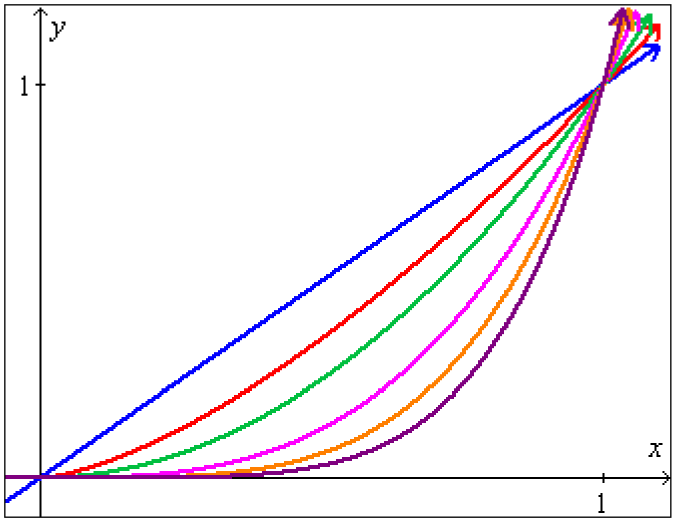
Graphs of Power Functions

For a power function $y = kx^p$ the greater the power of p, the faster the outputs grow. Below are the graphs of six power functions. Notice that as the power increases, the outputs increase more and more quickly. As x increases without bound (written "$x \to \infty$"), higher powers of x get a lot larger than (i.e., *dominate*) lower powers of x. (Note that we are discussing the **long-term** behavior of the function.)



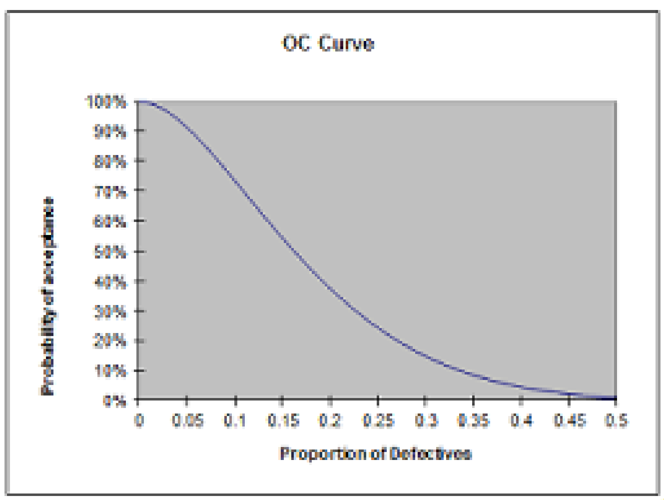
The graphs of $y = x$, $y = x^{3/2}$, $y = x^2$, $y = x^3$, $y = x^4$, $y = x^5$.

As x approaches zero (written "$x \to 0$"), the story is completely different. If x is between 0 and 1, x^3 is larger than x^4, which is larger than x^5. (Try $x = 0.1$ to confirm this). For values of x near zero, *smaller* powers dominate. On the graph below, notice how on the interval $(0, 1)$ the linear power function $y = x$ dominates power functions of larger power.



The graphs of $y = x$, $y = x^{3/2}$, $y = x^2$, $y = x^3$, $y = x^4$, $y = x^5$.

Operating Characteristic Curve
An operating characteristic (OC) curve is a chart that displays the probability of acceptance versus percentage of defective items (or lots). With no defects, we'll surely have 100% acceptance! But, take a look at 0.05 (5% defective).



Most Powerful Test
In statistical hypothesis testing, a uniformly most powerful (UMP) test is a **hypothesis test** which has the greatest power among all possible tests of a given size α. For example, according to the Neyman–Pearson lemma, the **likelihood-ratio test** is UMP for testing simple (point) hypotheses

In Most Powerful test both the null hypothesis and the alternative hypothesis is composite (the null hypothesis is composite because σ can take on any positive value). There is no UMP test because the z-test for known σ is more powerful than the t-test for unknown σ .

