MADHYA PRADESH BHOJ (OPEN) UNIVERSITY
Raja Bhoj Marg (Kolar Road), Bhopal (M.P.)-462016

पाठ्य सामग्री

M.A. Previous (Economics)

Quantitative Methods

Self learning Material

M.A. Previous(Economics)

Self Learning Material

QUANTITATIVE METHODS

PAPER-III

Madhya Pradesh Bhoj (Open) University, Bhopal

MADHYA PRADESH BHOJ (OPEN) UNIVERSITY
Raja Bhoj Marg (Kolar Raod) Bhopal - 462016. Tel: (0755) 2492095.
Fax: (0755)-2424640.
email: bedspc@rediffmail.com or bed@bhojvirtualuniversity.com
website : http://www.bhojvirtualuniversity.com

PAPER-III QUANTITATIVE METHODS

(Questions will be set from each Unit/Section)

UNIT-I **Mathematical Methods** - Concept of of function and types of functions Limit continuity and derivative; Simple ru7les of integration. Deteminants and their basic properties; Solution of simultaneous equaticns theough cramers' rule Concept fo matrix- their thpes, simple operations on matrices, inversion and rank of a matrix.

UNIT--II **Mathematical Methods** - Linear prcgramming- Basic concept; Formulation of a linear programming problem- optimal solution of linear programming through rgaphigical method; Concept of a game ; Strategies - simple and mixed; value of a saddle point solution;

UNIT-III **Statistical Methods** - Meaning assumptions and limtations of simple correlation and regression analysis; pearons' Concept of the least squares and the lines of regression; standard error of estimate Partial and multiple correlation

UNIT--IV **statistical Methods** - Deterministic and non- deterministic experiments: Various types of events - classical and empirical definition of probability: Laws of addition and mutiplication probability and concept of interdependence: Byas' therem and its application; Elementary concept of random variable probability Expectations, moments and generating functions; properties (without derivations) ot Binomial poisson and Normal Distributions.

UNIT-V **Basic Conecept of sampling** - random non-random sampling; Simple random and p.p.s. sampling; Concept of an estimator and its sampling distribution, Desirable properties of and estimator; Formulation of Statistical hypotheses - Null and altemative Goodness of fit Confidence inervals and leave of significance; Hypothesis testing based on Z, t, χ^2 (Chi-square) and F tests.

CONTENTS

BLOCK 1

BLOCK 2

BLOCK 3

BLOCK 4

BLOCK 1 MATHEMATICAL METHODS

The block comprising three units discussed comprehensively the basic mathematics which is of wide application in day to day life of decision makers in economic parlance.

The first unit deals systematically with various aspects of types of functional relationships among economics variables and their applicability in economic concepts. It also throws light on very useful concepts of integration and related rules.

The second unit gives you an insight into Basic calculus-Limits, continuity and derivatives and acquaints you with some very frequently used methods to find out derivatives with different techniques.

Subsequently the third unit explains the basic concepts, theoretical operations and various applications of matrix algebra in quantitative analysis of decisions pertaining to decision making process.

UNIT 1

FUNCTIONS AND INTEGRATION

Objectives

After studying this unit, you should be able to understand and appreciate:

- The need to identify or define the relationships that exists among variables.
- how to define functional relationships
- the various types of functional relationships
- concept of integration
- different rules of integration

Structure

1.1 INTRODUCTION

The concept of a function expresses dependence between two quantities, one of which is known and the other which is produced. A function associates a single output to each input element drawn from a fixed set, such as the real numbers, although different inputs may have the same output.

There are many ways to give a function: by a formula, by a plot or graph, by an algorithm that computes it, or by a description of its properties. Sometimes, a function is described through its relationship to other functions (see, for example, inverse function). In applied disciplines, functions are frequently specified by their tables of values or by a formula. Not all types of description can be given for every possible function, and one must make a firm distinction between the function itself and multiple ways of presenting or visualizing it.

1.2 CONCEPT OF FUNCTIONS

Functions in algebra are usually expressed in terms of algebraic operations. Functions studied in analysis, such as the exponential function, may have additional properties arising from continuity of space, but in the most general case cannot be defined by a single formula. Analytic functions in complex analysis may be defined fairly concretely through their series expansions. On the other hand, in lambda calculus, function is a primitive concept, instead of being defined in terms of set theory. The terms transformation and mapping are often synonymous with function. In some contexts, however, they differ slightly. In the first case, the term transformation usually applies to functions whose inputs and outputs are elements of the same set or more general structure. Thus, we speak of linear transformations from a vector space into itself and of symmetry transformations of a geometric object or a pattern. In the second case, used to describe sets whose nature is arbitrary, the term mapping is the most general concept of function.

In traditional calculus, a function is defined as a relation between two terms called variables because their values vary. Call the terms, for example, x and y. If every

value of x is associated with exactly one value of y, then y is said to be a function of x. It is customary to use x for what is called the "independent variable," and y for what is called the "dependent variable" because its value depends on the value of x.

Restated, mathematical functions are denoted frequently by letters, and the standard notation for the output of a function f with the input x is $f(x)$. A function may be defined only for certain inputs, and the collection of all acceptable inputs of the function is called its domain. The set of all resulting outputs is called the image of the function. However, in many fields, it is also important to specify the codomain of a function, which contains the image, but need not be equal to it. The distinction between image and co domain lets us ask whether the two happen to be equal, which in particular cases may be a question of some mathematical interest. The term range often refers to the co domain or to the image, depending on the preference of the author.
For example:

The expression $f(x) = x2$ describes a function f of a variable x, which, depending on the context, may be an integer, a real or complex number or even an element of a group. Let us specify that x is an integer; then this function relates each input, x, with a single output, x2, obtained from x by squaring. Thus, the input of 3 is related to the output of 9, the input of 1 to the output of 1, and the input of –2 to the output of 4, and we write $f(3) = 9$, $f(1)=1$, $f(-2)=4$. Since every integer can be squared, the domain of this function consists of all integers, while its image is the set of perfect squares. If we choose integers as the co domain as well, we find that many numbers, such as 2, 3, and 6, are in the co domain but not the image.

It is a usual practice in mathematics to introduce functions with temporary names like f; in the next paragraph we might define $f(x) = 2x+1$, and then $f(3) = 7$. When a name for the function is not needed, often the form y = x2 is used.
If we use a function often, we may give it a more permanent name as, for example,

$$\text{Square}(x) = x^2.$$

The essential property of a function is that for each input there must be a unique output.

Thus, for example, the formula

$$\text{Root}(x) = \pm\sqrt{x}$$

Does not define a real function of a positive real variable, because it assigns two outputs to each number: the square roots of 9 are 3 and −3. To make the square root a real function, we must specify, which square root to choose. The definition

$$\text{Posroot}(x) = \sqrt{x}$$

For any positive input chooses the positive square root as an output.

As mentioned above, a function need not involve numbers. By way of examples, consider the function that associates with each word its first letter or the function that associates with each triangle its area.

1.3 TYPES OF FUNCTIONS

In this section some different types of functions are introduced which are particularly useful in calculus.

1.3.1 LINEAR FUNCTIONS

These are names for functions of first, second and third order polynomial functions, respectively. What this means is that the highest order of x (the variable) in the function is 1, 2 or 3.

The generalized form for a linear function (1 is highest power):

f(x) = ax+b, where a and b are constants, and a is not equal to 0

The generalized form for a quadratic function (2 is highest power):

f(x) = ax2+bx+c, where a, b and c are constants, and a is not equal to 0

The generalized form for a cubic function (3 is highest power):

f(x) = ax3+bx2+cx+d, where a, b, c and d are constants, and a is not equal to 0

The roots of a function are defined as the points where the function f(x)=0. For linear and quadratic functions, this is fairly straight-forward, but the formula for a cubic is quite complicated and higher powers get even more involved.
a system of linear equations (or linear system) is a collection of linear equations involving the same set of variables.

For example,

$$\begin{aligned} 3x + 2y - \ z &= \ 1 \\ 2x - 2y + 4z &= -2 \\ -x + \tfrac{1}{2}y - \ z &= \ 0 \end{aligned}$$

is a system of three equations in the three variables x, y, z. A solution to a linear system is an assignment of numbers to the variables such that all the equations are simultaneously satisfied. A solution to the system above is given by

$$\begin{aligned} x &= \ 1 \\ y &= -2 \\ z &= -2 \end{aligned}$$

since it makes all three equations valid.

In mathematics, the theory of linear systems is a branch of linear algebra, a subject which is fundamental to modern mathematics. Computational algorithms for finding the solutions are an important part of numerical linear algebra, and such methods play a prominent role in engineering, physics, chemistry, computer science, and economics. A system of non-linear equations can often be

approximated by a linear system (see linearization), a helpful technique when making a mathematical model or computer simulation of a relatively complex system.

1.3.2 POLYNOMIAL FUNCTIONS

Stated quite simply, polynomial functions are functions with x as an input variable, made up of several terms, each term is made up of two factors, the first being a real number coefficient, and the second being x raised to some non-negative integer power. Actually, it's a bit more complicated than that. Please refer to the following links to get a deeper understanding.

Here a few examples of polynomial functions:

$f(x) = 4x3 + 8x2 + 2x + 3$

$g(x) = 2.5x5 + 5.2x2 + 7$

$h(x) = 3x2$

$i(x) = 22.6$

Polynomial functions are functions that have this form:

$f(x) = anxn + an-1xn-1 + ... + a1x + a0$

The value of n must be an nonnegative integer. That is, it must be whole number; it is equal to zero or a positive integer.

The coefficients, as they are called, are an, an-1, ..., a1, a0. These are real numbers.

The degree of the polynomial function is the highest value for n where an is not equal to 0.

So, the degree of g(x) = 2.5x5 + 5.2x2 + 7 is 5.

Notice that the second to the last term in this form actually has x raised to an exponent of

1, as in:

f(x) = anxn + an-1xn-1 + ... + a1x1 + a0

Of course, usually we do not show exponents of 1.

Notice that the last term in this form actually has x raised to an exponent of 0, as in:

f(x) = anxn + an-1xn-1 + ... + a1x + a0x0
Of course, x raised to a power of 0 makes it equal to 1, and we usually do not show multiplications by 1.

So, in its most formal presentation, one could show the form of a polynomial function as:

f(x) = anxn + an-1xn-1 + ... + a1x1 + a0x0

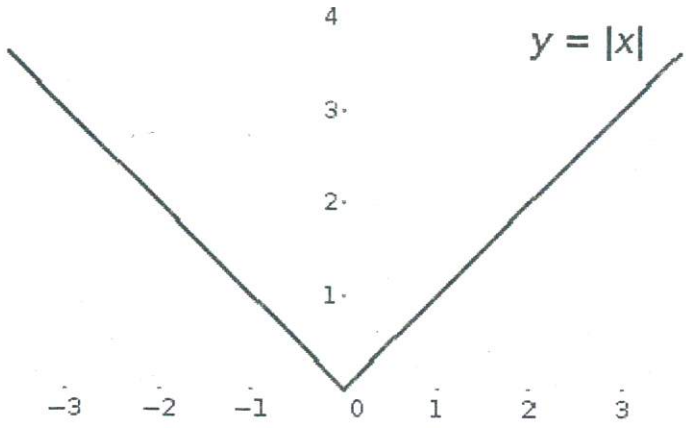
Here are some polynomial functions; notice that the coefficients can be positive or negative real numbers.

f(x) = 2.4x5 + 1.7x2 - 5.6x + 8.1
f(x) = 4x3 + 5.6x
f(x) = 3.7x3 - 9.2x2 + 0.1x - 5.2

1.3.3 ABSOLUTE VALUE FUNCTION

The absolute value (or modulus) of a real number is its numerical value without regard to its sign. So, for example, 3 is the absolute value of both 3 and –3.
The absolute value of a number a is denoted by | a | .

Generalizations of the absolute value for real numbers occur in a wide variety of mathematical settings. For example an absolute value is also defined for the complex numbers, the quaternions, ordered rings, fields and vector spaces. The absolute value is closely related to the notions of magnitude, distance, and norm in various mathematical and physical contexts.



The graph of the absolute value functions for real numbers.

More precisely, if D is an integral domain, then an absolute value is any mapping $|\cdot|$ from D to the real numbers R satisfying:

$|x| \geq 0$,

$|x| = 0$ if and only if $x = 0$,

$|xy| = |x||y|$,

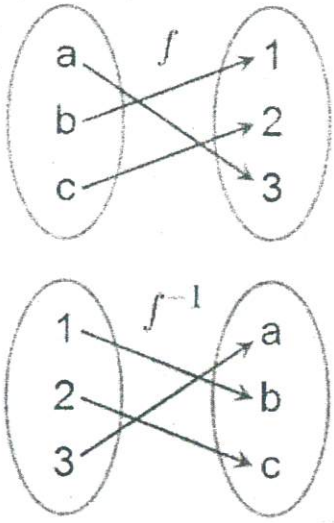
$|x + y| \leq |x| + |y|$.

Note that some authors use the term valuation or norm instead of "absolute value".

1.3.4 INVERSE FUNCTION

If f is a function from A to B then an inverse function for f is a function in the opposite direction, from B to A, with the property that a round trip (a composition) from A to B to A (or from B to A to B) returns each element of the initial set to itself. Thus, if an input x into the function f produces an output y, then inputting y into the inverse function f–1 (read f inverse, not to be confused

with exponentiation) produces the output x. Not every function has an inverse; those that do are called invertible.



A function f and its inverse f–1. Because f maps a to 3, the inverse f–1 maps 3 back to a.

For example, let f be the function that converts a temperature in degrees Celsius to a temperature in degrees Fahrenheit:

$f(C) = \frac{9}{5}C + 32;$

then its inverse function converts degrees Fahrenheit to degrees Celsius:

$f^{-1}(F) = \frac{5}{9}(F - 32).$

Or, suppose f assigns each child in a family of three the year of its birth. An inverse function would tell us which child was born in a given year. However, if the family has twins (or triplets) then we cannot know which to name for their common birth year. As well, if we are given a year in which no child was born then we cannot name a child. But if each child was born in a separate year, and if

we restrict attention to the three years in which a child was born, then we do have an inverse function. For example,

$$f(\text{Alan}) = 2005, \qquad f(\text{Brad}) = 2007, \qquad f(\text{Cary}) = 2001$$
$$f^{-1}(2001) = \text{Cary}. \qquad f^{-1}(2005) = \text{Alan}. \qquad f^{-1}(2007) = \text{Brad}$$

1.3.5 STEP FUNCTION

A step function is a special type of relationship in which one quantity increases in steps in relation to another quantity.

For example,

Postage cost increases as the weight of a letter or package increases. In the year 2001 a letter weighing between 0 and 1 ounce required a 34-cent stamp. When the weight of the letter increased above 1 ounce and up to 2 ounces, the postage amount increased to 55 cents, a step increase.

A graph of a step function *f* gives a visual picture to the term "step function." A step function exhibits a graph with steps similar to a ladder.

The domain of a step function *f* is divided or partitioned into a number of intervals. In each interval, a step function *f(x)* is constant. So within an interval, the value of the step function does not change. In different intervals, however, a step function *f* can take different constant values.

One common type of step function is the greatest-integer function. The domain of the greatest-integer function *f* is the real number set that is divided into intervals of the form ...[2, 1), [1, 0), [0, 1), [1, 2), [2, 3),... The intervals of the greatest-integer function are of the form [*k*, *k* 1), where *k* is an integer. It is constant on every interval and equal to *k*.

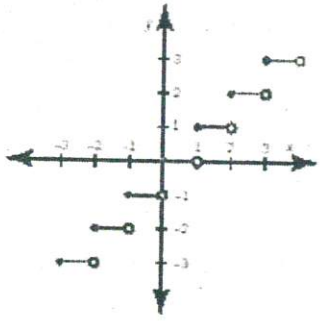
f(x) = 0 on [0, 1), or $0 \le x < 1$

f(x) = 1 on [1, 2), or $1 \le x < 2$

$f(x) = 2$ on [2, 3), or $2 \leq x < 3$

For instance, in the interval [2, 3), or $2 \leq x < 3$, the value of the function is 2. By definition of the function, on each interval, the function equals the greatest integer less than or equal to all the numbers in the interval. Zero, 1, and 2 are all integers that are less than or equal to the numbers in the interval [2, 3), but the greatest integer is 2.

Therefore, in general, when the interval is of the form $[k, k + 1)$, where k is an integer, the function value of greatest-integer function is k. So in the interval [5, 6), the function value is 5. The graph of the greatest integer function is similar to the graph shown below.



There are many examples where step functions apply to real-world situations. The price of items that are sold by weight can be presented as a cost per ounce (or pound) graphed against the weight. The average selling price of a corporation's stock can also be presented as a step function with a time period for the domain.

1.3.6 ALGEBRAIC AND TRANSCENDENTAL FUNCTIONS

An algebraic function is a function $f(x)$ which satisfies $p(x, f(x)) = 0$, where $p(x, y)$ is a polynomial in x and y with integer coefficients. Functions that can be constructed using only a finite number of elementary operations together with the inverses of functions capable of being so constructed are examples of algebraic functions. Nonalgebraic functions are called transcendental functions.

An algebraic equation in $n$variables is an polynomial equation of the form

$$f(x_1, x_2, \ldots x_n) = \sum_{c_1 \ldots c_n} c_{c_1 c_2 \ldots c_n} x_1^{c_1} x_2^{c_2} \cdots x_n^{c_n} = 0,$$

where the coefficients $c_{c_1 c_2 \ldots c_n}$ are integers (where the exponents c_i are nonnegative integers and the sum is finite).

A function which is not an algebraic function. In other words, a function which "transcends," i.e., cannot be expressed in terms of, algebra. Examples of transcendental functions include the exponential function, the logarithmic functions and the inverse functions of both.

The exponential function is the entire function defined by

$$\exp(z) \equiv e^z,$$

where e is the solution of the equation $\int_1^e dt/t$ so that $e = x = 2.718 \ldots$. $\exp(z)$ Is also the unique solution of the equation $df/dz = f(z)$ with $f(0) = 1$.
The exponential function is implemented in Mathematica as Exp[z].
It satisfies the identity

$$\exp(x + y) = \exp(x) \exp(y).$$

If $z \equiv x + i y$,

$$e^z = e^{x+iy} = e^x e^{iy} = e^x (\cos y + i \sin y).$$

The exponential function satisfies the identities

$$\begin{aligned} e^x &= \cosh x + \sinh x \\ &= \sec(\mathrm{gd}\, x) + \tan(\mathrm{gd}\, x) \\ &= \tan\left(\tfrac{1}{4}\pi + \tfrac{1}{2}\,\mathrm{gd}\, x\right) \end{aligned}$$

$$= \frac{1+\sin(\mathrm{gd}\,x)}{\cos(\mathrm{gd}\,x)},$$

where $^{\mathrm{gd}}x$ is the Gudermannian (Beyer 1987, p. 164; Zwillinger 1995, p. 485). The exponential function has Maclaurin series

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!},$$

and satisfies the limit

$$\exp(x) = \lim_{n\to\infty} \left(1+\frac{x}{n}\right)^n.$$

If

$$a+bi = e^{x+iy},$$

then

$$\begin{aligned} y &= \tan^{-1}\left(\frac{b}{a}\right) \\ x &= \ln\left\{b\csc\left|\tan^{-1}\left(\frac{b}{a}\right)\right|\right\} \\ &= \ln\left\{a\sec\left|\tan^{-1}\left(\frac{b}{a}\right)\right|\right\}. \end{aligned}$$

The exponential function has continued fraction

$$e^z = \cfrac{1}{1-\cfrac{z}{1+\cfrac{z}{2-\cfrac{z}{3+\cfrac{z}{2-\cfrac{z}{5+\cfrac{z}{2-\cdots}}}}}}}$$

(Wall 1948, p. 348).

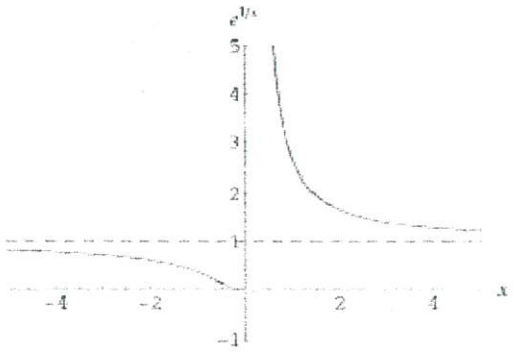


Fig 1.2

1.3.7 LOGARITHMIC FUNCTION

In mathematics, the logarithm of a number to a given base is the power or exponent to which the base must be raised in order to produce the number.

For example,

the logarithm of 1000 to the base 10 is 3, because 3 is how many 10s you must multiply to get 1000: thus 10 × 10 × 10 = 1000; the base 2 logarithm of 32 is 5 because 5 is how many 2s one must multiply to get 32: thus 2 × 2 × 2 × 2 × 2 = 32. In the language of exponents: 103 = 1000, so log101000 = 3, and 25 = 32, so log232 = 5.
The logarithm of x to the base b is written logb(x) or, if the base is implicit, as log(x). So, for a number x, a base b and an exponent y,

$$\text{if } x = b^y, \text{ then } y = \log_b(x).$$

An important feature of logarithms is that they reduce multiplication to addition, by the formula:

$$\log(xy) = \log x + \log y.$$

That is, the logarithm of the product of two numbers is the sum of the logarithms of those numbers. The use of logarithms to facilitate complicated calculations was a significant motivation in their original development.

1.4 INTEGRATION

Integration is an important concept in mathematics, specifically in the field of calculus and, more broadly, mathematical analysis. Given a function f of a real variable x and an interval $[a, b]$ of the real line, the **integral**

$$\int_a^b f(x)\,dx\,.$$

is defined informally to be the net signed area of the region in the xy-plane bounded by the graph of f, the x-axis, and the vertical lines $x = a$ and $x = b$.

The term "integral" may also refer to the notion of antiderivative, a function F whose derivative is the given function f. In this case it is called an **indefinite integral**, while the integrals discussed in this article are termed **definite integrals**. Some authors maintain a distinction between antiderivatives and indefinite integrals.

The principles of integration were formulated independently by Isaac Newton and Gottfried Leibniz in the late seventeenth century. Through the fundamental theorem of calculus, which they independently developed, integration is connected with differentiation: if f is a continuous real-valued function defined on a closed interval $[a, b]$, then, once an antiderivative F of f is known, the definite integral of f over that interval is given by

$$\int_a^b f(x)\,dx = F(b) - F(a)\,.$$

1.5 RULES FOR INTEGRATION

Although integration is the inverse of differentiation and we were given rules for differentiation, we are required to determine the answers in integration by trial and error. However, there are some rules to aid us in the determination of the answer.

In this section we will discuss four of these rules and how they are used to integrate standard elementary forms. In the rules we will let u and v denote a

differentiable function of a variable such as x. We will let C, n, and a denote constants.

Our proofs will involve searching for a function F(x) whose derivative is :

$f(x)\,dx$.

Rule 1. $\int du = u + C$

The integral of a differential of a function is the function plus a constant.

PROOF: If

$$\frac{d}{du}(u + C) = 1$$

then

$$d(u + C) = du$$

and

$$\int du = u + C$$

Example1

Evaluate the integral

$$\int dx$$

Solution: By Rule 1, we have

$$\int dx = x + C$$

Rule 2. $\int a\,du = a \int du = au + C$

A constant may be moved across the integral sign. NOTE: A variable may NOT be moved across the integral sign.

PROOF: If

$$\frac{d}{du}(au + C) = (a)\frac{d}{du}(u + C) = a$$

then

$$d(au + C) = a\, d(u + C) = a\, du$$

and

$$\int a\, du = a\int du = au + C$$

Example 2: Evaluate the integral

$$\int 4\, dx$$

Solution: By Rule 2,

$$\int 4\, dx = 4 \int dx$$

and by Rule 1,

$$\int dx = x + C$$

therefore,

$$\int 4\, dx = 4x + C$$

Rule 3. $\int u^n\, du = \frac{u^{n+1}}{n+1} + C$

The integral of u^n du may be obtained by adding 1 to the exponent and then dividing by this new exponent. NOTE: If n is minus 1, this rule is not valid and another method must be used.

PROOF.- If

$$d\left(\frac{u^{n+1}}{n+1}+C\right)=\frac{(n+1)u^n}{n+1}\,du$$

$$=u^n\,du$$

then

$$\int u^n\,du=\frac{u^{n+1}}{n+1}+C$$

Example 3: Evaluate the integral

$$\int x^3\,dx$$

Solution: By Rule 3,

$$\int x^3\,dx=\frac{x^{3+1}}{3+1}+C$$

$$=\frac{x^4}{4}+C$$

Example 4: Evaluate the integral

$$\int \frac{7}{x^3}\,dx$$

Solution: First write the integral

$$\int \frac{7}{x^3}\,dx$$

as

$\int 7x^{-3}\,dx$

Then, by Rule 2,

$7\int x^{-3}\,dx$

and by Rule 3,

$$7\int x^{-3}\,dx = 7\left(\frac{x^{-2}}{-2}\right) + C = -\frac{7}{2x^2} + C$$

Rule 4. $\int (du + dv + dw) = \int du + \int dv + \int dw$

$$= u + v + w + C$$

The integral of a sum is equal to the sum of the integrals.

PROOF: If

$d(u + v + w + C) = du + dv + dw$

then

$$\int (du + dv + dw) = (u + C_1) + (v + C_2) + (w + C_3)$$

such that

$\int (du + dv + dw) = u + v + w + C$

where

$C = C_1 + C_2 + C_3$

Example 5: Evaluate the integral

$\int (2x - 5x + 4)\,dx$

Solution: We will not combine 2x and -5x.

$$\int (2x - 5x + 4)\, dx$$

$$= \int 2x\, dx - \int 5x\, dx + \int 4\, dx$$

$$= 2\int x\, dx - 5\int x\, dx + 4\int dx$$

$$= \frac{2x^2}{2} + C_1 - \frac{5x^2}{2} + C_2 + 4x + C_3$$

$$= x^2 - \frac{5}{2}x^2 + 4x + C$$

where C is the sum of C_1, C_2, and C_3.

Example 6: Evaluate the integral

$$\int (x^{1/2} + x^{2/3})\, dx$$

Solution:

$$\int (x^{1/2} + x^{2/3})\, dx$$

$$= \int x^{1/2}\, dx + \int x^{2/3}\, dx$$

$$= \frac{x^{3/2}}{3/2} + C_1 + \frac{x^{5/3}}{5/3} + C_2$$

$$= \frac{2x^{3/2}}{3} + \frac{3x^{5/3}}{5} + C$$

Now we will discuss the evaluation of the constant of integration.

If we are to find the equation of a curve whose first derivative is 2 times the independent variable x, we may write

$$\frac{dy}{dx} = 2x$$

or

$dy = 2x\ dx$ (1)

We may obtain the desired equation for the curve by integrating the expression for d y ; that is, by integrating both sides of equation (1). If

$dy = 2x\ dx$

then,

$\int dy = \int 2x\ dx$

But, since

$\int dy = y$

and

$\int 2x\ dx = x^2 + C$

then

$y = x^2 + C$

We have obtained only a general equation of the curve because a different curve results for each value we assign to C. This is shown in figure 6 - 7 . If we specify that

x=0

And

y=6

we may obtain a specific value for C and hence a particular curve.

Suppose that

$y = x^2 + C$, $x = 0$, and $y = 6$

then,

$6 = 0^2 + C$

or

C=6

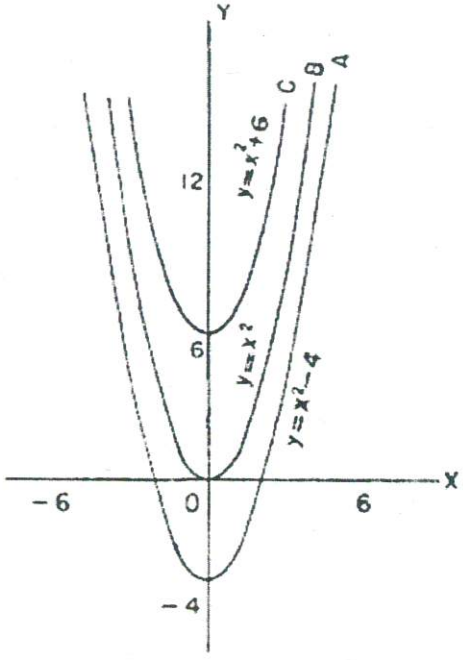


Figure 1.1-Family of curves.

By substituting the value 6 into the general equation, we find that the equation for the particular curve is

$y = x^2 + 6$

which is curve C of figure 6 - 7 .

The values for x and y will determine the value for C and also determine the particular curve of the family of curves.

In figure 6-7, curve A has a constant equal to - 4, curve B has a constant equal to 0, and curve C has a constant equal to 6.

Example 7: Find the equation of the curve if its first derivative is 6 times the independent variable, y equals 2, and x equals 0.

Solution: We may write

$$\frac{dy}{dx} = 6x$$

or

$$\int dy = \int 6x\, dx$$

such that,

$$y = 3x^2 + C$$

Solving for C when

x=0

and

y=2

We have

$$2 = 3(0^2) + C$$

or

C=2

so that the equation of the curve is

$y = 3x^2 + 2$

Activity 1

1. Consider the quadratic equation $2x^2-8x+c=0$. for what value of c, the equation has

   I. Real roots
   II. Equal roots
   III. Imaginary roots

2. Draw the graph of the following functions
   a) $Y = 3x-5$
   b) $Y = x^2$
   c) $C = \log 2x$
   d)

1.9 SUMMARY

The objective of this unit was to provide you exposure to functional relationship among decision variables. We started with the mathematical concept of function and defined terms such as constant, parameter, independent and dependent variable. Different types of function are discussed in depth with the description of their applications.

Attention is then directed to defining the concept of Integration. Further different rules of integration are discussed along with suitable examples.

1.10 FURTHER READINGS

- Alle, R.G.D (1974). Mathematical Analysis for Economists, Macmillan press and ELBS, London.

UNIT 2

BASIC CALCULUS
LIMITS, CONTINUITY AND DERIVATIVES

Objectives

After studying this unit, you should be able to understand:

- Concept of the term 'calculus'
- Concept of limit and slope which are fundamental to understanding of calculus
- Meaning of differentiation
- Derivatives and various ways to compute them

Structure

2.1 INTRODUCTION

Calculus (Latin, *calculus*, a small stone used for counting) is a branch of mathematics that includes the study of limits, derivatives, integrals, and infinite series, and constitutes a major part of modern university education. Historically, it has been referred to as "the calculus of infinitesimals", or "infinitesimal calculus". Most basically, calculus is the study of change, in the same way that geometry is the study of space.

Calculus has widespread applications in science, economics, and engineering and is used to solve problems for which algebra alone is insufficient. Calculus builds on algebra, trigonometry, and analytic geometry and includes two major branches, **differential calculus** and **integral calculus**, that are related by the fundamental theorem of calculus. In more advanced mathematics, calculus is usually called analysis and is defined as the study of functions.

2.2 LIMITS

The *limit of a function* $f(x)$ at some point x_0 exists and is equal to L if and only if every "small" interval about the limit L, say the interval $(L - \varepsilon, L + \varepsilon)$, means you can find a "small" interval about x_0, say the interval $(x_0 - \delta, x_0 + \delta)$, which has all values of $f(x)$ existing in the former "small" interval about the limit L, except possibly at x_0 itself.

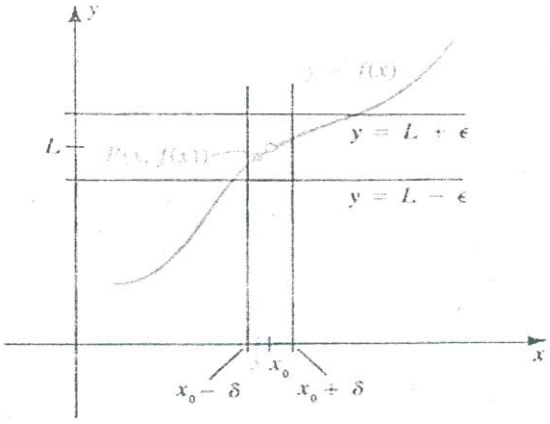


Figure 2. 1

This is a difficult concept to fully appreciate. However, you should be able to grasp the idea through several examples.

Examples:

1. Consider $f(x) = x^2 - x - 6$. Find the limit as x approaches 1. It is not hard to see from either the graph or from the way you have always evaluated this quadratic function that as x approaches 1, $f(x)$ approaches -6, since $f(1) = -6$.

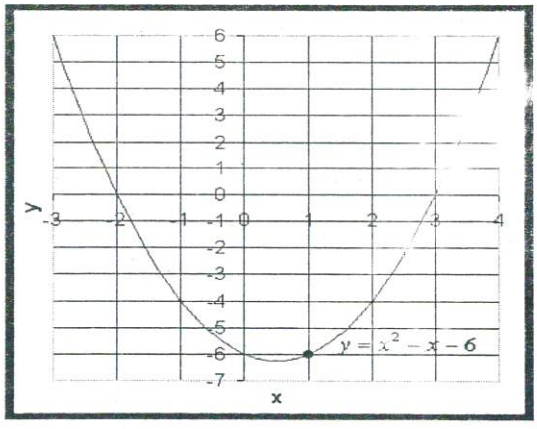


Figure 2. 2

Fact: Any polynomial, $p(x)$, has as its limit at some x_0, the value of $p(x_0)$.

2. Consider the rational function $r(x) = (x^2 - x - 6)/(x - 3)$. Find the limit as x approaches 1. If x is not 3, then this rational function reduces to $r(x) = x + 2$. So as x approaches 1, this function simply goes to 3.

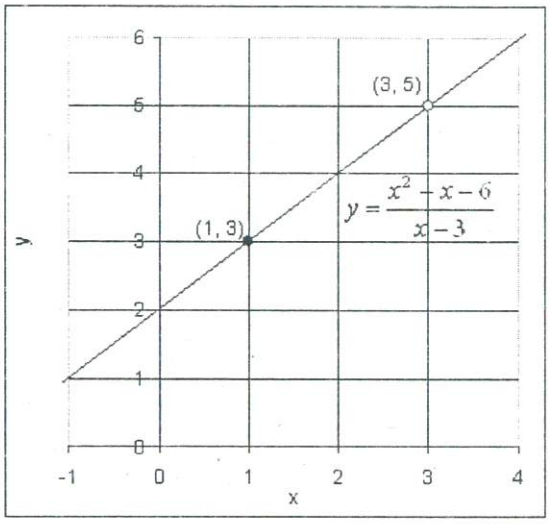


Figure 2. 3

Fact: Any rational function, $r(x) = p(x)/q(x)$, where $p(x)$ and $q(x)$ are polynomials with $q(x_0)$ not zero, then the limit exists with the limit being $r(x_0)$.

3. Consider the rational function in Example 2. Now f ind the limit as x approaches 3. Though $r(x)$ is not defined at $x_0 = 3$, we can see that arbitrarily "close" to 3, $r(x) = x + 2$. So as x approaches 3, this function simply goes to 5. Its limit exists though the function is not defined at $x_0 = 3$.

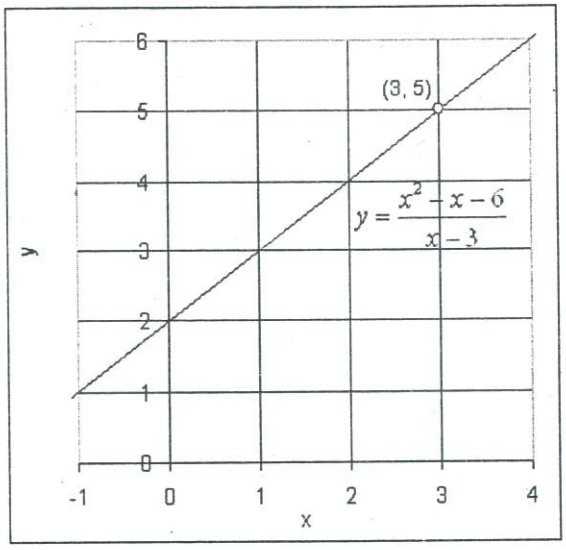


Figure 2. 4

4. Consider the rational function $f(x) = 1/x^2$. Find the limit as x approaches 0, if it exists. From our statement above on rational functions, this function has a limit for any value of x_0 where the denominator is not zero. However, at $x_0 = 0$, this function is undefined. Thus, the graph has a vertical asymptote at $x_0 = 0$. This means that no limit exists for $f(x)$ at $x_0 = 0$.

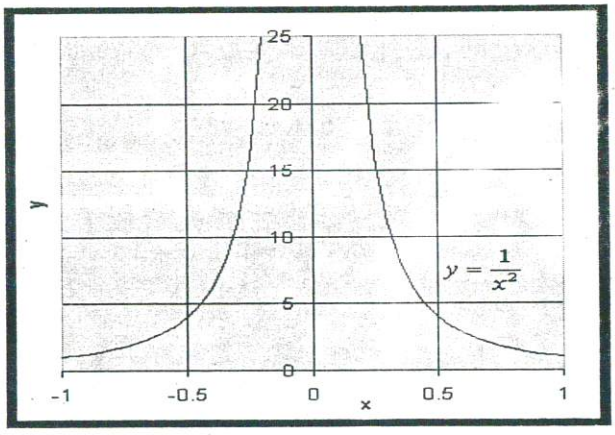


Figure 2. 5

Fact: Whenever you have a vertical asymptote at some x_0, then the limit fails to exist at that point.

2.3 CONTINUITY

Closely connected to the concept of a limit is that of continuity. Intuititvely, the idea of a continuous function is what you would expect. If you can draw the function without lifting your pencil, then the function is continuous. Most practical examples use functions that are continuous or at most have a few points of discontinuity.

