Statistics

Sample Readings

Anusha Illukkumbura
MSc. Business Statistics (Sri Lanka)
B.A. Social Statistics (Sri Lanka)
PhD Candidate at Southern Illinois University (United States)

Acknowledgement

First Edition

First Edition: January 2023
ISBN: 9798373806213

Preface

This sample reading book includes the introducing chapters of my books on Probability and Distributions, Introduction to Regression Analysis and Introduction to Hypothesis Testing.

The purpose of this book is to provide a free sample reading for my readers who have requested so.

Since I am currently pursuing my Doctoral degree, it will take me a while to publish a new book. I thank my readers for their support and encouragement since 2020.

Any suggestions to further improve the contents of this edition would be warmly appreciated. For any further suggestions, please contact me via website anushabooks.com

Anusha Illukkumbura©
MSc. Business Statistics (University of Moratuwa, Sri Lanka)
B.A. Social Statistics (University of Kelaniya, Sri Lanka)
January 2023

Book : Probability and Distribution

CHAPTER ONE: SET THEORY

1.1 Introduction

Set theory is a statistical theory, which is based on collection of well-defined elements. It is important to understand about set theory before learning probability.

1.2 Experiment

Experiment is a collection of well defined events which can be done for unlimited and repetitive times. Experiment is random and it has one or more possible outcomes. Experiment is also called as a trial. Outcomes of the trial are events.

Events

Set of results of an experiment are called events. It is the collection of one or more outcome of an experiment. A specific outcome of an experiment is called as a sample point; furthermore it is the most basic outcome. Each outcome of sample space is a sample point. Simple event is an event with possibility of only one outcome. When there is more than one possible outcome that event is called a compound or joint event. Favorable events are the number of desired outcomes in an experiment. When two or more events have an equal opportunity of happening, they are called equally likely events.

Incident of a child selecting a numbered ball out of a box

There are 100 balls numbered from 1 to 100 in a black bag. This bag belongs to a child named Peter. In this bag, balls are colored in 5 colors: red, blue, yellow, green and white.

Number 1-20 are colored in yellow, 21-40 in green, 41-60 in blue , 61-80 in red and 81-100 in white.

The child can take balls out of the bag for any number of times with replacement. Taking out the balls out of bag is the experiment.

Event: A child randomly takes out number 10 ball from the box. This is the event of taking number 10 out of the bag.

Sample point: number of the ball he took out for the *first time*. He can take only one ball for the first time. Therefore there is only one sample point in this event.

Simple event: A child takes out only one ball from the box.

Compound event: Child is taking out two balls simultaneously out of the bag.

Favorable events: Child wishes to take an even numbered ball out of the box. There are 50 even numbers between 1 to100. So the favorable event is taking out any of those 50 even numbers out of the box.

Equally likely events: there are 50 odd numbers and 50 even numbers from 1 to 100. Taking out an odd number or an even number are equally likely events.

Sample Space

Sample space is the set of all possible outcomes in an experiment. Sample space is represented by "S".

All the balls numbered from 1 to 100 are included in sample space.

Event Space

Event space is the all possible sets and subsets of outcomes in an event. This is different from sample space. Members of sample space are sample points. But members of event space are sets and subsets.

Independent events and Dependent events

If one event doesn't have an effect on another event then those two events are independent. But when one event has an impact on another event it is called a dependent event.

The child takes one ball out of the box randomly without looking inside the bag. This doesn't have any effect from any other event. This is an independent event. He doesn't like the color of the ball. He puts it back; take another ball again without looking inside the bag (randomly). This is also an independent event as there are same amount of balls in the bag (100) as the child put it back (the first event doesn't affect the second event). This event is also called as an event with replacement.

Child is taking one ball out of the bag, he likes the color keep it with him, and without replacing it he takes another ball. This second event is dependent. The child doesn't replace the first ball. So there are only 99 balls in the bag in the second trial. The first event has influence on the outcome of the second event.

Mutually Exclusive Events

When two or more events cannot happen at the same time it is called mutually exclusive events. If A and B are two mutually exclusive events then $P(A \cap B) = 0$, which means intersection is null. (This concept will be discussed in coming sections.)

When the child takes only one ball out of the bag, there is no possibility that the number of this ball is both even and odd. Therefore the probability of taking out both even and odd number is zero. Therefore event of taking a ball with both even and odd numbers is a mutually exclusive event.

Collectively Exhaustive Events

At least one event out of set of events must happen.

When the child selects the ball, the ball should either be an even or an odd number. Therefore at least one outcome of odd or even numbered ball should happen. Taking a ball with an odd or an even number is a collectively exhaustive event.

1.3 Sets

Set is a well defined collection of elements. Each object of the set is an element.
Sets can be named using English capital letters (Ex: A,B,C).
In the example of child selecting a ball, set of all the numbers written on the balls inside the bag, has elements from 1 to 100.
This can be written as A= {1,2,3,…,99,100} , where A is the set of numbers written on balls inside the bag.

Element

Element is any member of a set.
Numbers from 1 to 100 are the members of the set of "numbers written on the balls inside the bag".

Mike, Saman, Nil, Ravi and Amal are five boys in a class who weight between 50kg to 60 kg. This group can be defined as a set; it can be named as "A".
Mike, Saman, Nil, Ravi and Amal are elements of set "A", but John who weights 70 kg is not an element of that set.

Element of a set can be written using element symbol "∈", Mike ∈ A, Saman ∈ A, Nil∈ A, Ravi∈ A and Author∈ A.