Unbiased Test

In **statistical** hypothesis **testing**, a **test** is said to be **unbiased** if for some alpha level (between 0 and 1), the probability the null is rejected is less than or equal to the alpha level for the entire parameter space defined by the null hypothesis, whilst the probability the null is rejected is greater than or equals.

Possible Questions
PART-B

1. Three processes A, B and C are tested to see whether their outputs are equivalent. The following observations of outputs are made:

A	10	12	13	11	10	14	15	13
B	9	11	10	12	13			
C	11	10	15	14	12	13		

Given, table value of F for (2,16) d.f at 5% level of significance is 3.63. Carry out the analysis of variance and state your conclusion.

2. Explain one-way classification in ANOVA.
3. What are the criterions for a uniformly most powerful test?
4. Describe power functions and OC functions
5. Three processes A, B and C are tested to see whether their outputs are equivalent. The following observations of outputs are made:

A	10	12	13	11	10	14	15	13
B	9	11	10	12	13			
C	11	10	15	14	12	13		

Given, table value of F for (2,16) d.f at 5% level of significance is 3.63. Carry out the analysis of variance and state your conclusion.

6. Describe power functions and OC functions

PART-C

1. Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the treatments (cars types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use $\alpha = 5\%$.

	Compact cars	**Midsize cars**	**Full-size cars**
	643	469	484
	655	427	456
	702	525	402
$\bar{X}$	666.67	473.67	447.33
S	31.18	49.17	41.68

DEPARTMENT OF MATHEMATICS						
MATHEMATICAL STATISTICS (18MMP304)						
UNIT-V						
Sl. No.	**Question**	**Option 1**	**Option 2**	**Option 3**	**Option 4**	**Answer**
1	The square of the S.D is ……………….	Variance	Coefficient of variation	Square of variance	Square of coefficient of variation	Variance
2	Analysis of variance is a statistical method of comparing the ________ of several populations.	Standard deviations	Means	Variances	Proportions	Means
3	The analysis of variance is a statistical test that is used to compare how many group means?	Three	More than three	Three or more	Two or more	Two or more
4	Analysis of variance utilizes:	F-test	Chi-Square test	Z-test	t-test	F-test
5	What is two-way ANOVA?	An ANOVA with two variables and one factor	An ANOVA with one variable and two factors	An ANOVA with one variable and three factors	An ANOVA with both categorical and scale variables	An ANOVA with one variable and two factors
6	Which of the following is the correct F ratio in the one-way ANOVA?	MSA/MSE	MSBL/MSE	MST/MSE	MSE/MST	MST/MSE
7	For validity of F-test in Anova, parent population should be……….	Binomial	Poisson	Normal	Exponential	Normal
8	______ sum of squares measures the variability of the observed values around their respective tabulated values	Treatment	Error	Interaction	Total	Error
9	The ________ sum of squares measures the variability of the sample treatment means around the overall mean.	Total	Treatment	Error	Interaction	Treatment

10	If the true means of the k populations are equal, then MST/MSE should be:	more than 1.00	Close to 1.00	Close to - 1.00	A negative value between 0 and - 1	Close to 1.00
11	If MSE of ANOVA for six treatment groups is known, you can compute…………	Degree of freedom	The standard deviation of each treatment group	Variance	The pooled standard deviation	The pooled standard deviation
12	To determine whether the test statistic of ANOVA is statistically significant,to determine critical value we need	Sample size, number of groups	Mean, sample standard deviation	Expected frequency, obtained frequency	MSTR, MSE	Sample size, number of groups
13	Which of the following is an assumption of one-way ANOVA comparing samples from 3 or more experimental treatments?	Variables follow F-distribution	Variables follow normal distribution	Samples are dependent each other	Variables have different variances	Variables follow normal distribution
14	The error deviations within the SSE statistic measure distances:	Within groups	Between groups	Between each value and the grand mean	Betweeen samples	Within groups
15	In one-way ANOVA, which of the following is used within the F-ratio as a measurement of the variance of individual observations?	SSTR	MSTR	SSE	MSE	SSE
16	When conducting a one-way ANOVA, the _______ the between-treatment variability is when compared to the within-treatment variability	More random larger	Smalller	Larger	More random smaller	Smaller
17	When conducting a one-way ANOVA,the value of F DATA will be tend to be ……..	More random larger	Smalller	More random smaller	Larger	Smaller