Definition: A *function* $f(x)$ is continuous at a point x_0 if the limit exists at x_0 and is equal to $f(x_0)$.

The examples above should also help you appreciate this concept. In all of the cases except Example 3, the existence of a limit also corresponds to points of continuity. Example 3 is not continuous at $x_0 = 3$ though a limit exists here, as the function is not defined at 3. Examples 3 and 5 are discontinuous only at $x_0 = 3$, while Examples 4, 6 and 7 are discontinuous only at $x_0 = 0$. At all other points in the domains of these examples are continuous.

Example 5

Comparing Limits and Continuity

An example is provided to show the differences between limits and continuity. Below is a graph of a function, $f(x)$, that is defined on the interval [-2, 2], except at $x = 0$, where there is a vertical asymptote.

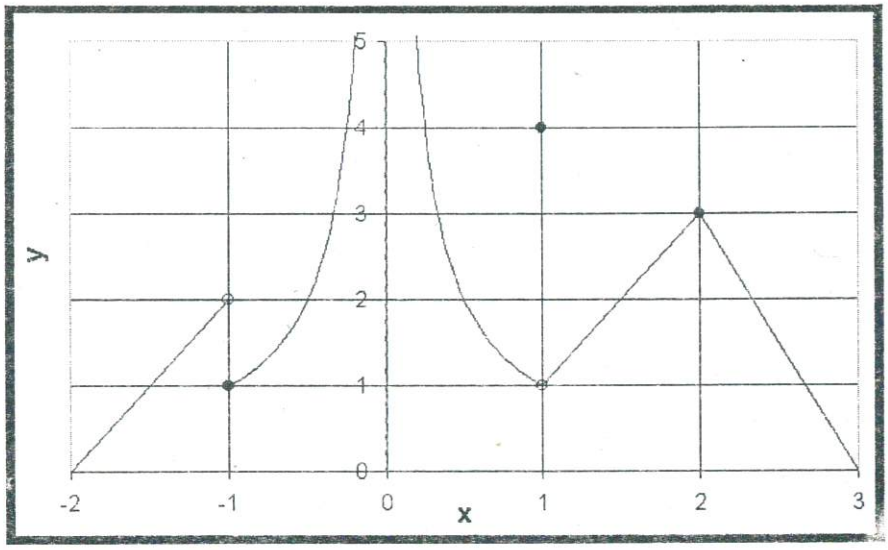


Figure 2. 6

It is clear that the difficulties with this function occur at integer values. At $x = -1$, the function has the value $f(-1) = 1$, but it is clear that the function is not continuous nor does a limit exist at this point. At $x = 0$, the function is not defined (not continuous nor has any limits) as there is a vertical asymptote. At $x = 1$, the function has the value $f(1) = 4$. The function is not continuous at $x = 1$, but the limit does exist with

$$\lim_{x \to 1} f(x) = 1$$

At $x = 2$, the function is continuous with $f(2) = 3$, which also means that the limit exists. At all non-integer values of x the function is continuous (hence its limit exists).

2.4 DERIVATIVES

INTRODUCTION

the **derivative** is a measure of how a function changes as its input changes. Loosely speaking, a derivative can be thought of as how much a quantity is

changing at a given point. For example, the derivative of the position (or distance) of a vehicle with respect to time is the instantaneous velocity (respectively, instantaneous speed) at which the vehicle is traveling. Conversely, the integral of the velocity over time is the vehicle's position.

The derivative of a function at a chosen input value describes the best linear approximation of the function near that input value. For a real-valued function of a single real variable, the derivative at a point equals the slope of the tangent line to the graph of the function at that point. In higher dimensions, the derivative of a function at a point is a linear transformation called the linearization.[1] A closely related notion is the differential of a function.

The process of finding a derivative is called **differentiation**. The fundamental theorem of calculus states that differentiation is the reverse process to integration.

2.4.1 DIFFERENTIATION AND THE DERIVATIVE

Differentiation is a method to compute the rate at which a dependent output y, changes with respect to the change in the independent input x. This rate of change is called the **derivative** of y with respect to x. In more precise language, the dependence of y upon x means that y is a function of x. If x and y are real numbers, and if the graph of y is plotted against x, the derivative measures the slope of this graph at each point. This functional relationship is often denoted $y = f(x)$, where f denotes the function.

The simplest case is when y is a linear function of x, meaning that the graph of y against x is a straight line. In this case, $y = f(x) = m\,x + c$, for real numbers m and c, and the slope m is given by

$$m = \frac{\text{change in } y}{\text{change in } x} = \frac{\Delta y}{\Delta x}$$

where the symbol Δ (the uppercase form of the Greek letter Delta) is an abbreviation for "change in." This formula is true because

$$y + \Delta y = f(x + \Delta x) = m\,(x + \Delta x) + c = m\,x + c + m\,\Delta x = y + m\Delta x.$$

It follows that $\Delta y = m\,\Delta x$.

This gives an exact value for the slope of a straight line. If the function f is not linear (i.e. its graph is not a straight line), however, then the change in y divided by the change in x varies: differentiation is a method to find an exact value for this rate of change at any given value of x.

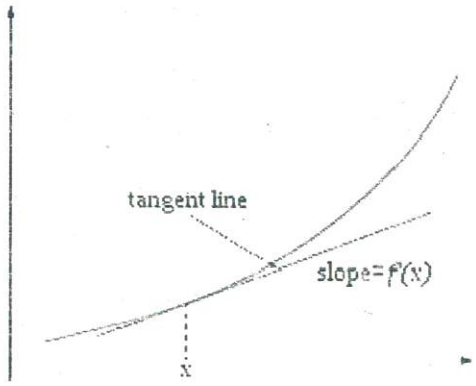


Figure 2.7 The tangent line at $(x, f(x))$

The idea, illustrated by Figures 1-3, is to compute the rate of change as the limiting value of the ratio of the differences Δy / Δx as Δx becomes infinitely small.

In Leibniz's notation, such an infinitesimal change in x is denoted by dx, and the derivative of y with respect to x is written

$$\frac{dy}{dx}$$

suggesting the ratio of two infinitesimal quantities. (The above expression is read as "the derivative of y with respect to x", "d y by d x", or "d y over d x". The oral form "d y d x" is often used conversationally, although it may lead to confusion.)

The most common approach[2] to turn this intuitive idea into a precise definition uses limits, but there are other methods, such as non-standard analysis.[3]

2.4.2 The derivative as a function

Let f be a function that has a derivative at every point a in the domain of f. Because every point a has a derivative, there is a function which sends the point a to the derivative of f at a. This function is written $f'(x)$ and is called the *derivative function* or the *derivative* of f. The derivative of f collects all the derivatives of f at all the points in the domain of f.

Sometimes f has a derivative at most, but not all, points of its domain. The function whose value at a equals $f'(a)$ whenever $f'(a)$ is defined and is undefined elsewhere is also called the derivative of f. It is still a function, but its domain is strictly smaller than the domain of f.

Using this idea, differentiation becomes a function of functions: The derivative is an operator whose domain is the set of all functions which have derivatives at every point of their domain and whose range is a set of functions. If we denote this operator by D, then $D(f)$ is the function $f'(x)$. Since $D(f)$ is a function, it can be evaluated at a point a. By the definition of the derivative function, $D(f)(a) = f'(a)$.

For comparison, consider the doubling function $f(x) = 2x$; f is a real-valued function of a real number, meaning that it takes numbers as inputs and has numbers as outputs:

$$1 \mapsto 2,$$
$$2 \mapsto 4,$$
$$3 \mapsto 6.$$

The operator D, however, is not defined on individual numbers. It is only defined on functions:

$$D(x \mapsto 1) = (x \mapsto 0),$$
$$D(x \mapsto x) = (x \mapsto 1),$$
$$D(x \mapsto x^2) = (x \mapsto 2 \cdot x).$$

Because the output of D is a function, the output of D can be evaluated at a point. For instance, when D is applied to the squaring function,

$$x \mapsto x^2,$$

D outputs the doubling function,

$$x \mapsto 2x,$$

which we named $f(x)$. This output function can then be evaluated to get $f(1) = 2$, $f(2) = 4$, and so on.

The derivative of a function can, in principle, be computed from the definition by considering the difference quotient, and computing its limit. For some examples, see Derivative (examples). In practice, once the derivatives of a few simple functions are known, the derivatives of other functions are more easily computed using *rules* for obtaining derivatives of more complicated functions from simpler ones.

Computation of derivatives of different functions is described as following:

2.4.3 Derivatives of elementary functions

Most derivative computations eventually require taking the derivative of some common functions. The following incomplete list gives some of the most frequently used functions of a single real variable and their derivatives.

If,

$$f(x) = x^r,$$

where r is any real number, then

$$f'(x) = rx^{r-1},$$

wherever this function is defined. For example, if $r = 1/2$, then

$$f'(x) = \frac{1}{2}x^{-\frac{1}{2}}$$

and the function is defined only for non-negative x. When $r = 0$, this rule recovers the constant rule.

- *Exponential and logarithmic functions*:

$$\frac{d}{dx}a^x = \ln(a)a^x$$

$$\frac{d}{dx}\ln(x) = \frac{1}{x}, \qquad x > 0$$
$$\frac{d}{dx}\log_a(x) = \frac{1}{x\ln(a)}$$

- *Trigonometric functions*:

$$\frac{d}{dx}\sin(x) = \cos(x).$$
$$\frac{d}{dx}\cos(x) = -\sin(x).$$
$$\frac{d}{dx}\tan(x) = \sec^2(x).$$

- *Inverse trigonometric functions*:

$$\frac{d}{dx}\arcsin(x) = \frac{1}{\sqrt{1-x^2}}.$$
$$\frac{d}{dx}\arccos(x) = -\frac{1}{\sqrt{1-x^2}}.$$
$$\frac{d}{dx}\arctan(x) = \frac{1}{1+x^2}.$$

2.4.4 Rules for finding the derivative

In many cases, complicated limit calculations by direct application of Newton's difference quotient can be avoided using differentiation rules. Some of the most basic rules are the following.

- *Constant rule*: if $f(x)$ is constant, then

  $$f' = 0$$

- *Sum rule*:

  $(af + bg)' = af' + bg'$ for all functions f and g and all real numbers a and b.

- *Product rule*:

  $(fg)' = f'g + fg'$ for all functions f and g.

- *Quotient rule*:

  $\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$ for all functions f and g where $g \neq 0$.

- *Chain rule*: If $f(x) = h(g(x))$, then

  $$f'(x) = h'(g(x)) \cdot g'(x).$$

Example 6

Computation

The derivative of

$$f(x) = x^4 + \sin(x^2) - \ln(x)e^x + 7$$

is

$$f'(x) = 4x^{(4-1)} + \frac{d(x^2)}{dx}\cos(x^2) - \frac{d(\ln x)}{dx}e^x - \ln x\frac{d(e^x)}{dx} + 0$$

$$= 4x^{(4-1)} + 2(x^{(2-1)})\cos(x^2) - \frac{1}{x}e^x - \ln x(e^x) + 0$$

Here the second term was computed using the chain rule and third using the product rule. The known derivatives of the elementary functions x^2, x^4, $\sin(x)$, $\ln(x)$ and $\exp(x) = e^x$, as well as the constant 7, were also used.

2.4.5 Derivatives of Inverse Trigonometric Functions

The following are the formulas for the derivatives of the inverse trigonometric functions:

$$\frac{d(\sin^{-1}u)}{dx} = \frac{1}{\sqrt{1-u^2}}\frac{du}{dx}$$

$$\frac{d(\cos^{-1}u)}{dx} = \frac{-1}{\sqrt{1-u^2}}\frac{du}{dx}$$

$$\frac{d(\tan^{-1}u)}{dx} = \frac{1}{1+u^2}\frac{du}{dx}$$

2.4.6 Quotient Rule for Derivatives

Let f and g be differentiable at x with $g(x) \neq 0$. Then f/g is differentiable at x and

$$\left[\frac{f(x)}{g(x)}\right]' = \frac{g(x)f'(x) - f(x)g'(x)}{[g(x)]^2}.$$

Example 7

If,

$$f(x) = \frac{2x+1}{x-3}$$

Then,

$$\begin{aligned} f'(x) &= \frac{(x-3)\frac{d}{dx}[2x+1] - (2x+1)\frac{d}{dx}[x-3]}{[x-3]^2} \\ &= \frac{(x-3)(2) - (2x+1)(1)}{(x-3)^2} \\ &= -\frac{7}{(x-3)^2}. \end{aligned}$$

2.4.7 Derivative of the Exponential Function

The importance of exponential functions in mathematics and the sciences stems mainly from properties of their derivatives. In particular,

That is, e^x is its own derivative and hence is a simple example of a pfaffian function. Functions of the form Ke^x for constant K are the only functions with that property. (This follows from the Picard-Lindelöf theorem, with $y(t) = e^t$, $y(0)=K$, and $f(t,y(t)) = y(t)$.) Other ways of saying the same thing include:

- The slope of the graph at any point is the height of the function at that point.
- The rate of increase of the function at x is equal to the value of the function at x.
- The function solves the differential equation $y' = y$.
- exp is a fixed point of derivative as a functional.

2.4.8 Formulas for Derivatives of Exponential Functions

If u is a function of x, we can obtain the derivative of an expression in the form e^u:

$$\frac{d(e^u)}{dx} = e^u \frac{du}{dx}$$

If we have an exponential function with some base b, we have the following derivative:

$$\frac{d(b^u)}{dx} = b^u \ln b \frac{du}{dx}$$

ACTIVITY 2

1. Find the derivative of $y = 10^{3x}$.
2. Suppose that $2x^2 + 6xy + y^2 = c$ for some constant c. Find dy/dx.
3. Suppose that the functions f and g are differentiable and $g(f(x)) = x$ for all values of x. Use implicit differentiation to find an expression for the derivative $f'(x)$ in terms of the derivative of g.

4. Find A which makes the function

$$f(x) = \begin{cases} x^2 - 2 & \text{if } x < 1 \\ Ax - 4 & \text{if } 1 \leq x \end{cases}$$

continuous at x=1.

5. Find the derivative of $y = \cos^{-1} 5x$.

2.5 SUMMARY

The objective of this unit was to provide you with some exposure to differential calculus. Differential calculus is useful to solve optimization problems in which the aim is either to maximize or minimize a given objective function. Applications of the derivative in both micro economics theory (cost, revenue, elasticity) and macro economic theory (income, consumption, savings) are good examples of its applications.

The unit begins with a discussion on the limit and continuity and then attention is directed to defining the slope of a linear functio and puoceeds with a discussion that extends this to include the slope of non linear function. Thes is followed by the difinition of the term derivative and rules for obraining the derivatives of the more commonly encountered functional forms.

2.6 FURTHER READINGS

Budnicks, F.S. 1983. Applied Mathematics for Business, Economics, and Social Sciences, McGraw Hill: New York.

Gulati, B.R. 1978. College Mathematics with Business Applications to Business and Social Sciences; Harper & Row: New York.

Hughes, A.J. 1983. Applied Mathematics for Business, Economics and the Social Sciences, Irwin: Homewood.

Weber, J.E. 1982. Mathematical Analysis: Business and Economics Applications, Harper & Row: New York.

UNIT 3

CONCEPTS OF MATRICES AND DETERMINANTS

Objectives

After studying this unit, you should know the:

- Basic concepts of the matrix
- Methods of representing large quantities of data in matrix form
- Various operations concerning matrices
- The solution method of simultaneous linear equations
- Concept and properties of determinants

Structure

3.1 INTRODUCTION

A matrix is a rectangular array of ordered numbers. The term ordered implies that the position of each number is significant and must be determined carefully to represent the information contained in the problem.

A matrix is defined as an ordered rectangular array of numbers. They can be used to represent systems of linear equations, as will be explained below

Here are a couple of examples of different types of matrices:

Symmetric

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & -5 \\ 3 & -5 & 6 \end{bmatrix}$$

Diagonal

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

Upper Triangular

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 7 & -5 \\ 0 & 0 & -4 \end{bmatrix}$$

Lower Triangular

$$\begin{bmatrix} 1 & 0 & 0 \\ -4 & 7 & 0 \\ 12 & 5 & 3 \end{bmatrix}$$

Zero

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Identity

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

And a fully expanded mxn matrix A, would look like this:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m1} & \cdots & a_{mn} \end{pmatrix}$$ or in a more compact form: $A = (a_{ij})$

3.2 MATRIX ADDITION AND SUBTRACTION

DEFINITION: Two matrices A and B can be added or subtracted if and only if their dimensions are the same (i.e. both matrices have the identical amount of rows and columns. Take:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 2 \end{pmatrix} \quad and \quad B = \begin{pmatrix} 2 & 1 & 2 \\ 1 & 0 & 3 \end{pmatrix}$$

Addition

If A and B above are matrices of the same type then the sum is found by adding the corresponding elements $a_{ij}+b_{ij}$

Here is an example of adding A and B together

$$A+B=\begin{pmatrix}1 & 2 & 3\\ 1 & 0 & 2\end{pmatrix}+\begin{pmatrix}2 & 1 & 2\\ 1 & 0 & 3\end{pmatrix}=\begin{pmatrix}3 & 3 & 5\\ 2 & 0 & 5\end{pmatrix}$$

Subtraction

If A and B are matrices of the same type then the subtraction is found by subtracting the corresponding elements $a_{ij}-b_{ij}$

Here is an example of subtracting matrices

$$A-B=\begin{pmatrix}1 & 2 & 3\\ 1 & 0 & 2\end{pmatrix}-\begin{pmatrix}2 & 1 & 2\\ 1 & 0 & 3\end{pmatrix}=\begin{pmatrix}-1 & 1 & 1\\ 0 & 0 & -1\end{pmatrix}$$

3.3 MATRIX MULTIPLICATION

DEFINITION: When the number of columns of the first matrix is the same as the number of rows in the second matrix then matrix multiplication can be performed.

Here is an example of matrix multiplication for two 2x2 matrices

$$\begin{pmatrix}a & b\\ c & d\end{pmatrix}\begin{pmatrix}e & f\\ g & h\end{pmatrix}=\begin{pmatrix}(ae+bg) & (af+bh)\\ (ce+dg) & (cf+dh)\end{pmatrix}$$

Here is an example of matrices multiplication for a 3x3 matrix

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}\begin{pmatrix} j & k & l \\ m & n & o \\ p & q & r \end{pmatrix} = \begin{pmatrix} (aj+bm+cp) & (ak+bn+cq) & (al+bo+cr) \\ (dj+em+fp) & (dk+en+fq) & (dl+eo+fr) \\ (gj+hm+ip) & (gk+hn+iq) & (gl+ho+ir) \end{pmatrix}$$

Now lets look at the nxn matrix case, Where A has dimensions mxn, B has dimensions nxp. Then the product of A and B is the matrix C, which has dimensions mxp. The ij^{th} element of matrix C is found by multiplying the entries of the i^{th} row of A with the corresponding entries in the j^{th} column of B and summing the n terms. The elements of C are:

$$c_{11} = a_{11}b_{11} + a_{12}b_{21} + \ldots + a_{1n}b_{n1} = \sum_{j=1}^{n} a_{1j}b_{j1}$$

$$c_{12} = a_{11}b_{12} + a_{12}b_{22} + \ldots + a_{1n}b_{n2}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$c_{mp} = a_{m1}b_{1p} + a_{m2}b_{2p} + \ldots + a_{mn}b_{np}$$

Note: That AxB is not the same as BxA

3.4 THE RANK OF MATRICES

The **column rank** of a matrix A is the maximal number of linearly independent **columns** of A. Likewise, the **row rank** is the maximal number of linearly independent **rows** of A.

Properties of rank of matrix

We assume that A is an m-by-n matrix over either the real numbers or the complex numbers, and we define the linear map f by $f(\mathbf{x}) = A\mathbf{x}$ as above.

- only a zero matrix has rank zero.
- $\operatorname{rank} A \leq \min(m,n)$

- f is injective if and only if A has rank n (in this case, we say that A has *full column rank*).
- f is surjective if and only if A has rank m (in this case, we say that A has *full row rank*).
- In the case of a square matrix A (i.e., $m = n$), then A is invertible if and only if A has rank n (that is, A has full rank).
- If B is any n-by-k matrix, then

$$\operatorname{rank}(AB) \leq \min(\operatorname{rank} A, \operatorname{rank} B)$$

As an example of the "<" case, consider the product

$$\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

Both factors have rank 1, but the product has rank 0.

- If B is an n-by-k matrix with rank n, then

$$\operatorname{rank}(AB) = \operatorname{rank}(A)$$

- If C is an l-by-m matrix with rank m, then

$$\operatorname{rank}(CA) = \operatorname{rank}(A)$$

- The rank of A is equal to r if and only if there exists an invertible m-by-m matrix X and an invertible n-by-n matrix Y such that

$$XAY = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}$$

where I_r denotes the r-by-r identity matrix.

- Sylvester's rank inequality: If A and B are any n-by-n matrices, then

$$\operatorname{rank}(A) + \operatorname{rank}(B) - n \leq \operatorname{rank}(AB)$$

- Subadditivity: $\operatorname{rank}(A+B) \leq \operatorname{rank}(A)+\operatorname{rank}(B)$ when A and B are of the same dimension. As a consequence, a rank-k matrix can be written as the sum of k rank-1 matrices, but not fewer.
- The rank of a matrix plus the nullity of the matrix equals the number of columns of the matrix (this is the "rank theorem" or the "rank-nullity theorem").
- The rank of a matrix and the rank of its corresponding Gram matrix are equal

  $\operatorname{rank}(A^TA) = \operatorname{rank}(AA^T) = \operatorname{rank}(A)$

  This can be shown by proving equality of their null spaces. Null space of the Gram matrix is given by vectors x for which $A^TAx = 0$. If this condition is fulfilled, also holds $0 = x^TA^TAx = |Ax|^2$. This proof was adapted from.[

Computation

The easiest way to compute the rank of a matrix A is given by the Gauss elimination method. The row-echelon form of A produced by the Gauss algorithm has the same rank as A, and its rank can be read off as the number of non-zero rows.

Consider for example the 4-by-4 matrix

$$A = \begin{bmatrix} 2 & 4 & 1 & 3 \\ -1 & -2 & 1 & 0 \\ 0 & 0 & 2 & 2 \\ 3 & 6 & 2 & 5 \end{bmatrix}$$

We see that the second column is twice the first column, and that the fourth column equals the sum of the first and the third. The first and the third columns are linearly independent, so the rank of A is two. This can be confirmed with the Gauss algorithm. It produces the following row echelon form of A:

$$A = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

which has two non-zero rows.

Example 1

$$A=\begin{bmatrix}1 & 2 & 1\\ 2 & 3 & 1\\ 1 & 1 & 2\end{bmatrix}.$$

1. Determine the row-rank of

Solution: To determine the row-rank of A we proceed as follows.

1. $\begin{bmatrix}1 & 2 & 1\\ 2 & 3 & 1\\ 1 & 1 & 2\end{bmatrix} \xrightarrow{R_{21}(-2),R_{31}(-1)} \begin{bmatrix}1 & 2 & 1\\ 0 & -1 & 1\\ 0 & -1 & 1\end{bmatrix}.$

2. $\begin{bmatrix}1 & 2 & 1\\ 0 & -1 & -1\\ 0 & -1 & 1\end{bmatrix} \xrightarrow{R_2(-1),R_{32}(1)} \begin{bmatrix}1 & 2 & 1\\ 0 & 1 & 1\\ 0 & 0 & 2\end{bmatrix}.$

3. $\begin{bmatrix}1 & 2 & 1\\ 0 & 1 & 1\\ 0 & 0 & 2\end{bmatrix} \xrightarrow{R_3(1/2),R_{12}(-2)} \begin{bmatrix}1 & 0 & -1\\ 0 & 1 & 1\\ 0 & 0 & 1\end{bmatrix}.$

4. $\begin{bmatrix}1 & 0 & 1\\ 0 & 1 & 1\\ 0 & 0 & 1\end{bmatrix} \xrightarrow{R_{23}(-1),R_{13}(1)} \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$

The last matrix in Step 1d is the row reduced form of Awhich has 3non-zero rows. Thus, row rank (A) =3. This result can also be easily deduced from the last matrix in Step 1b.

$$A=\begin{bmatrix}1 & 2 & 1\\ 2 & 3 & 1\\ 1 & 1 & 0\end{bmatrix}.$$

2. Determine the row-rank

Solution: Here we have

1. $\begin{bmatrix}1 & 2 & 1\\ 2 & 3 & 1\\ 1 & 1 & 0\end{bmatrix} \xrightarrow{R_{21}(-2),R_{31}(-1)} \begin{bmatrix}1 & 2 & 1\\ 0 & -1 & -1\\ 0 & 1 & 1\end{bmatrix}.$

2. $\begin{bmatrix} 1 & 2 & 1 \\ 0 & -1 & -1 \\ 0 & -1 & -1 \end{bmatrix} \xrightarrow{R_2(-1),\, R_{32}(1)} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$.

$\text{row-rank}(A) = 2$.

3.5 TRANSPOSE OF MATRICES

DEFINITION: The transpose of a matrix is found by exchanging rows for columns i.e. Matrix A = (a_{ij}) and the transpose of A is: $A^T=(a_{ji})$ where j is the column number and i is the row number of matrix A.

For example, The transpose of a matrix would be:

$$A=\begin{pmatrix} 5 & 2 & 3 \\ 4 & 7 & 1 \\ 8 & 5 & 9 \end{pmatrix} \quad A^T=\begin{pmatrix} 5 & 4 & 8 \\ 2 & 7 & 5 \\ 3 & 1 & 9 \end{pmatrix}$$

In the case of a square matrix (m=n), the transpose can be used to check if a matrix is symmetric. For a symmetric matrix $A = A^T$

$$A=\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}=A^T=\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}=A$$

3.6 SOLVING SYSTEMS OF EQUATIONS USING MATRICES

DEFINITION: A system of linear equations is a set of equations with n equations and n unknowns, is of the form of

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\
&\vdots \\
a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n
\end{aligned}
$$

The unknowns are denoted by $x_1, x_2, \ldots x_n$ and the coefficients (a's and b's above) are assumed to be given. In matrix form the system of equations above can be written as:

$$
\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}
$$

A simplified way of writing above is like this ; Ax = b
After looking at this we will now look at two methods used to solve matrices these are

- Inverse Matrix Method
- Cramer's Rule

3.6.1 Inverse Matrix Method

DEFINITION: The inverse matrix method uses the inverse of a matrix to help solve a system of equations, such like the above Ax = b. By pre-multiplying both sides of this equation by A^{-1} gives:

$$A^{-1}(Ax) = A^{-1}b$$
$$(A^{-1}A)x = A^{-1}b$$

or alternatively this gives

$$x = A^{-1}b$$

So by calculating the inverse of the matrix and multiplying this by the vector b we can find the solution to the system of equations directly. And from earlier we found that the inverse is given by

$$A^{-1} = \frac{adj(A)}{\det(A)}$$

From the above it is clear that the existence of a solution depends on the value of the determinant of A. There are three cases:

1. If the det(A) does not equal zero then solutions exist using $x = A^{-1}b$
2. If the det(A) is zero and b=0 then the solution will be not be unique or does not exist.
3. If the det(A) is zero and b=0 then the solution can be x = 0 but as in 2. is not unique or does not exist.

Looking at two equations we might have that

$$ax + by = c$$
$$dx + ey = f$$

Written in matrix form would look like

$$\begin{pmatrix} a & b \\ d & e \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c \\ f \end{pmatrix}$$

and by rearranging we would get that the solution would look like

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ d & e \end{pmatrix}^{-1}\begin{pmatrix} c \\ f \end{pmatrix}$$

Similarly for three simultaneous equations we would have:

$$a_{11}x + a_{12}y + a_{13}z = b_1$$
$$a_{21}x + a_{22}y + a_{23}z = b_2$$
$$a_{31}x + a_{32}y + a_{33}z = b_3$$

Written in matrix form would look like

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

and by rearranging we would get that the solution would look like

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}^{-1} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

The inverse of a 2x2 matrix

Take for example a arbitury 2x2 Matrix A whose determinant (ad-bc) is not equal to **zero**

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

where a,b,c,d are numbers, The inverse is:

$$A^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

The inverse of a nxn matrix

The inverse of a general nxn matrix A can be found by using the following equation:

$$A^{-1} = \frac{adj(A)}{\det(A)}$$

Where the adj(A) denotes the adjoint (or adjugate) of a matrix. It can be calculated by the following method

- Given the nxn matrix A, define
  $B = (b_{ij})$
  to be the matrix whose coefficients are found by taking the determinant of the (n-1) x (n-1) matrix obtained by deleting the i^{th} row and j^{th} column of A. The terms of B (i.e. B = b_{ij}) are known as the cofactors of A.
- And define the matrix C, where
  $c_{ij} = (-1)^{i+j} b_{ij}$
- The transpose of C (i.e C^T) is called the adjoint of matrix A.

Lastly to find the inverse of A divide the matrix C^T by the determinant of A to give its inverse.

3.6.2 Cramer's Rule to solve simultaneous equations

Cramer's rule is a theorem in linear algebra, which gives the solution of a system of linear equations or corresponding square matrices in terms of determinants Cramer's rules uses a method of determinants to solve systems of equations. Starting with equation below,

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\
&\vdots \\
a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n
\end{aligned}
$$

The first term x_1 above can be found by replacing the first column of A by $(b_1 \quad b_2 \quad \ldots \quad b_n)^T$. Doing this we obtain:

$$
x_1 = \frac{1}{|A|}\begin{vmatrix} b_1 & a_{12} & a_{13} & \ldots & a_{1n} \\ b_2 & a_{22} & a_{23} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_n & a_{n2} & a_{n3} & \ldots & a_{nn} \end{vmatrix}
$$

Similarly for the general case for solving x_r we replace the r^{th} column of A by $(b_1 \quad b_2 \quad \ldots \quad b_n)^T$ and expand the determinant.

This method of using determinants can be applied to solve systems of linear equations. We will illustrate this for solving two simultaneous equations in x and y and three equations with 3 unknowns x, y and z.

Two simultaneous equations in x and y

$$
\begin{aligned}
ax + by &= p \\
cx + dy &= q
\end{aligned}
$$

To solve use the following:

$$
x = \frac{\text{Det}\left(\begin{bmatrix} p & b \\ q & d \end{bmatrix}\right)}{\text{Det}\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right)} \quad \text{and} \quad y = \frac{\text{Det}\left(\begin{bmatrix} a & p \\ c & q \end{bmatrix}\right)}{\text{Det}\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right)}
$$

or
simplified:

$$x = \frac{pd - bq}{ad - bc} \quad \text{and} \quad y = \frac{aq - cp}{ad - bc}$$

Explicit formulas for Cramer's rule

Consider the linear system

$$\begin{aligned} ax + by &= e \\ cx + dy &= f \end{aligned}$$

which in matrix format is

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} e \\ f \end{bmatrix}.$$

Then, x and y can be found with Cramer's rule as:

$$x = \frac{\begin{vmatrix} e & b \\ f & d \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}} = \frac{ed - bf}{ad - bc}$$

and

$$y = \frac{\begin{vmatrix} a & e \\ c & f \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}} = \frac{af - ec}{ad - bc}.$$

The rules for 3×3 are similar. Given:

$$ax + by + cz = j,$$
$$dx + ey + fz = k,$$
$$gx + hy + iz = l,$$

which in matrix format is

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} j \\ k \\ l \end{bmatrix}$$

x, y and z can be found as follows:

$$x = \frac{\begin{vmatrix} j & b & c \\ k & e & f \\ l & h & i \end{vmatrix}}{\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}}, \quad y = \frac{\begin{vmatrix} a & j & c \\ d & k & f \\ g & l & i \end{vmatrix}}{\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}}, \text{ and } z = \frac{\begin{vmatrix} a & b & j \\ d & e & k \\ g & h & l \end{vmatrix}}{\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}}.$$

3.7 THE DETERMINANT OF A MATRIX

DEFINITION: Determinants play an important role in finding the inverse of a matrix and also in solving systems of linear equations. In the following we assume we have a square matrix (m=n). The determinant of a matrix A will be denoted by det(A) or |A|. Firstly the determinant of a 2x2 and 3x3 matrix will be introduced then the nxn case will be shown

For a 2×2 matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ the number ad - bc is called the **determinant** of A. We write it as det(A), or $|A|$ or $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$.

More generally, associated with any $n \times n$ matrix $A = (a_{ij})$ we have a number, called the **determinant** of A, denoted as above.

The definition of this number is rather complicated. I have given it for 2×2 matrices. The definition for 3×3 matrices is given in terms of 2×2 matrices as follows:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}.$$

For an $n \times n$ matrix A the determinant of the $(n - 1)\times(n - 1)$ matrix obtained by deleting the i^{th} row and j^{th} column of A is called the (i, j)-**minor** of A. We denote it by M_{ij}.

We can now write the above definition of the determinant of a 3×3 matrix as

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}M_{11} - a_{12}M_{12} + a_{13}M_{13},$$

which looks a bit more tidy.

I can now give you the definition of the determinant of an $n \times n$ matrix A. It is just the same as the above, expressing det(A) in terms of the minors of the top row of A.

$\det(A) = a_{11}M_{11} - a_{12}M_{12} + a_{13}M_{13} - \cdots \pm a_{1n}M_{1n}$.
Note that the signs are alternating + - + - + - etc.

That's the definition. We don't often work out determinants in this way if we can help it. It gets to be very hard work if n is much bigger than 4. It can be shown that, using the above method, it takes in all about $(e - 1)n!$ multiplications to work out an $n \times n$ determinant. The number of multiplications needed to evaluate a 20×20 determinant is 4, 180, 411, 311, 071, 440, 000. If a computer can do a million multiplications per second, and we don't count the time for the additions etc., then the evaluation of a 20×20 determinant will take about 130, 000 years by

this method. This is not practical! There are better methods which will reduce the time to a matter of seconds. These methods are consequences of the basic properties of determinants that I will now explain.

Properties of Determinants

Here are some rules:

1. Interchanging two rows of A just changes the sign of $\det(A)$.
2. Interchanging two columns of A just changes the sign of $\det(A)$.
3. If A has a complete row, or column, of zeroes then $\det(A) = 0$.
4. $\det(A) = \det(A^T)$.
5. To any row of A we can add any multiple of any other row without changing $\det(A)$.
6. To any column of A we can add any multiple of any other column without changing $\det(A)$.
7. A common factor of all the elements of a row of A can be `taken outside the determinant', in the following sense:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ p.a_{21} & p.a_{22} & p.a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = p \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}.$$

8. The same applies to columns.
9. If all the elements of A below (or above) the diagonal are zero then the determinant is equal to the product of the diagonal elements. In particular, the determinant of a diagonal matrix is equal to the product of the diagonal elements. For example

$$\begin{vmatrix} a & b & c & d \\ 0 & p & q & r \\ 0 & 0 & s & t \\ 0 & 0 & 0 & u \end{vmatrix} = a.p.s.u.$$

10. The determinant of a product is the product of the determinants. In symbols,

$$\det(AB) = \det(A)\det(B).$$

These give us ways to manipulate a determinant into a more manageable form for calculation.

Example 2: Show that

$$\Delta = \begin{vmatrix} 1 & 0 & 1 & 1 \\ 2 & 1 & 2 & 1 \\ 3 & 2 & 1 & 2 \\ 1 & 1 & 2 & 1 \end{vmatrix} = 3.$$

Solution: We aim to produce as many zeros as possible and, ideally to produce a matrix in which all the elements below (or above) the diagonal are zero.

$$\begin{vmatrix} 1 & 0 & 1 & 1 \\ 2 & 1 & 2 & 1 \\ 3 & 2 & 1 & 2 \\ 1 & 1 & 2 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & -1 \\ 3 & 2 & -2 & -1 \\ 1 & 1 & 1 & 0 \end{vmatrix}, \qquad [C_3' = C_3 - C_1];\ [C_4' = C_4 - C_1]$$

Here we have subtracted column 1 from column 3 and from column 4.

$$\Delta = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & -1 \\ 1 & 1 & -2 & 0 \\ 1 & 1 & 1 & 0 \end{vmatrix}, \qquad [R_3' = R_3 - R_2]$$

Here we have taken row 2 from row 3. Now switch over rows 2 and 4, which

changes the sign:

$$\Delta = -\begin{vmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & -2 & 0 \\ 2 & 1 & 0 & -1 \end{vmatrix} \qquad [R_2' = R_4]; [R_4' = R_2]$$

Finally, subtract column 2 from column 3 to get:

$$\Delta = -\begin{vmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & -3 & 0 \\ 2 & 1 & -1 & -1 \end{vmatrix} \qquad [C_3' = C_3 - C_1]$$

Now all the elements above the diagonal are zero, so the value of the determinant is the product of the diagonal elements. So

$\det(A) = -(1\times1\times -3\times -1) = -3.$

Example 3 Prove that $\begin{vmatrix} 1 & x & x^2 \\ 1 & y & y^2 \\ 1 & z & z^2 \end{vmatrix} = (x - y)(y - z)(z - x).$

Solution 3: Before we start, remember that $a^2 - b^2 = (a - b)(a + b)$. We are going to use this a lot.