∉ is the symbol of "not an element of". When an event is not an element of a set, it is written as John ∉ A.

Null-set

When there is no element in a set, that set is called a null set.

The bag, child has only contained of balls. So getting cards out of that bag is not happening. Set of getting a card out of Peter's bag is a null set.

Sub Set

When all the elements of one set (Set A) is included in another set (Set B) , then set A is a sub set of set B. It is symbolized as, A ⊂ B.

In Peter's bag, there are 50 even numbers. Set of these even numbers are included in the set of numbers from 1 to 100.

A- Set of even numbers
B- Set of numbers from 1 to 100.

Therefore $A \subset B$.

Equal Sets

When two or more sets have the same elements, then those sets are called equal sets (A=B)
When $A \subset B$ and $B \subset A$ then they are A=B

Balls from number 1 to 20 of Peter's bag are colored in yellow. They are the only ones colored in the yellow in the bag.

A- Set of yellow colored balls in the bag
B- Set of all numbers from 1 to 20

In this case ,$A \subset B$ and $B \subset A$, therefore A=B

Universal Set

Set of all the possible elements of a set is called a universal set. This is represented by "U". This is similar to the sample space. Collection of all the colored balls with numbers written on, is the universal set of balls in Peter's bag.

Set Union

When there are two sets called A and B, the set includes the elements, which are belonged to either set A or B or both, is called set union. It is represented by $A \cup B$

A- Yellow colored balls in Peter's bag
B- Red colored balls in Peter's bag
$A \cup B$ –yellow **or** red colored balls in Peter's bag (There are 40 balls)

Set Intersection

When there are two sets called A and B, the set includes the elements, which are belonged to both A and B, is called set intersection. It is represented by
$A \cap B$

A- numbers from 11 to 50 in Peter's bag

B- Yellow colored balls in Peter's bag

A∩B –Set of both yellow colored **and** numbers from 11to 50 in Peter's bag

Questions with set union have the word "or" meaning of "or". Questions with intersection have the word "and" or its meaning.

Venn Diagrams

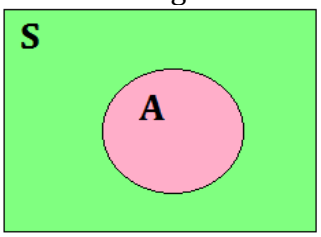
Venn is type of a diagram, which is used to illustrate the details of sets.

If S = Birds, A = Migratory birds and B = Hummingbirds

1.1 Venn Diagram

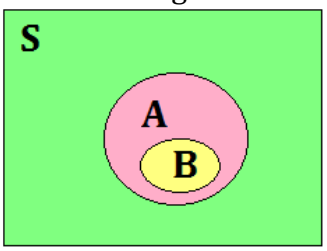
S is the universal set

1.2 Venn Diagram



A is a set

1.3 Venn Diagram



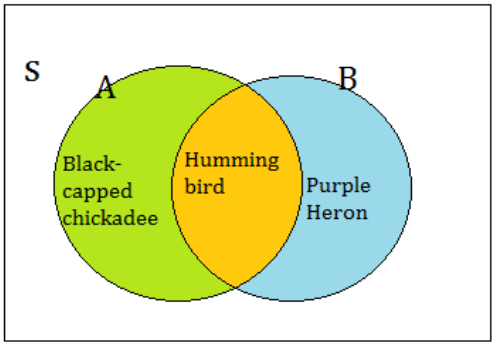
In this Venn diagram 1.3 ,

$B \subset A$

When A = Small Birds, B= Migratory birds

In Venn diagram 1.4, universal set (S) is birds. Black-capped chickadee is a non migratory small bird, so it belongs to set of small birds (A). Purple heron is a migratory big bird, so it belongs to set of migratory birds (B). Humming bird is both small and migratory bird so it belongs to the both category. Humming bird is an element of set intersection. If someone wants to take all the migrant or small or both small and migrant birds as a sample then that person should take the all elements of the set union of this example.

1.4 Venn Diagram



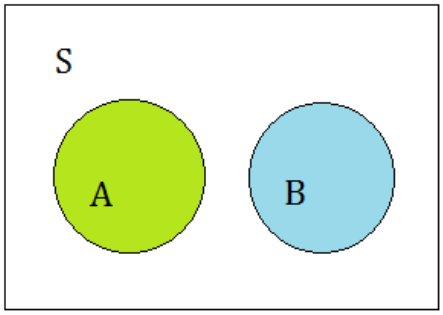
A- Small birds , Migrant birds
A = {Black-capped chickadee, Humming bird}
B= {Purple Heron, Humming bird}
A∩B – Small and Migrant birds
A∩B ={Humming birds}
A∪B – Small or migrant birds
A∪B = {Black-capped chickadee, Humming bird, Purple Heron}

Disjoint Set

When there are no common elements among two sets, it is called a disjoint set.
It is represented by A∩B =⌀

When S= Animals , A= Birds and B = Centipedes
then A∩B =⌀ (Venn diagram is illustrated in 1.5 Venn diagram of disjoint sets).

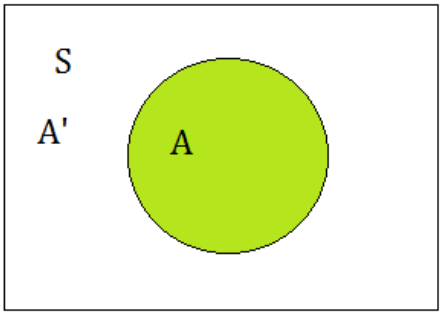
1.5 Venn Diagram of disjoint sets



Complement of a set

Complement of a set is the elements that are not included in a specific set, but they are included in universal set. It is represented by $A' = \{x \in S \mid x \notin A\}$.