18	When conducting an ANOVA, F DATA will always fall within what range?	Between negative infinity and infinity	Between 0 and 1	Between 0 and infinity	Between 1 and infinity	Between 0 and infinity
19	*If* F DATA = 5, the result is statistically significant	Always	Sometimes	Never	Is impossible	Sometimes
20	*If* F DATA= 0.9, the result is statistically significant	Always	Sometimes	Never	Is impossible	Never
21	When comparing three treatments in a one-way ANOVA ,the alternate hypothesis is…..	All three treatments have different effect on the mean response.	Exactly two of the three treatments have the same effect on the mean response.	At least two treatments are different from each other in terms of their effect on the mean response	All the treatments have same effect	At least two treatments are different from each other in terms of their effect on the mean response
22	If the sample means for each of k treatment groups were identical,the observed value of the ANOVA test statistic?	1	0	A value between 0.0 and 1.0	A negative value	0
23	If the null hypothesis is rejected, the probability of obtaining a F- ratio > the value in the F table as the 95th % is:	0.5	>0.5	<0.5	1	<0.5
24	ANOVA was used to test the outcomes of three drug treatments. Each drug was given to 20 individuals. If MSE =16, What is the standard deviation for all 60 individuals sampled for this study?	6.928	48	16	4	4
25	Analysis of variance technique originated in the field of………………	Agriculture	Industry	Biology	Genetics	Agriculture

26	With 90, 35, 25 as TSS, SSR and SSC , in case of two way classification, SSE is…..	50	40	30	20	30
27	Variation between classes or variation due to different basis of classification is commonly known as ……………...	Treatments	Total sum of squares	Sum of squares	Sum of squares due to error	Treatments
28	The total variation in observations in Anova is classified as:	Treatments and inherent variation	SSE and SST	MSE and MST	TSS and SSE	Treatments and inherent variation
29	In Anova, variance ratio is given by……………	MST/MSE	MSE/MST	SSE/SST	TSS/SSE	MST/MSE
30	Degree of freedom for TSS is……………..	N-1	k-1	h-1	(k-1)(h-1)	N-1
31	For Anova, MST stands for……………………...	Mean sum of squares of treatment	Mean sum of squares of varieties	Mean sum of squares of tables	Mean sum of sources of treatment	Mean sum of squares of treatment
32	An ANOVA procedure is applied to data of 4 samples,where each sample contains 10 observations.Then degree of freedom for critical value of F are	4 numerator and 9 denominator	3 numerator and 40 denominator	3 numerator and 36 denominator	4 numerator and 10 denominator	3 numerator and 36 denominator
33	The power function of a test is denoted by……………………	M(w,Q)	M(Q,Qo)	P(w,Q)	P(w,Qo)	M(w,Q)
34	Sum of power function and operation characteristic is…………	Unity	Zero	two	Negative	Unity
35	Operation characteristic is denoted by…………………	L(w,Q)	M(w,Q)	L(w,Qo)	M(w,Qo)	L(w,Q)
36	Operation characteristic is also known as …………………	Test characteristic	Power function	best characteristic	unique characteristic	Test characteristic
37	The formula to find OC is L(w,Q)=	1-Power Function	2xPower Function	Power Funtion -1	2xConfidance Interval	1-Power Function
38	Operation Characteristic is of a test is related to…	Power Function	Best Test	Unique Test	Uniformally Best Test	Power Function