Start by subtracting row 1 from both row 2 and row 3 to get:

$$\det(A) = \begin{vmatrix} 1 & x & x^2 \\ 0 & y-x & y^2-x^2 \\ 0 & z-x & z^2-x^2 \end{vmatrix}$$

All the terms in the second row now have common factor $(y - x)$ and al
in the third row have common factor $(z - x)$. So use the rules to pull thes

$$\det(A) = (y - x)(z - x) \begin{vmatrix} 1 & x & x^2 \\ 0 & 1 & y+x \\ 0 & 1 & z+x \end{vmatrix}$$

Next we subtract row 2 from row 3 and get a matrix in which all the te
the diagonal are zero:

$$\det(A) = (y - x)(z - x) \begin{vmatrix} 1 & x & x^2 \\ 0 & 1 & y+x \\ 0 & 0 & z-y \end{vmatrix} = (y - x)(z - x).1.1.(z - y)$$
$$= (x - y)(y - z)(z - x).$$

Determinant of a 2x2 matrix

Assuming A is an arbitrary 2x2 matrix A, where the elements are given

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

then the determinant of a this matrix is as follows:

$$\det(A) = |A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}$$

Now try an example of finding the determinant of a 2x2 matrix

Determinant of a 3x3 matrix

The determinant of a 3x3 matrix is a little more tricky and is found as f

$$\det(A) = \begin{vmatrix} 1 & x & x^2 \\ 0 & y-x & y^2-x^2 \\ 0 & z-x & z^2-x^2 \end{vmatrix}$$

All the terms in the second row now have common factor $(y - x)$ and all the terms in the third row have common factor $(z - x)$. So use the rules to pull these out:

$$\det(A) = (y - x)(z - x) \begin{vmatrix} 1 & x & x^2 \\ 0 & 1 & y+x \\ 0 & 1 & z+x \end{vmatrix}$$

Next we subtract row 2 from row 3 and get a matrix in which all the terms below the diagonal are zero:

$$\det(A) = (y - x)(z - x) \begin{vmatrix} 1 & x & x^2 \\ 0 & 1 & y+x \\ 0 & 0 & z-y \end{vmatrix} = (y - x)(z - x).1.1.(z - y)$$

$$= (x - y)(y - z)(z - x).$$

Determinant of a 2x2 matrix

Assuming A is an arbitrary 2x2 matrix A, where the elements are given by:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

then the determinant of a this matrix is as follows:

$$\det(A) = |A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}$$

Now try an example of finding the determinant of a 2x2 matrix

Determinant of a 3x3 matrix

The determinant of a 3x3 matrix is a little more tricky and is found as follows (for

this case assume A is an arbitrary 3x3 matrix A, where the elements are given below)

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

then the determinant of a this matrix is as follows:

$$\det(A) = |A| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

Determinant of a nxn matrix

For the general case, where A is an nxn matrix the determinant is given by:

$$\det(A) = |A| = a_{11}\alpha_{11} + a_{12}\alpha_{12} + \ldots + a_{1n}\alpha_{1n}$$

Where the coefficients α_{ij} are given by the relation

$$\alpha_{ij} = (-1)^{i+j}\beta_{ij}$$

where β_{ij} is the determinant of the (n-1) x (n-1) matrix that is obtained by deleting row i and column j. This coefficient α_{ij} is also called the cofactor of a_{ij}.

Activity 3

Let, $A = \begin{matrix} 4 & -1 \\ 6 & 9 \end{matrix}$

and $B = \begin{matrix} 0 & 3 \\ 3 & -2 \end{matrix}$

Find (*i*) $A + B$, (*ii*) $2A - B$, (*iii*) AB, (*iv*) BA, and (*v*) A' (the transpose of A).

Let $A = \begin{matrix} 4 & -1 \\ 6 & 9 \\ 2 & 3 \end{matrix}$

and $B = \begin{matrix} 0 & 3 \\ 3 & -2 \end{matrix}$

(a) Is AB defined? If so, find it. (*ii*) Is BA defined? If so, find it.

Use Cramer's rule to find the values of x and y that solve the following two equations simultaneously.

$$3x - 2y = 11$$
$$2x + y = 12$$

1. Use Cramer's rule to find the values of x, y, and z that solve the following three equations simultaneously.

$$4x + 3y - 2z = 7$$
$$x + y = 5$$
$$3x + z = 4$$

Solve the three equations by using matrix inversion

3.8 SUMMARY

Matrices provide a very convenient and compact system of writing a system of linear simultaneous equations and methods of solving them.

A number of basic matrix operations (such as matrix addition, subtraction and multiplication) were discussed in this unit. This was followed by a discussion on matrix inversion and procedure for finding matrix inverse. Numbers of examples were given in support of the above said operations and inverse of a matrix.

Finally Cramer's rule and determinants of matrix also have discussed in depth in order to give readers the full exposure of the concepts.

3.10 FURTHER READINGS

Budnicks, F.S. 1983. Applied Mathematics for Business, Economics, and Social Sciences, McGraw Hill: New York.

Hughes, A.J. 1983. Applied Mathematics for Business, Economics and the Social Sciences, Irwin: Homewood

Weber, J.E. 1982. Mathematical Analysis: Business and Economics Applications, Harper & Row: New York.

SOLUTIONS TO ACTIVITIES

ACTIVITY 1

1. ii

ACTIVITY 2

1.

$$y = 10^{3x}$$

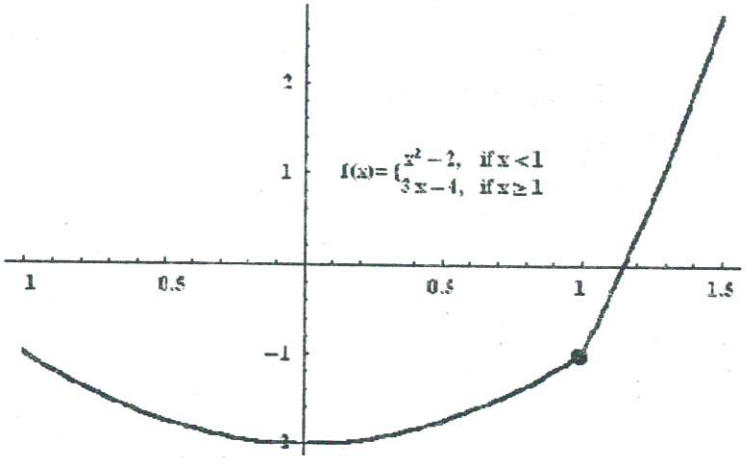
$$\frac{dy}{dx} = 10^{3x}(\ln 10)\left(\frac{d(3x)}{dx}\right)$$

$$= 3\ln 10(10^{3x})$$

2. $dy/dx = -(2x + 3y)/(3x + y)$.

3. $g'(f(x))f'(x) = 1$, so $f'(x) = 1/g'(f(x))$.

4. $A - 4 = -1$ or equivalently if $A = 3$.



5.

Put $u = 5x$ so $y = \cos^{-1}u$

$$y = \cos^{-1}5x$$

$$\frac{dy}{dx} = -\frac{1}{\sqrt{1-u^2}}\frac{du}{dx}$$

$$= -\frac{1}{\sqrt{1-(5x)^2}}\frac{d(5x)}{dx}$$

$$= \frac{-5}{\sqrt{1-25x^2}}$$

ACTIVITY 3

i. $\begin{matrix} 4 & 2 \\ 9 & 7 \end{matrix}$

ii. $\begin{matrix} 8 & -5 \\ 9 & 20 \end{matrix}$

iii. $\begin{matrix} -3 & 14 \\ 27 & 0 \end{matrix}$

iv. $\begin{matrix} 18 & 27 \\ 0 & -21 \end{matrix}$

v. $\begin{matrix} 4 & 6 \\ -1 & 9 \end{matrix}$

2. i. Yes; $\begin{matrix} -3 & 14 \\ 27 & 0 \\ 9 & 0 \end{matrix}$

ii. No

3. $(1/7)\begin{matrix} 1 & 2 \\ -2 & 3 \end{matrix}\begin{matrix} 11 \\ 12 \end{matrix} = \begin{matrix} 5 \\ 2 \end{matrix}$

4. $(1/7)\begin{matrix} 1 & -3 & 2 \\ -1 & 10 & -2 \\ -3 & 9 & 1 \end{matrix}\begin{matrix} 7 \\ 5 \\ 4 \end{matrix} = \begin{matrix} 0 \\ 5 \\ 4 \end{matrix}$

BLOCK 2

MATHEMATICAL METHODS

BLOCK 2 MATHEMATICAL METHODS

This block consists of two units. The first unit deals with basic concepts of linear programming, its uses, forms, slackness and concepts related to duality. The second unit thoroughly discusses solutions to linear programming problems including optimal solution of linear programming through graphical method. The unit also throws light on concept of games, different strategies of game and the saddle point solution.

UNIT 1

BASIC CONCEPTS OF LINEAR PROGRAMMING

Objectives

After studying this unit you should be able to:

- Understand the basic concept of Linear Programming
- Analyze the uses of Linear Programming
- Know the different concepts of standard form and augmented form problems.
- Have the knowledge of complementary slackness theorem

Structure

1.1 INTRODUCTION

In mathematics, **linear programming** (LP) is a technique for optimization of a linear objective function, subject to linear equality and linear inequality constraints. Informally, linear programming determines the way to achieve the best outcome (such as maximum profit or lowest cost) in a given mathematical model and given some list of requirements represented as linear equations.

More formally, given a polytope (for example, a polygon or a polyhedron), and a real-valued affine function

$$f(x_1, x_2, \ldots, x_n) = c_1x_1 + c_2x_2 + \cdots + c_nx_n + d$$

defined on this polytope, a linear programming method will find a point in the polytope where this function has the smallest (or largest) value. Such points may not exist, but if they do, searching through the polytope vertices is guaranteed to find at least one of them.

Linear programs are problems that can be expressed in canonical form:

Maximize $\mathbf{c}^T\mathbf{x}$
Subject to $A\mathbf{x} \leq \mathbf{b}$.

X represents the vector of variables (to be determined), while c and b are vectors of (known) coefficients and A is a (known) matrix of coefficients. The expression to be maximized or minimized is called the objective function ($\mathbf{c}^T\mathbf{x}$ in this case). The equations $A\mathbf{x} \leq \mathbf{b}$ are the constraints which specify a convex polyhedron over which the objective function is to be optimized.

Linear programming can be applied to various fields of study. Most extensively it is used in business and economic situations, but can also be utilized for some engineering problems. Some industries that use linear programming models include transportation, energy, telecommunications, and manufacturing. It has proved useful in modeling diverse types of problems in planning, routing, scheduling, assignment, and design.

Theory

Geometrically, the linear constraints define a convex polyhedron, which is called the feasible region. Since the objective function is also linear, hence a convex function, all local optima are automatically global optima (by the KKT theorem). The linearity of the objective function also implies that the set of optimal solutions is the convex hull of a finite set of points - usually a single point.

There are two situations in which no optimal solution can be found. First, if the constraints contradict each other (for instance, $x \geq 2$ and $x \leq 1$) then the feasible region is empty and there can be no optimal solution, since there are no solutions at all. In this case, the LP is said to be *infeasible*.

Alternatively, the polyhedron can be unbounded in the direction of the objective function (for example: maximize $x_1 + 3\ x_2$ subject to $x_1 \geq 0$, $x_2 \geq 0$, $x_1 + x_2 \geq 10$), in which case there is no optimal solution since solutions with arbitrarily high values of the objective function can be constructed.

Barring these two pathological conditions (which are often ruled out by resource constraints integral to the problem being represented, as above), the optimum is always attained at a vertex of the polyhedron. However, the optimum is not necessarily unique: it is possible to have a set of optimal solutions covering an edge or face of the polyhedron, or even the entire polyhedron (This last situation would occur if the objective function were constant).

1.2 USES OF LINEAR PROGRAMMING

Linear programming is a considerable field of optimization for several reasons. Many practical problems in operations research can be expressed as linear programming problems. Certain special cases of linear programming, such as *network flow* problems and *multicommodity flow* problems are considered important enough to have generated much research on specialized algorithms for their solution. A number of algorithms for other types of optimization problems work by solving LP problems as sub-problems. Historically, ideas from linear programming have inspired many of the central concepts of optimization theory,

such as *duality, decomposition,* and the importance of *convexity* and its generalizations. Likewise, linear programming is heavily used in microeconomics and company management, such as planning, production, transportation, technology and other issues. Although the modern management issues are ever-changing, most companies would like to maximize profits or minimize costs with limited resources. Therefore, many issues can boil down to linear programming problems.

1.3 STANDARD FORM

Standard form is the usual and most intuitive form of describing a linear programming problem. It consists of the following three parts:

- A **linear function to be maximized**

  e.g. maximize $c_1x_1 + c_2x_2$

- **Problem constraints** of the following form

  e.g.
  $a_{11}x_1 + a_{12}x_2 \leq b_1$
  $a_{21}x_1 + a_{22}x_2 \leq b_2$
  $a_{31}x_1 + a_{32}x_2 \leq b_3$

- **Non-negative variables**

  e.g. $x_1 \geq 0$
  $x_2 \geq 0$

The problem is usually expressed in *matrix form*, and then becomes:

maximize $\mathbf{c^T x}$
subject to $\mathbf{Ax \leq b, x \geq 0}$

Other forms, such as minimization problems, problems with constraints on alternative forms, as well as problems involving negative variables can always be rewritten into an equivalent problem in standard form.

Example 1

Suppose that a farmer has a piece of farm land, say A square kilometres large, to be planted with either wheat or barley or some combination of the two. The farmer has a limited permissible amount F of fertilizer and P of insecticide which can be used, each of which is required in different amounts per unit area for wheat (F_1, P_1) and barley (F_2, P_2). Let S_1 be the selling price of wheat, and S_2 the price of barley. If we denote the area planted with wheat and barley by x_1 and x_2 respectively, then the optimal number of square kilometres to plant with wheat vs barley can be expressed as a linear programming problem:

maximize $S_1x_1 + S_2x_2$ (maximize the revenue — revenue is the "objective function")

subject to
$x_1 + x_2 \leq A$ (limit on total area)

$F_1x_1 + F_2x_2 \leq F$ (limit on fertilizer)

$P_1x_1 + P_2x_2 \leq P$ (limit on insecticide)

$x_1 \geq 0,\ x_2 \geq 0$ (cannot plant a negative area)

Which in matrix form becomes:

$$\text{maximize} \begin{bmatrix} S_1 & S_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\text{subject to} \begin{bmatrix} 1 & 1 \\ F_1 & F_2 \\ P_1 & P_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} A \\ F \\ P \end{bmatrix}, \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \geq 0$$

1.4 AUGMENTED FORM (SLACK FORM)

Linear programming problems must be converted into *augmented form* before being solved by the simplex algorithm. This form introduces non-negative *slack variables* to replace inequalities with equalities in the constraints. The problem can then be written in the following form:

Maximize Z in:

$$\begin{bmatrix} 1 & -\mathbf{c}^T & 0 \\ 0 & \mathbf{A} & \mathbf{I} \end{bmatrix} \begin{bmatrix} Z \\ \mathbf{x} \\ \mathbf{x}_s \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix}$$

$$\mathbf{x}, \mathbf{x}_s \geq 0$$

where $\mathbf{x}_s$are the newly introduced slack variables, and Z is the variable to be maximized.

Example 2

The example above becomes as follows when converted into augmented form:

$$\begin{aligned} &\text{maximize } S_1x_1 + S_2x_2 && \text{(objective function)} \\ &\text{subject to } x_1 + x_2 + x_3 = A && \text{(augmented constraint)} \\ &\qquad F_1x_1 + F_2x_2 + x_4 = F && \text{(augmented constraint)} \\ &\qquad P_1x_1 + P_2x_2 + x_5 = P && \text{(augmented constraint)} \\ &\qquad x_1, x_2, x_3, x_4, x_5 \geq 0 \end{aligned}$$

where $x_3, x_4, x_5$are (non-negative) slack variables.

Which in matrix form becomes:

Maximize Z in:

$$\begin{bmatrix} 1 & -S_1 & -S_2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & F_1 & F_2 & 0 & 1 & 0 \\ 0 & P_1 & P_2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Z \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 0 \\ A \\ F \\ P \end{bmatrix}, \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \geq 0$$

1.5 DUALITY

Every linear programming problem, referred to as a primal problem, can be converted into a dual problem, which provides an upper bound to the optimal value of the primal problem. In matrix form, we can express the *primal problem* as:

$$\text{maximize } \mathbf{c}^T\mathbf{x}$$
$$\text{subject to } \mathbf{A}\mathbf{x} \leq \mathbf{b},\ \mathbf{x} \geq 0$$

The corresponding *dual problem* is:

$$\text{minimize } \mathbf{b}^T\mathbf{y}$$
$$\text{subject to } \mathbf{A}^T\mathbf{y} \geq \mathbf{c},\ \mathbf{y} \geq 0$$

where y is used instead of x as variable vector.

There are two ideas fundamental to duality theory. One is the fact that the dual of a dual linear program is the original primal linear program. Additionally, every feasible solution for a linear program gives a bound on the optimal value of the objective function of its dual. The weak duality theorem states that the objective function value of the dual at any feasible solution is always greater than or equal to the objective function value of the primal at any feasible solution. The strong duality theorem states that if the primal has an optimal solution, x*, then the dual also has an optimal solution, y*, such that $c^Tx^*=b^Ty^*$.

A linear program can also be unbounded or infeasible. Duality theory tells us that if the primal is unbounded then the dual is infeasible by the weak duality theorem. Likewise, if the dual is unbounded, then the primal must be infeasible. However, it is possible for both the dual and the primal to be infeasible (See also Farkas' lemma).

Example 3

Revisit the above example of the farmer who may grow wheat and barley with the set provision of some A land, F fertilizer and P insecticide. Assume now that unit prices for each of these means of production (inputs) are set by a planning board. The planning board's job is to minimize the total cost of procuring the set amounts of inputs while providing the farmer with a floor on the unit price of each of his crops (outputs), S1 for wheat and S2 for barley. This corresponds to the following linear programming problem:

minimize	$Ay_A + Fy_F + Py_P$	(minimize the total cost of the means of production as the "objective function")
subject to	$y_A + F_1 y_F + P_1 y_P \geq S_1$	(the farmer must receive no less than S_1 for his wheat)
	$y_A + F_2 y_F + P_2 y_P \geq S_2$	(the farmer must receive no less than S_2 for his barley)
	$y_A \geq 0,\ y_F \geq 0,\ y_P \geq 0$	(prices cannot be negative)

Which in matrix form becomes:

$$\text{minimize} \quad \begin{bmatrix} A & F & P \end{bmatrix} \begin{bmatrix} y_A \\ y_F \\ y_P \end{bmatrix}$$

$$\text{subject to} \quad \begin{bmatrix} 1 & F_1 & P_1 \\ 1 & F_2 & P_2 \end{bmatrix} \begin{bmatrix} y_A \\ y_F \\ y_P \end{bmatrix} \geq \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}, \quad \begin{bmatrix} y_A \\ y_F \\ y_P \end{bmatrix} \geq 0$$

The primal problem deals with physical quantities. With all inputs available in limited quantities, and assuming the unit prices of all outputs is known, what quantities of outputs to produce so as to maximize total revenue? The dual problem deals with economic values. With floor guarantees on all output unit prices, and assuming the available quantity of all inputs is known, what input unit pricing scheme to set so as to minimize total expenditure?

To each variable in the primal space corresponds an inequality to satisfy in the dual space, both indexed by output type. To each inequality to satisfy in the primal space corresponds a variable in the dual space, both indexed by input type.

The coefficients which bound the inequalities in the primal space are used to compute the objective in the dual space, input quantities in this example. The coefficients used to compute the objective in the primal space bound the inequalities in the dual space, output unit prices in this example.

Both the primal and the dual problems make use of the same matrix. In the primal space, this matrix expresses the consumption of physical quantities of inputs necessary to produce set quantities of outputs. In the dual space, it expresses the creation of the economic values associated with the outputs from set input unit prices.

Since each inequality can be replaced by an equality and a slack variable, this means each primal variable corresponds to a dual slack variable, and each dual variable corresponds to a primal slack variable. This relation allows us to complementary slackness.

1.6 SPECIAL CASES

A *packing LP* is a linear program of the form

$$\begin{aligned} &\text{maximize } \mathbf{c}^T\mathbf{x} \\ &\text{subject to } \mathbf{A}\mathbf{x} \leq \mathbf{b},\ \mathbf{x} \geq 0 \end{aligned}$$

such that the matrix A and the vectors b and c are non-negative.

The dual of a packing LP is a *covering LP*, a linear program of the form

$$\begin{aligned} &\text{minimize } \mathbf{b}^T\mathbf{y} \\ &\text{subject to } \mathbf{A}^T\mathbf{y} \geq \mathbf{c},\ \mathbf{y} \geq 0 \end{aligned}$$

such that the matrix A and the vectors b and c are non-negative.

Example 4

Covering and packing LPs commonly arise as a linear programming relaxation of a combinatorial problem. For example, the LP relaxation of set packing problem, independent set problem, or matching is a packing LP. The LP relaxation of set cover problem, vertex cover problem, or dominating set problem is a covering LP.

Finding a fractional coloring of a graph is another example of a covering LP. In this case, there is one constraint for each vertex of the graph and one variable for each independent set of the graph.

1.7 COMPLEMENTARY SLACKNESS

It is possible to obtain an optimal solution to the dual when only an optimal solution to the primal is known using the complementary slackness theorem. The theorem states:

Suppose that $x = (x_1, x_2, \ldots, x_n)$ is primal feasible and that $y = (y_1, y_2, \ldots, y_m)$ is dual feasible. Let $(w_1, w_2, \ldots, w_m)$ denote the corresponding primal slack variables, and let $(z_1, z_2, \ldots, z_n)$ denote the corresponding dual slack variables. Then x and y are optimal for their respective problems if and only if $x_j z_j = 0$, for $j = 1, 2, \ldots, n$, $w_i y_i = 0$, for $i = 1, 2, \ldots, m$.

So if the ith slack variable of the primal is not zero, then the ith variable of the dual is equal zero. Likewise, if the jth slack variable of the dual is not zero, then the jth variable of the primal is equal to zero.

Activity 1

1. Discuss the uses of Linear Programming.
2. Explain briefly the concept of Duality.

1.8 SUMMARY

Linear programming is an important field of optimization for several reasons. Many practical problems in operations research can be expressed as linear programming problems. Followed by the basic concept the concepts of duality, standard form and augmented form have described in the chapter.
The different kind of problems can be solved using Linear Programming approach is discussed in special case section. Further the theorem of Complementary slackness was discussed in brief to have more clear understanding of solution to Linear programming problems.

1.9 FURTHER READINGS

- Mark de Berg, Marc van Kreveld, Mark Overmars, and Otfried Schwarzkopf (2000). *Computational Geometry* (2nd revised edition ed.). Springer-Verlag.
- V. Chandru and M.R.Rao, Linear Programming, Chapter 31 in *Algorithms and Theory of Computation Handbook*, edited by M.J.Atallah, CRC Press
- V. Chandru and M.R.Rao, Integer Programming, Chapter 32 in *Algorithms and Theory of Computation Handbook*, edited by M.J.Atallah,
- Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*, Second Edition. MIT Press and McGraw-Hill

UNIT 2

SOLUTIONS TO LINEAR PROGRAMMING PROBLEMS AND THEORY OF GAME

Objectives

After studying this unit you should be able to:

- Understand the basic concepts of computation of linear programming problems.
- Know the approaches to solve prototype and general linear programming problems.
- Solve the linear programming problems using graphical method.
- Appreciate the concept and strategies pertaining to game theory.
- Be aware about the saddle point solution.

Structure

2.1 INTRODUCTION

A *Linear Programming* problem is a special case of a *Mathematical Programming* problem. From an analytical perspective, a mathematical program tries to identify an *extreme* (i.e., minimum or maximum) point of a function $f(x_1, x_2, \ldots, x_n)$, which furthermore satisfies a set of constraints, e.g., $g(x_1, x_2, \ldots, x_n) \geq b$. Linear programming is the specialization of mathematical programming to the case where both, function f - to be called the *objective function* - and the problem constraints are *linear*.

Solution procedure used when a Linear Programming (LP) problem has two (or at most three) decision variables. The graphical method follows these steps:

(1) Change inequalities to equalities.

(2) Graph the equalities.

(3) Identify the correct side for the original inequalities.

(4) Then identify the *feasible region*, the area of Feasible Solution.

(5) Determine the Contribution Margin (CM) or cost at each of the *corner points* (*basic feasible solutions*) of the feasible region.

(6) Pick either the most profitable or least cost combination, which is an Optimal Solution.

2.2 A PROTOTYPE LP PROBLEM

Consider a company which produces two types of products P1 and P2 . Production of these products is supported by two workstations W1 and W2 , with each station visited by both product types. If workstation W1 is dedicated completely to the production of product type P1, it can process 40 units per day, while if it is dedicated to the production of product P2, it can process 60 units per

day. Similarly, workstation W2 can produce daily 50 units of product P1 and 50 units of product P2 , assuming that it is dedicated completely to the production of the corresponding product. If the company's profit by disposing one unit of product P1 is $200 and that of disposing one unit of P2 is $400, and assumning that the company can dispose its entire production, how many units of each product should the company produce on a daily basis to maximize its profit?

Solution: First notice that this problem is an *optimization* problem. Our *objective* is to *maximize* the company's profit, which under the problem assumptions, is equivalent to maximizing the company's *daily* profit. Furthermore, we are going to maximize the company profit by adjusting the levels of the daily production for the two items P1 and P2 . Therefore, these daily production levels are the control/decision factors, the values of which we are called to determine. In the analytical formulation of the problem, the role of these factors is captured by modeling them as the problem *decision variables*:

- X1 := number of units of product P1 to be produced daily
- X2:= number of units of product P2 to be produced daily

In the light of the above discussion, the problem objective can be expressed analytically as:

$$\max f(X_1, X_2) := 200X_1 + 400X_2 \qquad (1)$$

Equation 1 will be called the *objective function* of the problem, and the coefficients 200 and 400 which multiply the decision variables in it, will be called the *objective function coefficients*.

Furthermore, any decision regarding the daily production levels for items P1 and P2 in order to be realizable in the company's operation context must observe the production capacity of the two worksations W1 and W2 . Hence, our next step in the problem formulation seeks to introduce these *technological* constraints in it. Let's focus first on the constraint which expresses the finite production capacity of workstation W1 . Regarding this constraint, we know that one day's work dedicated to the production of item P1 can result in 40 units of that item, while

the same period dedicated to the production of item P2 will provide 60 units of it. *Assuming that production of one unit of product type Pi,i=1,2 requires a constant amount of processing time at workstation W1* , it follows that: $\tau_{11} = \frac{1}{40}$ and $\tau_{12} = \frac{1}{60}$. *Under the further assumption that the combined production of both items has no side-effects, i.e., does not impose any additional requirements for production capacity of workstation W1(e.g., zero set-up times)*, the total capacity (in terms of time length) required for producing X1units of product P1and X2 units of product P2 is equal to $\frac{1}{40}X_1 + \frac{1}{60}X_2$. Hence, the technological constraint imposing the condition that our total daily processing requirements for workstation W1 should not exceed its production capacity, is analytically expressed by:

$$\frac{1}{40}X_1 + \frac{1}{60}X_2 \leq 1.0 \qquad (2)$$

Notice that in Equation 2 time is measured in days.

Following the same line of reasoning (and under similar assumptions), the constraint expressing the finite processing capacity of workstation W_2 is given by:

$$\frac{1}{50}X_1 + \frac{1}{50}X_2 \leq 1.0 \qquad (3)$$

Constraints 2 and 3 are known as the *technological constraints* of the problem. In particular, the coefficients of the variables $X_i,\ i = 1,2$ in them, $\frac{1}{\tau_{ji}},\ j,i = 1,2$, are known as the *technological coefficients* of the problem formulation, while the values on the right-hand-side of the two inequalities define the *right-hand side (rhs)* vector of the constraints.

Finally, to the above constraints we must add the requirement that any permissible value for variables $X_i,\ i = 1,2$ must be nonnegative, i.e.,

$$X_i \geq 0 \quad i = 1,2 \qquad (4)$$

since these values express production levels. These constraints are known as the variable *sign restrictions.*

Combining Equations 1 to 4, the analytical formulation of our problem is as follows:

$$\max f(X_1, X_2) := 200X_1 + 400X_2$$

2.3 THE GENERAL LP FORMULATION

Generalizing formulation 5, the general form for a Linear Programming problem is as follows:

Objective Function:

$$\max / \min f(X_1, X_2, \ldots, X_n) := c_1X_1 + c_2X_2 + \ldots + c_nX_n \qquad (6)$$

s.t.

Technological Constraints:

$$a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{in}X_n \left(\begin{array}{c} \leq \\ = \\ \geq \end{array} \right) b_i, \quad i = 1, \ldots, m \qquad (7)$$

Sign Restrictions:

$$(X_j \geq 0) \text{ or } (X_j \leq 0) \text{ or } (X_j \text{ urs}), \quad j = 1, \ldots, n \qquad (8)$$

where ``urs" implies *unrestricted in sign.*

The formulation of Equations 6 to 8 has the general structure of a mathematical programming problem, presented in the introduction of this section, but it is further characterized by the fact that the functions involved in the problem objective and the left-hand-side of the technological constraints are *linear.* It is the

assumptions implied by linearity that to a large extent determine the applicability of the above model in real-world applications.

To provide a better feeling of the linearity concept, let us assume that the different decision variables $X_1, \ldots, X_n$ correspond to various activities from which any solution will be eventually synthesized, and the values assigned to the variables by any given solution indicate the activity level in the considered plan(s). For instance, in the above example, the two activities are the production of items P1 and P2 , while the activity levels correspond to the daily production volume. Furthermore, let us assume that each technological constraint of Equation 7 imposes some restriction on the consumption of a particular resource. Referring back to the prototype example, the two problem resources are the daily production capacity of the two workstations W1 and W2. Under this interpretation, the linearity property implies that:

Additivity assumption:

the total consumption of each resource, as well as the overall objective value are the aggregates of the resource consumptions and the contributions to the problem objective, resulting by carrying out each activity independently, and

Proportionality assumption:

these consumptions and contributions for each activity are proportional to the actual activity level.

It is interesting to notice how the above statement reflects to the logic that was applied when we derived the technological constraints of the prototype example: (i) Our assumption that the processing of each unit of product at every station requires a constant amount of time establishes the *proportionality* property for our model. (ii) The assumption that the total processing time required at every station to meet the production levels of both products is the aggregate of the processing times required for each product if the corresponding activity took place independently, implies that our system has an *additive* behavior. It is also interesting to see how the linearity assumption restricts the modeling capabilities of the LP framework: As an example, in the LP paradigm, we cannot immediately model effects like economies of scale in the problem cost structure, and/or

situations in which resource consumption by one activity depends on the corresponding consumption by another complementary activity. In some cases, one can approach these more complicated problems by applying some *linearization* scheme. The resulting approximations for many of these cases have been reported to be quite satisfactory.

Another approximating element in many real-life LP applications results from the so called *divisibility* assumption. This assumption refers to the fact that for LP theory and algortihms to work, the problem variables must be *real*. However, in many LP formulations, meaningful values for the levels of the activities involved can be only *integer*. This is, for instance, the case with the production of items P_1 and P_2 in our prototype example. Introducing integrality requirements for some of the variables in an LP formulation turns the problem to one belonging in the class of *(Mixed) Integer Programming (MIP)*. The complexity of a MIP problem is much higher than that of LP's. Actually, the general IP formulation has be shown to belong to the notorious class of *NP-complete* problems. (This is a class of problems that have been ``formally" shown to be extremely ``hard" computationally). Given the increased difficulty of solving IP problems, sometimes in practice, near optimal solutions are obtained by solving the LP formulation resulting by relaxing the integrality requirements - known as the *LP relaxation* of the corresponding IP - and (judiciously) rounding off the fractional values for the integral variables in the optimal solution. Such an approach can be more easily justified in cases where the typical values for the integral variables are in the order of tens or above, since the errors introduced by the rounding-off are rather small, in a relative sense.

We conclude our discussion on the general LP formulation, by formally defining the solution search space and optimality. Specifically, we shall define as the *feasible region* of the LP of Equations 6 to 8, the entire set of vectors $< X_1, X_2, \ldots, X_n >^T$ that satisfy the technological constraints of Eq. 7 and the sign restrictions of Eq. 8. An *optimal* solution to the problem is any feasible vector that further satisfies the optimality requirement expressed by Eq. 6. In the next section, we provide a geometric characterization of the feasible region and

the optimality condition, for the special case of LP's having only two decision variables.

2.4 OPTIMAL SOLUTION THROUGH GRAPHICAL METHOD

This section develops a solution approach for LP problems, which is based on a geometrical representation of the feasible region and the objective function. In particular, the space to be considered is the *n*-dimensional space with each dimension defined by one of the LP variables Xj . The objective function will be described in this *n*-dim space by its *contour plots*, i.e., the sets of points that correspond to the same objective value. To the extent that the proposed approach requires the visualization of the underlying geometry, it is applicable only for LP's with upto three variables. Actually, to facilitate the visualization of the concepts involved, in this section we shall restrict ourselves to the two-dimensional case, i.e., to LP's with two decision variables. In the next section, we shall generalize the geometry introduced here for the 2-var case, to the case of LP's with *n* decision variables, providing more analytic (algebraic) characterizations of these concepts and properties.

Graphical solution of the prototype example 1: 2-var LP with a unique optimal solution

The `` sliding motion" described suggests a way for identifying the optimal values for, let's say, a max LP problem. The underlying idea is to keep ``sliding" the isoprofit line $c_1X_1 + c_2X_2 = \alpha_0$ in the direction of increasing α's, until we cross the boundary of the LP feasible region. The implementation of this idea on the prototype LP of Equation 5 is depicted in Figure 3.

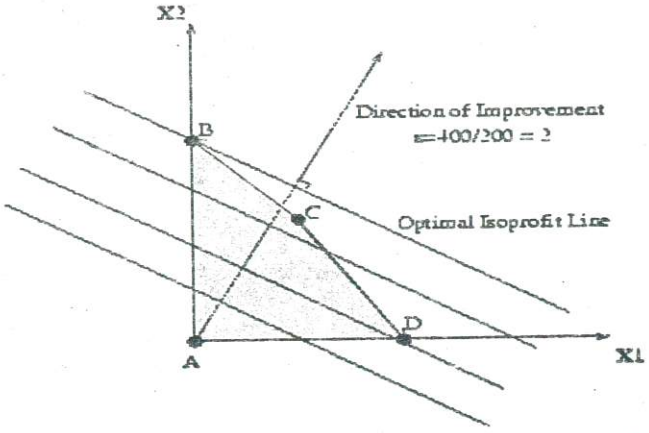


Figure 2.1: Graphical solution of the prototype example LP

From this figure, it follows that the optimal daily production levels for the protoype LP are given by the coordinates of the point corresponding to the intersection of line $\frac{1}{50}X_1 + \frac{1}{50}X_2 = 0$ with the X_2-axis, i.e., $X_1^{opt} = 0;\ X_2^{opt} = 50$. The maximal daily profit is $f(X_1^{opt}, X_2^{opt}) = 200 \cdot 0 + 400 \cdot 50 = 20,000\ (\$)$. Notice that the optimal point is one of the ``*corner*'' points of the feasible region depicted in Figure 3. Can you argue that for the geometry of the feasible region for 2-var LP's described above, if there is a bounded optimal solution, then there will be one which corresponds to one of the corner points? (This argument is developed for the broader context of n-var LP's in the next section.)

2-var LP's with many optimal solutions

Consider our prototype example with the unit profit of item P_1 being \$600 instead of \$200. Under this modification, the problem isoprofit lines become:

$$600X_1 + 400X_2 = \alpha \Leftrightarrow X_2 = -\frac{3}{2}X_1 + \frac{\alpha}{400}$$

and they are parallel to the line corresponding to the first problem constraint:

$$\frac{1}{40}X_1 + \frac{1}{60}X_2 = 1 \Leftrightarrow X_2 = -\frac{3}{2}X_1 + 60.$$

Therefore, if we try to apply the optimizing technique of the previous paragraph in this case, we get the situation depicted below (Figure 4), i.e., every point in the line segment *CD* is an optimal point, providing the optimal objective value of \$24,000.

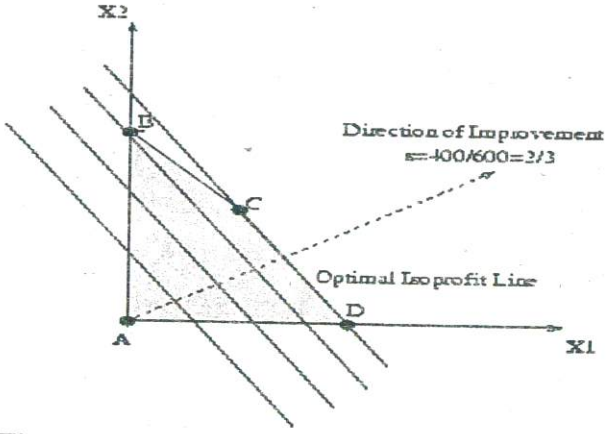


Figure 2.2 : An LP with many optimal solutions

It is worth-noticing that even in this case of many optimal solutions, we have two of them corresponding to ``corner" points of the feasible region, namely points *C* and *D*.

2.5 CONCEPT OF A GAME

Game theory attempts to mathematically capture behavior in *strategic situations*, in which an individual's success in making choices depends on the choices of others. While initially developed to analyze competitions in which one individual does better at another's expense (zero sum games), it has been expanded to treat a wide class of interactions, which are classified according to several criteria. Today, "game theory is a sort of umbrella or 'unified field' theory for the rational

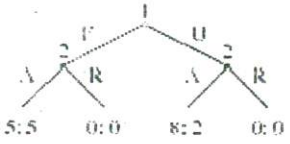
side of social science, where 'social' is interpreted broadly, to include human as well as non-human players (computers, animals, plants)" *(Aumann 1987).*

Representation of games

The games studied in game theory are well-defined mathematical objects. A game consists of a set of players, a set of moves (or strategies) available to those players, and a specification of payoffs for each combination of strategies. Most cooperative games are presented in the characteristic function form, while the extensive and the normal forms are used to define noncooperative games.