1.6 Venn Diagram of complement of a set



Number of elements in the set is called the **order of the set.** It is described by "n". Order of the set A is represented by n(A).

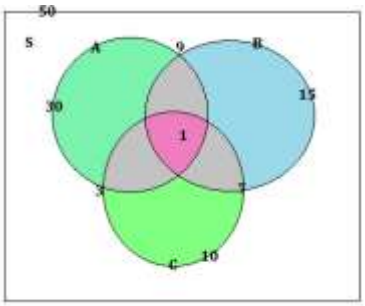
If a set is countable and finite it is called **finite set** and if it is uncountable it is called **infinite set**. Number of students in a classroom is a finite set. Number of stars on the sky is infinite set.

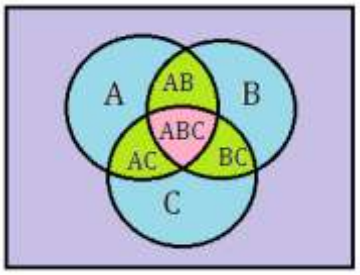
Example 1.1: Venn diagram

Bakery products manufacturing company tested 50 samples on chocolate cakes, coffee cakes and butter cakes using 50 customers. The company found that 30 of them liked chocolate cakes, 15 liked coffee cakes, 10 liked butter cakes, 7 liked both coffee and butter cakes and also 9 liked both coffee and chocolate cakes. 3 liked both chocolate and butter cakes. One liked all three types of cakes.

If A- People who like chocolate cakes
B- People who like coffee cakes
C- People who like butter cake

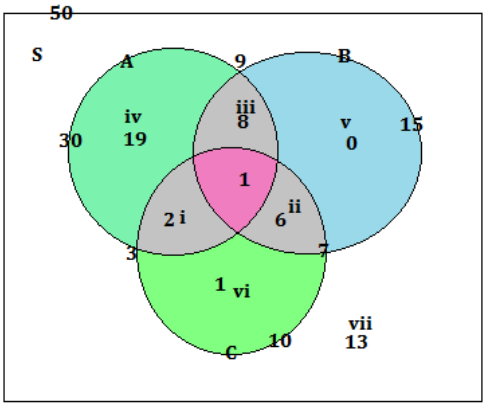
Above information can be illustrated in a Venn diagram as below





Calculate the missing information using the steps indicated by roman numbers (refer next page)

i. 3-1 =2
ii. 7-1 =6
iii. 9-1=8
iv. 30-8-2-1= 19
v. 15-8-1-6 =0
vi. 10-6-1-2
vii. 50-30-0-6-1 =13



Using the given information calculate

I) Number of people who like only two types of cake
= 8+6+2=16

II) Number of people who like only one type of cake
= 19+0+1 =20

III) Number of people who likes at least one type of cake
=8+2+6+1+19+0+1=37

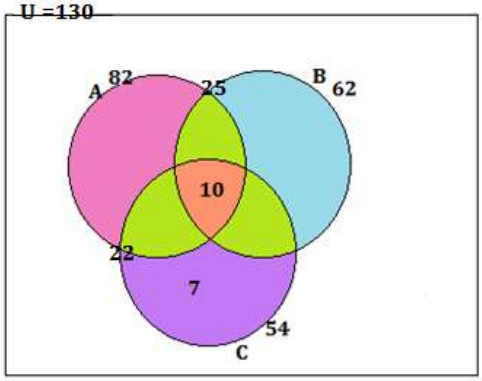
IV) Ones who like chocolate or coffee cake but not butter cake. = 19+ 8+0 =27

V) People who likes none of them
50-30-0-6-1 =13

Example 1.2: Venn diagram

130 tourists were questioned on three luxury restaurants in a popular tourist destination, "Sigiriya". Luxury restaurants are called A,B and C. 82 of them have visited A and 62 have visited B. 54 has visited C. 22 of them have visited both A and C , 25 of them have visited A and B.7 has visited C but not A or B.10 has visited all three restaurants.

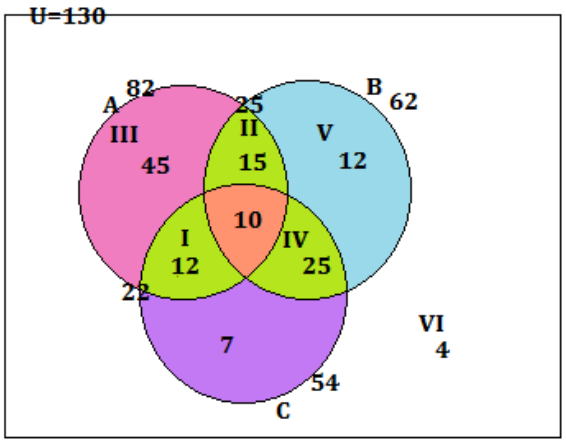
This information can be drawn in a Venn diagram as below



How to calculate the missing information?