39	If the Hypothesis is correct the operation charectristics will be…………	1	0	-1	0.5	1
40	If the Hypothesis is wrong the operation charectristics will be……….	0	1	0.5	0.333333	0
41	In which test we verify a null hypothesis against any other definite alternate hypothesis?	Best Test	Unique Test	Uniformally Best Test	Unbiased Test	Best Test
42	A Best Test is a Test such that the critical region for which _______ attains least value for a given α.	Beta	1-Beta	Alpha	1-Alpha	1-Beta
43	A Test whose power funtion attains its mean at point Q = Qo is called ____ Test	Unique	Unbiased	Power	Operation Characteristic	Unique
44	A Best Unique Test exist ____________	Always	Never	Sometimes	When Q not = to Qo	Sometimes
45	Operation Characteristic is related to	Power Function	Unique Test	Best Test	Uniformally Best Test	Power Function
46	Power is the ability to detect:	A statistically significant effect where one exists	A psychologically important effect where one exists	Both (a) and (b) above	Design flaws	A statistically significant effect where one exists
47	Calculating how much of the total variance is due to error and the experimental manipulation is called:	Calculating the variance	Partitioning the variance	Producing the variance	Summarizing the variance	Partitioning the variance

48	ANOVA is useful for:	Teasing out the individual effects of factors on an Independent Variables	Analyzing data from research with more than one Independent Variable and one Dependent Variable	Analyzing correlational data	Individual effects of factors on an Dependent Variables	Analyzing data from research with more than one Independent Variable and one Dependent Variable
49	What is the definition of a simple effect?	The effect of one variable on another	The difference between two conditions of one Independent Variable at one level of another Independent Variable	The easiest way to get a significant result	Difference between two Dependent Variables	The difference between two conditions of one Independent Variable at one level of another Independent Variable
50	In a study with gender as the manipulated variable, the Independent Variable is:	Within participants	Correlational	Between participants	Regressional	Between participants
51	Which of the following statements are true of experiments?	The Independent Variable is manipulated by the experimenter	The Dependent Variable is assumed to be dependent upon the IV	They are difficult to conduct	both (a) and (b)	both (a) and (b)

52	All other things being equal, repeated-measures designs:	Have exactly the same power as independent designs	Are often less powerful than independent designs	Are often more powerful than independent designs	Are rarely less powerful when compare to than independent designs	Are often more powerful than independent designs
53	Professor P. Nutt is examining the differences between the scores of three groups of participants. If the groups show homogeneity of variance, this means that the variances for the groups:	Are similar	Are dissimilar	Are exactly the same	Are enormously different	Are similar
54	Differences between groups, which result from our experimental manipulation, are called:	Individual differences	Treatment effects	Experiment error	Within-participants effects	Treatment effects
55	Herr Hazelnuss is thinking about whether he should use a related or unrelated design for one of his studies. As usual, there are advantages and disadvantages to both. He has four conditions. If, in a related design, he uses 10 participants, how many would he need for an unrelated design?	40	20	10	100	40
56	Individual differences within each group of participants are called:	Treatment effects	Between-participants error	Within-participants error	Individual biases	Within-participants error
57	Calculating how much of the total variance is due to error and the experimental manipulation is called:	Calculating the variance	Partitioning the variance	Producing the variance	Summarizing the variance	Partitioning the variance
58	The decision on how many factors to keep is decided on:	Statistical criteria	Theoretical criteria	Both (a) and (b)	Neither (a) nor (b)	Both (a) and (b)

59	It is possible to extract:	As many factors as variables	More factors than variables	More variables than factors	Correlation between the actual and predicted variables	As many factors as variables
60	Four groups have the following means on the covariate: 35, 42, 28, 65. What is the grand mean?	43.5	42.5	56.7	58.9	42.5
61	You can perform ANCOVA on:	Two groups	Three groups	Four groups	All of the above	All of the above
62	When carrying out a pretest--posttest study, researchers often wish to:	Partial out the effect of the dependent variable	Partial out the effect of the pretest	Reduce the correlation between the pretest and posttest scores	Correlation between the two tests scores	Partial out the effect of the pretest
63	Using difference scores in a pretest--posttest design does not partial out the effect of the pretest for the following reason:	The pretest scores are not normally correlated with the posttest scores	The pretest scores are normally correlated with the different scores	The posttest scores are normally correlated with the different scores	Up normal relationship with the different scores	The pretest scores are normally correlated with the different scores
64	Experimental designs are characterized by:	Two conditions	No control condition	Random allocation of participants to conditions	More than two conditions	Random allocation of participants to conditions