Extensive form

Main article: Extensive form game



An extensive form game

The extensive form can be used to formalize games with some important order. Games here are often presented as trees (as pictured to the left). Here each vertex (or node) represents a point of choice for a player. The player is specified by a number listed by the vertex. The lines out of the vertex represent a possible action for that player. The payoffs are specified at the bottom of the tree.

In the game pictured here, there are two players. *Player 1* moves first and chooses either *F* or *U*. *Player 2* sees *Player 1*'s move and then chooses *A* or *R*. Suppose that *Player 1* chooses *U* and then *Player 2* chooses *A*, then *Player 1* gets 8 and *Player 2* gets 2.

The extensive form can also capture simultaneous-move games and games with imperfect information. To represent it, either a dotted line connects different vertices to represent them as being part of the same information set (i.e., the

players do not know at which point they are), or a closed line is drawn around them.

Normal form

The normal (or strategic form) game is usually represented by a matrix which shows the players, strategies, and payoffs (see the example to the right). More generally it can be represented by any function that associates a payoff for each player with every possible combination of actions. In the accompanying example there are two players; one chooses the row and the other chooses the column. Each player has two strategies, which are specified by the number of rows and the number of columns. The payoffs are provided in the interior. The first number is the payoff received by the row player (Player 1 in our example); the second is the payoff for the column player (Player 2 in our example). Suppose that Player 1 plays *Up* and that Player 2 plays *Left*. Then Player 1 gets a payoff of 4, and Player 2 gets 3.

	Player 2 chooses *Left*	Player 2 chooses *Right*
Player 1 chooses *Up*	**4, 3**	**–1, –1**
Player 1 chooses *Down*	**0, 0**	**3, 4**

Normal form or payoff matrix of a 2-player, 2-strategy game

When a game is presented in normal form, it is presumed that each player acts simultaneously or, at least, without knowing the actions of the other. If players have some information about the choices of other players, the game is usually presented in extensive form.

Characteristic function form

Main article: Cooperative game

In cooperative games with transferable utility no individual payoffs are given. Instead, the characteristic function determines the payoff of each coalition. The standard assumption is that the empty coalition obtains a payoff of 0.

The origin of this form is to be found in the seminal book of von Neumann and Morgenstern who, when studying coalitional normal form games, assumed that when a coalition C forms, it plays against the complementary coalition ($N \setminus C$) as if they were playing a 2-player game. The equilibrium payoff of C is *characteristic*. Now there are different models to derive coalitional values from normal form games, but not all games in characteristic function form can be derived from normal form games.

Formally, a characteristic function form game (also known as a TU-game) is given as a pair (N,v), where N denotes a set of players and $v : 2^N \longrightarrow \mathbb{R}$ is a characteristic function.

The characteristic function form has been generalised to games without the assumption of transferable utility.

Partition function form

The characteristic function form ignores the possible externalities of coalition formation. In the partition function form the payoff of a coalition depends not only on its members, but also on the way the rest of the players are partitioned *(Thrall & Lucas 1963)*.

2.6 GAME STRATEGIES

The particular behavior or suite of behaviors that a player uses is termed a **strategy** (see **important note**). Strategies can be behaviors that are on some continuum (e.g., how long to wait or display) or they may represent discrete behavior types (e.g., display, fight, or flee). Sometimes the terms **pure strategy** and **mixed strategy** are used.

A **simple or pure strategy** provides a complete definition of how a player will play a game. In particular, it determines the move a player will make for any situation they could face. A player's **strategy set** is the set of pure strategies available to that player. A **pure strategy** in fact, is a strategy that is not defined in terms of other strategies present in the game.

A **mixed strategy** is an assignment of a probability to each pure strategy. This allows for a player to randomly select a pure strategy. Since probabilities are continuous, there are infinitely many mixed strategies available to a player, even if their strategy set is finite.

Mixed strategy

Suppose the payoff matrix pictured to the right (known as a coordination game). Here one player chooses the row and the other chooses a column. The row player receives the first payoff, the column the second. If row opts to play **A** with probability 1 (i.e. play **A** for sure), then he is said to be playing a pure strategy. If column opts to flip a coin and play **A** if the coin lands heads and **B** if the coin lands tails, then she is said to be playing a mixed strategy, and not a pure strategy.

	A	B
A	1, 1	0, 0
B	0, 0	1, 1

Pure coordination game

Significance

In his famous paper John Forbes Nash proved that there is an equilibrium for every finite game. One can divide Nash equilibria into two types. *Pure strategy Nash equilibria* are Nash equilibria where all players are playing pure strategies. *Mixed strategy Nash equilibria* are equilibria where at least one player is playing a mixed strategy. While Nash proved that every finite game has a Nash equilibrium, not all have pure strategy Nash equilibria. For an example of a game that does not have a Nash equilibrium in pure strategies see Matching pennies. However, many games do have pure strategy Nash equilibria (e.g. the Coordination game, the Prisoner's dilemma, the Stag hunt). Further, games can have both pure strategy and mixed strategy equilibria.

The Nash equilibrium concept is used to analyze the outcome of the strategic interaction of several decision makers. In other words, it is a way of predicting what will happen if several people or several institutions are making decisions at the same time, and if the decision of each one depends on the decisions of the others. The simple insight underlying John Nash's idea is that we cannot predict

the result of the choices of multiple decision makers if we analyze those decisions in isolation. Instead, we must ask what each player would do, *taking into account* the decision-making of the others.

Formal definition

Let (S, f) be a game with n players, where S_i is the strategy set for player i, $S=S_1 X S_2 ... X S_n$ is the set of strategy profiles and $f=(f_1(x), ..., f_n(x))$ is the payoff function. Let x_{-i} be a strategy profile of all players except for player i. When each player $i \in \{1, ..., n\}$ chooses strategy x_i resulting in strategy profile $x = (x_1, ..., x_n)$ then player i obtains payoff $f_i(x)$. Note that the payoff depends on the strategy profile chosen, i.e. on the strategy chosen by player i as well as the strategies chosen by all the other players. A strategy profile $x^* \in S$ is a Nash equilibrium (NE) if no unilateral deviation in strategy by any single player is profitable for that player, that is

$$\forall i, x_i \in S_i, x_i \neq x_i^* : f_i(x_i^*, x_{-i}^*) \geq f_i(x_i, x_{-i}^*).$$

A game can have a pure strategy NE or an NE in its mixed extension (that of choosing a pure strategy stochastically with a fixed frequency). Nash proved that if we allow mixed strategies, then every n-player game in which every player can choose from finitely many strategies admits at least one Nash equilibrium.

When the inequality above holds strictly (with > instead of $\geq$) for all players and all feasible alternative strategies, then the equilibrium is classified as a **strict Nash equilibrium**. If instead, for some player, there is exact equality between x_i^* and some other strategy in the set S, then the equilibrium is classified as a **weak Nash equilibrium**.

Coordination game

	Player 2 adopts strategy A	**Player 2 adopts strategy B**
Player 1 adopts strategy A	*4, 4*	*1, 3*
Player 1 adopts strategy B	*3, 1*	*3, 3*

A sample coordination game showing relative payoff for player1 / player2 with each combination

The *coordination game* is a classic (symmetric) two player, two strategy game, with an example payoff matrix shown to the right. The players should thus coordinate, both adopting strategy A, to receive the highest payoff, i.e., 4. If both players chose strategy B though, there is still a Nash equilibrium. Although each player is awarded less than optimal payoff, neither player has incentive to change strategy due to a reduction in the immediate payoff (from 3 to 1). An example of a coordination game is the setting where two technologies are available to two firms with compatible products, and they have to elect a strategy to become the market standard. If both firms agree on the chosen technology, high sales are expected for both firms. If the firms do not agree on the standard technology, few sales result. Both strategies are Nash equilibria of the game.

Driving on a road, and having to choose either to drive on the left or to drive on the right of the road, is also a coordination game. For example, with payoffs 100 meaning no crash and 0 meaning a crash, the coordination game can be defined with the following payoff matrix:

	Drive on the Left	**Drive on the Right**
Drive on the Left	100, 100	0, 0
Drive on the Right	0, 0	100, 100

The driving game

In this case there are two pure strategy Nash equilibria, when both choose to either drive on the left or on the right. If we admit mixed strategies (where a pure strategy is chosen at random, subject

to some fixed probability), then there are three Nash equilibria for the same case: two we have seen from the pure-strategy form, where the probabilities are (0%,100%) for player one, (0%, 100%) for player two; and (100%, 0%) for player one, (100%, 0%) for player two respectively. We add another where the probabilities for each player is (50%, 50%).

2.7 THE SADDLE POINT SOLUTION

a saddle point is a point in the domain of a function of two variables which is a stationary point but not a local extremum. At such a point, in general, the surface resembles a saddle that *curves up* in one direction, and *curves down* in a different direction (like a mountain pass). In terms of contour lines, a saddle point can be recognized, in general, by a contour that appears to intersect itself. For example, two hills separated by a high pass will show up a saddle point, at the top of the pass, like a figure-eight contour line.

A simple criterion for checking if a given stationary point of a real-valued function $F(x,y)$ of two real variables is a saddle point is to compute the function's Hessian matrix at that point: if the Hessian is indefinite, then that point is a saddle point. For example, the Hessian matrix of the function $z = x^2 - y^2$ at the stationary point (0,0) is the matrix

$$\begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

which is indefinite. Therefore, this point is a saddle point. This criterion gives only a sufficient condition. For example, the point (0,0) is a saddle point for the function $z = x^4 - y^4$, but the Hessian matrix of this function at the origin is the null matrix, which is not indefinite.

In the most general terms, a **saddle point** for a smooth function (whose graph is a curve, surface or hypersurface) is a stationary point such that the curve/surface/etc. in the neighborhood of that point is not entirely on any side of the tangent space at that point.

In one dimension, a saddle point is a point which is both a stationary point and a point of inflection. Since it is a point of inflection, it is not a local extremum.

THE VALUE AND METHOD OF SADDLE POINT

In mathematics, the **steepest descent method** or **saddle-point approximation** is a method used to approximate integrals of the form

$$\int_a^b e^{Mf(x)}\,dx$$

where $f(x)$ is some twice-differentiable function, M is a large number, and the integral endpoints a and b could possibly be infinite. The technique is also often referred to as **Laplace's method**, which in fact concerns the special case of real-valued functions f admitting a maximum at a real point.

Further, In dynamical systems, a *saddle point* is a periodic point whose stable and unstable manifolds have a dimension which is not zero. If the dynamic is given by a differentiable map f then a point is hyperbolic if and only if the differential of f^n (where n is the period of the point) has no eigenvalue on the (complex) unit circle when computed at the point.

In a two-player zero sum game defined on a continuous space, the equilibrium point is a saddle point.

A saddle point is an element of the matrix which is both the smallest element in its column and the largest element in its row.

For a second-order linear autonomous systems, a critical point is a saddle point if the characteristic equation has one positive and one negative real eigenvalue [1].

simple discussion (where the method is termed *steepest descents*).

The idea of Laplace's method

Assume that the function $f(x)$ has a unique global maximum at x_0. Then, the value $f(x_0)$ will be larger than other values $f(x)$. If we multiply this function by a large number M, the gap between $Mf(x_0)$ and $Mf(x)$ will only increase, and then it will grow exponentially for the function

$$e^{Mf(x)}.$$

As such, significant contributions to the integral of this function will come only from points x in a neighborhood of x_0, which can then be estimated.

General theory of Laplace's method

To state and prove the method, we need several assumptions. We will assume that x_0 is not an endpoint of the interval of integration, that the values $f(x)$ cannot be very close to $f(x_0)$ unless x is close to x_0, and that $f''(x_0) < 0$.

We can expand $f(x)$ around x_0 by Taylor's theorem,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + O\left((x - x_0)^3\right).$$

Since f has a global maximum at x_0, and since x_0 is not an endpoint, it is a stationary point, the derivative of f vanishes at x_0. Therefore, the function $f(x)$ may be approximated to quadratic order

$$f(x) \approx f(x_0) - \frac{1}{2}|f''(x_0)|(x - x_0)^2$$

for x close to x_0 (recall that the second derivative is negative at the global maximum $f(x_0)$). The assumptions made ensure the accuracy of the approximation

$$\int_a^b e^{Mf(x)}\,dx \approx e^{Mf(x_0)} \int_a^b e^{-M|f''(x_0)|(x-x_0)^2/2} dx$$

where the integral is taken in a neighborhood of x_0. This latter integral is a Gaussian integral if the limits of integration go from $-\infty$ to $+\infty$ (which can be

assumed so because the exponential decays very fast away from x_0), and thus it can be calculated. We find

$$\int_a^b e^{Mf(x)}\,dx \approx \sqrt{\frac{2\pi}{M|f''(x_0)|}}e^{Mf(x_0)} \text{ as } M \to \infty.$$

A generalization of this method and extension to arbitrary precision is provided by Fog (2008).

Steepest descent

In extensions of Laplace's method, complex analysis, and in particular Cauchy's integral formula, is used to find a contour *of steepest descent* for an (asymptotically with large M) equivalent integral, expressed as a line integral. In particular, if no point x_0 where the derivative of f vanishes exists on the real line, it may be necessary to deform the integration contour to an optimal one, where the above analysis will be possible. Again the main idea is to reduce, at least asymptotically, the calculation of the given integral to that of a simpler integral that can be explicitly evaluated. See the book of Erdelyi (1956) for a Other uses

Activity 2

1. Make a linear programming graph from the following LP model and find out the most profitable solution.

Maximize CM = \$25A + \$40B

Subject to: $2A + 4B \leq 100$ hours

$3A + 2B \leq 90$

$A \geq 0, B \geq 0$

2. Discuss briefly the saddle point solution and find out its applications.

2.8 SUMMARY

It has been discussed in this chapter that linear programming problem is basically a type of mathematical programming problem, which was discussed with the help of a prototype LP problem. Graphical method to optimal solution was discussed with suitable examples on 1 and 2 variable case solutions.
Concept o game with the representation of different games was explained in depth. Further different strtegies of game theory were discussed.
Finally the saddle point solution was explained briefly.

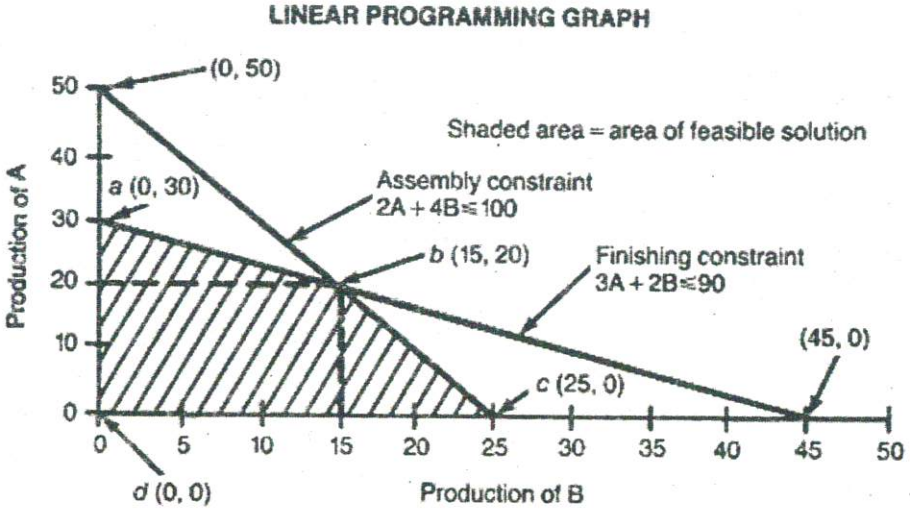
2.9 FURTHER READINGS

- Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*, Second Edition. MIT Press and McGraw-Hill
- Michael R. Garey and David S. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman.
- Bernd Gärtner, Jiří Matoušek (2006). *Understanding and Using Linear Programming*, Berlin: Springer
- Jalaluddin Abdullah, *Optimization by the Fixed-Point Method, Version 1.97.* [3].
- Alexander Schrijver, *Theory of Linear and Integer Programming*. John Wiley & sons
- Michael J. Todd (February 2002). "The many facets of linear programming". *Mathematical Programming*

ANSWERS TO ACTIVITIES

ACTIVITY 2

1.After going through steps 1 through 4, the feasible region (shaded area) is obtained, as shown in the following exhibit. Then all the corner points in the feasible region are evaluated in terms of their CM as follows:



	Corner Points		CM
	A	B	$25A + $40B
(a)	30	0	$25(30) + $40(0) = $750
(b)	20	15	$25(20) + $40(15) = $1100
(c)	0	25	$25(0) + $40(25) = $1000
(d)	0	0	$25(0) + $40(0) = $0

The corner 20A, 15B produces the most profitable solution.

The corner 20A, 15B produces the most profitable solution.

BLOCK 3

STATISTICAL METHODS

BLOCK 3 STATISTICAL METHODS

This block on Statistical methods consists of three units.

Unit 1 presents the concepts of correlation which is central in model development for forecasting. Various measures of association between variables are described with potential applications. The unit also discusses a very important technique for establishing relationships between variables, namely Regression. Fundamentals of linear regression are presented with applications and interpretations of statistical computation.

Unit 2 deals with basic concepts of probability including classical and empirical definitions of probability, basic concepts of Experiments, sample space and events, random variable, probability expectations and generating functions.

Unit 3 explains various probability laws and distributions. Laws of addition, multiplication are discussed with discussing normal, poison and binomial distributions.

UNIT 1

CORRELATION AND REGRESSION

Objectives

After completing this unit, you should be able to:

- Understand the meaning of correlation
- Compute the coefficient of correlation between two variables from sample observations
- Become aware of the concept of Pearson Product moment correlation
- Understand the role of regression in establishing mathematical relationships between dependent and independent variables from given data
- Determine the standard errors of estimate of the estimated parameters
- Know the basic concepts of partial and multiple correlation and their applications

Structure

1.1 INTRODUCTION

CORRELATION

correlation (often measured as a correlation coefficient) indicates the strength and direction of a *linear* relationship between two random variables. That is in contrast with the usage of the term in colloquial speech, denoting any relationship, not necessarily linear. In general statistical usage, *correlation* or co-relation refers to the departure of two random variables from independence. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of the data.

Correlation in fact, is a statistical measurement of the relationship between two variables. Possible correlations range from +1 to –1. A zero correlation indicates that there is no relationship between the variables. A correlation of –1 indicates a perfect negative correlation, meaning that as one variable goes up, the other goes down. A correlation of +1 indicates a perfect positive correlation, meaning that both variables move in the same direction together. Some of the basic assumptions of correlation are discussed as follows:

Mathematical properties

The correlation coefficient $\rho_{X,Y}$ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y},$$

where E is the expected value operator and cov means covariance. A widely used alternative notation is

$$\operatorname{corr}(X,Y) = \rho_{X,Y}.$$

Since $\mu_X = E(X)$, $\sigma_X^2 = E[(X - E(X))^2] = E(X^2) - E^2(X)$ and likewise for Y, we may also write

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\,\sqrt{E(Y^2) - E^2(Y)}}.$$

The correlation is defined only if both of the standard deviations are finite and both of them are nonzero. It is a corollary of the Cauchy-Schwarz inequality that the correlation cannot exceed 1 in absolute value.

The correlation is 1 in the case of an increasing linear relationship, −1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the variables. The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables.

If the variables are independent then the correlation is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables. Here is an example: Suppose the random variable X is uniformly distributed on the interval from −1 to 1, and $Y = X^2$. Then Y is completely determined by X, so that X and Y are dependent, but their correlation is zero; they are uncorrelated. However, in the special case when X and Y are jointly normal, uncorrelatedness is equivalent to independence.

A correlation between two variables is diluted in the presence of measurement error around estimates of one or both variables, in which case disattenuation provides a more accurate coefficient.

Geometric Interpretation of correlation

For centered data (i.e., data which have been shifted by the sample mean so as to have an average of zero), the correlation coefficient can also be viewed as the cosine of the angle between the two vectors of samples drawn from the two random variables.

Some practitioners prefer an uncentered (non-Pearson-compliant) correlation coefficient. See the example below for a comparison.

As an example, suppose five countries are found to have gross national products of 1, 2, 3, 5, and 8 billion dollars, respectively. Suppose these same five countries

(in the same order) are found to have 11%, 12%, 13%, 15%, and 18% poverty. Then let **x** and **y** be ordered 5-element vectors containing the above data: **x** = (1, 2, 3, 5, 8) and **y** = (0.11, 0.12, 0.13, 0.15, 0.18).

By the usual procedure for finding the angle between two vectors (see dot product), the *uncentered* correlation coefficient is:

$$\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|} = \frac{2.93}{\sqrt{103}\sqrt{0.0983}} = 0.920814711.$$

Note that the above data were deliberately chosen to be perfectly correlated: $y = 0.10 + 0.01\ x$. The Pearson correlation coefficient must therefore be exactly one. Centering the data (shifting **x** by E(**x**) = 3.8 and **y** by E(**y**) = 0.138) yields **x** = (−2.8, −1.8, −0.8, 1.2, 4.2) and **y** = (−0.028, −0.018, −0.008, 0.012, 0.042), from which

$$\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|} = \frac{0.308}{\sqrt{30.8}\sqrt{0.00308}} = 1 = \rho_{xy},$$

as expected.

1.2 THE SAMPLE CORRELATION

If we have a series of n measurements of X and Y written as x_i and y_i where i = 1, 2, ..., n, then the Pearson product-moment correlation coefficient can be used to estimate the correlation of X and Y. The Pearson coefficient is also known as the "sample correlation coefficient". The Pearson correlation coefficient is then the best estimate of the correlation of X and Y. The Pearson correlation coefficient is written:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}.$$

where $\bar{x}$and $\bar{y}$are the sample means of X and Y, s_x and s_y are the sample standard deviations of X and Y and the sum is from $i = 1$ to n. As with the population correlation, we may rewrite this as

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

Again, as is true with the population correlation, the absolute value of the sample correlation must be less than or equal to 1. Though the above formula conveniently suggests a single-pass algorithm for calculating sample correlations, it is notorious for its numerical instability (see below for something more accurate).

The square of the sample correlation coefficient, which is also known as the coefficient of determination, is the fraction of the variance in y_i that is accounted for by a linear fit of x_i to y_i. This is written

$$r_{xy}^2 = 1 - \frac{s_{y|x}^2}{s_y^2},$$

where $s_{ytx}{}^2$ is the square of the error of a linear regression of x_i on y_i by the equation $y = a + bx$:

$$s_{y|x}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - a - bx_i)^2,$$

and $s_y{}^2$ is just the variance of y:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Note that since the sample correlation coefficient is symmetric in x_i and y_i, we will get the same value for a fit of y_i to x_i:

$$r_{xy}^2 = 1 - \frac{s_{x|y}^2}{s_x^2}.$$

This equation also gives an intuitive idea of the correlation coefficient for higher dimensions. Just as the above described sample correlation coefficient is the fraction of variance accounted for by the fit of a 1-dimensional linear submanifold to a set of 2-dimensional vectors (x_i, y_i), so we can define a correlation coefficient for a fit of an m-dimensional linear submanifold to a set of n-dimensional vectors. For example, if we fit a plane $z = a + bx + cy$ to a set of data (x_i, y_i, z_i) then the correlation coefficient of z to x and y is

$$r^2 = 1 - \frac{s_{z|xy}^2}{s_z^2}.$$

The distribution of the correlation coefficient has been examined by R. A. Fisher[2][3] and A. K. Gayen.[4]

1.3 PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT

In statistics, the **Pearson product-moment correlation coefficient** (sometimes referred to as the **MCV** or **PMCC**, and typically denoted by r) is a common measure of the correlation (linear dependence) between two variables X and Y. It is very widely used in the sciences as a measure of the strength of linear dependence between two variables, giving a value somewhere between +1 and -1 inclusive. It was first introduced by Francis Galton in the 1880s, and named after Karl Pearson.[1]

In accordance with the usual convention, when calculated for an entire population, the Pearson product-moment correlation is typically designated by the analogous Greek letter, which in this case is ρ (rho). Hence its designation by the Latin letter r implies that it has been computed for a sample (to provide an estimate for that of the underlying population). For these reasons, it is sometimes called "Pearson's r."

The statistic is defined as the sum of the products of the standard scores of the two measures divided by the degrees of freedom.[2] If the data comes from a sample, then

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

where

$$\frac{X_i - \bar{X}}{s_X}, \bar{X}, \text{ and } s_X$$

are the standard score, sample mean, and sample standard deviation (calculated using $n - 1$ in the denominator).[2]

If the data comes from a population, then

$$\rho = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \mu_X}{\sigma_X}\right)\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right)$$

where

$$\frac{X_i - \mu_X}{\sigma_X}, \mu_X, \text{ and } \sigma_X$$

are the standard score, population mean, and population standard deviation (calculated using n in the denominator).

The result obtained is equivalent to dividing the covariance between the two variables by the product of their standard deviations.

Interpretation

The coefficient of correlation ranges from −1 to 1. A value of 1 shows that a linear equation describes the relationship perfectly and positively, with all data points lying on the same line and with Y increasing with X. A score of −1 shows

that all data points lie on a single line but that Y increases as X decreases. A value of 0 shows that a linear model is not needed – that there is no linear relationship between the variables.[2]

The linear equation that best describes the relationship between X and Y can be found by linear regression. This equation can be used to "predict" the value of one measurement from knowledge of the other. That is, for each value of X the equation calculates a value which is the best estimate of the values of Y corresponding the specific value. We denote this predicted variable by Y'.

Any value of Y can therefore be defined as the sum of Y' and the difference between Y and Y':

$$Y = Y' + (Y - Y').$$

The variance of Y is equal to the sum of the variance of the two components of Y:

$$s_y^2 = s_{y'}^2 + s_{y.x}^2.$$

Since the coefficient of determination implies that $s_{y.x}^2 = s_y^2(1 - r^2)$ we can derive the identity

$$r^2 = \frac{s_{y'}^2}{s_y^2}.$$

The square of r is conventionally used as a measure of the association between X and Y. For example, if r^2 is 0.90, then 90% of the variance of Y can be "accounted for" by changes in X and the linear relationship between X and Y.

Example 1

Let's calculate the correlation between Reading (X) and Spelling (Y) for the 10 students whose scores appeared in Table 3. There is a fair amount of calculation required as you can see from the table below. First we have to sum up the X

values (55) and then divide this number by the number of subjects (10) to find the mean for the X values (5.5). Then we have to do the same thing with the Y values to find their mean (10.3).

Correlation Between Reading and Spelling for Data in Table 3 Using the Definitional Formula					
Student	**Reading (X)**	**Spelling (Y)**	$X - \mu_X$	$Y - \mu_Y$	$(X - \mu_X)(Y - \mu_Y)$
1	3	11	a	0.7	-1.75
2	7	1	1.5	-9.3	-13.95
3	2	19	-3.5	8.7	-30.45
4	9	5	3.5	-5.3	-18.55
5	8	17	2.5	6.7	16.75
6	4	3	-1.5	-7.3	10.95
7	1	15	-4.5	4.7	-21.15
8	10	9	4.5	-1.3	-5.85
9	6	15	0.5	4.7	2.35
10	5	8	-0.5	-2.3	1.15
Sum	55	103	0.0	0.0	-60.5
Mean	5.5	10.3			
Standard Deviation	2.872	5.832			

Then we have to take each X score and subtract the mean from it to find the X deviation score. We can see that subject 1's X deviation score is -2.5, subject 2's X deviation score is 1.5 etc. We could make another column of the squares of the X deviation scores and sum up this column to use to calculate the standard deviation

of X using the definitional formula for the standard deviation of a population as we did in Lesson 6.

We can then find each subject's Y-deviation score. Subject 1's Y deviation score is 0.7 (11 - 10.3) and subject 2's Y deviation score is -9.3 (1 - 10.3). We could then add another column to square the Y-deviation scores and use the sum of this column to find the standard deviation for the Y scores.

We can then fill in the last column in which we multiply each subject's X deviation score times the same subject's Y deviation score. For subject 1 this is -1.75 (-2.5 times 0.7) and for subject 2 this is -13.95 (1.5 times -9.3). Finally if we sum up the last column (X deviation score times Y deviation score) we can use that quantity (-60.5), along with the standard deviations of the two variables and N, the number of subjects, to calculate the correlation coefficient.

$$r = \frac{\Sigma(X - \mu_x)(Y - \mu_y)}{N\sigma_x\sigma_y}$$

$$= \frac{-60.5}{(10)(2.872)(5.832)} = \frac{-60.5}{167.495} = -0.36$$

We have calculated the Pearson Product Moment Correlation Coefficient for the association between Reading and Spelling for the 10 subjects in Table 3. The correlation we obtained was -.36, showing us that there is a small negative correlation between reading and spelling. The correlation coefficient is a number that can range from -1 (perfect negative correlation) through 0 (no correlation) to 1 (perfect positive correlation).

You can see that it is fairly difficult to calculate the correlation coefficient using the definitional formula. In real practice we use another formula that is mathematically identical but is much easier to use. This is the computational or raw score formula for the correlation coefficient. The computational formula for the Pearsonian r is

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

By looking at the formula we can see that we need the following items to calculate r using the raw score formula:

1. The number of subjects, N
2. The sum of each subjects X score times the Y score, summation XY
3. The sum of the X scores, summation X
4. The sum of the Y scores, summation Y
5. The sum of the squared X scores, summation X squared
6. The sum of the squared Y scores, summation Y squared

Each of these quantities can be found as show in the computation table below:

Correlation Between Reading and Spelling for Data in Table 3 Using Computational Formula					
Student	**Reading (X)**	**Spelling (Y)**	X^2	Y^2	XY
1	3	11	9	121	33
2	7	1	49	1	7
3	2	19	4	361	38
4	9	5	81	25	45
5	8	17	64	289	136
6	4	3	16	9	12
7	1	15	1	225	15
8	10	9	100	81	90
9	6	15	36	225	90
10	5	8	25	64	40
Sum	55	103	385	1401	506

In we plug each of these sums into the raw score formula we can calculate the correlation coefficient.

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{(10)(506)-(55)(103)}{\sqrt{(10)(385)-(55)^2}\sqrt{(10)(1401)-(103)^2}}$$

$$= \frac{5060-5665}{\sqrt{3850-3025}\sqrt{14010-10609}} = \frac{-605}{\sqrt{825}\sqrt{3401}}$$

$$= \frac{-605}{(28.723)(58.318)} = \frac{-605}{1675.0679} = -0.36$$

We can see that we got the same answer for the correlation coefficient (-.36) with the raw score formula as we did with the definitional formula.

It is still computationally difficult to find the correlation coefficient, especially if we are dealing with a large number of subjects. In practice we would probably use a computer to calculate the correlation coefficient. We will consider just that (Using the Excel Spreadsheet Program to Calculate the Correlation Coefficient) after we have considered the Spearman Rank Order Correlation Coefficient.

1.4 REGRESSION

Regression Equation

Regression analysis is most often used for prediction. The goal in regression analysis is to create a mathematical model that can be used to predict the values of a dependent variable based upon the values of an independent variable. In other words, we use the model to predict the value of Y when we know the value of X. (The dependent variable is the one to be predicted). Correlation analysis is often used with regression analysis because correlation analysis is used to measure the strength of association between the two variables X and Y.

In regression analysis involving one independent variable and one dependent variable the values are frequently plotted in two demensions as a scatter plot. The scatter plot allows us to visually inspect the data prior to running a regression analysis. Often this step allows us to see if the relationship between the two variables is increasing or decreasing and gives only a rough idea of the relationship

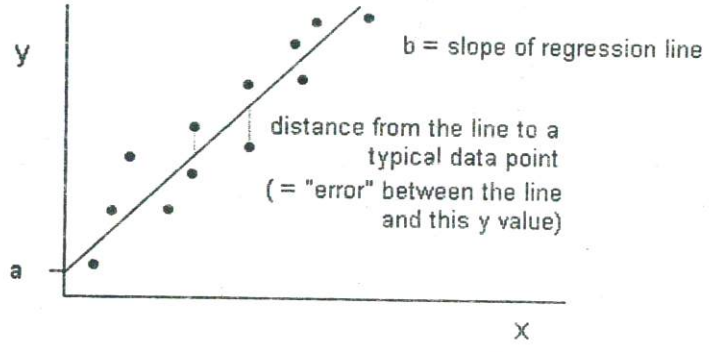
A regression equation allows us to express the relationship between two (or more) variables algebraically. It indicates the nature of the relationship between two (or more) variables. In particular, it indicates the extent to which you can predict some variables by knowing others, or the extent to which some are associated with others.

Regression analysis is any statistical method where the mean of one or more random variables is predicted based on other measured random variables [Wikipedia2006R]. There are two types of *regression analysis*, chosen according to whether the data approximate a straight line, when linear regression is used, or not, when non-linear regression is used.

A regression line is a line drawn through a scatterplot of two variables. The line is chosen so that it comes as close to the points as possible. *Regression analysis*, on the other hand, is more than curve fitting. It involves fitting a model with both deterministic and stochastic components. The deterministic component is called the predictor and the stochastic component is called the error term.

The simplest form of a regression model contains a dependent variable, also called the "Y-variable" and a single independent variable, also called the "X-variable".

It's customary to use "a" or "alpha" for the intercept of the line, and "b" or "beta" for the slope; so linear regression gives you a formula of the form: $y = bx + a$



1.5 LINEAR REGRESSION

Introduction to linear regression

Linear regression analyzes the relationship between two variables, X and Y. For each subject (or experimental unit), you know both X and Y and you want to find the best straight line through the data. In some situations, the slope and/or intercept have a scientific meaning. In other cases, you use the linear regression line as a standard curve to find new values of X from Y, or Y from X.

The term "regression", like many statistical terms, is used in statistics quite differently than it is used in other contexts. The method was first used to examine the relationship between the heights of fathers and sons. The two were related, of course, but the slope is less than 1.0. A tall father tended to have sons shorter than himself; a short father tended to have sons taller than himself. The height of sons regressed to the mean. The term "regression" is now used for many sorts of curve fitting.

Prism determines and graphs the best-fit linear regression line, optionally including a 95% confidence interval or 95% prediction interval bands. You may also force the line through a particular point (usually the origin), calculate residuals, calculate a runs test, or compare the slopes and intercepts of two or more regression lines.

In general, the goal of linear regression is to find the line that best predicts Y from X. Linear regression does this by finding the line that minimizes the sum of the squares of the vertical distances of the points from the line.

Note that linear regression does not *test* whether your data are linear (except via the runs test). It assumes that your data are linear, and finds the slope and intercept that make a straight line best fit your data.

Minimizing sum-of-squares

The goal of linear regression is to adjust the values of slope and intercept to find the line that best predicts Y from X. More precisely, the goal of regression is to minimize the sum of the squares of the vertical distances of the points from the line. Why minimize the sum of the squares of the distances? Why not simply minimize the sum of the actual distances?

If the random scatter follows a Gaussian distribution, it is far more likely to have two medium size deviations (say 5 units each) than to have one small deviation (1 unit) and one large (9 units). A procedure that minimized the sum of the absolute value of the distances would have no preference over a line that was 5 units away from two points and one that was 1 unit away from one point and 9 units from another. The sum of the distances (more precisely, the sum of the absolute value of the distances) is 10 units in each case. A procedure that minimizes the sum of the squares of the distances prefers to be 5 units away from two points (sum-of-squares = 50) rather than 1 unit away from one point and 9 units away from another (sum-of-squares = 82). If the scatter is Gaussian (or nearly so), the line determined by minimizing the sum-of-squares is most likely to be correct.

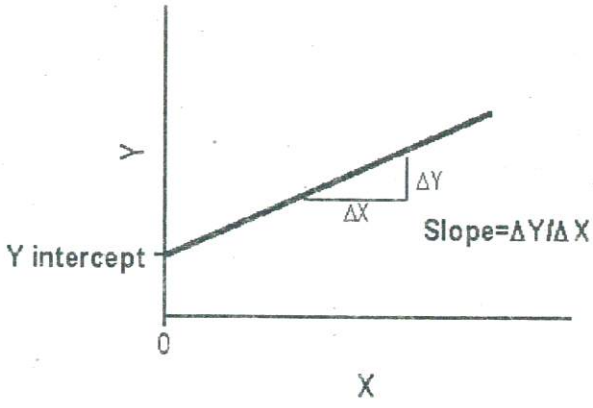
The calculations are shown in every statistics book, and are entirely standard.

Slope and intercept

Prism reports the best-fit values of the slope and intercept, along with their standard errors and confidence intervals.

The slope quantifies the steepness of the line. It equals the change in Y for each unit change in X. It is expressed in the units of the Y-axis divided by the units of the X-axis. If the slope is positive, Y increases as X increases. If the slope is negative, Y decreases as X increases.

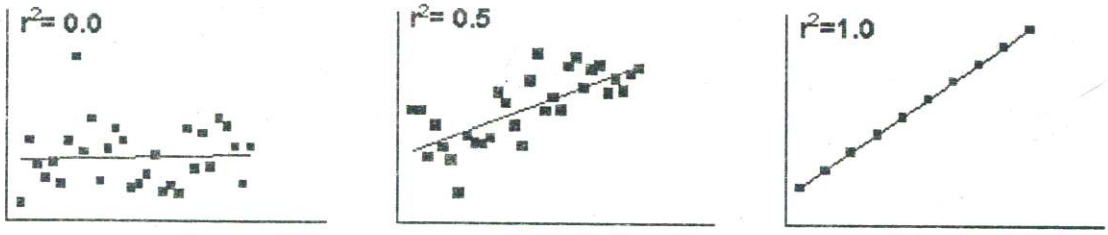
The Y intercept is the Y value of the line when X equals zero. It defines the elevation of the line.