Calculation steps are presented in roman numbers

I. 22-10=12
II. 25-10=15
III. 82- 12-10-15 =45
IV. 54-7-12-10=25
V. 62-15-10-25=12
VI. 130 –82-12-25-7= 4



I) Number of people who have visited both B and C but not A =25

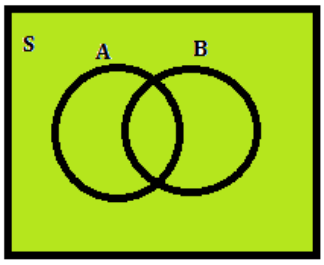
II) Number of people who have visited only B
=12

III) Number of people who have visited only has visited A
=45

IV) Number of visited who have not visited any of these three places
=4

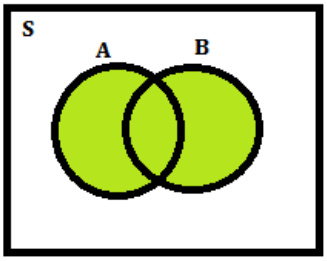
Below are some definitions in Set calculations illustrated using Venn diagrams

Universal Set

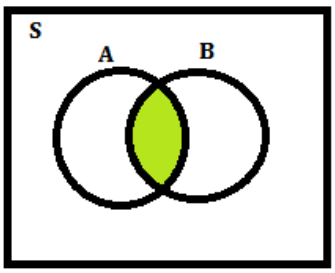


P(A∩B)

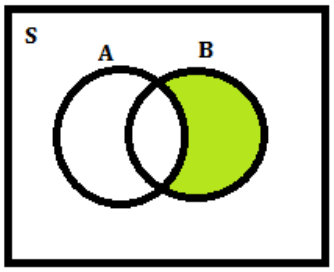
P(A∪B)



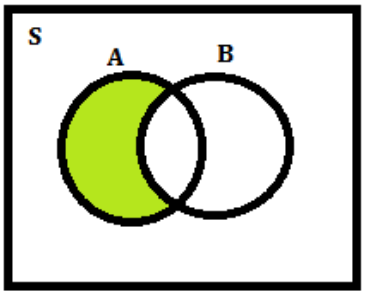
P(B-A)= P(B)- P(A∩B)



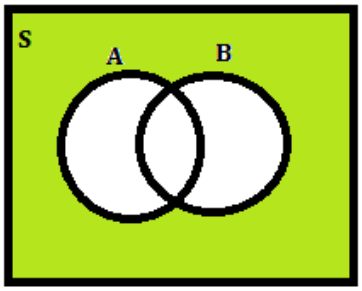
P(A-B)= P(A)- P(A∩B)



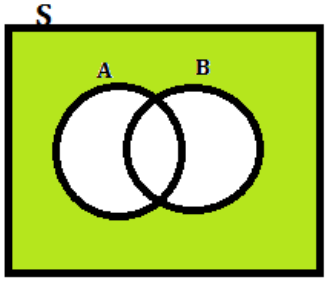
P(AUB)'



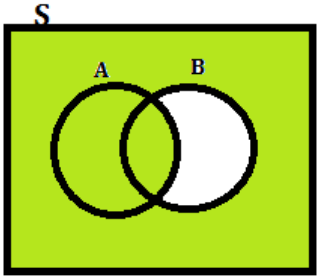
A'∩B' = 1-(AUB)



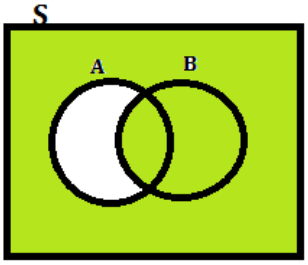
(AUB')

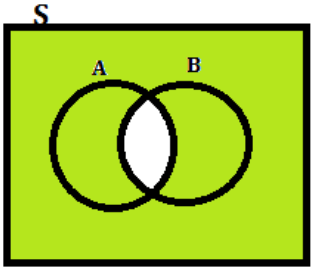


(A'UB)



(A'UB')=1- (A∩B)





1.4 Factorial Notation

Factorial notation is appeared as an exclamation mark following the number. Example: 10!, 21!.

5! is similar to multiplication of all the consecutive integers from 1 to 5.

Example: 5! = 5*4*3*2*1 =120

n! is similar to multiplication of all the consecutive integers from 1 to n.
Example: n! = 1 *2*3*…*n

Example 1.3: Factorial Notation

Solve below questions

I) 2!7!
$= (1*2)* (1*2*3*4*5*6*7)$
$= 2* 5040$
$=1080$

II) $\frac{8!}{4!}$
$=\frac{(1*2*3*4*5*6*7*8)}{(1*2*3*4)}$
$=\frac{40,320}{24}$
$=1680$

III) $\frac{8!+3!}{4!-2!}$
$= \frac{40,320+6}{24-2}$
$= \frac{40326}{22}$
$= 1833$

IV) $\frac{1}{4!}+\frac{2}{3!}$
$=\frac{1}{(1*2*3*4)}+\frac{2}{1*2*3}$
$=\frac{1}{24}+\frac{1}{3}$
$= 0.042 + 0.33$
$=0.372$

V) $\frac{\frac{5!}{2!}}{4!*3!}$
$= \frac{\frac{1*2*3*4*5}{1*2}}{(1*2*3*4)*(1*2*3)}$
$= \frac{3*4*5}{24*6}$
$=\frac{5}{12}$

Example 1.4: Factorial Notation

Write below numbers in factorial notation

I) 12
$=\frac{12!}{11!}$

II) 8*7*6
$=\frac{8!}{5!}$

III) $\frac{7*6*5}{2*3*4}$
$=\frac{7!}{4!}$

1.5 Permutation

“r” number of items which are taken out of total of “n” number of items that can be arranged in an ordered way is permutation.