65	Between-participants designs can be:	Either quasi-experimental or experimental	Only experimental	Only quasi-experimental	Only correlational	Either quasi-experimental or experimental
66	A continuous variable can be described as:	Able to take only certain discrete values within a range of scores	Able to take any value within a range of scores	Being made up of categories	Being made up of variables	Able to take any value within a range of scores
67	In a within-participants design with two conditions, if you do not use counterbalancing of the conditions then your study is likely to suffer from:	Order effects	Effects of time of day	Lack of participants	Effects of participants	Order effects
68	Demand effects are possible confounding variables where:	Participants behave in the way they think the experimenter wants them to behave	Participants perform poorly because they are tired or bored	Participants perform well because they have practiced the experimental task	Participants perform strongly	Participants behave in the way they think the experimenter wants them to behave
69	Power can be calculated by a knowledge of:	The statistical test, the type of design and the effect size	The statistical test, the criterion significance level and the effect size	The criterion significance level, the effect size and the type of design	The criterion significance level, the effect size and the sample size	The criterion significance level, the effect size and the sample size
70	Relative to large effect sizes, small effect sizes are:	Easier to detect	Harder to detect	As easy to detect	As difficult to detect	As difficult to detect

71	Differences between groups, which result from our experimental manipulation, are called:	Individual differences	Treatment effects	Experiment error	Within-participants effects	Treatment effects

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established under Section 3 of UGC Act, 1956)

Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

Subject : Mathematical Statistics **Semester III** **L T P C**
Subject Code : 18MMP304 **Class : II M.Sc Mathematics** **4 0 0 4**

Sl. No.	Reg. No.	Name of the Student	Seminar Topic
1	18MMP001	Ajithkumar K	Contribution of Statistical Tests in Insurance Industry
2	18MMP002	Aruthra B	Application of Statistical Concepts in Food Processing Industry
3	18MMP003	Asaithambi C	A recent development and applications of t-distribution in Telecommunication
4	18MMP004	Baby Salini R	Applications of Kolmogorow - Smirnov test
5	18MMP005	Deventhiran T	ANOVA for examine the relationship between Security and Safety for Human Development
6	18MMP006	Elavarasan A	A review of most powerful test in decision making process
7	18MMP007	Kalaikumar K	Application of Chi-Square test in Psychology
8	18MMP008	Kamalnath M	Estimation Theory and Information Technology
9	18MMP009	Karthick V	Application of ANOVA in medical research
10	18MMP010	Kokilamani S	Bayes Theorem and its contribution to Hospital Industry
11	18MMP011	Muthupandi K	Consistency and limitation of paired t-test and Wilcoxon test
12	18MMP012	Nandhini R	A study on Normalization of data for decision making
13	18MMP014	Parveen S	Contribution of F-distribution in agricultural industry
14	18MMP015	Prakash R	Application of Mann Whitney U - test for different sample size
15	18MMP016	Ragapriya M	How one way ANOVA is used as decision making tool in industry
16	18MMP017	Santhosh Robinson P	A study on how statistical tests are used in Defense Department
17	18MMP018	Saranya S	Application of Chi-Square test in manufacturing industry.
18	18MMP019	Sridhar G	OC curve as decision making tool in production industry
19	18MMP020	Sudha S	How two way ANOVA is used as decision making tool in industry
20	18MMP021	Tharani S	Application of OC curve in production process as quality control measure
21	18MMP023	Vignesh G	A study on different estimators and their applications
22	18MMP024	Vijay K	A study on acceptance sampling plans and its applications

Signature of the Course Faculty

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

Subject : Mathematical Statistics **Semester III** **L T P C**
Subject Code : 18MMP304 **Class : II M.Sc Mathematics** **4 0 0 4**