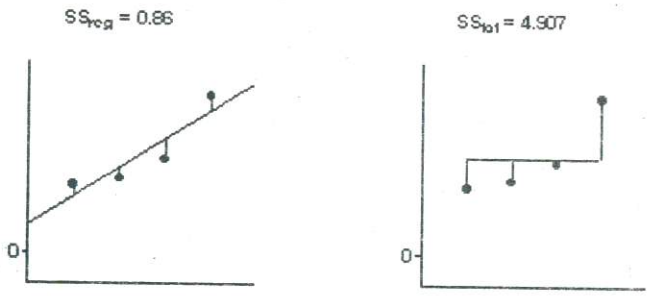
The standard error values of the slope and intercept can be hard to interpret, but their main purpose is to compute the 95% confidence intervals. If you accept the assumptions of linear regression, there is a 95% chance that the 95% confidence interval of the slope contains the true value of the slope, and that the 95% confidence interval for the intercept contains the true value of the intercept.

r2, a measure of goodness-of-fit of linear regression

The value r2 is a fraction between 0.0 and 1.0, and has no units. An r2 value of 0.0 means that knowing X does not help you predict Y. There is no linear relationship between X and Y, and the best-fit line is a horizontal line going through the mean of all Y values. When r2 equals 1.0, all points lie exactly on a straight line with no scatter. Knowing X lets you predict Y perfectly.



This figure demonstrates how Prism computes r2.



$$r^2 = 1 - \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{0.86}{4.91} = 0.83$$

The left panel shows the best-fit linear regression line This lines minimizes the sum-of-squares of the vertical distances of the points from the line. Those vertical distances are also shown on the left panel of the figure. In this example, the sum of squares of those distances (SSreg) equals 0.86. Its units are the units of the Y-axis squared. To use this value as a measure of goodness-of-fit, you must compare it to something.

The right half of the figure shows the null hypothesis -- a horizontal line through the mean of all the Y values. Goodness-of-fit of this model (SStot) is also calculated as the sum of squares of the vertical distances of the points from the line, 4.907 in this example. The ratio of the two sum-of-squares values compares the regression model with the null hypothesis model. The equation to compute r2 is shown in the figure. In this example r2 is 0.8248. The regression model fits the data much better than the null hypothesis, so SSreg is much smaller than SStot, and r2 is near 1.0. If the regression model were not much better than the null hypothesis, r2 would be near zero.

You can think of r2 as the fraction of the total variance of Y that is "explained" by variation in X. The value of r2 (unlike the regression line itself) would be the same if X and Y were swapped. So r2 is also the fraction of the variance in X that is "explained" by variation in Y. In other words, r2 is the fraction of the variation that is shared between X and Y.

In this example, 84% of the total variance in Y is "explained" by the linear regression model. That leaves the rest of the vairance (16% of the total) as variability of the data from the model (SStot)

Example 2

The following data set gives the average heights and weights for American women aged 30–39 (source: *The World Almanac and Book of Facts, 1975*).

Height (m)	1.47	1.5	1.52	1.55	1.57	1.60	1.63	1.65	1.68	1.7	1.73	1.75	1.78	1.8	1.83
Weight (kg)	52.21	53.12	54.48	55.84	57.2	58.57	59.93	61.29	63.11	64.47	66.28	68.12	69.92	72.19	74.46

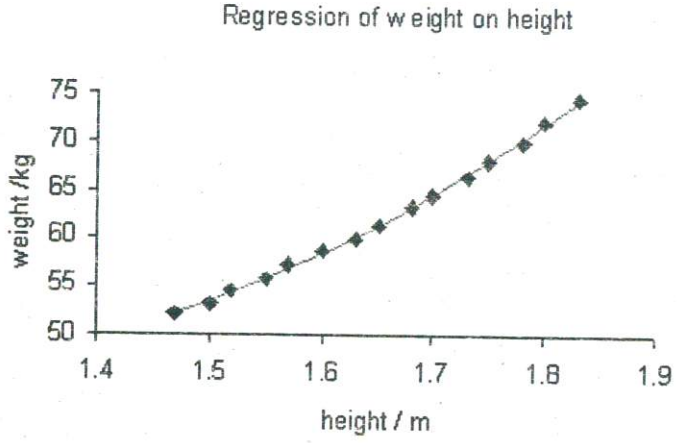
A plot of weight against height (see below) shows that it cannot be modeled by a straight line, so a regression is performed by modeling the data by a parabola.

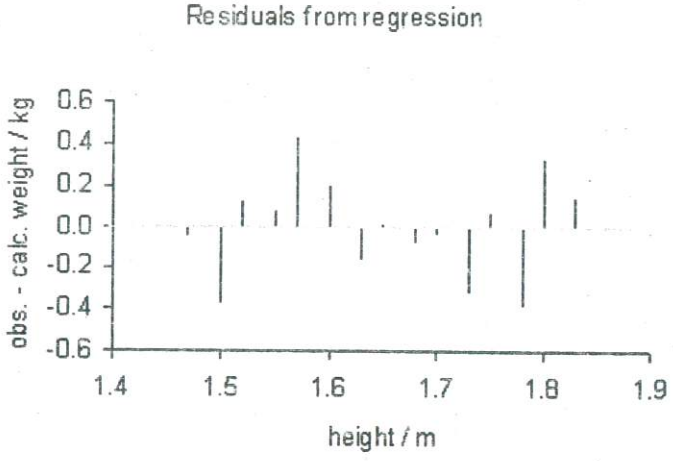
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

where the dependent variable Y_i is weight and the independent variable X_i is height.

Place the observations $x_i,\ x_i^2,\ i = 1, \ldots, n$, in the matrix **X**.

$$= \begin{pmatrix} 1 & 1.47 & 2.16 \\ 1 & 1.50 & 2.25 \\ 1 & 1.52 & 2.31 \\ 1 & 1.55 & 2.40 \\ 1 & 1.57 & 2.46 \\ 1 & 1.60 & 2.56 \\ 1 & 1.63 & 2.66 \\ 1 & 1.65 & 2.72 \\ 1 & 1.68 & 2.82 \\ 1 & 1.70 & 2.89 \\ 1 & 1.73 & 2.99 \\ 1 & 1.75 & 3.06 \\ 1 & 1.78 & 3.17 \\ 1 & 1.80 & 3.24 \\ 1 & 1.83 & 3.35 \end{pmatrix}$$





The values of the parameters are found by solving the normal equations

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Element ij of the normal equation matrix, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$is formed by summing the products of column i and column j of $\mathbf{X}$.

$$x_{ij} = \sum_{k=1}^{15} x_{ki}x_{kj}$$

Element i of the right-hand side vector $\mathbf{X}^{\mathrm{T}}\mathbf{y}$is formed by summing the products of column i of $\mathbf{X}$ with the column of dependent variable values.

$$(\mathbf{X}^{\mathrm{T}}\mathbf{y})_i = \sum_{k=1}^{15} x_{ki}y_k$$

Thus, the normal equations are

$$\begin{pmatrix} 15 & 24.76 & 41.05 \\ 24.76 & 41.05 & 68.37 \\ 41.05 & 68.37 & 114.35 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 931 \\ 1548 \\ 2586 \end{pmatrix}$$

$\beta_0 = 129 \pm 16$ (value $\pm$standard deviation)

$\beta_1 = -143 \pm 20$

$\beta_2 = 62 \pm 6$

The calculated values are given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$$

The observed and calculated data are plotted together and the residuals, $\overline{y_i - \hat{y}_i}$, are calculated and plotted. Standard deviations are calculated using the sum of squares, $S = 0.76$.

1.6 STANDARD ERROR OF THE ESTIMATE

The standard error of the estimate is a measure of the accuracy of predictions made with a regression line.

Example 2

Consider the following data.

X	Y	Y'	Y-Y'	$(Y-Y')^2$
3.25	18.71	17.79	0.92	0.86
3.96	18.15	20.11	-1.96	3.83
4.35	19.72	21.38	-1.66	2.77
4.40	23.02	21.55	1.47	2.17
4.42	22.26	21.61	0.65	0.42
4.51	19.61	21.91	-2.30	5.28
4.87	27.74	23.09	4.65	21.67
5.65	24.89	25.64	-0.75	0.56
5.68	27.83	25.74	2.09	4.39
5.71	23.09	25.83	-2.74	7.53
6.28	24.25	27.70	-3.45	11.89
6.52	31.55	28.48	3.07	9.40
				$\Sigma = 70.77$

The second column (Y) is predicted by the first column (X). The slope and Y intercept of the regression line are 3.2716 and 7.1526 respectively. The third column, (Y'), contains the predictions and is computed according to the formula:

Y' = 3.2716X + 7.1526.

The fourth column (Y-Y') is the error of prediction. It is simply the difference between what a subject's actual score was (Y) and what the predicted score is (Y').

The sum of the errors of prediction is zero. The last column, $(Y-Y')^2$, contains the squared errors of prediction.

The regression line seeks to minimize the sum of the squared errors of prediction. The square root of the average squared error of prediction is used as a measure of the accuracy of prediction. This measure is called the standard error of the estimate and is designated as σ_{est}. The formula for the standard error of the estimate is:

$$\sigma_{est} = \sqrt{\frac{\sum(Y-Y')^2}{N}}$$

where N is the number of pairs of (X,Y) points. For this example, the sum of the squared errors of prediction (the numerator) is 70.77 and the number of pairs is 12. The standard error of the estimate is therefore equal to:

$$\sigma_{est} = \sqrt{\frac{70.77}{12}} = 2.43.$$

An alternate formula for the standard error of the estimate is:

$$\sigma_{est} = \sigma_y \sqrt{1-\rho^2}$$

where is σ_y is the population standard deviation of Y and ρ is the population correlation between X and Y. For this example,

$$\sigma_{est} = 3.95\sqrt{-0.788^2} = 2.43.$$

One typically does not know the population parameters and therefore has to estimate from a sample. The symbol s_{est} is used for the estimate of σ_{est}. The relevant formulas are:

$$s_{est} = \sqrt{\frac{\sum(Y-Y')^2}{N-2}}$$

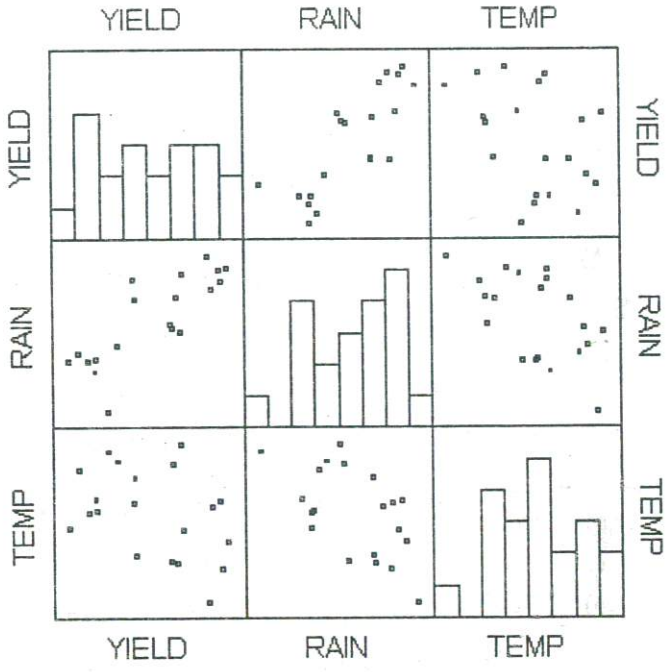
and

$$s_{est} = S_y\sqrt{1-r^2}\sqrt{\frac{N}{N-2}}$$

where r is the sample correlation and S_y is the sample standard deviation of Y. (Note that S_y has a capital rather than a small "S" so it is computed with N in the denominator). The similarity between the standard error of the estimate and the standard deviation should be noted: The standard deviation is the square root of the average squared deviation from the mean; the standard error of the estimate is the square root of the average squared deviation from the regression line. Both statistics are measures of unexplained variation.

1.7 PARTIAL CORRELATION

The partial correlation between X and Y given a set of n controlling variables $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_n\}$, written $\rho_{XY \cdot \mathbf{Z}}$, is the correlation between the residuals R_X and R_Y resulting from the linear regression of X with $\mathbf{Z}$ and of Y with $\mathbf{Z}$, respectively. In fact, the first-order partial correlation is nothing else than a difference between a correlation and the product of the removable correlations divided by the product of the coefficients of alienation of the removable correlations.



Scatterplots, correlation coefficients, and simple linear regression coefficients are inter-related. The scattterplot shows the data. The correlation coefficient measures of linear association between the variables. The regression coefficient describes the linear association through a number that gives the expected change in the response per unit change in the predictor.

The coefficients of a multiple regression equation give the change in response per unit change in a predictor when all other predictors are held fixed. This raises the question of whether there are analogues to the correlation coefficient and the scatterplot to summarize the relation and display the data after adjusting for the effects of other variables.

This note answers these questions and illustrates them by using the crop yield example of Hooker reported by Kendall and Stuart in volume 2 of their *Advanced*

Theory of Statistics, Vol, 2, 3 rd ed.(example 27.1) Neither Hooker nor Kendall & Stuart provide the raw data, so I have generated a set of random data with means, standard deviations, and correlations identical to those given in K&S. These statistics are sufficient for all of the methods that will be discussed here (*sufficient* is a technical term meaning nothing else to do with the data has any effect on the analysis. All data sets with the same values of the sufficient statistics are equivalent for our purposes), so the random data will be adequate.

The variables are yields of "seeds' hay" in cwt per acre, spring rainfall in inches and the accumulated temperature above 42 F in the spring for an English area over 20 years. The plots suggest yield and rainfall are positively correlated, while yield and temperature are negatively correlated! This is borne out by the correlation matrix itself.

Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

	YIELD	RAIN	TEMP
YIELD	1.00000	0.80031 <.0001	-0.39988 0.0807
RAIN	0.80031 <.0001	1.00000	-0.55966 0.0103
TEMP	-0.39988 0.0807	-0.55966 0.0103	1.00000

Just as the simple correlation coefficient between Y and X describes their joint behavior, the partial correlation describes the behavior of Y and X_1 when $X_2..X_p$ are held fixed. The partial correlation between Y and X_1 holding $X_2..X_p$ fixed is denoted $r_{X_1 Y \cdot X_2 \ldots X_p}$ or $r_{X_1 Y | X_2 \ldots X_p}$.

A partial correlation coefficient can be written in terms of simple correlation coefficients

$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

Thus, $r_{XY|Z} = r_{XY}$ if X & Y are both uncorrelated with Z.

A partial correlation between two variables can differ substantially from their simple correlation. Sign reversals are possible, too. For example, the partial correlation between YIELD and TEMPERATURE holding RAINFALL fixed is 0.09664. While it does not reach statistical significance (P = 0.694), the sample value is positive nonetheless.

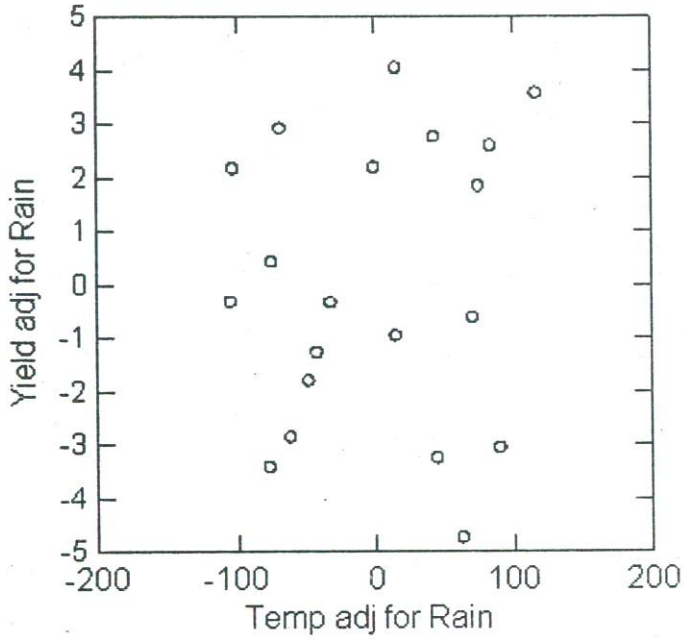
The partial correlation between X & Y holding a set of variables fixed will have the same sign as the multiple regression coefficient of X when Y is regressed on X and the set of variables being held fixed. Also,

$$r_{XY \cdot list} = \frac{t}{\sqrt{t^2 + \text{Res } df}}$$

where *t* is the t statistic for the coefficient of X in the multiple regression of Y on X and the variables in the list.

Just as the simple correlation coefficient describes the data in an ordinary scatterplot, the partial correlation coefficient describes the data in the partial regression residual plot.

Let Y and X_1 be the variables of primary interest and let $X_2..X_p$ be the variables held fixed. First, calculate the residuals after regressing Y on $X_2..X_p$. These are the parts of the Ys that cannot be predicted by $X_2..X_p$. Then, calculate the residuals after regressing X_1 on $X_2..X_p$. These are the parts of the X_1s that cannot be predicted by $X_2..X_p$. The partial correlation coefficient between Y and X_1 adjusted for $X_2..X_p$ is the correlation between these two sets of residuals. Also, the regression coefficient when the Y residuals are regressed on the X_1 residuals is equal to the regression coefficient of X_1 in the multiple regression equation when Y is regressed on the entire set of predictors.



For example, the partial correlation of YIELD and TEMP adjusted for RAIN is the correlation between the residuals from regressing YIELD on RAIN and the residuals from regressing TEMP on RAIN. In this partial regression residual plot, the correlation is 0.09664. The regression coefficient of TEMP when the YIELD residuals are regessed on the TEMP residuals is 0.003636. The multiple regression equation for the original data set is

YIELD = 9.298850 + 3.373008 RAIN + 0.003636 TEMP

Because the data are residuals, they are centered around zero. The values, then, are not similar to the original values. However, perhaps this is an advantage. It stops them from being misinterpreted as Y or X_1 values "adjusted for $X_2..X_p$".

While the regression of Y on $X_2..X_p$ seems reasonable, it is not uncommon to hear questions about adjusting X_1, that is, some propose comparing the residuals of Y on $X_2..X_p$ with X_1directly.

This approach has been suggested many times over the years. Lately, it has been used in the field of nutrition by Willett and Stampfer (AJE, 124(1986):17-22) to produce "calorie-adjusted nutrient intakes", which are the residuals obtained by regressing nutrient intakes on total energy intake. These adjusted intakes are used as predictors in other regression equations. However, total energy intake does not appear in the equations and the response is not adjusted for total energy intake. Willett and Stampfer recognize this, but propose using calorie-adjusted intakes nonetheless. They suggest "calorie-adjusted values in multivariate models will overcomethe problem of high-collinearity frequently observed between nutritional factors", but this is just an artifact of adjusting only some of the factors. The correlation between an adjusted factor and an unadjusted factor is always smaller in magnitude than the correlation between two adjusted factors.

This method was first proposed before the ready availability of computers as a way to approximate multiple regression with two independent variables (regress Y on X1, regress the residuals on X2) and was given the name two-stage regression. Today, however, it is a mistake to use the approximation when the correct answer is easily obtained. If the goal is to report on two variables after adjusting for the effects of another set of variables, then both variables must be adjusted.

1.8 MULTIPLE CORRELATION

In statistics, regression analysis is a method for explanation of phenomena and prediction of future events. In the regression analysis, a coefficient of correlation r between random variables X and Y is a quantitative index of association between these two variables. In its squared form, as a coefficient of determination r^2, indicates the amount of variance in the criterion variable Y that is accounted for by the variation in the predictor variable X. In the multiple regression analysis, the set of predictor variables X_1, X_2, ... is used to explain variability of the criterion variable Y. A multivariate counterpart of the coefficient of determination r^2 is the *coefficient of multiple determination*, R^2. The square root of the coefficient of multiple determination is the **coefficient of multiple correlation, *R***.

Let Y be one variable, and (X_1, X_2 , ..., X_n) a set of other variables. Let X be a linear combination of the X_i's :

$$X = \sum_i a_i X_i$$

and consider the correlation coefficient $\rho(X, Y)$.

When the coefficients a_i are made to vary in every possible way, the value of ρ changes. It can be shown that, in general, there is a single set of values of the coefficients that maximizes ρ. This largest possible value of $\rho(X, Y)$ is usually denoted **R**, and is called the **Multiple Correlation Coefficient** between Y and the set of variables (X_1, X_2 , ..., X_n).

The Multiple Correlation Coefficient plays a central role in Multiple Linear Regression, as $\mathbf{R^2}$ is then equal to the ratio of the explained variance to the total variance, and is therefore a measure of the quality of the regression. So, in this respect, there is a complete similarity between Simple and Multiple Linear Regression.

Conceptualization of multiple correlation

An intuitive approach to the multiple regression analysis is to sum the squared correlations between the predictor variables and the criterion variable to obtain an index of the over-all relationship between the predictor variables and the criterion variable. However, such a sum is often greater than one, suggesting that simple summation of the squared coefficients of correlations is not a correct procedure to employ. In fact, a simple summation of squared coefficients of correlations between the predictor variables and the criterion variable is the correct procedure, but only in the special case when the predictor variables are not correlated. If the predictors are related, their inter-correlations must be removed so that only the unique contributions of each predictor toward explanation of the criterion.

Fundamental equation of multiple regression analysis

Initially, a matrix of correlations R is computed for all variables involved in the analysis. This matrix can be conceptualized as a supermatrix, consisting of the vector of cross-correlations between the predictor variables and the criterion variable c, its transpose c' and the matrix of intercorrelations between predictor variables R_{xx}. The fundamental equation of the multiple regression analysis is

$$R^2 = c' R_{xx}^{-1} c.$$

The expression on the left side signifies the coefficient of multiple determination (squared coefficient of multiple correlation). The expressions on the right side are the transposed vector of cross-correlations c', the matrix of inter-correlations R_{xx} to be inverted (cf., matrix inversion), and the vector of cross-correlations, c. The premultiplication of the vector of cross-correlations by its transpose changes the coefficients of correlation into coefficients of determination. The inverted matrix of the inter-correlations removes the redundant variance from the of inter-correlations of the predictor set of variables. These not-redundant cross-correlations are summed to obtain the multiple coefficient of determination R^2. The square root of this coefficient is the coefficient of multiple correlations R.

Activity 1

1. with the following data in 6 cities calculate the coefficient of correlation by Pearson's method between the density of population and death rate:

City	Area in kilometres	Population in '000	No. of deaths
A	150	30	300
B	180	90	1440
C	100	40	560
D	60	42	840

E	120	72	1224
F	80	24	312

2. Obtain regression equation of Y and X and estimate Y when X=55 from the following:

X:	40	50	38	60	65	50	35
Y:	38	60	55	70	60	48	30

3. Heights of fathers and sons are given below. Find height of the son when the height of father is 70 inches.

Father (inches):	71	68	66	67	70	71	70	73	72
		65	66						
Son (inches):	69	64	65	63	65	62	65	64	66
		59	62						

1.9 SUMMARY

In this chapter the concept of correlation or the association between two variables has been discussed. It has been made clear that the value of Pearson correlation coefficient r quantifies this association. The correlation coefficient r may assume values between -1 and 1.

Further the concepts of regression and linear regression was also discussed in this chapter. Broadly speaking, the fitting of any chosen mathematical function to given data is termed as regression analysis. The estimation of the parameters of this model is accomplished by the least squares criterion which tries to minimize the sum of squares for all the data points. There are simultaneous linear equations equal in number to the number of parameters to be estimated, obtained by partially differentiating the sum of squares of errors with respect to the individual parameters.

Finally the concepts of partial and multiple correlations have been discussed with the help of suitable examples.

1.10 FURTHER READINGS

- Altman D.G. (1991) Practical Statistics. Chapman & Hall, London.
- Campbell M.J. & Machin D. (1993) Medical Statistics a Commonsense Approach. 2nd edn.Wiley, London.
- Draper, N and H.Smith, 1966. Applied Regression Analysis. John Willey: New York.
- Edwards, B. 1980. The Readable Maths and Statistics Book, George Allen and Unwin. London

UNIT 2

BASIC CONCEPTS OF PROBABILITY

Objectives

After reading this unit, you should be able to:

- Appreciate the relevance of probability theory in decision making
- Understand the different approaches of probability
- Identify the basic concepts of probability including experiments, sample space and events.
- Calculate probabilities using random variable approach
- Have deep understanding of various kind of generating functions.

Structure

2.1 INTRODUCTION

Probability, or chance, is a way of expressing knowledge or belief that an event will occur or has occurred. In mathematics the concept has been given an exact meaning in probability theory, that is used extensively in such areas of study as mathematics, statistics, finance, gambling, science, and philosophy to draw conclusions about the likelihood of potential events and the underlying mechanics of complex systems.

Probability theory in fact, is the branch of mathematics concerned with analysis of random phenomena. The central objects of probability theory are random variables, stochastic processes, and events: mathematical abstractions of non-deterministic events or measured quantities that may either be single occurrences or evolve over time in an apparently random fashion. Although an individual coin toss or the roll of a die is a random event, if repeated many times the sequence of random events will exhibit certain statistical patterns, which can be studied and predicted. Two representative mathematical results describing such patterns are the law of large numbers and the central limit theorem.

As a mathematical foundation for statistics, probability theory is essential to many human activities that involve quantitative analysis of large sets of data. Methods of probability theory also apply to descriptions of complex systems given only partial knowledge of their state, as in statistical mechanics. A great discovery of twentieth century physics was the probabilistic nature of physical phenomena at atomic scales, described in quantum mechanics..

Discrete probability distributions

Main article: Discrete probability distribution

Discrete probability theory deals with events that occur in countable sample spaces.

Examples: Throwing dice, experiments with decks of cards, and random walk.

Classical definition: Initially the probability of an event to occur was defined as number of cases favorable for the event, over the number of total outcomes possible in an equiprobable sample space.

For example, if the event is "occurrence of an even number when a die is rolled", the probability is given by $\frac{3}{6} = \frac{1}{2}$, since 3 faces out of the 6 have even numbers and each face has the same probability of appearing.

Modern definition: The modern definition starts with a set called the **sample space**, which relates to the set of all *possible outcomes* in classical sense, denoted by $\Omega = \{x_1, x_2, \ldots\}$. It is then assumed that for each element $x \in \Omega$, an intrinsic "probability" value $f(x)$is attached, which satisfies the following properties:

1. $f(x) \in [0, 1]$ for all $x \in \Omega$;
2. $\sum_{x \in \Omega} f(x) = 1$.

That is, the probability function *f*(*x*) lies between zero and one for every value of x in the sample space Ω, and the sum of *f*(*x*) over all values x in the sample space Ω is exactly equal to 1. An **event** is defined as any subset $E$of the sample space Ω. The **probability** of the event Edefined as

$$P(E) = \sum_{x \in E} f(x).$$

So, the probability of the entire sample space is 1, and the probability of the null event is 0.

The function $f(x)$mapping a point in the sample space to the "probability" value is called a **probability mass function** abbreviated as **pmf**. The modern definition does not try to answer how probability mass functions are obtained; instead it builds a theory that assumes their existence.

Continuous probability distributions

Main article: Continuous probability distribution

Continuous probability theory deals with events that occur in a continuous sample space.

Classical definition: The classical definition breaks down when confronted with the continuous case. See Bertrand's paradox.

Modern definition: If the outcome space of a random variable X is the set of real numbers ($\mathbb{R}$) or a subset thereof, then a function called the **cumulative distribution function** (or **cdf**) F exists, defined by $F(x) = P(X \leq x)$. That is, $F(x)$ returns the probability that X will be less than or equal to x.

The cdf necessarily satisfies the following properties.

1. F is a monotonically non-decreasing, right-continuous function;
2. $\lim_{x \to -\infty} F(x) = 0$;
3. $\lim_{x \to \infty} F(x) = 1$.

If F is absolutely continuous, i.e., its derivative exists and integrating the derivative gives us the cdf back again, then the random variable X is said to have a **probability density function** or **pdf** or simply **density** $f(x) = \frac{dF(x)}{dx}$.

Measure-theoretic probability theory

The *raison d'être* of the measure-theoretic treatment of probability is that it unifies the discrete and the continuous, and makes the difference a question of which measure is used. Furthermore, it covers distributions that are neither discrete nor continuous nor mixtures of the two.

An example of such distributions could be a mix of discrete and continuous distributions, for example, a random variable which is 0 with probability 1/2, and takes a random value from a normal distribution with probability 1/2. It can still

be studied to some extent by considering it to have a pdf of $(\delta[x] + \varphi(x))/2$, where $\delta[x]$ is the Dirac delta function.

Other distributions may not even be a mix, for example, the Cantor distribution has no positive probability for any single point, neither does it have a density. The modern approach to probability theory solves these problems using measure theory to define the probability space:

Given any set Ω, (also called **sample space**) and a σ-algebra $\mathcal{F}$ on it, a measure P defined on $\mathcal{F}$ is called a **probability measure** if $P(\Omega) = 1$.

If $\mathcal{F}$ is the Borel σ-algebra on the set of real numbers, then there is a unique probability measure on $\mathcal{F}$ for any cdf, and vice versa. The measure corresponding to a cdf is said to be **induced** by the cdf. This measure coincides with the pmf for discrete variables, and pdf for continuous variables, making the measure-theoretic approach free of fallacies.

The *probability* of a set E in the σ-algebra $\mathcal{F}$ is defined as

$$P(E) = \int_{\omega \in E} \mu_F(d\omega)$$

where the integration is with respect to the measure μ_F induced by F.

Along with providing better understanding and unification of discrete and continuous probabilities, measure-theoretic treatment also allows us to work on probabilities outside $\mathbb{R}^n$, as in the theory of stochastic processes. For example to study Brownian motion, probability is defined on a space of functions.

Mathematical treatment

In mathematics, a probability of an event *A* is represented by a real number in the range from 0 to 1 and written as P(*A*), p(*A*) or Pr(*A*). An impossible event has a probability of 0, and a certain event has a probability of 1. However, the converses are not always true: probability 0 events are not always impossible, nor

probability 1 events certain. The rather subtle distinction between "certain" and "probability 1" is treated at greater length in the article on "almost surely".

The *opposite* or *complement* of an event A is the event [not A] (that is, the event of A not occurring); its probability is given by P(not A) = 1 - P(A). As an example, the chance of not rolling a six on a six-sided die is 1 - (chance of rolling a six) = $1 - \frac{1}{6} = \frac{5}{6}$. See Complementary event for a more complete treatment.

If both the events A and B occur on a single performance of an experiment this is called the intersection or joint probability of A and B, denoted as $P(A \cap B)$. If two events, A and B are independent then the joint probability is

$$P(A \text{ and } B) = P(A \cap B) = P(A)P(B),$$

for example, if two coins are flipped the chance of both being heads is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

If either event A or event B or both events occur on a single performance of an experiment this is called the union of the events A and B denoted as $P(A \cup B)$. If two events are mutually exclusive then the probability of either occurring is

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B).$$

For example, the chance of rolling a 1 or 2 on a six-sided die is $P(1 \text{ or } 2) = P(1) + P(2) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

If the events are not mutually exclusive then

$$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \text{ and } B).$$

For example, when drawing a single card at random from a regular deck of cards, the chance of getting a heart or a face card (J,Q,K) (or one that is both) is $\frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{11}{26}$, because of the 52 cards of a deck 13 are hearts, 12 are face cards, and 3 are both: here the possibilities included in the "3 that are both" are

included in each of the "13 hearts" and the "12 face cards" but should only be counted once.

Conditional probability is the probability of some event A, given the occurrence of some other event B. Conditional probability is written $P(A|B)$, and is read "the probability of A, given B". It is defined by

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

If $P(B) = 0$ then $P(A \mid B)$ is undefined.

Above discussed treatments will be discussed deeply in next chapter.

2.2 THE CLASSICAL DEFINITION OF PROBABILITY

The classical definition of probability is identified with the works of Pierre Simon Laplace. As stated in his Théorie analytique des probabilités,

"The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible".

This definition is essentially a consequence of the principle of indifference. If elementary events are assigned equal probabilities, then the probability of a disjunction of elementary events is just the number of events in the disjunction divided by the total number of elementary events.
The classical definition of probability was called into question by several writers of the nineteenth century, including John Venn and George Boole. The frequentist definition of probability became widely accepted as a result of their criticism, and especially through the works of R.A. Fisher. The classical definition enjoyed a revival of sorts due to the general interest in Bayesian probability

The classical definition of probability came from the gambling games theory: the ratio of a number of successful outcomes to the number of all possible outcomes, presuming they are all equally likely. Imagine a certain experiment that can equally likely produce some known finite number n of outcomes, for example, a die is rolled, n = 6. Imagine also that we bet on some kind of outcomes, for example, that a die comes up an even number, here 2, 4, or 6. Intuitively, the probability P of us winning is the number of k successful outcomes divided by the total number of outcomes n: P = k/n

2.3 THE EMPIRICAL DEFINITION OF PROBABILITY

Empirical probability, also known as **relative frequency**, or **experimental probability**, is the ratio of the number favorable outcomes to the total number of trials, not in a sample space but in an actual sequence of experiments. In a more general sense, empirical probability estimates probabilities from experience and observation. The phrase **a posteriori probability** has also been used an alternative to empirical probability or relative frequency. This unusual usage of the phrase is not directly related to Bayesian inference and not to be confused with its equally occasional use to refer to posterior probability, which is something else.

In statistical terms, the empirical probability is an estimate of a probability. If modelling using a binomial distribution is appropriate, it is the maximum likelihood estimate. It is the Bayesian estimate for the same case if certain assumptions are made for the prior distribution of the probability

An advantage of estimating probabilities using empirical probabilities is that this procedure is relatively free of assumptions. For example, consider estimating the probability among a population of men that they satisfy two conditions: (i) that they are over 6 feet in height; (ii) that they prefer strawberry jam to raspberry jam. A direct estimate could be found by counting the number of men who satisfy both conditions to give the empirical probability the combined condition. An alternative estimate could be found by multiplying the proportion of men who are over 6 feet in height with the proportion of men who prefer strawberry jam to

raspberry jam, but this estimate relies on the assumption that the two conditions are statistically independent.

The classical interpretation of probability is a theoretical probability based on the physics of the experiment, but does not require the experiment to be performed. For example, we know that the probability of a balanced coin turning up heads is equal to 0.5 without ever performing trials of the experiment. Under the classical interpretation, the probability of an event is defined as the ratio of the number of outcomes favorable to the event divided by the total number of possible outcomes. Sometimes a situation may be too complex to understand the physical nature of it well enough to calculate probabilities. However, by running a large number of trials and observing the outcomes, we can estimate the probability. This is the empirical probability based on long-run relative frequencies and is defined as the ratio of the number of observed outcomes favorable to the event divided by the total number of observed outcomes. The larger the number of trials, the more accurate the estimate of probability. If the system can be modeled by computer, then simulations can be performed in place of physical trials.

A manager frequently faces situations in which neither classical nor empirical probabilities are useful. For example, in a one-shot situation such as the launch of a unique product, the probability of success can neither be calculated nor estimated from repeated trials. However, the manager may make an educated guess of the probability. This subjective probability can be thought of as a person's degree of confidence that the event will occur. In absence of better information upon which to rely, subjective probability may be used to make logically consistent decisions, but the quality of those decisions depends on the accuracy of the subjective estimate.

2.4 BASIC CONCEPTS: EXPERIMENTS, SAMPLE SPACE AND EVENT

2.4.1 EXPERIMENT

Any activity that yields a result or an outcome is called as experiment. There are two types of experiment we observe in our natural phenomenon.

1. Deterministic Experiments

2. Non-deterministic Experiments (or Random Experiments)

Deterministic Experiments

The experiment in which the outcome can be predicted in advance under essentially homogeneous conditions is known as deterministic experiment. For example, in a physics laboratory if we insert a battery into a simple circuit, we can predict the current flow
(C) by Ohm's law:

$C = E/R$

where E (potential difference between the two ends of the conductor) is the known value and R is the resistance.

Non-Deterministic (or Random) Experiments

The experiment in which the outcome cannot be predicted in advance is known as nondeterministic experiment. For example, if we toss an unbiased coin, the outcome may be either 'head' or 'tail'. But we cannot predict in advance which one will occur exactly. Similarly, throwing a die is also a nondeterministic experiment. The probability theory is associated with this type of random experiments only.

Probability theory is based on the paradigm of a *random experiment*; that is, an experiment whose outcome cannot be predicted with certainty, before the experiment is run. We usually assume that the experiment can be repeated indefinitely under essentially the same conditions. This assumption is important because probability theory is concerned with the long-term behavior as the experiment is replicated. Naturally, a complete definition of a random experiment requires a careful definition of precisely what information about the experiment is being recorded, that is, a careful definition of what constitutes an *outcome*.

2.4.2 SAMPLE SPACE

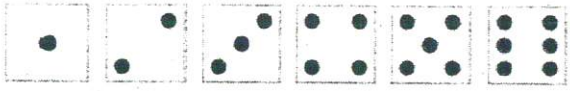
A *sample space* is a collection of all possible outcomes of a random experiment. A *random variable* is a function defined on a sample space. We shall consider several examples shortly. Later on we shall introduce probability functions on the sample spaces. A sample space may be *finite* or *infinite*. Infinite sample spaces may be *discrete* or *continuous*.

Finite Sample Spaces

Tossing a coin. The experiment is tossing a coin (or any other object with two distinct sides.) The coin may land and stay on the edge, but this event is so enormously unlikely as to be considered impossible and be disregarded. So the coin lands on either one or the other of its two sides. One is usually called *head*, the other *tail*. These are two possible outcomes of a toss of a coin. In the case of a single toss, the sample space has two elements that interchangeably, may be denoted as, say,

{Head, Tail}, or {H, T}, or {0, 1}, ...