It can also be interpreted as an ordered arrangement of "r" numbers of different elements that are selected from total of "n" numbers of elements. Order is very important in permutation. r should be less than or equal to n.

It is represented by ${}^{n}P_{r}$, $P_{n,r}$ and P (n,r).

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

Let's see how many permutations we can arrange by selecting 2 out of 3 geometrical shapes

Permutations

Below is the mathematical calculation.

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

$${}^{3}P_{2} = \frac{3!}{(1)!}$$

$${}^{3}P_{2} = 3*2*1 = 6$$

Example 1.5: Permutation

I) There are 12 books. A student was asked to select three of them randomly. How many different ways the student can select the books?

$${}^{12}P_{3} = \frac{12!}{9!}$$

$${}^{3}P_{2} = \frac{12*11*10*9!}{9!}$$

$${}^{3}P_{2} = 12*11*10$$

$${}^{3}P_{2} = 12*11*10$$

$${}^{3}P_{2} = 1320$$

II) 20 competitors participate in a race. How many ways are there that the competitors can win gold, silver and bronze medals?
Gold medal can be won in 20 ways

Silver medal can be won in 19 ways

Bronze medal can be won in 18 ways

Therefore at one race the ways of models can be won is 20*19*18 =6840

III) There are 4 doors to a cinema. In how many ways can a customer enter the room through a door and leave the room by a different door?
There are 4 doors and if a customer can enter the supermarket in 4 ways. Now one door is used. Now he has to use other 3 doors to leave the super market without using the first door. Therefore the calculation is 4*3= 12

There are twelve ways that a customer can enter and leave the supermarket using different doors.

IV) How many passwords can be formed by re-arranging the letters of the word STATISTICS . Meaning of the word is not relevant.
There are 10 letters in the word. There are 3 letters S and T. There are 2 “I”s. There is only one “A”and C

$\frac{10!}{3!\,3!2!1!1!}$ When this number is solved answer is 50400.

This type of problem solving is called MISSISSIPI rule in statistics.

V) How many passwords can be formed by re-arranging letters of the word SMILE such that S and E occupy the first and last positions respectively? Meaning of the word is not relevant.
Here S and E are in fixed places. Only permutation of M,I,L should be calculated.

$1^* 3^*2^*1^*1 = 6$

S is fixed so; it is used in only one way. Second letter can be selected out of 3 letters. Third letter can be selected out of 2 letters. For fourth letter only one choice is left. Fifth letter E can also be used in only one way.

VI) How many passwords can be formed by re-arranging letters of the word SMILE such that S and E occupy the first and last position? Meaning of the word is not relevant.

In this arrangement it is told S and E can be used as first or last letters. So there are two choices for the first letter. But for the last letter there is only one letter is left out of S and E.

$2^* 3^*2^*1^*1 = 12$

VII) 6 teachers, the principle and the vice principle are to be seated around a circular table. If the principle should sit between a teacher and the vice principle, in how many ways can they be seated?

There is only one principle and a vice principle, principle should be in between a teacher and a vice principle.

Take them as one set. . there are 6 sets now (five of them consists of single teachers) . A teacher and Vice principle can be seated in two ways ; teacher and vice principle or the other way round (2 ways) . Teachers can be seated in 6! Ways. Therefore, $6! *2 = 720*2 = 1440$

VIII) Find the number of permutations of the letters of the word 'PROBLEM' such that the vowels always occur in odd places.

There are 7 different letters in the word PROBLEM

There are 6 constants and 2 vowels.

There are four odd places and three even places.

Number of ways 2 vowels can appear in 4 different places = ${}^4P_2 =$ 12.
After 2 vowels take 2 places, 5 more spaces are left. They can be arranged in 5! =120 ways.
Therefore, total number of permutations possible is 120*12 = 1440

1.6 Combination

Any subsets that can be created by selecting "r" number of different items out of total of "n" numbers of items are called combination. Order is not important in permutation. "r" should be less than or equal to "n".
It is represented by ${}^nC_r$, $C_{n,r}$ $\binom{n}{r}$and C (n,r)

$${}^nC_r = \frac{n!}{r!\,(n-r)!}$$

Let's see how many combinations we can create by selecting 2 out of 3 geometrical shapes

Combinations

Below is the mathematical calculation.

$${}^nC_r = \frac{n!}{r!\,(n-r)!}$$

$${}^3C_2 = \frac{3!}{2!\,(1)!}$$

$${}^3C_2 = 3$$

Since the order does not matter in combination, only 3 combinations can be selected.

Example 1.6: Combinations

I) There are 3 books and 4 magazines. A student is told to choose 1 book and 2 magazines. How many ways can he select 1 book and 2 magazines?

$^{n}C_r * {}^{n}C_r = \frac{3!}{1!\,(2)!} * \frac{4!}{2!\,(2)!}$

$^{3}C_1 * {}^{4}C_2 = 3 * 6$

$^{3}C_1 * {}^{4}C_2 = 18$ (There are total of 18 ways)

II) There is a set of 6 blue balls and 3 red balls, a child is told to select 5 such that at least 3 of them are blue balls. How many selections can he make?

There are 9 balls, 5 balls are to be selected, but there should be at least 3 blue balls. So these 5 balls can consists of 3 blue balls and 2 red balls ($^{6}C_3 * {}^{3}C_2$), 4 blue balls ($^{6}C_4 * {}^{3}C_1$) and 1 red balls and only 5 blue balls ($^{6}C_5 * {}^{3}C_0$).