Glossary of Statistical Terms

2 X 5 factorial design	A factorial design with one variable having two levels and the other having five levels.
Alpha	The probability of a Type I error.
Abscissa	Horizontal axis.
Additive law of probability	The rule giving the probability of the occurrence of one or more mutually exclusive events.
Adjacent values	Actual data points that are no more extreme than the inner fences.
Alternative hypothesis (H_1)	The hypothesis that is adopted when H_0 is rejected. Usually the same as the research hypothesis.
Analysis of variance (ANOVA)	A statistical technique for testing for differences in the means of several groups.
Analytic view	Definition of probability in terms of analysis of possible outcomes.
ß (Beta)	The probability of a Type II error.
Categorical data	Data representing counts or number of observations in each category.
Cell	The combination of a particular row and column ‹the set of observations obtained under identical treatment conditions.
Central limit theorem	The theorem that specifies the nature of the sampling distribution of the mean.
Chi-square test	A statistical test often used for analyzing categorical data.
Conditional probability	The probability of one event *given* the occurrence of some other event.
Confidence interval	An interval, with limits at either end, with a specified probability of including the parameter being estimated.
Confidence limits	An interval, with limits at either end, with a specified probability of including the parameter being estimated.
Constant	A number that does not change in value in a given situation.
Contingency table	A two dimensional table in which each observation is classified on the basis of two variables simultaneously.
Continuous variables	Variables that take on *any* value.
Correlation	Relationship between variables.
Correlation coefficient	A measure of the relationship between variables.
Count data	Data representing counts or number of observations in each category.
Covariance	A statistic representing the degree to which two variables vary together.

Criterion variable	The variable to be predicted.
Critical value	The value of a test statistic at or beyond which we will reject H0 .
Decision making	A procedure for making logical decisions on the basis of sample data.
Degrees of freedom (*df*)	The number of independent pieces of information remaining after estimating one or more parameters.
Density	Height of the curve for a given value of X- closely related to the probability of an observation in an interval around X.
Dependent variables	The variable being measured. The data or score.
df_{error}	Degrees of freedom associated with $SS_{error} = k(n - 1)$.
df_{group}	Degrees of freedom associated with $SS_{group} = k - 1$.
df_{total}	Degrees of freedom associated with $SS_{total} = N - 1$.
Dichotomous variables	Variables that can take on only two different values.
Directional test	A test that rejects extreme outcomes in only one specified tail of the distribution.
Discrete variables	Variables that take on a small set of possible values.
Dispersion	The degree to which individual data points are distributed around the mean.
Distribution free tests	Statistical tests that do not rely on parameter estimation or precise distributional assumptions.
Effect size	The difference between two population means divided by the standard deviation of either population.
Efficiency	The degree to which repeated values for a statistic cluster around the parameter.
Error variance	The square of the standard error of estimate.
Event	The outcome of a trial.
Exhaustive	A set of events that represents all possible outcomes.
Expected value	The average value calculated for a statistic over an infinite number of samples.
Expected frequencies	The expected value for the number of observations in a cell if H_0 is true.
Experimental hypothesis	Another name for the research hypothesis.
Exploratory data analysis (EDA)	A set of techniques developed by Tukey for presenting data in visually meaningful ways.
External Validity	The ability to generalize the results from this experiment to a larger population.
Frequency distribution	A distribution in which the values of the dependent variable are tabled or plotted against their frequency of occurrence.
Frequency data	Data representing counts or number of observations in each category.
Friedman's rank test for *k* correlated samples	A nonparametric test analogous to a standard one-way repeated measures analysis of variance.
Goodness of fit	A test for comparing observed frequencies with theoretically predicted

test	frequencies.
Grand total (□X)	The sum of all of the observations.
Heterogeneity of variance	A situation in which samples are drawn from populations having different variances.
Hypothesis testing	A process by which decisions are made concerning the values of parameters.
Independent variables	Those variables controlled by the experimenter.
Independent events	Events are independent when the occurrence of one has no effect on the probability of the occurrence of the other.
Interaction	A situation in a factorial design in which the effects of one independent variable depend upon the level of another independent variable.
Intercept	The value of Y when X is 0.
Interval scale	Scale on which equal intervals between objects represent equal differences ‹differences are meaningful.
Interval estimate	A range of values estimated to include the parameter.
Joint probability	The probability of the co-occurrence of two or more events.
Kruskal Wallis one-way analysis of variance	A nonparametric test analogous to a standard one-way analysis of variance.
Kurtosis	A measure of the peakedness of a distribution.
Leading digits (most significant digits)	Left-most digits of a number.
Least significant difference test	A technique in which we run *t* tests between pairs of means only if the analysis of variance was significant.
Leptokurtic	A distribution that has relatively more scores in the center and in the tails.
Linear relationship	A situation in which the best-fitting regression line is a straight line.
Linear regression	Regression in which the relationship is linear.
Mann-Whitney test	A nonparametric test for comparing the central tendency of two independent samples.
Marginal totals	Totals for the levels of one variable summed across the levels of the other variable.
Matched samples	An experimental design in which the same subject is observed under more than one treatment.
Mean absolute deviation (m.a.d.)	Mean of the absolute deviations about the mean.
Mean	The sum of the scores divided by the number of scores.
Measurement	The assignment of numbers to objects.
Measurement data	Data obtained by measuring objects or events.
Measures of central tendency	Numerical values referring to the center of the distribution.