Rolling a die. The experiment is rolling a die. A common die is a small cube whose faces shows numbers 1, 2, 3, 4, 5, 6 one way or another. These may be the real digits or arrangements of an appropriate number of dots, e.g. like these



There are six possible outcomes and the sample space consists of six elements:

{1, 2, 3, 4, 5, 6}.

Many random variables may be associated with this experiment: the square of the outcome $f(x) = x^2$, with values from

{1, 4, 9, 16, 25, 36},

centered values from

{-2.5, -1.5, -0.5, 0.5, 1.5, 2.5},

with the variable defined by f(x) = x - 3.5, etc.

Drawing a card. The experiment is drawing a card from a standard deck of 52 cards. The cards are of two colors - black (spades and clubs) and red (diamonds and hearts), four suits (spades, clubs, diamonds, hearts), 13 values (2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King, Ace). (Some decks use 4 colors, others use different names. For example, a Jack may be called a Knave. We shall abbreviate the named designations as J, Q, K, A.) There are 52 possible outcomes with the sample space

{2♠, 2♣, 2♦, 2♥, 3♠, 3♣, 3♦, 3♥, ..., A♠, A♣, A♦, A♥}.

Of course, if we are only interested in the color of a drawn card, or its suite, or perhaps the value, then it would be as natural to consider other sample spaces:

{b, r},
{♠, ♣, ♦, ♥} or
{2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A}.

Choosing a birthday. The experiment is to select a single date during a given year. This can be done, for example, by picking a random person and inquiring for his or her birthday. Disregarding leap years for the sake of simplicity, there are 365 possible birthdays, which may be enumerated

{1, 2, 3, 4, ..., 365}.

Tossing two coins. The experiment is tossing two coins. One may toss two coins simultaneously, or one after the other. The difference is in that in the second case we can easily differentiate between the coins: one is the first, the other second. If

the two indistinguishable coins are tossed simultaneously, there are just three possible outcomes, {H, H}, {H, T}, and {T, T}. If the coins are different, or if they are thrown one after the other, there are four distinct outcomes: (H, H), (H, T), (T, H), (T, T), which are often presented in a more concise form: HH, HT, TH, TT. Thus, depending on the nature of the experiment, there are 3 or 4 outcomes, with the sample spaces

Indistinguishable coins

{{H, H}, {H, T}, {T, T}}.

Distinct coins

{HH, HT, TH, TT}

Rolling two dice. The experiment is rolling two dice. If the dice are distinct or if they are rolled successively, there are 36 possible outcomes: 11, 12, ..., 16, 21, 22, ..., 66. If they are indistinguishable, then some outcomes, like 12 and 21, fold into one. There are $6 \times 5/2 = 15$ such pairs giving the total number of possible outcomes as $36 - 15 = 21$. In the first case, the sample space is

{11, 12, ..., 16, 21, 22, ..., 66}.

When we throw two dice we are often interested not in individual numbers that show up, but in their sum. The sum of the two top numbers is an example of a random variable, say $Y(ab) = a + b$ (where a, b range from 1 through 6), that takes values from the set {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}. It is also possible to think of this set of a sample space of a random experiment. However, there is a point in working with random variables. It is often a convenience to be able to consider several random variables related to the same experiment, i.e., to the same sample space. For example, besides Y, we may be interested in the product (or some other function) of the two numbers. (Concept of Random variable will be described in depth in the later section.)

Infinite Discrete Sample Spaces

First tail. The experiment is to repeatedly toss a coin until first tail shows up. Possible outcomes are sequences of H that, if finite, end with a single T, and an infinite sequence of H:

$$\{T, HT, HHT, HHHT, ..., \{HHH...\}\}.$$

As we shall see elsewhere, this is a remarkable space that contains a not impossible event whose probability is 0. One random variable is defined most naturally as the length of an outcome. It draws values from the set of whole numbers augmented by the symbol of infinity:

$$\{1, 2, 3, 4, ..., \infty\}.$$

Continuous Sample Spaces

Arrival time. The experimental setting is a metro (underground) station where trains pass (ideally) with equal intervals. A person enters the station. The experiment is to note the time of arrival past the departure time of the last train. If T is the interval between two consecutive trains, then the sample space for the experiment is the interval [0, T], or

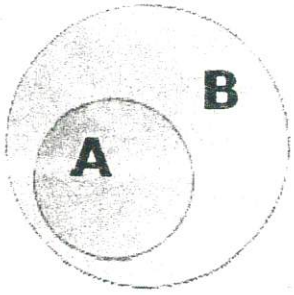
$$[0, T] = \{t: 0 \leq y \leq T\}.$$

2.4.3 EVENT

In probability theory, an **event** is a set of outcomes (a subset of the sample space) to which a probability is assigned. Typically, when the sample space is finite, any subset of the sample space is an event (*i.e.* all elements of the power set of the sample space are defined as events). However, this approach does not work well in cases where the sample space is infinite, most notably when the outcome is a real number. So, when defining a probability space it is possible, and often necessary, to exclude certain subsets of the sample space from being events.

Example 1

If we assemble a deck of 52 playing cards and no jokers, and draw a single card from the deck, then the sample space is a 52-element set, as each individual card is a possible outcome. An event, however, is any subset of the sample space, including any single-element set (an elementary event, of which there are 52, representing the 52 possible cards drawn from the deck), the empty set (an impossible event, defined to have probability zero) and the sample space itself (the entire set of 52 cards), which is defined to have probability one. Other events are proper subsets of the sample space that contain multiple elements. So, for example, potential events include:



A Venn diagram of an event. B is the sample space and A is an event. By the ratio of their areas, the probability of A is approximately 0.4.

- "Red and black at the same time without being a joker" (0 elements),
- "The 5 of Hearts" (1 element),
- "A King" (4 elements),
- "A Face card" (12 elements),
- "A Spade" (13 elements),
- "A Face card or a red suit" (32 elements),
- "A card" (52 elements).

Since all events are sets, they are usually written as sets (e.g. {1, 2, 3}), and represented graphically using Venn diagrams. Venn diagrams are particularly useful for representing events because the probability of the event can be identified with the ratio of the area of the event and the area of the sample space.

(Indeed, each of the axioms of probability, and the definition of conditional probability can be represented in this fashion.)

Types of events: Independent and Dependent Events

By **independent** we mean that the first event does not affect the probability of the second event. Coin tosses are independent. They cannot affect each other's probabilities; the probability of each toss is independent of a previous toss and will always be 1/2. Separate drawings from a deck of cards are independent events if you put the cards back.

An example of a **dependent event**, one in which the probability of the second event is affected by the first, is drawing a card from a deck but not returning it. By not returning the card, you've decreased the number of cards in the deck by 1, and you've decreased the number of whatever kind of card you drew. If you draw an ace of spades, there are 1 fewer aces and 1 fewer spades. This affects our simple probability: (number of favorable outcomes)/ (total number of outcomes. This type of probability is formulated as follows:
If A and B are not independent, then the probability of A and B is
$P(A \text{ and } B) = P(A) \times P(B|A)$
where $P(B|A)$ is the conditional probability of B given A.

Example 2

If someone draws a card at random from a deck and then, without replacing the first card, draws a second card, what is the probability that both cards will be aces?

Solution

Event A is that the first card is an ace. Since 4 of the 52 cards are aces, $P(A) = 4/52 = 1/13$. Given that the first card is an ace, what is the probability that the second card will be an ace as well? Of the 51 remaining cards, 3 are aces. Therefore, $p(B|A) = 3/51 = 1/17$, and the probability of A and B is $1/13 \times 1/17 = 1/221$. The same reasoning is applied to marbles in a jar. (assume event was successful)

Example 3

If there are 30 red and blue marbles in a jar, and the ratio of red to blue marbles is 2:3, what is the probability that, drawing twice, you will select two red marbles if you return the marbles after each draw?

Solution

First, let's determine the number of red and blue marbles respectively. The ratio 2:3 tells us that the total of 30 marbles must be broken into 5 groups of 6 marbles, each with 2 groups of red marbles and 3 groups of blue marbles. Setting up the equation $2x + 3x = 5x = 30$ employs the same reasoning. Solving, we find that there are 12 red marbles and 18 blue marbles. We are asked to draw twice and return the marble after each draw. Therefore, the first draw does not affect the probability of the second draw. We return the marble after the draw, and therefore, we return the situation to the initial conditions before the second draw. Nothing is altered in between draws, and therefore, the events are independent.

Now let's examine the probabilities. Drawing a red marble would be $12/30 = 2/5$. The same is true for the second draw. Since we want two red marbles in a row, the question is really saying that we want a red marble on the first draw and a red marble on the second draw. The "and" means we should expect a lower probability than 2/5. Understanding that the "and" is implicit can help you eliminate choices d and e which are both too big. Therefore, our total probability is:

$$P(A \text{ and } B) = P(A) \times P(B) = 2/5 \times 2/5 = 4/25.$$

Now consider the same question with the condition that you do not return the marbles after each draw. The probability of drawing a red marble on the first draw remains the same, $12/30 = 2/5$. The second draw, however, is different. The initial conditions have been altered by the first draw. We now have only 29 marbles in the jar and only 11 red. Don't panic! We simply use those numbers to figure our new probability of drawing a red marble the second time, 11/29. The events are dependent and the total probability is:

P(A and B) = P(A) ×. P(B) = 2/5 × 11/29 = 132/870 = 22/145.

If you return every marble you select, the probability of drawing another marble is unaffected; the events are independent. If you do not return the marbles, the number of marbles is affected and therefore dependent.

2.5 RANDOM VARIABLE

random variables are used in the study of chance and probability. They were developed to assist in the analysis of games of chance, stochastic events, and the results of scientific experiments by capturing only the mathematical properties necessary to answer probabilistic questions. Further formalizations have firmly grounded the entity in the theoretical domains of mathematics by making use of measure theory.

Broadly, there are two types of random variables — discrete and continuous. Discrete random variables take on one of a set of specific values, each with some probability greater than zero. Continuous random variables can be realized with any of a range of values (e.g., a real number between zero and one), and so there are several ranges (e.g. 0 to one half) that have a probability greater than zero of occurring.

A random variable can be thought of as an unknown value that may change every time it is inspected. Thus, a random variable can be thought of as a function mapping the sample space of a random process to the real numbers. A few examples will highlight this.

Example 4

For a coin toss, the possible events are heads or tails. The number of heads appearing in one fair coin toss can be described using the following random variable:

$$X = \begin{cases} 1, & \text{if heads,} \\ 0, & \text{if tails.} \end{cases}$$

with probability mass function given by:

$$\rho_X(x) = \begin{cases} \frac{1}{2}, & \text{if } x = 0, \\ \frac{1}{2}, & \text{if } x = 1, \\ 0, & \text{otherwise.} \end{cases}$$

A random variable can also be used to describe the process of rolling a fair die and the possible outcomes. The most obvious representation is to take the set {1, 2, 3, 4, 5, 6} as the sample space, defining the random variable X as the number rolled. In this case ,

$$X = \begin{cases} 1, & \text{if a 1 is rolled,} \\ 2, & \text{if a 2 is rolled,} \\ 3, & \text{if a 3 is rolled,} \\ 4, & \text{if a 4 is rolled,} \\ 5, & \text{if a 5 is rolled,} \\ 6, & \text{if a 6 is rolled.} \end{cases}$$

$$\rho_X(x) = \begin{cases} \frac{1}{6}, & \text{if } x = 1, 2, 3, 4, 5, 6, \\ 0, & \text{otherwise.} \end{cases}$$

A random variable has either an associated probability distribution (discrete random variable) or probability density function (continuous random variable).

Discrete Probability Distributions

If a random variable is a discrete variable, its probability distribution is called a discrete probability distribution.

Example 5

Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the random variable X represent the number of Heads that result from this experiment. The random variable X can only take on the values 0, 1, or 2, so it is a discrete random variable.

The probability distribution for this statistical experiment appears below.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

The above table represents a discrete probability distribution because it relates each value of a discrete random variable with its probability of occurrence.

Continuous Probability Distributions

If a random variable is a continuous variable, its probability distribution is called a continuous probability distribution.

A continuous probability distribution differs from a discrete probability distribution in several ways.

The probability that a continuous random variable will assume a particular value is zero.

As a result, a continuous probability distribution cannot be expressed in tabular form.

Instead, an equation or formula is used to describe a continuous probability distribution.

Most often, the equation used to describe a continuous probability distribution is called a probability density function. Sometimes, it is referred to as a density function, a PDF, or a pdf. For a continuous probability distribution, the density function has the following properties:

Since the continuous random variable is defined over a continuous range of values (called the domain of the variable), the graph of the density function will also be continuous over that range.

The area bounded by the curve of the density function and the x-axis is equal to 1, when computed over the domain of the variable.

The probability that a random variable assumes a value between a and b is equal to the area under the density function bounded by a and b.

For example, consider the probability density function shown in the graph below. Suppose we wanted to know the probability that the random variable X was less than or equal to a. The probability that X is less than or equal to a is equal to the area under the curve bounded by a and minus infinity - as indicated by the shaded area.

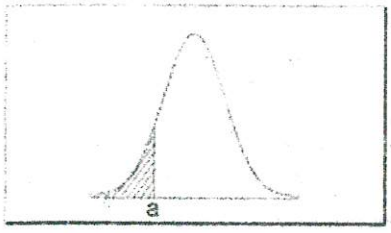


Figure 2.1

Note: The shaded area in the graph represents the probability that the random variable X is less than or equal to a. This is a cumulative probability. However, the probability that X is exactly equal to a would be zero. A continuous random variable can take on an infinite number of values. The probability that it will equal a specific value (such as a) is always zero.

2.6 PROBABILITY EXPECTATIONS

In probability theory and statistics, the **expected value** (or **expectation value**, or **mathematical expectation**, or **mean**, or **first moment**) of a random variable is the integral of the random variable with respect to its probability measure. For discrete random variables this is equivalent to the probability-weighted sum of the possible values, and for continuous random variables with a density function it is the probability density -weighted integral of the possible values.

The term "expected value" can be misleading. It must not be confused with the "most probable value." The expected value is in general not a typical value that the random variable can take on. It is often helpful to interpret the expected value of a random variable as the long-run average value of the variable over many independent repetitions of an experiment.

The expected value may be intuitively understood by the law of large numbers: The expected value, when it exists, is almost surely the limit of the sample mean as sample size grows to infinity. The value may not be expected in the general sense – the "expected value" itself may be unlikely or even impossible (such as having 2.5 children), just like the sample mean. The expected value does not exist for all distributions, such as the Cauchy distribution.

It is possible to construct an expected value equal to the probability of an event by taking the expectation of an indicator function that is one if the event has occurred and zero otherwise. This relationship can be used to translate properties of expected values into properties of probabilities, e.g. using the law of large numbers to justify estimating probabilities by frequencies.

Example 6

The expected value from the roll of an ordinary six-sided die is

$$\mathrm{E(Roll\ With\ 6\ Sided\ Die)} = \frac{1+2+3+4+5+6}{6} = 3.5$$

which is not among the possible outcomes.

A common application of expected value is gambling. For example, an American roulette wheel has 38 places where the ball may land, all equally likely. A winning bet on a single number pays 35-to-1, meaning that the original stake is not lost, and 35 times that amount is won, so you receive 36 times what you've bet. Considering all 38 possible outcomes, the expected value of the profit resulting from a dollar bet on a single number is the sum of what you may lose times the odds of losing and what you will win times the odds of winning, that is,

$$\mathrm{E(winnings\ \$1\ bet)} = \left(-\$1 \times \frac{37}{38}\right) + \left(\$35 \times \frac{1}{38}\right) = -\$\frac{1}{19} \approx -\$0.0526.$$

The change in your financial holdings is –$1 when you lose, and $35 when you win. Thus one may expect, on average, to lose about five cents for every dollar bet, and the **expected value** of a one-dollar bet is $0.9474. In gambling, an event

of which the expected value equals the stake (of which the bettor's expected profit is zero) is called a "fair game."

2.7 GENERATING FUNCTIONS

2.7.1 The ordinary generating function

We define the ordinary generating function of a sequence. This is by far the most common type of generating function and the adjective "ordinary" is usually not used.

But we will need a different type of generating function below (the exponential gen-

erating function) so we have added the adjective "ordinary" for this first type of generating function.

2.7.2 Moment-generating function

The moment-generating function is so called because, if it exists on an open interval around $t = 0$, then it is the ordinary generating function of the moments of the probability distribution:

$$E(X^n) = M_X^{(n)}(0) = \frac{d^n M_X}{dt^n}(0).$$

In probability theory and statistics, the **moment-generating function** of a random variable X is

$$M_X(t) := E\left(e^{tX}\right), \quad t \in \mathbb{R},$$

wherever this expectation exists.

A key problem with moment-generating functions is that moments and the moment-generating function may not exist, as the integrals need not converge. By contrast, the characteristic function always exists (because the integral is a bounded function on a space of finite measure), and thus may be used instead.

More generally, where $\mathbf{X} = (X_1, \ldots, X_n)$, an n-dimensional random vector, one uses $\mathbf{t} \cdot \mathbf{X} = \mathbf{t}^{\mathrm{T}}\mathbf{X}$ instead of tX:

$$M_{\mathbf{X}}(\mathbf{t}) := E\left(e^{\mathbf{t}^{\mathrm{T}}\mathbf{X}}\right).$$

Calculation

If X has a continuous probability density function $f(x)$ then the moment generating function is given by

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \left(1 + tx + \frac{t^2x^2}{2!} + \cdots\right) f(x)\,\mathrm{d}x \\ &= 1 + tm_1 + \frac{t^2 m_2}{2!} + \cdots, \end{aligned}$$

where m_i is the ith moment. $M_X(-t)$ is just the two-sided Laplace transform of $f(x)$.

Regardless of whether the probability distribution is continuous or not, the moment-generating function is given by the Riemann-Stieltjes integral

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx}\, dF(x)$$

where F is the cumulative distribution function.

If X_1, X_2, ..., X_n is a sequence of independent (and not necessarily identically distributed) random variables, and

$$S_n = \sum_{i=1}^{n} a_i X_i,$$

where the a_i are constants, then the probability density function for S_n is the convolution of the probability density functions of each of the X_i and the moment-generating function for S_n is given by

$$M_{S_n}(t) = M_{X_1}(a_1 t) M_{X_2}(a_2 t) \cdots M_{X_n}(a_n t).$$

For vector-valued random variables X with real components, the moment-generating function is given by

$$M_X(t) = E\left(e^{\langle t, X \rangle}\right)$$

where t is a vector and $\langle t, X \rangle$ is the dot product.

Relation to other functions

Related to the moment-generating function are a number of other transforms that are common in probability theory:

Characteristic function

The characteristic function $\varphi_X(t)$ is related to the moment-generating function via $\varphi_X(t) = M_{iX}(t) = M_X(it)$ the characteristic function is the moment-generating function of iX or the moment generating function of X evaluated on the imaginary axis.

cumulant-generating function

The cumulant-generating function is defined as the logarithm of the moment-generating function; some instead define the cumulant-generating function as the logarithm of the characteristic function, while others call this latter the second cumulant-generating function.

probability-generating function

The probability-generating function is defined as $G(z) = E[z^X]$. This immediately implies that $G(e^t) = E[e^{tX}] = M_X(t)$.

Activity 2

1. Distinguish between the classical and empirical approaches of probability.
2. Try to find out two main events in your life where you faced uncertainty in taking decisions. Elaborate how you dealt.

3. under an employment promotion program it is puoposed to allow sale of newspapers on the buses during off-peak hours. The vendor can purchase the newspaper at a special rate of 25 paise per copy against the selling price of 40 paise per copy. Any unsold copies are however, a dead loss. A vendor has estimated the following probability distribution for the number of copies demanded.

No of copies	15	16	17	18	19	20
Probability	.04	.19	.33	.26	.11	.07

How many copies should be ordered so that his expected profit will be maximum?

2.8 SUMMARY

Probability in common parlance means the chance of occurrence of event. The need to develop a formal and precise expression for uncertainty in decision making, has led to different approaches to probability measurement. These approaches, namely classical and empirical arose mainly to cater to different types of situations where we face uncertainties. In this unit followed by main definitions we discussed the basic concepts of probability including the two types of experiments – Deterministic and Non Deterministic, Sample space and events and their types. Further concepts of random variable and probability expectations were explained using suitable examples. Finally the generating functions, ordinary and moment generating functions were described in brief.

2.9 FURTHER READINGS

- Mood A.M., Graybill F.A., Boes D.C. (1974) Introduction to the Theory of Statistics (3rd Edition). McGraw-Hill.
- Kallenberg, O., *Foundations of Modern Probability*, 2nd edition. Springer-Verlag, New York, Berlin, Heidelberg (2001).
- Patrick Billingsley (1979). *Probability and Measure*. New York, Toronto, London: John Wiley and Sons.

UNIT 3

PROBABILITY LAWS AND DISTRIBUTIONS

Objectives

After reading this unit, you should be able to:

- Use the laws of addition, subtraction and multiplication in decision making.
- Understand the Baye's theorem and its applicability
- Identify situations where discrete and continuous probability distributions can be applied.
- Find or assess probability distributions for different uncertain situations.

Structure

3.1 INTRODUCTION

There are indefinite numbers of ways which can be used in solving probability problems. These methods include the tree diagrams, laws of probability, sample space, insight, and contingency table. Because of the individuality and variety of probability problems, some approaches apply more readily in certain cases than in others. There is no best method for solving all probability problems. Three laws of probability the additive law, the multiplication law, and Baye's theorem are discussed in this chapter.

Discrete probability distributions – Binomial and Poisson distributions and continuous distribution - Normal distribution are also explained in detail.

3.2THE LAW OF ADDITION

As we have already noted, the sample space S is the set of all possible outcomes of a given experiment. Certain events A and B are subsets of S. In the previous Section we defined what was meant by P(A), P(B) and their complements in the particular case in which the experiment had equally likely outcomes.Events, like sets, can be combined to produce new events.

• A U B denotes the event that event A or event B (or both) occur when the experiment is
performed.

• A ∩ B denotes the event that both A and B occur together.

In this Section we obtain expressions for determining the probabilities of these combined events,
which are written P(A U B) and P(A ∩ B) respectively.

The law of addition can be bifurcated into two following rules as:

A. General Rule of Addition:

when two or more events will happen at the same time, and the events are not mutually exclusive, then:

P(X or Y) = P(X) + P(Y) - P(X and Y)

For example, what is the probability that a card chosen at random from a deck of cards will either be a king or a heart?
P(King or Heart) = P(X or Y) = 4/52 + 13/52 - 1/52 = 30.77%

B. Special Rule of Addition:

when two or more events will happen at the same time, and the events are mutually exclusive, then:

P(X or Y) = P(X) + P(Y)

Example 1

Suppose we have a machine that inserts a mixture of beans, broccoli, and other types of vegetables into a plastic bag. Most of the bags contain the correct weight, but because of slight variation in the size of the beans and other vegetables, a package might be slightly underweight or overweight. A check of many packages in the past indicate that:

Weight.....Event.....No. of Packages.....Probability

Underweight.....X.....100.....0.025
Correct weight.....Y.....3600.....0.9
Overweight.....Z.....300.....0.075
Total.....4000.....1.00

What is the probability of selecting a package at random and having the package be under weight or over weight? Since the events are mutually exclusive, a package cannot be underweight and overweight at the same time. The answer is: P(X or Z) = P(0.025 + 0.075) = 0.1

3.3 MULTIPLICATION RULE OF PROBABILITY

The addition rule helped us solve problems when we performed one task and wanted to know the probability of two things happening during that task. This lesson deals with the multiplication rule. The multiplication rule also deals with two events, but in these problems the events occur as a result of more than one task (rolling one die then another, drawing two cards, spinning a spinner twice, pulling two marbles out of a bag, etc).

When asked to find the probability of A and B, we want to find out the probability of events A and B happening.

The Multiplication Rule:

Consider events A and B. $P(A \cap B) = P(A) \cdot P(B)$.

Note: Some books will say to take care that A and B are independent, but the rule can also be used with dependent events, you just have to be more careful in find P(A) and P(B).

What The Rule Means:

Suppose we roll one die followed by another and want to find the probability of rolling a 4 on the first die and rolling an even number on the second die. Notice in this problem we are not dealing with the sum of both dice. We are only dealing with the probability of 4 on one die only and then, as a separate event, the probability of an even number on one die only.

P(4) = 1/6
P(even) = 3/6

So $P(4 \cap even) = (1/6)(3/6) = 3/36 = 1/12$

While the rule can be applied regardless of dependence or independence of

events, we should note here that rolling a 4 on one die followed by rolling an even number on the second die are independent events. Each die is treated as a separate thing and what happens on the first die does not influence or effect what happens on the second die. This is our basic definition of independent events: the outcome of one event does not influence or effect the outcome of another event.

We'll look at examples later that deal with dependent events. Just keep in mind that what happens on one event will effect the other event.

Example 2

Suppose you have a box with 3 blue marbles, 2 red marbles, and 4 yellow marbles. You are going to pull out one marble, record its color, put it back in the box and draw another marble. What is the probability of pulling out a red marble followed by a blue marble?

The multiplication rule says we need to find P(red) •P(blue).

P(red) = 2/9
P(blue) = 3/9

$P(red \cap blue) = (2/9)(3/9) = 6/81 = 2/27$

The events in this example were independent. Once the first marble was pulled out and its color recorded, it was returned to the box. Therefore, the probability for the second marble was not effected by what happened on the first marble.

Notice that the final answer is always simplified. Some students find it helpful to simplify before multiplying, but the final answer must always be simplified.

Consider the same box of marbles as in the previous example. However in

this case, we are going to pull out the first marble, leave it out, and then pull out another marble. What is the probability of pulling out a red marble followed by a blue marble?

We can still use the multiplication rule which says we need to find P(red) • P(blue). But be aware that in this case when we go to pull out the second marble, there will only be 8 marbles left in the bag.

P(red) = 2/9
P(blue) = 3/8

P(red∩blue) = (2/9)(3/8) = 6/72 = 1/12

The events in this example were dependent. When the first marble was pulled out and kept out, it effected the probability of the second event. This is what is meant by dependent events.

Suppose you are going to draw two cards from a standard deck. What is the probability that the first card is an ace and the second card is a jack (just one of several ways to get "blackjack" or 21).

Using the multiplication rule we get

P(ace) •P(jack) = (4/52)(4/51) = 16/2652 = 4/663

Notice that this will be the same probability even if the question had asked for the probability of a jack followed by an ace.

3.4 BAYES' THEOREM

In probability theory, ***Bayes'* theorem** (often called **Bayes' law** and named after Rev Thomas Bayes; IPA:/'beɪz/) relates the conditional and marginal probabilities of two random events. It is often used to compute posterior probabilities given observations. For example, a patient may be observed to have certain symptoms.

Bayes' theorem can be used to compute the probability that a proposed diagnosis is correct, given that observation. (See example 2) Bayes' Theorem states that judgements should be influenced by two main factors: the base rate, and the likelihood ratio.

As a formal theorem, Bayes' theorem is valid in all common interpretations of probability. However, it plays a central role in the debate around the foundations of statistics: frequentist and Bayesian interpretations disagree about the ways in which probabilities should be assigned in applications. Frequentists assign probabilities to random events according to their frequencies of occurrence or to subsets of populations as proportions of the whole, while Bayesians describe probabilities in terms of beliefs and degrees of uncertainty.

Bayes' theorem relates the conditional and marginal probabilities of events A and B, where B has a non-vanishing probability:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

Each term in Bayes' theorem has a conventional name:

- P(A) is the prior probability or marginal probability of A. It is "prior" in the sense that it does not take into account any information about B.
- P(A|B) is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.
- P(B|A) is the conditional probability of B given A.
- P(B) is the prior or marginal probability of B, and acts as a normalizing constant.

Intuitively, Bayes' theorem in this form describes the way in which one's beliefs about observing 'A' are updated by having observed 'B'.

Example 3

Suppose there is a co-ed school having 60% boys and 40% girls as students. The girl students wear trousers or skirts in equal numbers; the boys all wear trousers. An observer sees a (random) student from a distance; all they can see is that this student is wearing trousers. What is the probability this student is a girl? The correct answer can be computed using Bayes' theorem.

The event A is that the student observed is a girl, and the event B is that the student observed is wearing trousers. To compute $P(A|B)$, we first need to know:

- $P(A)$, or the probability that the student is a girl regardless of any other information. Since the observers sees a random student, meaning that all students have the same probability of being observed, and the fraction of girls among the students is 40%, this probability equals 0.4.
- $P(A')$, or the probability that the student is a boy regardless of any other information (A' is the complementary event to A). This is 60%, or 0.6.
- $P(B|A)$, or the probability of the student wearing trousers given that the student is a girl. As they are as likely to wear skirts as trousers, this is 0.5.
- $P(B|A')$, or the probability of the student wearing trousers given that the student is a boy. This is given as 1.
- $P(B)$, or the probability of a (randomly selected) student wearing trousers regardless of any other information. Since $P(B) = P(B|A)P(A) + P(B|A')P(A')$, this is $0.5 \times 0.4 + 1 \times 0.6 = 0.8$.

Given all this information, the probability of the observer having spotted a girl given that the observed student is wearing trousers can be computed by substituting these values in the formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.5 \times 0.4}{0.8} = 0.25.$$

Another, essentially equivalent way of obtaining the same result is as follows. Assume, for concreteness, that there are 100 students, 60 boys and 40 girls. Among these, 60 boys and 20 girls wear trousers. All together there are 80

trouser-wearers, of which 20 are girls. Therefore the chance that a random trouser-wearer is a girl equals 20/80 = 0.25.

It is often helpful when calculating conditional probabilities to create a simple table containing the number of occurrences of each outcome, or the relative frequencies of each outcome, for each of the independent variables. The table below illustrates the use of this method for the above girl-or-boy example

	Girls	**Boys**	**Total**
Trousers	20	60	80
Skirts	20	0	20
Total	40	60	100

3.5 PROBABILITY DISTRIBUTIONS

An example will make clear the relationship between random variables and probability distributions. Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the variable X represent the number of Heads that result from this experiment. The variable X can take on the values 0, 1, or 2. In this example, X is a random variable; because its value is determined by the outcome of a statistical experiment.

A **probability distribution** is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence. Consider the coin flip experiment described above. The table below, which associates each outcome with its probability, is an example of a probability distribution.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

The above table represents the probability distribution of the random variable X.

3.5.1 Cumulative Probability Distributions

A **cumulative probability** refers to the probability that the value of a random variable falls within a specified range.

Let us return to the coin flip experiment. If we flip a coin two times, we might ask: What is the probability that the coin flips would result in one or fewer heads? The answer would be a cumulative probability. It would be the probability that the coin flip experiment results in zero heads plus the probability that the experiment results in one head.

$P(X < 1) = P(X = 0) + P(X = 1) = 0.25 + 0.50 = 0.75$

Like a probability distribution, a cumulative probability distribution can be represented by a table or an equation. In the table below, the cumulative probability refers to the probability than the random variable X is less than or equal to x.

Number of heads: x	Probability: $P(X = x)$	Cumulative Probability: $P(X < x)$
0	0.25	0.25
1	0.50	0.75
2	0.25	1.00

3.5.2 Uniform Probability Distribution

The simplest probability distribution occurs when all of the values of a random variable occur with equal probability. This probability distribution is called the **uniform distribution.**

Uniform Distribution. Suppose the random variable X can assume k different values. Suppose also that the $P(X = x_k)$ is constant. Then,

$P(X = x_k) = 1/k$

Example 4

Suppose a die is tossed. What is the probability that the die will land on 6 ?

Solution: When a die is tossed, there are 6 possible outcomes represented by: S = { 1, 2, 3, 4, 5, 6 }. Each possible outcome is a random variable (X), and each outcome is equally likely to occur. Thus, we have a uniform distribution. Therefore, the $P(X = 6) = 1/6$.

Example 5

Suppose we repeat the dice tossing experiment described in Example 1. This time, we ask what is the probability that the die will land on a number that is smaller than 5 ?

Solution: When a die is tossed, there are 6 possible outcomes represented by: S = { 1, 2, 3, 4, 5, 6 }. Each possible outcome is equally likely to occur. Thus, we have a uniform distribution.

This problem involves a cumulative probability. The probability that the die will land on a number smaller than 5 is equal to:

$P(X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1/6 + 1/6 + 1/6 + 1/6 = 2/3$

Three main types of probability distributions are discussed in next section.

3.6 BINOMIAL DISTRIBUTION

To understand binomial distributions and binomial probability, it helps to understand binomial experiments and some associated notation; so we cover those topics first.

Binomial Experiment

A **binomial experiment** (also known as a **Bernoulli trial**) is a statistical experiment that has the following properties:

- The experiment consists of n repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
- The probability of success, denoted by P, is the same on every trial.
- The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.

Consider the following statistical experiment. You flip a coin 2 times and count the number of times the coin lands on heads. This is a binomial experiment because:

- The experiment consists of repeated trials. We flip a coin 2 times.
- Each trial can result in just two possible outcomes - heads or tails.
- The probability of success is constant - 0.5 on every trial.
- The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

Notation

The following notation is helpful, when we talk about binomial probability.

- x: The number of successes that result from the binomial experiment.
- n: The number of trials in the binomial experiment.
- P: The probability of success on an individual trial.

- Q: The probability of failure on an individual trial. (This is equal to 1 - P.)
- b(x; n, P): Binomial probability - the probability that an n-trial binomial experiment results in exactly x successes, when the probability of success on an individual trial is P.
- ${}_nC_r$: The number of combinations of n things, taken r at a time.

Binomial Distribution

A binomial random variable is the number of successes x in n repeated trials of a binomial experiment. The probability distribution of a binomial random variable is called a binomial distribution (also known as a Bernoulli distribution).

Suppose we flip a coin two times and count the number of heads (successes). The binomial random variable is the number of heads, which can take on values of 0, 1, or 2. The binomial distribution is presented below.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

The binomial distribution has the following properties:

The mean of the distribution (μx) is equal to n * P .
The variance ($\sigma 2x$) is n * P * (1 - P).
The standard deviation (σx) is sqrt[n * P * (1 - P)].

Binomial Probability

The binomial probability refers to the probability that a binomial experiment results in exactly x successes. For example, in the above table, we see that the binomial probability of getting exactly one head in two coin flips is 0.50.

Binomial Formula. Suppose a binomial experiment consists of n trials and results in x successes. If the probability of success on an individual trial is P, then the binomial probability is:

$b(x; n, P) = nCx * Px * (1 - P)n - x$

Example 6

Suppose a die is tossed 5 times. What is the probability of getting exactly 2 fours?
Solution: This is a binomial experiment in which the number of trials is equal to 5, the number of successes is equal to 2, and the probability of success on a single trial is 1/6 or about 0.167. Therefore, the binomial probability is:

$b(2; 5, 0.167) = 5C2 * (0.167)2 * (0.833)3$

$b(2; 5, 0.167) = 0.161$

Cumulative Binomial Probability

A cumulative binomial probability refers to the probability that the binomial random variable falls within a specified range (e.g., is greater than or equal to a stated lower limit and less than or equal to a stated upper limit).

Example 7

we might be interested in the cumulative binomial probability of obtaining 45 or fewer heads in 100 tosses of a coin (see Example 1 below). This would be the sum of all these individual binomial probabilities.

$b(x < 45; 100, 0.5) = b(x = 0; 100, 0.5) + b(x = 1; 100, 0.5) + ... + b(x = 44; 100, 0.5) + b(x = 45; 100, 0.5)$

Example 8

What is the probability of obtaining 45 or fewer heads in 100 tosses of a coin?
Solution: To solve this problem, we compute 46 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer we seek. Thus,

$b(x < 45; 100, 0.5) = b(x = 0; 100, 0.5) + b(x = 1; 100, 0.5) + \ldots + b(x = 45; 100, 0.5)$

$b(x < 45; 100, 0.5) = 0.184$

Example 9

What is the probability that the world series will last 4 games? 5 games? 6 games? 7 games? Assume that the teams are evenly matched.

Solution: This is a very tricky application of the binomial distribution. If you can follow the logic of this solution, you have a good understanding of the material covered in the tutorial, to this point.

In the world series, there are two baseball teams. The series ends when the winning team wins 4 games. Therefore, we define a success as a win by the team that ultimately becomes the world series champion.

For the purpose of this analysis, we assume that the teams are evenly matched. Therefore, the probability that a particular team wins a particular game is 0.5.

Let's look first at the simplest case. What is the probability that the series lasts only 4 games. This can occur if one team wins the first 4 games. The probability of the National League team winning 4 games in a row is:

$b(4; 4, 0.5) = 4C4 * (0.5)4 * (0.5)0 = 0.0625$

Similarly, when we compute the probability of the American League team winning 4 games in a row, we find that it is also 0.0625. Therefore, probability that the series ends in four games would be 0.0625 + 0.0625 = 0.125; since the series would end if either the American or National League team won 4 games in a row.

Now let's tackle the question of finding probability that the world series ends in 5 games. The trick in finding this solution is to recognize that the series can only end in 5 games, if one team has won 3 out of the first 4 games. So let's first find the probability that the American League team wins exactly 3 of the first 4 games.

b(3; 4, 0.5) = 4C3 * (0.5)3 * (0.5)1 = 0.25

Given that the American League team has won 3 of the first 4 games, the American League team has a 50/50 chance of winning the fifth game to end the series. Therefore, the probability of the American League team winning the series in 5 games is 0.25 * 0.50 = 0.125. Since the National League team could also win the series in 5 games, the probability that the series ends in 5 games would be

0.125 + 0.125 = 0.25.