$=(^{6}C_3 * {}^{3}C_2) + (^{6}C_4 * {}^{3}C_1) + (^{6}C_5 * {}^{3}C_0).$

$=(20 * 3) + (15 * 3) + (6 * 1).$

$= 60+45+6$

$=111$ (There are total of 111 ways)

III) Three directors and one CEO should be selected out of 7 director candidates and 3 CEO candidates of a company. How many ways can 3 directors and 1 CEO be selected?

$^{7}C_3 * {}^{3}C_1$

$= 35 * 3$

$=105$ (There are 105 ways)

IV) Passwords can be created using 3 numbers and 7 letters of English language. How many combinations of passwords can be created?

$^{10}C_3 * {}^{26}C_7$

$=120*657800$

$=78936000$

Example 1.7: Permutation & Combination

"PCIKTIVVATIKG" is a word in a language. In how many rearrangements of the letters can be done with no two 'I's appear together?

There are 13 letters including one P ,C ,A and G, two K, T and V s and three I s.

First remove the three I s. There will be 10 letters. Calculate the permutation of them it will be $10!/(2!*2!*2!) = 453{,}600$

There will be 11 spaces where three of the I s can be added between other 10 letters.

1 P 2 C 3 I 4 K 5 T 6 I 7 V 8 V 9 A 10 T 11 I 12 K 13 G

Find out how can be three "I"s be placed in 11 space using combination ${}^{11}C_3 = 165$

So the total will be $165*453{,}600 = 74{,}844{,}000$.

Example 1.8: Permutation & Combination

Lottery has a 3 digit vowels and one number.

I) A person can win if those vowels and the number appear in any order. How many ways can it happen?
   ${}^{6}C_3 * {}^{10}C_1$
   $=20*10$
   $=200$
II) If the vowels should appear in first three places and the number should appear the last, how many ways can it happen?
   ${}^{6}P_3 * {}^{10}P_1$
   $=6*5*4*10$
   $=1200$

BOOK : Introduction to Regression Analysis

CHAPTER ONE: CORRELATION

1.1 Introduction

Correlation is used to measure the mutual relationship between two or more variables. Correlation coefficient is the numerical measure of the relationship between two variables. Correlation shows the presence of a relationship between variables, strength of the relationship and its direction. It can be illustrated in graphs which simplify the interpretation. Correlation demonstrates the relationship between the measurable variables, but it doesn't identify the cause of the relationship. Therefore there can be underlying variables which affects the relationship.

Scatter diagrams, Karl Pearson's correlation coefficient and Spearsman's Rank correlation are few methods which used measure the correlation. Correlation coefficient of sample is represented by "r", Correlation coefficient of population is represented by "ρ". Correlation Coefficient ranges from +1 to -1. If the correlation coefficient is +1, there is a perfect positive correlation between two variables. When correlation coefficient is -1 there is a perfect negative correlation between two variables. When there is no correlation at all the correlation coefficient is 0. When it is below $\pm$ 0.5 relationship considered to be not strong. On the other hand when it is above $\pm$ 0.75 relationship considered to be strong. Between $\pm$ 0.5 to $\pm$ 0.75 it is considered to have a moderate correlation between variables.

1.1 Scatter Diagram

Scatter diagram is the method of drawing a graph with (x,y) coordinates. According to the spread of the coordinates, one can easily guess the direction and the strength of the relationship. X is used for independent variables and Y is used for dependent variables. Dependent variable is the variable affected by one or more independent variables. Independent variables are the cause of the dependent variable. When there are multiple independent variables they are represented by $x_1, x_2, x_3, \ldots, x_n$.

Followings are few examples for scatter plot diagrams.