Median location	The location of the median in an ordered series.
Median (Med)	The score corresponding to the point having 50% of the observations below it when observations are arranged in numerical order.
Mesokurtic	A distribution with a neutral degree of kurtosis.
Midpoints	Center of interval -- average of upper and lower limits.
Mode (Mo)	The most commonly occurring score.
Monotonic relationship	A relationship represented by a regression line that is continually increasing (or decreasing), but perhaps not in a straight line.
$MS_{between\ groups}$ (MS_{group})	Variability among group means.
MS_{within} (MS_{error})	Variability among subjects in the same treatment group.
Multiplicative law of probability	The rule giving the probability of the joint occurrence of independent events.
Mutually exclusive	Two events are mutually exclusive when the occurrence of one precludes the occurrence of the other.
Negative relationship	A relationship in which increases in one variable are associated with decreases in the other.
Negatively skewed	A distribution that trails off to the left.
Nominal scale	Numbers used only to distinguish among objects.
Nonparametric tests	Statistical tests that do not rely on parameter estimation or precise distributional assumptions.
normal distribution	A specific distribution having a characteristic bell-shaped form.
Null hypothesis (H_0)	The statistical hypothesis tested by the statistical procedure. Usually a hypothesis of no difference or no relationship.
One-tailed test	A test that rejects extreme outcomes in only one specified tail of the distribution.
One-way ANOVA	An analysis of variance where the groups are defined on only one independent variable.
Ordinal scale	Numbers used only to place objects in order.
Ordinate	Vertical axis.
Outlier	An extreme point that stands out from the rest of the distribution.
p level	The probability that a particular result would occur by chance if H_0 is true. The exact probability of a Type I error.
Parameters	Numerical values summarizing population data.
Parametric tests	Statistical tests that involve assumptions about, or estimation of, population parameters.
Pearson product-moment correlation coefficient (*r*)	The most common correlation coefficient.
Percentile	The point below which a specified percentage of the observations fall.
Phi	The correlation coefficient when both of the variables are measured as dichotomies.
Platykurtic	A distribution that is relatively thick in the "shoulders."
Point estimate	The specific value taken as the estimate of a parameter.

Pooled variance	A weighted average of the separate sample variances.
Population variance	Variance of the population ‹usually estimated, rarely computed.
Population	Complete set of events in which you are interested.
Positively skewed	A distribution that trails off to the right.
Power	The probability of correctly rejecting a false H_0.
Predictor variable	The variable from which a prediction is made.
Protected *t*	A technique in which we run *t* tests between pairs of means only if the analysis of variance was significant.
Quantitative data	Data obtained by measuring objects or events.
Random sample	A sample in which each member of the population has an equal chance of inclusion.
Random Assignment	Assigning participants to groups or cells on a random basis.
Range	The distance from the lowest to the highest score.
Range restrictions	Refers to cases in which the range over which X or Y varies is artificially limited.
Ranked data	Data for which the observations have been replaced by their numerical ranks from lowest to highest.
Rank - randomization tests	A class of nonparametric tests based on the theoretical distribution of randomly assigned ranks.
Ratio scale	A scale with a true zero point -- ratios are meaningful.
Real lower limit	The points halfway between the top of one interval and the bottom of the next.
Real upper limit	The points halfway between the top of one interval and the bottom of the next.
Rectangular distribution	A distribution in which all outcomes are equally likely.
Regression	The prediction of one variable from knowledge of one or more other variables.
Regression equation	The equation that predicts Y from X.
Regression coefficients	The general name given to the slope and the intercept ‹most often refers just to the slope.
Rejection region	The set of outcomes of an experiment that will lead to rejection of H_0 .
Rejection level	The probability with which we are willing to reject H0 when it is in fact correct.
Related samples	An experimental design in which the same subject is observed under more than one treatment.
Relative frequency view	Definition of probability in terms of past performance.
Research hypothesis	The hypothesis that the experiment was designed to investigate.
Sample	Set of actual observations. Subset of the population.
Sample statistics	Statistics calculated from a sample and used primarily to describe the