The rest of the problem would be solved in the same way. You should find that the probability of the series ending in 6 games is 0.3125; and the probability of the series ending in 7 games is also 0.3125.

While this is statistically correct in theory, over the years the actual world series has turned out differently, with more series than expected lasting 7 games.

3.7 POISSON DISTRIBUTION

A **Poisson experiment** is a statistical experiment that has the following properties:

- The experiment results in outcomes that can be classified as successes or failures.
- The average number of successes (μ) that occurs in a specified region is known.
- The probability that a success will occur is proportional to the size of the region.
- The probability that a success will occur in an extremely small region is virtually zero.

Note that the specified region could take many forms. For instance, it could be a length, an area, a volume, a period of time, etc.

Notation

The following notation is helpful, when we talk about the Poisson distribution.

- e: A constant equal to approximately 2.71828. (Actually, e is the base of the natural logarithm system.)
- μ: The mean number of successes that occur in a specified region.
- x: The actual number of successes that occur in a specified region.
- $P(x; \mu)$: The **Poisson probability** that exactly x successes occur in a Poisson experiment, when the mean number of successes is μ.

Poisson distribution

A **Poisson random variable** is the number of successes that result from a Poisson experiment. The probability distribution of a Poisson random variable is called a **Poisson distribution**.

Given the mean number of successes (μ) that occur in a specified region, we can compute the Poisson probability based on the following formula:

Poisson Formula. Suppose we conduct a Poisson experiment, in which the average number of successes within a given region is μ. Then, the Poisson probability is:

$$P(x; \mu) = (e^{-\mu}) (\mu^{x}) / x!$$

where x is the actual number of successes that result from the experiment, and e is approximately equal to 2.71828.

The Poisson distribution has the following properties:

- The mean of the distribution is equal to μ .
- The variance is also equal to μ .

Example 10

The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Solution: This is a Poisson experiment in which we know the following:

- $\mu = 2$; since 2 homes are sold per day, on average.
- $x = 3$; since we want to find the likelihood that 3 homes will be sold tomorrow.
- $e = 2.71828$; since *e* is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:

$P(x; \mu) = (e^{-\mu}) (\mu^x) / x!$
$P(3; 2) = (2.71828^{-2}) (2^3) / 3!$
$P(3; 2) = (0.13534) (8) / 6$
$P(3; 2) = 0.180$

Thus, the probability of selling 3 homes tomorrow is 0.180 .

Cumulative Poisson Probability

A **cumulative Poisson probability** refers to the probability that the Poisson random variable is greater than some specified lower limit and less than some specified upper limit.

Example 11

Suppose the average number of lions seen on a 1-day safari is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari?

Solution: This is a Poisson experiment in which we know the following:

- $\mu = 5$; since 5 lions are seen per safari, on average.
- $x = 0, 1, 2,$ or 3; since we want to find the likelihood that tourists will see fewer than 4 lions; that is, we want the probability that they will see 0, 1, 2, or 3 lions.
- $e = 2.71828$; since *e* is a constant equal to approximately 2.71828.

To solve this problem, we need to find the probability that tourists will see 0, 1, 2, or 3 lions. Thus, we need to calculate the sum of four probabilities: P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5). To compute this sum, we use the Poisson formula:

P(x < 3, 5) = P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5)
P(x < 3, 5) = [(e-5)(50) / 0!] + [(e-5)(51) / 1!] + [(e-5)(52) / 2!] + [(e-5)(53) / 3!]
P(x < 3, 5) = [(0.006738)(1) / 1] + [(0.006738)(5) / 1] + [(0.006738)(25) / 2] + [(0.006738)(125) / 6]
P(x < 3, 5) = [0.0067] + [0.03369] + [0.084224] + [0.140375]
P(x < 3, 5) = 0.2650

Thus, the probability of seeing at no more than 3 lions is 0.2650.

3.8 NORMAL DISTRIBUTION

The **normal distribution** refers to a family of continuous probability distributions described by the normal equation.

The Normal Equation

The normal distribution is defined by the following equation:

Normal equation. The value of the random variable *Y* is:

$$Y = [\ 1/\sigma * \text{sqrt}(2\pi)\ ] * e^{-(x - \mu)2/2\sigma 2}$$

where *X* is a normal random variable, μ is the mean, σ is the standard deviation, π is approximately 3.14159, and *e* is approximately 2.71828.

The random variable *X* in the normal equation is called the **normal random variable**. The normal equation is the probability density function for the normal distribution.

The Normal Curve

The graph of the normal distribution depends on two factors - the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph. When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow. All normal distributions look like a symmetric, bell-shaped curve, as shown below.

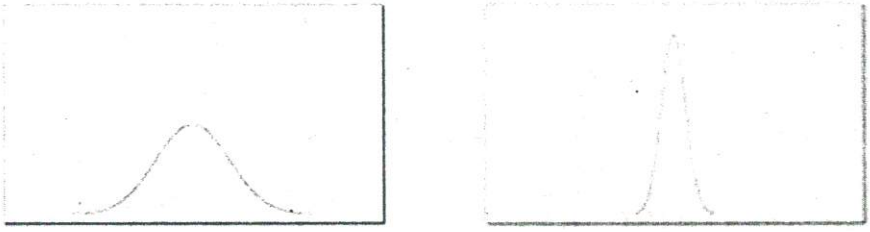


Figure 3.1

The curve on the left is shorter and wider than the curve on the right, because the curve on the left has a bigger standard deviation.

Probability and the Normal Curve

The normal distribution is a continuous probability distribution. This has several implications for probability.

- The total area under the normal curve is equal to 1.
- The probability that a normal random variable X equals any particular value is 0.
- The probability that X is greater than a equals the area under the normal curve bounded by a and plus infinity (as indicated by the *non-shaded* area in the figure below).
- The probability that X is less than a equals the area under the normal curve bounded by a and minus infinity (as indicated by the *shaded* area in the figure below).

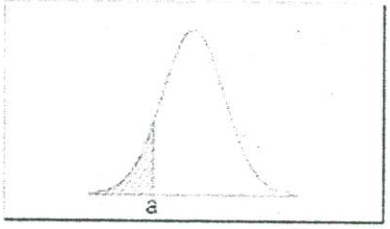


Figure 3.2

Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following "rule".

- About 68% of the area under the curve falls within 1 standard deviation of the mean.
- About 95% of the area under the curve falls within 2 standard deviations of the mean.
- About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

Collectively, these points are known as the **empirical rule** or the **68-95-99.7 rule**. Clearly, given a normal distribution, most outcomes will be within 3 standard deviations of the mean.

Example 12

An average light bulb manufactured by the Acme Corporation lasts 300 days with a standard deviation of 50 days. Assuming that bulb life is normally distributed, what is the probability that an Acme light bulb will last at most 365 days?

Solution: Given a mean score of 300 days and a standard deviation of 50 days, we want to find the cumulative probability that bulb life is less than or equal to 365 days. Thus, we know the following:

- The value of the normal random variable is 365 days.
- The mean is equal to 300 days.
- The standard deviation is equal to 50 days.

We enter these values into the Normal Distribution Calculator and compute the cumulative probability. The answer is: P(X < 365) = 0.90. Hence, there is a 90% chance that a light bulb will burn out within 365 days.

Example 13

Suppose scores on an IQ test are normally distributed. If the test has a mean of 100 and a standard deviation of 10, what is the probability that a person who takes the test will score between 90 and 110?

Solution: Here, we want to know the probability that the test score falls between 90 and 110. The "trick" to solving this problem is to realize the following:

P(90 < *X* < 110) = P(X < 110) - P(X < 90)

We use the Normal Distribution Calculator to compute both probabilities on the right side of the above equation.

- To compute P(X < 110), we enter the following inputs into the calculator: The value of the normal random variable is 110, the mean is 100, and the standard deviation is 10. We find that P(X < 110) is 0.84.
- To compute P(X < 90), we enter the following inputs into the calculator: The value of the normal random variable is 90, the mean is 100, and the standard deviation is 10. We find that P(X < 90) is 0.16.

We use these findings to compute our final answer as follows:

P(90 < *X* < 110) = P(X < 110) - P(X < 90)
P(90 < *X* < 110) = 0.84 - 0.16
P(90 < *X* < 110) = 0.68

Thus, about 68% of the test scores will fall between 90 and 110.

Activity 3

1. An urn contains 6 red marbles and 4 black marbles. Two marbles are drawn *with replacement* from the urn. What is the probability that both of the marbles are black?

(A) 0.16
(B) 0.32
(C) 0.36
(D) 0.40
(E) 0.60

2. A card is drawn randomly from a deck of ordinary playing cards. You win $10 if the card is a spade or an ace. What is the probability that you will win the game?

(A) 1/13
(B) 13/52
(C) 4/13
(D) 17/52
(E) None of the above.

3. What is the probability of drawing a Heart and a Club from a deck without replacement?

4.Write down the sample space for the following experiments:
(a) tossing a coin;
(b) rolling a dice;
(c) answering a true-false question; and
(d) tossing two coins.

5.The probability that a student is accepted to a prestigeous college is 0.3. If 5 students from the same school apply, what is the probability that at most 2 are accepted?

3.9 SUMMARY

We have introduced the laws of addition and multiplication of probability in this chapter. Certain results in probability which are helpful in making day to day decisions have been presented. Baye's theorem and its applications are discussed in detail in next section which was about the conditional probability approach. Probability distribution which can be known as an equation that links each outcome of a statistical experiment with its probability of occurrence was discussed with its three forms viz. binomial distribution, poisson distribution and normal distribution. We also have looked into situations that give rise to these type of probability distribution and discussed how these distributions are helpful in decision making.

3.10 FURTHER READINGS

- Glenn Shafer; Vladimir Vovk, *The origins and legacy of Kolmogorov's Grundbegriffe*
- Isaac Todhunter (1865). *A History of the Mathematical Theory of Probability from the time of Pascal to that of Laplace*, Macmillan. Reprinted 1949, 1956 by Chelsea and 2001 by Thoemmes.
- Stephen M. Stigler (1982). "*Thomas Bayes' Bayesian Inference*," Journal of the Royal Statistical Society
- Von Plato, Jan, 2005, "Grundbegriffe der Wahrscheinlichtkeitsrechnung" in Grattan-Guinness, I., ed., *Landmark Writings in Western Mathematics.* Elsevier: (in English)

Answers to activities

Activity 1

1. .988
2. Y = 57.89
3. Height of son when height of father is 70 inches shall be 64.53 inches

Activity 2

4. 17 copies will give the maximum expected profit of 84 paise

Activity 3

1. The correct answer is A. Let A = the event that the first marble is black; and let B = the event that the second marble is black. We know the following:

- In the beginning, there are 10 marbles in the urn, 4 of which are black. Therefore, P(A) = 4/10.
- After the first selection, we replace the selected marble; so there are still 10 marbles in the urn, 4 of which are black. Therefore, P(B|A) = 4/10.

Therefore, based on the rule of multiplication:

$P(A \cap B) = P(A)\ P(B|A)$
$P(A \cap B) = (4/10)*(4/10) = 16/100 = 0.16$

Therefore, based on the rule of addition:

$P(S \cup A) = P(S) + P(A) - P(S \cap A)$
$P(S \cup A) = 13/52 + 4/52 - 1/52 = 16/52 = 4/13$

2. The correct answer is C.

3. .064

4. (a) If a coin is tossed, the sample space is S = {H, T}.
   (b) In rolling a dice, the sample space is S = {1, 2, 3, 4, 5, 6}.
   (c) In answering a true-false question, the sample space is S ={T, F}.
   (d) In tossing two coins, the sample space is S = {HH, HT, TH, TT}.

5.b(x < 2; 5, 0.3) = b(x = 0; 5, 0.3) + b(x = 1; 5, 0.3) + b(x = 2; 5, 0.3)
b(x < 2; 5, 0.3) = 0.1681 + 0.3601 + 0.3087
b(x < 2; 5, 0.3) = 0.8369

BLOCK 4

BASIC CONCEPTS OF SAMPLING

BLOCK 4 BASIC CONCEPTS OF SAMPLING

This block comprises two units. The first unit deals with basic concepts of sampling, concepts of distribution of some commonly used statistics with specific applications of the same.

The second unit gives you the understanding of hypothesis and it's testing using commonly used tests namely the chi-square test, Z test, T test and F test. The unit also acquaints you with the concepts of goodness of fit, confidence interval and level of significance.

UNIT 1

BASIC CONCEPTS OF SAMPLING AND SAMPLING METHODS

Objectives

On successful completion of this unit, you should be able to:

- Appreciate the concept of sampling.
- Identify the potential sampling frame.
- List the various sampling methods with their applications.
- Distinguish between probability and non probability sampling.
- Know when to use the probability proportional sampling.
- Recognize the factors which affect the sample size decisions.

Structure

1.1 INTRODUCTION

Sampling is that part of statistical practice concerned with the selection of individual observations intended to yield some knowledge about a population of concern, especially for the purposes of statistical inference. Each observation measures one or more properties (weight, location, etc.) of an observable entity enumerated to distinguish objects or individuals. Survey weights often need to be applied to the data to adjust for the sample design. Results from probability theory and statistical theory are employed to guide practice.

The sampling process comprises several stages:

- Defining the population of concern
- Specifying a sampling frame, a set of items or events possible to measure
- Specifying a sampling method for selecting items or events from the frame
- Determining the sample size
- Implementing the sampling plan
- Sampling and data collecting
- Reviewing the sampling process

1.2 POPULATION

Successful statistical practice is based on focused problem definition. In sampling, this includes defining the population from which our sample is drawn. A population can be defined as including all people or items with the characteristic one wishes to understand. Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population.

Sometimes that which defines a population is obvious. For example, a manufacturer needs to decide whether a batch of material from production is of high enough quality to be released to the customer, or should be sentenced for scrap or rework due to poor quality. In this case, the batch is the population.
Although the population of interest often consists of physical objects, sometimes we need to sample over time, space, or some combination of these dimensions.

For instance, an investigation of supermarket staffing could examine checkout line length at various times, or a study on endangered penguins might aim to understand their usage of various hunting grounds over time. For the time dimension, the focus may be on periods or discrete occasions.

In other cases, our 'population' may be even less tangible. For example, Joseph Jagger studied the behaviour of roulette wheels at a casino in Monte Carlo, and used this to
identify a biased wheel. In this case, the 'population' Jagger wanted to investigate was the overall behaviour of the wheel (i.e. the probability distribution of its results over infinitely many trials), while his 'sample' was formed from observed results from that wheel. Similar considerations arise when taking repeated measurements of some physical characteristic such as the electrical conductivity of copper.

This situation often arises when we seek knowledge about the cause system of which the observed population is an outcome. In such cases, sampling theory may treat the observed population as a sample from a larger 'super population'. For example, a researcher might study the success rate of a new 'quit smoking' program on a test group of 100 patients, in order to predict the effects of the program if it were made available nationwide. Here the super population is "everybody in the country, given access to this treatment" - groups which does not yet exist, since the program isn't yet available to all.

Note also that the population from which the sample is drawn may not be the same as the population about which we actually want information. Often there is large but not complete overlap between these two groups due to frame issues etc (see below). Sometimes they may be entirely separate - for instance, we might study rats in order to get a better understanding of human health, or we might study records from people born in 2008 in order to make predictions about people born in 2009.

Time spent in making the sampled population and population of concern precise is often well spent, because it raises many issues, ambiguities and questions that would otherwise have been overlooked at this stage.

1.3 SAMPLING FRAME

In the most straightforward case, such as the sentencing of a batch of material from production (acceptance sampling by lots), it is possible to identify and measure every single item in the population and to include any one of them in our sample. However, in the more general case this is not possible. There is no way to identify all rats in the set of all rats. Where voting is not compulsory, there is no way to identify which people will actually vote at a forthcoming election (in advance of the election).

These imprecise populations are not amenable to sampling in any of the ways below and to which we could apply statistical theory.

As a remedy, we seek a sampling frame which has the property that we can identify every single element and include any in our sample. The most straightforward type of frame is a list of elements of the population (preferably the entire population) with appropriate contact information. For example, in an opinion poll, possible sampling frames include:

Electoral register

Telephone directory

Not all frames explicitly list population elements. For example, a street map can be used as a frame for a door-to-door survey; although it doesn't show individual houses, we can select streets from the map and then visit all houses on those streets. (One advantage of such a frame is that it would include people who have recently moved and are not yet on the list frames discussed above.)

The sampling frame must be representative of the population and this is a question outside the scope of statistical theory demanding the judgment of experts in the particular subject matter being studied. All the above frames omit some people who will vote at the next election and contain some people who will not; some frames will contain multiple records for the same person. People not in the frame have no prospect of being sampled. Statistical theory tells us about the uncertainties in extrapolating from a sample to the frame. In extrapolating from frame to population, its role is motivational and suggestive.

"To the scientist, however, representative sampling is the only justified procedure for choosing individual objects for use as the basis of generalization, and is therefore usually the only acceptable basis for ascertaining truth." (Andrew A.

Marino) [1]. It is important to understand this difference to steer clear of confusing prescriptions found in many web pages.

In defining the frame, practical, economic, ethical, and technical issues need to be addressed. The need to obtain timely results may prevent extending the frame far into the future.

The difficulties can be extreme when the population and frame are disjoint. This is a particular problem in forecasting where inferences about the future are made from historical data. In fact, in 1703, when Jacob Bernoulli proposed to Gottfried Leibniz the possibility of using historical mortality data to predict the probability of early death of a living man, Gottfried Leibniz recognized the problem in replying:

"Nature has established patterns originating in the return of events but only for the most part. New illnesses flood the human race, so that no matter how many experiments you have done on corpses, you have not thereby imposed a limit on the nature of events so that in the future they could not vary."

A frame may also provide additional 'auxiliary information' about its elements; when this information is related to variables or groups of interest, it may be used to improve survey design. For instance, an electoral register might include name and sex; this information can be used to ensure that a sample taken from that frame covers all demographic categories of interest. (Sometimes the auxiliary information is less explicit; for instance, a telephone number may provide some information about location.)

Having established the frame, there are a number of ways for organizing it to improve efficiency and effectiveness.

It's at this stage that the researcher should decide whether the sample is in fact to be the whole population and would therefore be a census.

1.4 PROBABILITY AND NONPROBABILITY SAMPLING

A probability sampling scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

Example 1

We want to estimate the total income of adults living in a given street. We visit each household in that street, identify all adults living there, and randomly select one adult from each household. (For example, we can allocate each person a random number, generated from a uniform distribution between 0 and 1, and select the person with the highest number in each household). We then interview the selected person and find their income.

People living on their own are certain to be selected, so we simply add their income to our estimate of the total. But a person living in a household of two adults has only a one-in-two chance of selection. To reflect this, when we come to such a household, we would count the selected person's income twice towards the total. (In effect, the person who is selected from that household is taken as representing the person who isn't selected.)

In the above example, not everybody has the same probability of selection; what makes it a probability sample is the fact that each person's probability is known. When every element in the population does have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

Probability sampling includes: Simple Random Sampling, Systematic Sampling, Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling. These various ways of probability sampling have two things in common: 1) Every element has a known nonzero probability of being sampled and 2) involves random selection at some point.

Nonprobability sampling is any sampling method where some elements of the population have no chance of selection (these are sometimes referred to as 'out of coverage'/'undercovered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, nonprobability sampling does not allow the estimation of sampling errors. These conditions place limits on how much information a sample can provide about the population. Information about the

relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population.

Example 2

We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant, this is a nonprobability sample, because some people are more likely to answer the door (e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities.
Nonprobability Sampling includes: Accidental Sampling, Quota Sampling and Purposive Sampling. In addition, nonresponse effects may turn any probability design into a nonprobability design if the characteristics of nonresponse are not well understood, since nonresponse effectively modifies each element's probability of being sampled.

1.5 SAMPLING METHODS

Within any of the types of frame identified above, a variety of sampling methods can be employed, individually or in combination. Factors commonly influencing the choice between these designs include:

1. Nature and quality of the frame
2. Availability of auxiliary information about units on the frame
3. Accuracy requirements, and the need to measure accuracy
4. Whether detailed analysis of the sample is expected
5. Cost/operational concerns

1.5.1 Simple random sampling

In a simple random sample ('SRS') of a given size, all such subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection: the frame is not subdivided or partitioned. Furthermore, any given pair of elements has the same chance of selection as any other such pair

(and similarly for triples, and so on). This minimises bias and simplifies analysis of results. In particular, the variance between individual results within the sample is a good indicator of variance in the overall population, which makes it relatively easy to estimate the accuracy of results.

However, SRS can be vulnerable to sampling error because the randomness of the selection may result in a sample that doesn't reflect the makeup of the population. For instance, a simple random sample of ten people from a given country will on average produce five men and five women, but any given trial is likely to overrepresent one sex and underrepresent the other. Systematic and stratified techniques, discussed below, attempt to overcome this problem by using information about the population to choose a more representative sample.

SRS may also be cumbersome and tedious when sampling from an unusually large target population. In some cases, investigators are interested in research questions specific to subgroups of the population. For example, researchers might be interested in examining whether cognitive ability as a predictor of job performance is equally applicable across racial groups. SRS cannot accommodate the needs of researchers in this situation because it does not provide subsamples of the population. Stratified sampling, which is discussed below, addresses this weakness of SRS.

Simple random sampling is always an EPS design, but not all EPS designs are simple random sampling.

1.5.2 Systematic sampling

Systematic sampling relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every kth element from then onwards. In this case, k=(population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the kth element in the list. A simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10').

As long as the starting point is randomized, systematic sampling is a type of probability sampling. It is easy to implement and the stratification induced can

make it efficient, if the variable by which the list is ordered is correlated with the variable of interest. 'Every 10th' sampling is especially useful for efficient sampling from databases.

Example 3

Suppose we wish to sample people from a long street that starts in a poor district (house #1) and ends in an expensive district (house #1000). A simple random selection of addresses from this street could easily end up with too many from the high end and too few from the low end (or vice versa), leading to an unrepresentative sample. Selecting (e.g.) every 10th street number along the street ensures that the sample is spread evenly along the length of the street, representing all of these districts. (Note that if we always start at house #1 and end at #991, the sample is slightly biased towards the low end; by randomly selecting the start between #1 and #10, this bias is eliminated.)

However, systematic sampling is especially vulnerable to periodicities in the list. If periodicity is present and the period is a multiple or factor of the interval used, the sample is especially likely to be unrepresentative of the overall population, making the scheme less accurate than simple random sampling.

Example 4

Consider a street where the odd-numbered houses are all on the north (expensive) side of the road, and the even-numbered houses are all on the south (cheap) side. Under the sampling scheme given above, it is impossible' to get a representative sample; either the houses sampled will all be from the odd-numbered, expensive side, or they will all be from the even-numbered, cheap side.

Another drawback of systematic sampling is that even in scenarios where it is more accurate than SRS, its theoretical properties make it difficult to quantify that accuracy. (In the two examples of systematic sampling that are given above, much of the potential sampling error is due to variation between neighbouring houses - but because this method never selects two neighbouring houses, the sample will not give us any information on that variation.)

As described above, systematic sampling is an EPS method, because all elements have the same probability of selection (in the example given, one in ten). It is not

'simple random sampling' because different subsets of the same size have different selection probabilities - e.g. the set {4,14,24,...,994} has a one-in-ten probability of selection, but the set {4,13,24,34,...} has zero probability of selection.
Systematic sampling can also be adapted to a non-EPS approach; for an example, see discussion of PPS samples below.

1.5.3 Stratified sampling

Where the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected (Pedhazur & Schmelkin, 1991). There are several potential benefits to stratified sampling.
First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample.
Second, utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). It is important to note that even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less efficiency than would simple random sampling, provided that each stratum is proportional to the group's size in the population.
Third, it is sometimes the case that data are more readily available for individual, pre-existing strata within a population than for the overall population; in such cases, using a stratified sampling approach may be more convenient than aggregating data across groups (though this may potentially be at odds with the previously noted importance of utilizing criterion-relevant strata).
Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.
There are, however, some potential drawbacks to using stratified sampling. First, identifying strata and implementing such an approach can increase the cost and complexity of sample selection, as well as leading to increased complexity of population estimates. Second, when examining multiple criteria, stratifying

variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata. Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods (although in most cases, the required sample size would be no larger than would be required for simple random sampling.

A stratified sampling approach is most effective when three conditions are met:

1) Variability within strata are minimized

2) Variability between strata are maximized

3) The variables upon which the population is stratified are strongly correlated with the desired dependent variable.

Poststratification

Stratification is sometimes introduced after the sampling phase in a process called "poststratification (Pedhazur & Schmelkin, 1991)." This approach is typically implemented due to a lack of prior knowledge of an appropriate stratifying variable or when the experimenter lacks the necessary information to create a stratifying variable during the sampling phase. Although the method is susceptible to the pitfalls of post hoc approaches, it can provide several benefits in the right situation. Implementation usually follows a simple random sample. In addition to allowing for stratification on an ancillary variable, poststratification can be used to implement weighting, which can improve the precision of a sample's estimates (Pedhazur & Schmelkin, 1991).

Oversampling

Choice-based sampling is one of the stratified sampling strategies. In choice-based sampling (Scott and Wild 1986), the data are stratified on the target and a sample is taken from each strata so that the rare target class will be more represented in the sample. The model is then built on this biased sample. The effects of the input variables on the target are often estimated with more precision with the choice-based sample even when a smaller overall sample size is taken, compared to a random sample. The results usually must be adjusted to correct for the oversampling.

1.5.4 Probability proportional to size sampling (PPS)

In some cases the sample designer has access to an "auxiliary variable" or "size measure", believed to be correlated to the variable of interest, for each element in the population. This data can be used to improve accuracy in sample design. One option is to use the auxiliary variable as a basis for stratification, as discussed above.

Another option is probability-proportional-to-size ('PPS') sampling, in which the selection probability for each element is set to be proportional to its size measure, up to a maximum of 1. In a simple PPS design, these selection probabilities can then be used as the basis for Poisson sampling. However, this has the drawbacks of variable sample size, and different portions of the population may still be over- or under-represented due to chance variation in selections. To address this problem, PPS may be combined with a systematic approach.

Example 5

suppose we have six schools with populations of 150, 180, 200, 220, 260, and 490 students respectively (total 1500 students), and we want to use student population as the basis for a PPS sample of size three. To do this, we could allocate the first school numbers 1 to 150, the second school 151 to 330 (=150+180), the third school 331 to 530, and so on to the last school (1011 to 1500). We then generate a random start between 1 and 500 (equal to 1500/3) and count through the school populations by multiples of 500. If our random start was 137, we would select the schools which have been allocated numbers 137, 637, and 1137, i.e. the first, fourth, and sixth schools.

The PPS approach can improve accuracy for a given sample size by concentrating sample on large elements that have the greatest impact on population estimates. PPS sampling is commonly used for surveys of businesses, where element size varies greatly and auxiliary information is often available - for instance, a survey attempting to measure the number of guest-nights spent in hotels might use each hotel's number of rooms as an auxiliary variable. In some cases, an older

measurement of the variable of interest can be used as an auxiliary variable when attempting to produce more current estimates.

1.5.5 Cluster sampling

Sometimes it is cheaper to 'cluster' the sample in some way e.g. by selecting respondents from certain areas only, or certain time-periods only. (Nearly all samples are in some sense 'clustered' in time - although this is rarely taken into account in the analysis.)

Cluster sampling is an example of 'two-stage sampling' or 'multistage sampling': in the first stage a sample of areas is chosen; in the second stage a sample of respondents within those areas is selected.

This can reduce travel and other administrative costs. It also means that one does not need a sampling frame listing all elements in the target population. Instead, clusters can be chosen from a cluster-level frame, with an element-level frame created only for the selected clusters. Cluster sampling generally increases the variability of sample estimates above that of simple random sampling, depending on how the clusters differ between themselves, as compared with the within-cluster variation.

Nevertheless, some of the disadvantages of cluster sampling are the reliance of sample estimate precision on the actual clusters chosen. If clusters chosen are biased in a certain way, inferences drawn about population parameters from these sample estimates will be far off from being accurate.

Multistage sampling: Multistage sampling is a complex form of cluster sampling in which two or more levels of units are imbedded one in the other. The first stage consists of constructing the clusters that will be used to sample from. In the second stage, a sample of primary units is randomly selected from each cluster (rather than using all units contained in all selected clusters). In following stages, in each of those selected clusters, additional samples of units are selected, and so on. All ultimate units (individuals, for instance) selected at the last step of this procedure are then surveyed.

This technique, thus, is essentially the process of taking random samples of preceding random samples. It is not as effective as true random sampling, but it probably solves more of the problems inherent to random sampling. Moreover, It

is an effective strategy because it banks on multiple randomizations. As such, it is extremely useful.

Multistage sampling is used frequently when a complete list of all members of the population does not exist and is inappropriate. Moreover, by avoiding the use of all sample units in all selected clusters, multistage sampling avoids the large, and perhaps unnecessary, costs associated traditional cluster sampling.

1.5.6 Matched random sampling

A method of assigning participants to groups in which pairs of participants are first matched on some characteristic and then individually assigned randomly to groups. (Brown, Cozby, Kee, & Worden, 1999, p.371).

The Procedure for Matched random sampling can be briefed with the following contexts,

a) Two samples in which the members are clearly paired, or are matched explicitly by the researcher. For example, IQ measurements or pairs of identical twins.

b) Those samples in which the same attribute, or variable, is measured twice on each subject, under different circumstances. Commonly called repeated measures. Examples include the times of a group of athletes for 1500m before and after a week of special training; the milk yields of cows before and after being fed a particular diet.

1.5.7 Quota sampling

In quota sampling, the population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion. For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.

It is this second step which makes the technique one of non-probability sampling. In quota sampling the selection of the sample is non-random. For example interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance

of selection. This random element is its greatest weakness and quota versus probability has been a matter of controversy for many years

1.5.8 Mechanical sampling

Mechanical sampling is typically used in sampling solids, liquids and gases, using devices such as grabs, scoops, thief probes, the COLIWASA and riffle splitter. Care is needed in ensuring that the sample is representative of the frame. Much work in the theory and practice of mechanical sampling was developed by Pierre Gy and Jan Visman.

1.5.9 Convenience sampling

Convenience sampling (sometimes known as grab or opportunity sampling) is the method of choosing items in an unstructured manner from the population frame. Though almost impossible to treat rigorously, it is the method most commonly employed in many practical situations. This is due largely to the fact that when most researchers aim to study the behaviors of human beings, very rarely does one find an ideal environment for carrying out that research. It may be that if not for the convenience sample, a particular type of research could simply not take place. Several important considerations for researchers using convenience samples include: 1. Are there controls within the research design or experiment which can serve to lessen the impact of a non-random, convenience sample whereby ensuring the results will be more representative of the population? 2. Is there good reason to believe that a particular convenience sample would or should respond or behave differently than a random sample from the same population? 3. Is the question being asked by the research one that can adequately be answered using a convenience sample?

In social science research, snowball sampling is a similar technique, where existing study subjects are used to recruit more subjects into the sample.

1.5.10 Line-intercept sampling

Line-intercept sampling is a method of sampling elements in a region whereby an element is sampled if a chosen line segment, called a "transect", intersects the element.

1.5.11 Panel sampling

Panel sampling is the method of first selecting a group of participants through a random sampling method and then asking that group for the same information again several times over a period of time. Therefore, each participant is given the same survey or interview at two or more time points; each period of data collection is called a "wave". This sampling methodology is often chosen for large scale or nation-wide studies in order to gauge changes in the population with regard to any number of variables from chronic illness to job stress to weekly food expenditures. Panel sampling can also be used to inform researchers about within-person health changes due to age or help explain changes in continuous dependent variables such as spousal interaction. There have been several proposed methods of analyzing panel sample data, including MANOVA, growth curves, and structural equation modeling with lagged effects. For a more thorough look at analytical techniques for panel data, see Johnson (1995).

1.5.12 Event Sampling Methodology

Event Sampling Methodology (ESM) is a new form of sampling method that allows researchers to study ongoing experiences and events that vary across and within days in its naturally-occurring environment. Because of the frequent sampling of events inherent in ESM, it enables researchers to measure the typology of activity and detect the temporal and dynamic fluctuations of work experiences. Popularity of ESM as a new form of research design increased over the recent years because it addresses the shortcomings of cross-sectional research, where once unable to, researchers can now detect intra-individual variances across time. In ESM, participants are asked to record their experiences and perceptions in a paper or electronic diary.

There are three types of ESM: 1) Signal contingent – random beeping notifies participants to record data. The advantage of this type of ESM is minimization of recall bias. 2) Event contingent – records data when certain events occur 3) Interval contingent – records data according to the passing of a certain period of time

ESM has several disadvantages. One of the disadvantages of ESM is it can sometimes be perceived as invasive and intrusive by participants. ESM also leads to possible self-selection bias. It may be that only certain types of individuals are willing to participate in this type of study creating a non-random sample. Another concern is related to participant cooperation. Participants may not be actually fill out their diaries at the specified times. Furthermore, ESM may substantively change the phenomenon being studied. Reactivity or priming effects may occur, such that repeated measurement may cause changes in the participants' experiences. This method of sampling data is also highly vulnerable to common method variance. (Alliger & Williams, 1993)

Further, it is important to think about whether or not an appropriate dependent variable is being used in an ESM design. For example, it might be logical to use ESM in order to answer research questions which involve dependent variables with a great deal of variation throughout the day. Thus, variables such as change in mood, change in stress level, or the immediate impact of particular events may be best studied using ESM methodology. However, it is not likely that utilizing ESM will yield meaningful predictions when measuring someone performing a repetitive task throughout the day or when dependent variables are long-term in nature (coronary heart problems).

Replacement of selected units

Sampling schemes may be without replacement ('WOR' - no element can be selected more than once in the same sample) or with replacement ('WR' - an element may appear multiple times in the one sample). For example, if we catch fish, measure them, and immediately return them to the water before continuing with the sample, this is a WR design, because we might end up catching and measuring the same fish more than once. However, if we do not return the fish to the water (e.g. if we eat the fish), this becomes a WOR design.

1.6 SAMPLE SIZE

Formulas, tables, and power function charts are well known approaches to determine sample size.

Where the frame and population are identical, statistical theory yields exact recommendations on sample size.[1] However, where it is not straightforward to define a frame representative of the population, it is more important to understand the cause system of which the population are outcomes and to ensure that all sources of variation are embraced in the frame. Large number of observations are of no value if major sources of variation are neglected in the study. In other words, it is taking a sample group that matches the survey category and is easy to survey. Bartlett, Kotrlik, and Higgins (2001) published a paper titled Organizational Research: Determining Appropriate Sample Size in Survey Research Information Technology, Learning, and Performance Journal that provides an explanation of Cochran's (1977) formulas. A discussion and illustration of sample size formulas, including the formula for adjusting the sample size for smaller populations, is included. A table is provided that can be used to select the sample size for a research problem based on three alpha levels and a set error rate.

Steps for using sample size tables

1. Postulate the effect size of interest, α, and β.
2. Check sample size table (Cohen, 1988)
    a. Select the table corresponding to the selected α
    b. Locate the row corresponding to the desired power
    c. Locate the column corresponding to the estimated effect size
    d. The intersection of the column and row is the minimum sample size required.

1.7 ESTIMATION AND SAMPLING DISTRIBUTIONS

We sometimes tend to concentrate on sample means and variances. This is natural as much of the basic statistical techniques are based on these statistics. However, if we wish to define a general theory of statistical methodology we need to

consider a more general framework. We have seen that a Binomial distribution depends only on N and p. The value of N is usually known from the design of the sampling process, so that we consider only p as the unknown. This is referred to as the parameter of the distribution, the
parameter controls the properties of the distribution. Knowing the value of p means we know everything about the distribution including its mean and variance (and all other moments). If we wished to know the variance of the Binomial distribution we would determine the value of p and find the variance from this. We would determine the value of p that matches (fits) the sample data in the "best" possible way. This process is referred to as estimating the parameter and the resulting value is the parameter estimate.
There are several statistical questions we might ask at this stage:
1. What do we mean by "best" estimate?
2. How accurate is the estimate?
3. If some theory specifies the value of the parameter how do we decide if the data provides evidence that the theory is wrong? It is the role of statistical theory to provide answers to these questions.

Sampling Distribution

An estimator is the mathematical formula (or algorithm) used to derive the estimate from data. In many of the examples we will consider the chosen estimator seems "obvious" or "common sense", but in more complex situations with more general distributions our common sense will desert us and we need a general framework. Even in simple cases we can make the wrong choice. If we were asked to estimate the median of a Normal distribution (the median is the value with 50% of the population less than this value) then statistical theory would recommend that we estimate this by the sample mean not the sample median. In the previous question we really meant "what do we mean by best estimator?" After all, in a particular sample we may obtain a very poor estimate, i.e. one that is far from the true value. This may not be the fault of the estimation process but rather an unfortunate sample. This will always happen occasionally. When speaking of a good estimator we are considering that it is good "in the long run" rather than on a single occasion. In order to establish what might be a good estimator we must establish its epeated sampling properties.

Repeated Sampling

We are considering taking a sample of size 100 and estimating the median of the distribution. We assume that the distribution is Normal. We are considering using the sample median as our estimator. We wish to establish whether this is a good estimator in the long run. Conceptually, we could proceed as follows.
Take a sample of 100 and calculate the median. Take another sample of 100 and calculate another median. Repeat this process ay infinitum. At the end of this repeated sampling we would have an infinite collection of sample medians. From this we can form a distribution. This distribution is called the sampling distribution of the estimator (the median in this case).