Figure 1.1 : Perfect Positive Correlation

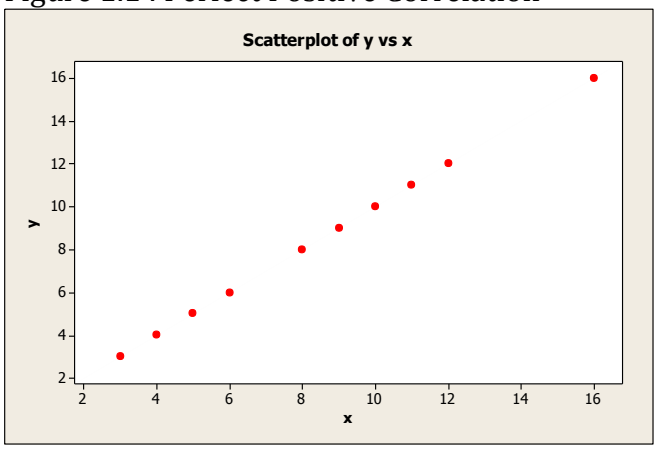


Figure 1.2: Perfect Negative Correlation

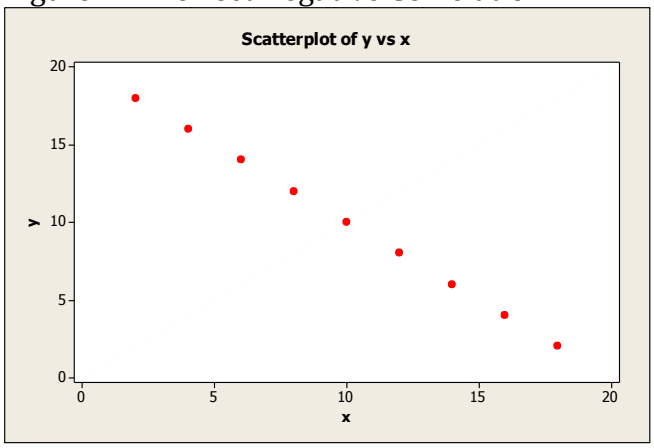


Figure 1.3 : Strong Positive Correlation

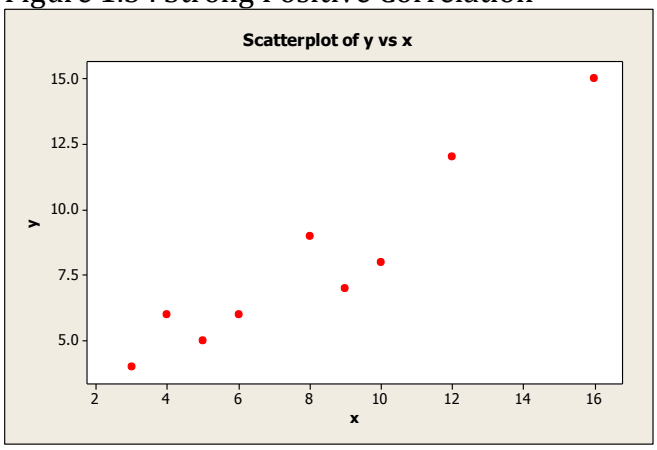


Figure 1.4: Strong Negative Correlation

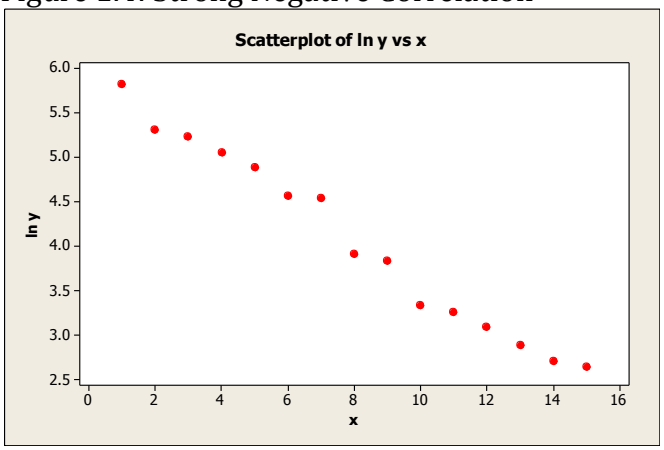


Figure 1.5: Moderate Positive Correlation

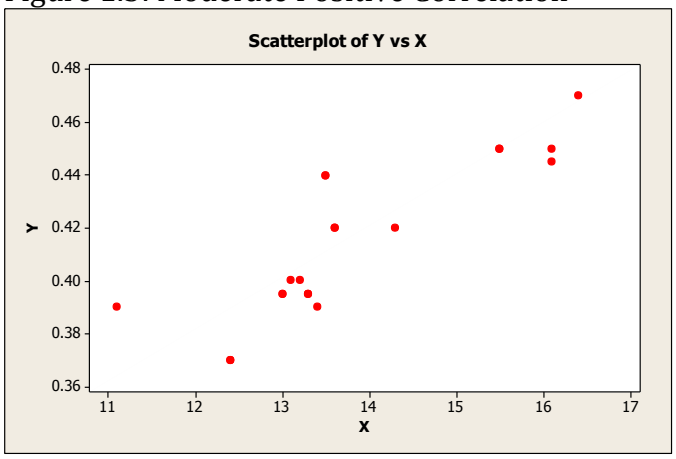


Figure 1.6: Moderate Negative Correlation

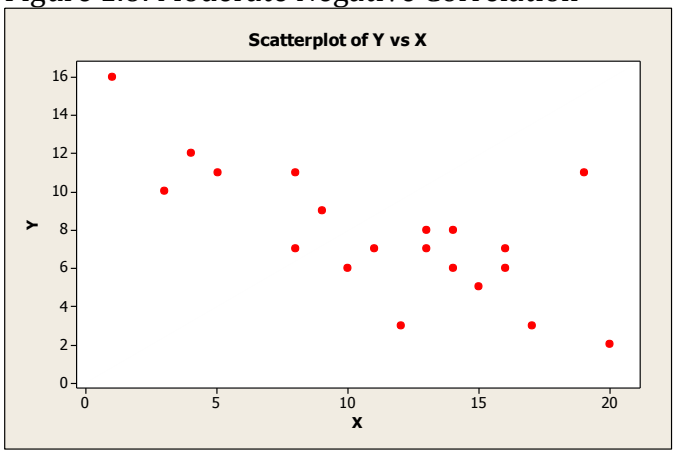


Figure 1.7: Non Linear Correlation

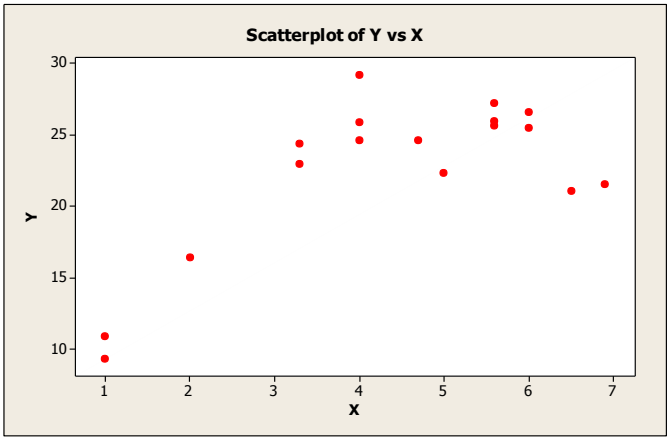
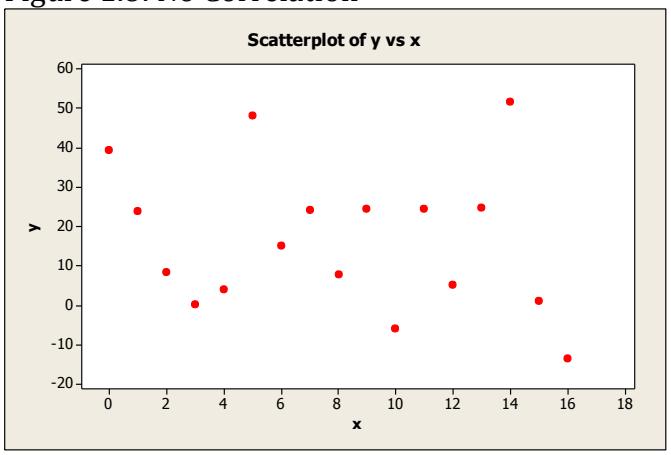


Figure 1.8: No Correlation



When there is more than one independent variable used in a regression, relationship among independent variables; relationship between independent variables and dependent variable should be analyzed using a correlation matrix.

1.3 Karl Pearson's Correlation Coefficient

Karl Pearson's correlation coefficient measures both strength and direction of the relationship. The coefficient gives a measurable value to the strength of relationship. Karl Pearson's coefficient is calculated assuming there are no other factors influencing the dependent variable other than one dependent variable.

Covariance of (x,y) or Cov(x,y) is equal to $E[(x-\bar{x})-(y-\bar{y})]$, on assumption of variances are positive, correlation of $(x,y) = \frac{cov(x,y)}{sd(x)sd(y)}$, where sd is standard deviation

Below equations which can be used to calculate the correlation are based on the covariance of relationship.

$$r = \frac{\sum xy - n * \bar{x} * \bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}}$$

$$r = \frac{\sum[(x - \bar{x})(y - \bar{y})]}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