	sample.
Sample variance (s^2)	Sum of the squared deviations about the mean divided by *N* - 1.
Sample with replacement	Sampling in which the item drawn on trial *N* is replaced before the drawing on trial *N* + 1.
Sampling distribution of differences between means	The distribution of the differences between means over repeated sampling from the same population(s).
Sampling distribution of the mean	The distribution of sample means over repeated sampling from one population.
Sampling distributions	The distribution of a statistic over repeated sampling from a specified population.
Sampling error	Variability of a statistic from sample to sample due to chance.
Scales of measurement	Characteristics of relations among numbers assigned to objects.
Scatter plot	A figure in which the individual data points are plotted in two-dimensional space.
Scatter diagram	A figure in which the individual data points are plotted in two-dimensional space.
Scattergram	A figure in which the individual data points are plotted in two-dimensional space.
Sigma	Symbol indicating summation.
Significance level	The probability with which we are willing to reject H_0 when it is in fact correct.
Simple effect	The effect of one independent variable at one level of another independent variable.
Skewness	A measure of the degree to which a distribution is asymmetrical.
Slope	The amount of change in Y for a one unit change in X.
Spearman's correlation coefficient for ranked data (r_s)	A correlation coefficient on ranked data.
SS_{cells}	The sum of squares assessing differences among cell totals.
SS_{error}	The sum of the squared residuals.
SS_{error}	The sum of the sums of squares within each group.
SS_{group}	The sum of squares of group totals divided by the number of scores per group minus □X^2/N.
SS_{total}	The sum of squares of all of the scores, regardless of group membership.
SS_Y	The sum of the squared deviations.
Standard deviation	Square root of the variance.
Standard error	The standard deviation of a sampling distribution.
Standard error of differences between means	The standard deviation of the sampling distribution of the differences between means.
Standard error	The average of the squared deviations about the regression line.

of estimate	
Standard scores	Scores with a predetermined mean and standard deviation.
Standard normal distribution	A normal distribution with a mean equal to 0 and variance equal to 1. Denoted *N* (0, 1).
Statistics	Numerical values summarizing sample data.
Student's *t* distribution	The sampling distribution of the *t* statistic.
Subjective probability	Definition of probability in terms of personal subjective belief in the likelihood of an outcome.
Sufficient statistic	A statistic that uses all of the information in a sample.
Sums of squares	The sum of the squared deviations around some point (usually a mean or predicted value).
Symmetric	Having the same shape on both sides of the center.
T scores	A set of scores with a mean of 50 and a standard deviation of 10.
Test statistics	The results of a statistical test.
Two-Tailed test	A test that rejects extreme outcomes in either tail of the distribution.
Type I error	The error of rejecting H_0 when it is true.
Type II error	The error of not rejecting H_0 when it is false.
Unbiased estimator	A statistic whose expected value is equal to the parameter to be estimated.
Unconditional probability	The probability of one event *ignoring* the occurrence or nonoccurrence of some other event.
Unimodal	A distribution having one distinct peak.
Variables	Properties of objects that can take on different values.
Weighted average	The mean of the form: $(a_1X_1 + a_2X_2)/(a_1 + a_2)$ where a_1 and a_2 are weighting factors and X_1 and X_2 are the values to be average.
Wilcoxon's matched-pairs signed ranks test	A nonparametric test for comparing the central tendency of two matched (related) samples.
z score	Number of standard deviations above or below the mean.