Properties of Estimators

In the early development of the theory there was much emphasis on unbiasedness i.e. that the expectation of the parameter estimate was equal to the true value in the repeated sampling sense. That is the mean of the sampling distribution is equal to the true parameter value. In many contexts this is too restrictive and we settle for consistency whereby the bias decreases to zero as we increase the sample size.

Another property of estimators is efficiency defined in terms of the variance of the sampling distribution with high efficiency associated with low variance. We have already seen the we can readily establish the properties of the sample mean as an estimator of the population mean, it is unbiased and has variance σ 2/n. It can be shown that of all possible unbiased estimators of the parameter μ of the Normal distribution the sample mean has the smallest variance. It is said to be MVUE, the minimum variance unbiased estimator. Note that since μ is also the median of the distribution we should estimate the median of the distribution by the sample mean 4-5 This measure of efficiency is only sensible when we also insist on unbiasedness so becomes redundant if we relax this insistence. One measure that can be useful is mean square error (MSE) defined as follows. Define $\theta\hat{}$ to be an estimator for θ then:

$MSE = E(\hat{\theta} - \theta)2$

Activity 1

1. list the various reasons that make sampling so attractive in drawing conclusions about the population

2. What is the major difference between probability and non probability sampling?

3. A study aims to quantify the organizational climate in any organization by administering a questionnaire to a sample of its employees. There are 1000 employees in a company with 100 executives. 200 supervisors and 700 workers. If the employees are stratified based on this classification and a sample of 100 employees is required, what should be the sample size be from each stratum, if proportional sampling is used?
Discuss in brief the concept of estimator and sampling distribution pertaining to it.

4. What do you understand by estimation in context of sampling distributions? discuss main properties of estimators.

1.8 SUMMARY

In this unit we dealt with the basic concepts of sampling, first started by sampling and population definitions. In further sections, concept of sample frame was discussed in depth. Further we looked at various sampling methods available when one wants to make some inferences about a population without enumerating it completely. We started by probability and non probability sampling and their differences and moved ahead to other methods like simple random sampling, systematic sampling, stratified sampling, probability proportional to size sampling, cluster sampling, matched random sampling, quota sampling, mechanical sampling, convenience sampling, line intercept sampling, panel sampling, event sampling along with discussion of areas of their applications.

Further the concept of sample size have discussed followed by the estimation approach and distributions pertaining to it.

1.10 FURTHER READINGS

- Brown, K.W., Cozby, P.C., Kee, D.W., & Worden, P.E. (1999). *Research Methods in Human Development*, 2d ed. Mountain View, CA: Mayfield
- Chambers, R L, and Skinner, C J (editors) (2003), *Analysis of Survey Data* Wiley
- Cochran, W G (1977) *Sampling Techniques*, Wiley
- Flyvbjerg, B (2006) *"Five Misunderstandings About Case Study Research."* Qualitative Inquiry, vol. 12, no. 2, April 2006
- Kish, L (1995) *Survey Sampling*, Wiley
- Korn, E L, and Graubard, B I (1999) *Analysis of Health Surveys*, Wiley
- Lohr, H (1999) *Sampling: Design and Analysis,* Duxbury

UNIT 2

TESTING OF HYPOTHESES

Objectives

Upon successful completion of this unit, you should be able to:

- Understand the meaning of statistical hypothesis.
- Absorb the basic concepts of hypothesis testing.
- Learn the method of hypothesis formulation.
- Appreciate the concept of hypothesis testing.
- Perform test using student's T-test, Z-test, F-test and chi-square tests.

Structure

2.1 INTRODUCTION

A statistical hypothesis test is a method of making statistical decisions using experimental data. It is sometimes called confirmatory data analysis, in contrast to exploratory data analysis. In frequency probability, these decisions are almost always made using null-hypothesis tests; that is, ones that answer the question Assuming that the null hypothesis is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed?[1] One use of hypothesis testing is deciding whether experimental results contain enough information to cast doubt on conventional wisdom.

Statistical hypothesis testing is a key technique of frequentist statistical inference, and is widely used,[citation needed] but also much criticized. The main alternative to statistical hypothesis testing is Bayesian inference.

The critical region of a hypothesis test is the set of all outcomes which, if they occur, cause the null hypothesis to be rejected in favor of the alternative hypothesis. The critical region is usually denoted by C.

Example 1

As an example, consider determining whether a suitcase contains some radioactive material. Placed under a Geiger counter, it produces 10 counts per minute. The null hypothesis is that no radioactive material is in the suitcase and that all measured counts are due to ambient radioactivity typical of the surrounding air and harmless objects. We can then calculate how likely it is that we would observe 10 counts per minute if the null hypothesis were true. If the null hypothesis predicts (say) on average 9 counts per minute and a standard deviation of 1 count per minute, then we say that the suitcase is compatible with the null hypothesis (this does not guarantee that there is no radioactive material, just that we don't have enough evidence to suggest there is). On the other hand, if the null hypothesis predicts 3 counts per minute and a standard deviation of 1 count per

minute, then the suitcase is not compatible with the null hypothesis, and there are likely other factors responsible to produce the measurements.
The test described here is more fully the null-hypothesis statistical significance test. The null hypothesis represents what we would believe by default, before seeing any evidence. Statistical significance is a possible finding of the test, declared when the observed sample is unlikely to have occurred by chance if the null hypothesis were true. The name of the test describes its formulation and its possible outcome. One characteristic of the test is its crisp decision: to reject or not reject the null hypothesis. A calculated value is compared to a threshold, which is determined from the tolerable risk of error.

Formulating a hypothesis

To formulate a research hypothesis we start with a research question and:

1. Generate operational definitions for all variables, and

2. Formulate a research hypothesis keeping in mind
   - expected relationships or differences
   - operational definitions

3. Hypothesis can also be classified in terms of how they were derived
   - inductive hypothesis - a generalization based on observation
   - deductive hypothesses - derived from theory

4. A hypothesis can be directional or non-directional.

5. Hypotheses can also be stated as research hypotheses (as we have considered them so far) or as statistical hypotheses.

6. The statistical hypotheses consist of the null hypothesis (H0), the hypothesis of no difference and the alternative hypothesis (H1 or HA) which is similar in form to the research hypothesis.

2.2 BASIC CONCEPTS OF HYPOTHESES TESTING

Alternative hypothesis

In statistical hypothesis testing, the alternative hypothesis (or maintained hypothesis or research hypothesis) and the null hypothesis are the two rival hypotheses which are compared by a statistical hypothesis test. An example might be where water quality in a stream has been observed over many years and a test is made of the null hypothesis that there is no change in quality between the first and second halves of the data against the alternative hypothesis that the quality is poorer in the second half of the record.

The concept of an alternative hypothesis in testing was devised by Jerzy Neyman and Egon Pearson, and is it used in the Neyman–Pearson lemma. It forms a major component modern statistical hypothesis testing. However it was not part of Ronald Fisher's formulation of statistical hypothesis testing, and he violently opposed its use.[1] In Fisher's approach to testing, the central idea is to assess whether the observed dataset could have resulted from chance if the null hypothesis were assumed to hold, notionally without preconceptions about what other model might hold. Modern statistical hypothesis testing accommodates this type of test since tyhe alternative hypothesis can be just the negation of the null hypothesis.

Null hypothesis

n statistical hypothesis testing, the null hypothesis (H0) formally describes some aspect of the statistical behaviour of a set of data; this description is treated as valid unless the actual behaviour of the data contradicts this assumption. Thus, the null hypothesis is contrasted against another hypothesis. Statistical hypothesis testing is used to make a decision about whether the data contradicts the null hypothesis: this is called significance testing. A null hypothesis is never proven by such methods, as the absence of evidence against the null hypothesis does not establish it. In other words, one may either reject, or not reject the null hypothesis; one cannot accept it. Failing to reject it gives no strong reason to change decisions

predicated on its truth, but it also allows for the possibility of obtaining further data and then re-examining the same hypothesis.

Example 2

Suppose we wanted to determine whether a coin was fair and balanced. A null hypothesis might be that half the flips would result in Heads and half, in Tails. The alternative hypothesis might be that the number of Heads and Tails would be very different. Symbolically, these hypotheses would be expressed as
H0: P = 0.5
Ha: P â‰ 0.5
Suppose we flipped the coin 50 times, resulting in 40 Heads and 10 Tails. Given this result, we would be inclined to reject the null hypothesis and accept the alternative hypothesis.

Type I Error

In a hypothesis test, a type I error occurs when the null hypothesis is rejected when it is in fact true; that is, H0 is wrongly rejected.
For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; i.e.
H0: there is no difference between the two drugs on average.
A type I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them.
The following table gives a summary of possible results of any hypothesis test:

		Decision	
		Reject H0	Don't reject H0
Truth	H0	Type I Error	Right decision
	H1	Right decision	Type II Error

A type I error is often considered to be more serious, and therefore more important to avoid, than a type II error. The hypothesis test procedure is therefore

adjusted so that there is a guaranteed 'low' probability of rejecting the null hypothesis wrongly; this probability is never 0. This probability of a type I error can be precisely computed as

P(type I error) = significance level = α

The exact probability of a type II error is generally unknown.

If we do not reject the null hypothesis, it may still be false (a type II error) as the sample may not be big enough to identify the falseness of the null hypothesis (especially if the truth is very close to hypothesis).

For any given set of data, type I and type II errors are inversely related; the smaller the risk of one, the higher the risk of the other.

A type I error can also be referred to as an error of the first kind.

Type II Error

In a hypothesis test, a type II error occurs when the null hypothesis H0, is not rejected when it is in fact false. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; i.e.

H0: there is no difference between the two drugs on average.

A type II error would occur if it was concluded that the two drugs produced the same effect, i.e. there is no difference between the two drugs on average, when in fact they produced different ones.

A type II error is frequently due to sample sizes being too small.

The probability of a type II error is generally unknown, but is symbolised by β and written

P(type II error) = β

A type II error can also be referred to as an error of the second kind.

Level of significance

The **significance level** of a test is a traditional frequentist statistical hypothesis testing concept. In simple cases, it is defined as the probability of making a decision to reject the null hypothesis when the null hypothesis is actually true (a decision known as a Type I error, or "false positive determination"). The decision

is often made using the p-value: if the p-value is less than the significance level, then the null hypothesis is rejected. The smaller the p-value, the more significant the result is said to be.

In more complicated, but practically important cases, the significance level of a test is a probability such that the probability of making a decision to reject the null hypothesis when the null hypothesis is actually true is *no more than* the stated probability. This allows for those applications where the probability of deciding to reject may be much smaller than the significance level for some sets of assumptions encompassed within the null hypothesis.

The significance level is usually represented by the Greek symbol, α (alpha). Popular levels of significance are 5%, 1% and 0.1%. If a *test of significance* gives a p-value lower than the α-level, the null hypothesis is rejected. Such results are informally referred to as 'statistically significant'. For example, if someone argues that "there's only one chance in a thousand this could have happened by coincidence," a 0.1% level of statistical significance is being implied. The lower the significance level, the stronger the evidence.

In some situations it is convenient to express the statistical significance as $1 - \alpha$. In general, when interpreting a stated significance, one must be careful to note what, precisely, is being tested statistically.

Different α-levels have different advantages and disadvantages. Smaller α-levels give greater confidence in the determination of significance, but run greater risks of failing to reject a false null hypothesis (a Type II error, or "false negative determination"), and so have less statistical power. The selection of an α-level inevitably involves a compromise between significance and power, and consequently between the Type I error and the Type II error.

Fixed significance levels such as those mentioned above may be regarded as useful in exploratory data analyses. However, modern statistical advice is that, where the outcome of a test is essentially the final outcome of an experiment or other study, the p-value should be quoted explicitly. And, importantly, it should be quoted whether or not the p-value is judged to be significant. This is to allow

maximum information to be transferred from a summary of the study into meta-analyses.

Goodness of fit

The **goodness of fit** of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, e.g. to test for normality of residuals, to test whether two samples are drawn from identical distributions (see Kolmogorov-Smirnov test), or whether outcome frequencies follow a specified distribution (see Pearson's chi-square test). In the analysis of variance, one of the components into which the variance is partitioned may be a lack-of-fit sum of squares.

Example 3

(The approach of chi-square will be discussed in later sections however in order to illustrate goodness of fit we use it in this section)

The chi-square statistic is a sum of differences between observed and expected outcome frequencies, each squared and divided by the expectation:

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

where:

O = an observed frequency

E = an expected (theoretical) frequency, asserted by the null hypothesis

The resulting value can be compared to the chi-square distribution to determine the goodness of fit.

In order to determine the degrees of Freedom of the Chi-Squared distribution, one takes the total number of observed frequencies and subtracts one. For example, if there are eight different frequencies, one would compare to a chi-squared with seven degrees of freedom.

Another way to describe the chi-squared statistic is with the differences weighted based on measurement error:

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

where σ^2 is the variance of the observation. This definition is useful when one has estimates for the error on the measurements.

The reduced chi-squared statistic is simply the chi-squared divided by the number of degrees of freedom:

$$\chi^2_{red} = \frac{\chi^2}{\nu} = \frac{1}{\nu} \sum \frac{(O - E)^2}{\sigma^2}$$

where ν is the number of degrees of freedom, usually given by $N - n$, where N is the number of data points, and n is the number of fit parameters. The advantage of the reduced chi-squared is that it already normalizes for the number of data points and model complexity. As a rule of thumb, a large χ^2_{red} indicates a poor model fit. However $\chi^2_{red} < 1$ indicates that the model is 'over-fitting' the data (either the model is improperly fitting noise, or the error bars have been over-estimated). A $\chi^2_{red} > 1$ indicates that the fit has not fully captured the data (or that the error bars have been under-estimated). In principle a $\chi^2_{red} = 1$ is the best-fit for the given data and error bars.

Confidence interval

In statistics, a **confidence interval (CI)** is an interval estimate of a population parameter. Instead of estimating the parameter by a single value, an interval likely to include the parameter is given. Thus, confidence intervals are used to indicate the reliability of an estimate. How likely the interval is to contain the parameter is determined by the **confidence level** or confidence coefficient. Increasing the desired confidence level will widen the confidence interval.

A confidence interval is always qualified by a particular **confidence level** (say, γ), usually expressed as a percentage; thus one speaks of a "95% confidence interval". The end points of the confidence interval are referred to as **confidence limits**. For a given estimation procedure in a given situation, the higher the value of γ, the wider the confidence interval will be.

The calculation of a confidence interval generally requires assumptions about the nature of the estimation process – it is primarily a *parametric* method – for example, it may depend on an assumption that the distribution of errors of estimation is normal. As such, confidence intervals as discussed below are not robust statistics, though modifications can be made to add robustness – see robust confidence intervals.

Confidence intervals are used within Neyman-Pearson (frequentist) statistics; in Bayesian statistics a similar role is played by the credible interval, but the credible interval and confidence interval have different conceptual foundations and in general they take different values. As part of the general debate between frequentism and Bayesian statistics, there is disagreement about which of these statistics is more useful and appropriate, as discussed in alternatives and critiques.

2.3 HYPOTHESIS TESTS

Statisticians follow a formal process to determine whether to accept or reject a null hypothesis, based on sample data. This process, called hypothesis testing, consists of four steps.

State the hypotheses. This involves stating the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.

Formulate an analysis plan. The analysis plan describes how to use sample data to accept or reject the null hypothesis. The accept/reject decision often focuses around a single test statistic.

Analyze sample data. Find the value of the test statistic (mean score, proportion, t-score, z-score, etc.) described in the analysis plan. Complete other computations, as required by the plan.

Interpret results. Apply the decision rule described in the analysis plan. If the test statistic supports the null hypothesis, accept the null hypothesis; otherwise, reject the null hypothesis.

There are a number of tests available to test the relevance of hypotheses. Important ones are discussed in this chapter.

2.4 STUDENT'S T-TEST

A ***t*-test** is any statistical hypothesis test in which the test statistic has a Student's t distribution if the null hypothesis is true. It is applied when the population is assumed to be normally distributed but the sample sizes are small enough that the statistic on which inference is based is not normally distributed because it relies on an uncertain estimate of standard deviation rather than on a precisely known value.

Independent one-sample *t*-test

In testing the null hypothesis that the population mean is equal to a specified value μ_0, one uses the statistic

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}},$$

where s is the sample standard deviation of the sample and n is the sample size. The degrees of freedom used in this test is $n - 1$.

Independent two-sample *t*-test

Equal sample sizes, equal variance

This test is only used when both:

- the two sample sizes (that is, the n or number of participants of each group) are equal;

- it can be assumed that the two distributions have the same variance.

Violations of these assumptions are discussed below.

The *t* statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 X_2} \cdot \sqrt{\frac{2}{n}}}$$

where

$$S_{X_1 X_2} = \sqrt{\frac{S^2_{X_1} + S^2_{X_2}}{2}}$$

Here $S_{X_1 X_2}$is the grand standard deviation (or pooled standard deviation), 1 = group one, 2 = group two. The denominator of *t* is the standard error of the difference between two means.

For significance testing, the degrees of freedom for this test is $2n - 2$ where *n* is the number of participants in each group.

Unequal sample sizes, equal variance

This test is used only when it can be assumed that the two distributions have the same variance. (When this assumption is violated, see below.) The *t* statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$S_{X_1 X_2} = \sqrt{\frac{(n_1 - 1)S^2_{X_1} + (n_2 - 1)S^2_{X_2}}{n_1 + n_2 - 2}}.$$

Note that the formulae above are generalizations for the case where both samples have equal sizes (substitute n_1 and n_2 for n and you'll see).

$S_{X_1X_2}$is an estimator of the common standard deviation of the two samples: it is defined in this way so that its square is an unbiased estimator of the common variance whether or not the population means are the same. In these formulae, n = number of participants, 1 = group one, 2 = group two. $n - 1$ is the number of degrees of freedom for either group, and the total sample size minus two (that is, $n_1 + n_2 - 2$) is the total number of degrees of freedom, which is used in significance testing.

Unequal sample sizes, unequal variance

This test is used only when the two sample sizes are unequal and the variance is assumed to be different. See also Welch's t test. The t statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

where

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where s^2 is the unbiased estimator of the variance of the two samples, n = number of participants, 1 = group one, 2 = group two. Note that in this case, $s^2_{\overline{X}_1 - \overline{X}_2}$is not a pooled variance. For use in significance testing, the distribution of the test statistic is approximated as being an ordinary Student's t distribution with the degrees of freedom calculated using

$$\text{D.F.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

This is called the Welch-Satterthwaite equation. Note that the true distribution of the test statistic actually depends (slightly) on the two unknown variances: see Behrens–Fisher problem.

Dependent *t*-test for paired samples

This test is used when the samples are dependent; that is, when there is only one sample that has been tested twice (repeated measures) or when there are two samples that have been matched or "paired".

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{N}}.$$

Example 4

Consider the following research problem: We have a random sample of 25 fifth grade pupils who can do 15 pushups on the average, with a standard deviation of 9, after completing a special physical education program. Does this value of 15 differ significantly from the population value of 12?

In this problem we are comparing a sample mean with a population mean but we do not know the population standard deviation. We can't use the Z-test in this case but we can use the one-sample t-test. The one sample t-test does not require the population standard deviation. The formula for the one-sample t-test is

$$t = \frac{\overline{X} - \mu}{S_{\overline{X}}}$$

Where $\overline{X}$ is the sample mean,

μ is the population mean,

and $S_{\overline{X}}$ is the sample estimate of the standard error of the mean.

In the problem we are considering, we do not know the population standard deviation (or the standard error of the mean) so we estimate it from the sample data. The sample estimate of the standard error of the mean is based on S (the sample standard deviation) and the square root of n (the sample size).

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

If you look back at the research problem you will see that we have all the data we need to calculate the value of t.

The sample mean, $\bar{X}$ is 15.

The population mean, μ is 12.

The sample standard deviation, S is 9.

The sample size, n is 25.

We can thus calculate the value of t as follows:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{9}{\sqrt{25}} = \frac{9}{5} = 1.8$$

$$t = \frac{\bar{X} - \mu}{S_{\bar{X}}} = \frac{15 - 12}{1.8} = \frac{3}{1.8} = 1.667$$

The t statistic is not distributed normally like the z statistic is but is distributed as (guess what) the t-distribution, also referred to as student's distribution. We will use this distribution when we do the six step process for testing statistical hypothesis. To use the table for the t-distribution we need to know one other piece of information and that is the degrees of freedom for the one sample t-test.

2.5 Z-TEST

A **Z-test** is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Since many test statistics are approximately normally distributed for large samples (due to the central limit theorem), many statistical tests can be performed as approximate Z-tests if the sample size is not too small. In addition, some statistical tests, such as comparisons of means between two samples, or a comparison of the mean of one sample to a given constant, are exact Z-tests under certain assumptions.

One sample location test

The term *Z-test* is often used to refer specifically to the one-sample location test comparing the mean of a set of measurements to a given constant. If the observed data X_1, ..., X_n are (i) uncorrelated, (ii) have a common mean μ, and (iii) have a common variance σ^2, then the sample average $\bar{X}$ has mean μ and variance σ^2/n. If our null hypothesis is that the mean value of the population is a given number μ_0, we can use $\bar{X} - \mu_0$ as a test-statistic, rejecting the null hypothesis if $\bar{X} - \mu_0$ is large.

In order to calculate the standardized statistic $Z = (\bar{X} - \mu_0)/s$, we need to either know or have an approximate value for σ^2, from which we can calculate $s^2 = \sigma^2 / n$. In some applications, σ^2 is known, but this is uncommon. If the sample size is moderate or large, we can substitute the sample variance for σ^2, giving a *plug-in* test. The resulting test will not be an exact Z-test since the uncertainty in the sample variance is not accounted for — however, it will be a good approximation unless the sample size is small. A t-test can be used to account for the uncertainty in the sample variance when the sample size is small and the data are exactly normal. There is no universal constant at which the sample size is generally considered large enough to justify use of the plug-in test. Typical rules of thumb range from 20 to 50 samples. For larger sample sizes, the t-test procedure gives almost identical p-values as the Z-test procedure

Example 5

Suppose that in a particular U.S. state, the mean and standard deviation of scores on a reading test are 100 points, and 12 points, respectively. Our interest is in the scores of 55 fifth grade students in a particular elementary school who received a mean score of 96. We can ask whether this mean score is significantly lower than the state level mean — that is, are the students in this school comparable to a simple random sample of 55 students from the state as a whole, or are their scores surprisingly low?

We begin by calculating the standard error of the mean:

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{55}} = \frac{12}{7.42} = 1.62$$

Next we calculate the z-score, which is the distance from the sample mean to the population mean in units of the standard error:

$$z = \frac{M - \mu}{\mathrm{SE}} = \frac{96 - 100}{1.62} = -2.47$$

In our example, the mean score of 96 is −2.47 standard error units from the population mean of 100. Looking up the z-score in a table of the standard normal distribution, we find that the probability of observing a standard normal value below -2.47 is approximately 0.0068. This is the one-sided p-value for the null hypothesis that the 55 students are a simple random sample from the population of test-takers in the state. The two-sided p-value is approximately 0.014 (twice the one-sided p-value).

Another way of stating things is that with probability 1-0.014=0.986, a simple random sample of 55 students would have a mean test score within 4 units of the population mean. We could also say that with 98% confidence we reject the null hypothesis that the the 55 test takers are a simple random sample from the population of test-takers.

The Z-test tells us that the 55 students of interest have an unusually low mean test score compared to most simple random samples of similar size from the population of test-takers. A deficiency of this analysis is that it does not consider

whether the effect size of 4 points is meaningful. If the average score of 900 students (say, all students in a county) were 99, nearly the same z-score and p-value would be observed, showing that if the sample size is large enough, very small differences from the null value can be highly statistically significant. See statistical hypothesis testing for further discussion of this issue.

2.6 CHI-SQUARE TEST

A **chi-square test** (also **chi-squared** or χ^2 **test**) is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true, or any in which this is *asymptotically* true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-square distribution as closely as desired by making the sample size large enough.

Some examples of chi-squared tests where the chi-square distribution is only approximately valid:

- **Pearson's chi-square test**, also known as the chi-square goodness-of-fit test or chi-square test for independence. When mentioned without any modifiers or without other precluding context, this test is usually understood.
- Yates' chi-square test, also known as Yates' correction for continuity.
- Mantel-Haenszel chi-square test.
- Linear-by-linear association chi-square test.
- The portmanteau test in time-series analysis, testing for the presence of autocorrelation
- Likelihood-ratio tests in general statistical modelling, for testing whether there is evidence of the need to move from a simple model to a more complicated one (where the simple model is nested within the complicated one).

One case where the distribution of the test statistic is an exact chi-square distribution is the test that the variance of a normally-distributed population has a

given value based on a sample variance. Such a test is uncommon in practice because values of variances to test against are seldom known exactly.

Pearson's chi-square (χ^2) **test** is the best-known of several chi-square tests – statistical procedures whose results are evaluated by reference to the chi-square distribution. Its properties were first investigated by Karl Pearson. In contexts where it is important to make a distinction between the test statistic and its distribution, names similar to **Pearson X-squared** test or statistic are used.

It tests a null hypothesis that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The events considered must be mutually exclusive and have total probability 1. A common case for this is where the events each cover an outcome of a categorical variable. A simple example is the hypothesis that an ordinary six-sided die is "fair", i.e., all six outcomes are equally likely to occur. Pearson's chi-square is the original and most widely-used chi-square test.

Test for fit of a distribution

In this case N observations are divided among n cells. A simple application is where it is required to test the hypothesis that, in the general population, values would occur in each cell with equal frequency. Then the "theoretical frequency" for any cell (under the null hypothesis of a discrete uniform distribution) is calculated as

$$E_i = N/n,$$

and the reduction in the degrees of freedom is p=1: notionally because the observed frequencies O_i are constrained to sum to N. When testing whether observations are random variables whose distribution belongs to a given family of distributions, the "theoretical frequencies" are calculated using a distribution from that family fitted in some standard way.

The reduction in the degrees of freedom is calculated as $p=s+1$, where s is the number of parameters used in fitting the distribution. For instance, when checking

a 3-parameter Weibull distribution, p=4, and when checking a normal distribution (where the parameters are mean and standard deviation), p=3.

In other words, there will be $(n - p)$ degrees of freedom, where n is the number of categories. It should be noted that the degrees of freedom are not based on the number of observations as with a Student's t or F-distribution. For example, if testing for a fair, six-sided die, there would be five degrees of freedom because there are six categories/parameters (each number). The number of times the die is rolled will have absolutely no effect on the number of degrees of freedom.

The value of the test-statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

where

X^2 = the test statistic that asymptotically approaches a χ^2 distribution.
O_i = an observed frequency;
E_i = an expected (theoretical) frequency, asserted by the null hypothesis;
n = the number of possible outcomes of each event.

The chi-square statistic can then be used to calculate a p-value by comparing the value of the statistic to a chi-square distribution. The number of degrees of freedom is equal to the number of cells n, minus the reduction in degrees of freedom, p.

The result about the number of degrees of freedom is valid when the original data was multinomial and hence the estimated parameters are efficient for minimizing the chi-square statistic. More generally however, when maximum likelihood estimation does not coincide with minimum chi-square estimation, the distribution will lie somewhere between a chi-square distribution with $n - 1 - p$ and $n - 1$ degrees of freedom (See for instance Chernoff and Lehmann 1954).

Bayesian method

In Bayesian statistics, one would instead use a Dirichlet distribution as conjugate prior. If one took a uniform prior, then the maximum likelihood estimate for the population probability is the observed probability, and one may compute a credible region around this or another estimate.

Test of independence

In this case, an "observation" consists of the values of two outcomes and the null hypothesis is that the occurrence of these outcomes is statistically independent. Each outcome is allocated to one cell of a two-dimensional array of cells (called a table) according to the values of the two outcomes. If there are r rows and c columns in the table, the "theoretical frequency" for a cell, given the hypothesis of independence, is

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \sum_{k=1}^{r} O_{k,j}}{N},$$

and fitting the model of "independence" reduces the number of degrees of freedom by $p = r + c - 1$. The value of the test-statistic is

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

The number of degrees of freedom is equal to the number of cells rc, minus the reduction in degrees of freedom, p, which reduces to $(r - 1)(c - 1)$.

For the test of independence, a chi-square probability of less than or equal to 0.05 (or the chi-square statistic being at or larger than the 0.05 critical point) is commonly interpreted by applied workers as justification for rejecting the null hypothesis that the row variable is unrelated (that is, only randomly related) to the column variable.[1] The alternative hypothesis corresponds to the variables having an association or relationship where the structure of this relationship is not specified.

Example 6

For example, to test the hypothesis that a random sample of 100 people has been drawn from a population in which men and women are equal in frequency, the observed number of men and women would be compared to the theoretical frequencies of 50 men and 50 women. If there were 45 men in the sample and 55 women, then

$$X^2 = \frac{(45-50)^2}{50} + \frac{(55-50)^2}{50} = 1.$$

If the null hypothesis is true (i.e., men and women are chosen with equal probability in the sample), the test statistic will be drawn from a chi-square distribution with one degree of freedom. Though one might expect two degrees of freedom (one each for the men and women), we must take into account that the total number of men and women is constrained (100), and thus there is only one degree of freedom (2 – 1). Alternatively, if the male count is known the female count is determined, and vice-versa.

Consultation of the chi-square distribution for 1 degree of freedom shows that the probability of observing this difference (or a more extreme difference than this) if men and women are equally numerous in the population is approximately 0.3. This probability is higher than conventional criteria for statistical significance (.001-.05), so normally we would not reject the null hypothesis that the number of men in the population is the same as the number of women (i.e. we would consider our sample within the range of what we'd expect for a 50/50 male/female ratio.)

2.7 F-TEST

An **F-test** is any statistical test in which the test statistic has an F-distribution if the null hypothesis is true. In the simplest case, it is used to examine the effect of some factor on some outcome. A factor groups events into any number of

different categories, and we'd like to know if these different categories can help predict different outcomes. The null hypothesis presumes the factor will NOT effect the outcome. Differences in the outcome's variation *between* factor groups and *within* factor groups should simply be due to chance. However, when the variation *between* factor groups is much greater than the variation *within* factor groups, then F will be high, and the probability that the factor has no effect (the p-value) will typically be lower (depending on the degrees of freedom). Then, the null hypothesis is highly unlikely, and we can conclude that the factor has a statistically significant effect (well, if our data actually adhere to gaussian presumptions).

The name was coined by George W. Snedecor, in honour of Sir Ronald A. Fisher. Fisher initially developed the statistic as the variance ratio in the 1920s.[1] Examples include:

- The hypothesis that the means of multiple normally distributed populations, all having the same standard deviation, are equal. This is perhaps the most well-known of hypotheses tested by means of an F-test, and the simplest problem in the analysis of variance (ANOVA).

- The hypothesis that a proposed regression model fits well. See Lack-of-fit sum of squares.

- The hypothesis that the standard deviations of two normally distributed populations are equal, and thus that they are of comparable origin.

Note that if it is equality of variances (or standard deviations) that is being tested, the F-test is extremely non-robust to non-normality. That is, even if the data displays only modest departures from the normal distribution, the test is unreliable and should not be used.

Formula and calculation

The value of the test statistic used in an F-test consists of the ratio two different estimates of quantities which are the same according to the null hypothesis being tested. If the null hypothesis were true and if estimated values were not being

used, this ratio would have a value of 1: however, because estimated values are used, F would sometimes be above or below 1. If the null hypothesis is not true the ratio would be rather different from 1. In the usual applications, statistical modelling assumptions are made founded on using the normal distribution to describe random errors and the estimates used in the ratio are statistically independent but are typically derived from the same data set.

In the case of multiple-comparison ANOVA problems, the F-test is used to test if the variance measuring the differences between groups in a certain pre-defined grouping of observations is large compared to the variance measuring the differences within the groups: a large value would tend to suggest that grouping is *good* or *valid* in some sense, or that there are real differences between the groups. The formula for an F-test is:

$$F = \frac{(\text{explained variance})}{(\text{unexplained variance})},$$

or:

$$F = \frac{(\text{between-group variability})}{(\text{within-group variability})},$$

where the quantities on the top and bottom of this ratio are each unbiased estimates of the within-group variance on the assumption that the between group variance is zero. Note that when there are only two groups for the F-test,

$$F = t^2,$$

where t is the Student's t statistic.

One-way ANOVA example

Consider an experiment to study the effect of three different levels of some factor on a response (e.g. three types of fertilizer on plant growth). If we had 6 observations for each level, we could write the outcome of the experiment in a table like this, where a_1, a_2, and a_3 are the three levels of the factor being studied.

a_1	a_2	a_3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

The null hypothesis, denoted H_0, for the overall F-test for this experiment would be that all three levels of the factor produce the same response, on average. To calculate the F-ratio:

Step 1: Calculate the A_i values where i refers to the number of the condition. So:

$$A_1 = \sum a_1 = 6 + 8 + 4 + 5 + 3 + 4 = 30$$
$$A_2 = \sum a_2 = 8 + 12 + 9 + 11 + 6 + 8 = 54$$
$$A_3 = \sum a_3 = 13 + 9 + 11 + 8 + 7 + 12 = 60$$

Step 2: Calculate $\bar{Y}_{Ai}$ being the average of the values of condition a_i

$$\overline{Y}_{A1} = \frac{A_1}{n} = \frac{30}{6} = 5$$
$$\overline{Y}_{A2} = \frac{A_2}{n} = \frac{54}{6} = 9$$
$$\overline{Y}_{A3} = \frac{A_3}{n} = \frac{60}{6} = 10$$

Step 3: Calculate these values:

Total:

$$T = \sum A_i = A_1 + A_2 + A_3 = 30 + 54 + 60 = 144$$

Average overall score:

$$\overline{Y}_T = \frac{T}{a(n)} = \frac{144}{3(6)} = 8$$

where a = the number of conditions and n = the number of participants in each condition.

$[Y] = \sum\left(Y^2\right) = 1304$

This is every score in every condition squared and then summed.

$$[A] = \frac{\sum(A_i^2)}{n} = 1236$$

$$[T] = \frac{T^2}{a(n)} = 1152$$

Step 4: Calculate the sum of squared terms:

$SS_A = [A] - [T] = 84$

$SS_{S/A} = [Y] - [A] = 68$

Step 5: The degrees of freedom are now calculated:

$df_a = a - 1 = 3 - 1 = 2$

$df_{S/A} = a(n - 1) = 3(6 - 1) = 15$

Step 6: The Means Squared Terms are calculated:

$$MS_A = \frac{SS_A}{df_A} = 42$$

$$MS_{S/A} = \frac{SS_{S/A}}{df_{S/A}} = 4.5$$

Step 7: Finally the ending F-ratio is now ready:

$$F = \frac{MS_A}{MS_{S/A}} = 9.27$$

Step 8: Look up the F_{crit} value for the problem:

$F_{crit}(2,15) = 3.68$ at $\alpha = 0.05$. Since $F = 9.27 \geq 3.68$, the results are significant at the 5% significance level. One would reject the null hypothesis, concluding that the three levels of the factor in this experiment do not all produce the same response on average.

Note $F(x, y)$ denotes an F-distribution with x degrees of freedom in the numerator and y degrees of freedom in the denominator.

Activity 2

1.A drug manufacturing company conducted a survey of customers. The research question is: Is there a significant relationship between packaging preference (size of the bottle purchased) and economic status? There were four packaging sizes: small, medium, large, and jumbo. Economic status was: lower, middle, and upper. The following data was collected.

	Lower	Middle	Upper
Small	24	22	18
Medium	23	28	19
Large	18	27	29
Jumbo	16	21	33

2. Discuss different tests that can be applied to hypotheses testing.

3. Here are the results of a public opinion poll broken down by gender. What is the exact probability that the difference between the observed and expected frequencies occurred by chance? (Hint: apply the f-test)

	Male	Female
Favor	10	14
Opposed	15	9

2.8 SUMMARY

A hypothesis is a specific statement of prediction. It describes in concrete (rather than theoretical) terms what you expect will happen in your study. A single study may have one or many hypotheses formulations. Basic concepts of hypotheses testing such as confidence interval, level of confidence, degrees of freedom, type I and II errors and concepts of alternative and null hypotheses must be kept in mind while testing the hypotheses. In order to prove the relevance of hypotheses, hypotheses testing are important. A number of tests are available to test the hypotheses important of which are discussed in detail in this chapter, the student's T test, chi-square test, Z-test and f-test.

2.9 FURTHER READINGS

- Cramer, Duncan; Dennis Howitt (2004). *The Sage Dictionary of Statistics*. p. 76.
- Lehmann, E.L.; Joseph P. Romano (2005). *Testing Statistical Hypotheses* (3E ed.). New York: Springer.
- Fisher, Sir Ronald A. (1956) [1935]. "Mathematics of a Lady Tasting Tea". in James Roy Newman.
- Box, Joan Fisher (1978). *R.A. Fisher, The Life of a Scientist*. New York: Wiley
- McCloskey, Deirdre (2008). *The Cult of Statistical Significance*. Ann Arbor: University of Michigan Press.
- Wallace, Brendan; Alastair Ross (2006). *Beyond Human Error*. Florida: CRC Press.
- Harlow, Lisa Lavoie; Stanley A. Mulaik; James H. Steiger (1997). *What If There Were No Significance Tests?*. Mahwah, N.J.: Lawrence Erlbaum Associates Publishers

Answers to activity

Activity 2

1. Chi-square statistic = 9.743
   Degrees of freedom = 6
   Probability of chance = .1359

3. Fisher's exact probability = .0828

वाग्देवी तानुपान्महे

MADHYA PRADESH BHOJ (OPEN) UNIVERSITY

RAJA BHOJ MARG (KOLAR ROAD), BHOPAL- 462016