Example 1.1:
Given below is the data set of marks scored for the year end mathematics examination and hours spent for mathematic homework per week. Find out if there is a relationship between these variables and describe the nature of the relationship.

x	1	2	3	4	5
y	35	60	75	85	95

x- hours spent for mathematic homework per day

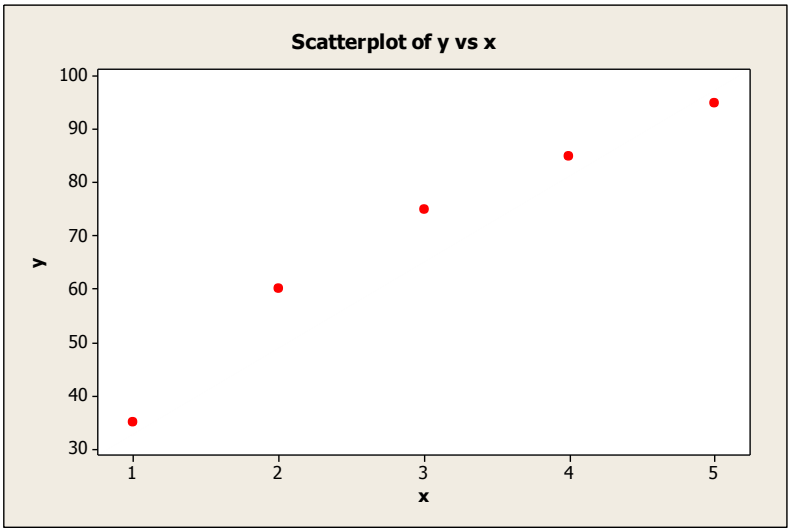
y- marks scored for the year end mathematics examination

Answer
Calculate $\sum x$, $\sum y$, $\sum xy$, $\sum x^2$, $\sum y^2$ as explained in table 1.1.

Table 1.1: Descriptive Statistics

	x	**y**	**xy**	**x^2**	**y^2**
	1	35	35	1	1225
	2	60	120	4	3600
	3	75	225	9	5625
	4	85	340	16	7225
	5	95	475	25	9025
Total	**15**	**350**	**1195**	**55**	**26700**

Figure 1.9: Scatter plot



$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

$$r = \frac{5 * 1195 - 15 * 350}{\sqrt{(5 * 55) - (15)^2}\sqrt{(5 * 26700) - (350)^2}}$$

$$r = 0.978$$

In above example correlation coefficient is $+0.978$ which is close to $+1$. It can be concluded that there is a strong positive relationship between hours a student spend for home work per week and year end mathematics exam results.

It can be calculated using below equation as well.

$$\bar{x} = \frac{15}{5} = 3 \quad \bar{y} = \frac{350}{5} = 70$$

$$r = \frac{\sum xy - n * \bar{x} * \bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}}$$

$$r = \frac{1195 - (5 * 3 * 70)}{\sqrt{(55 - 5 * 9)}(26700 - 5 * 4900)}$$

$$r = \frac{145}{148.3239}$$

$$r = 0.9775$$

Example 1.2:
Calculate the correlation coefficient between marks for English and French of 20 students in a class room.

X	5	15	20	1	4	11	17	9	14	8	3	16	10	13	19	16	12
Y	7	11	15	3	6	11	18	13	10	10	2	14	7	10	16	12	16
X	14	8	13														
y	10	11	8														

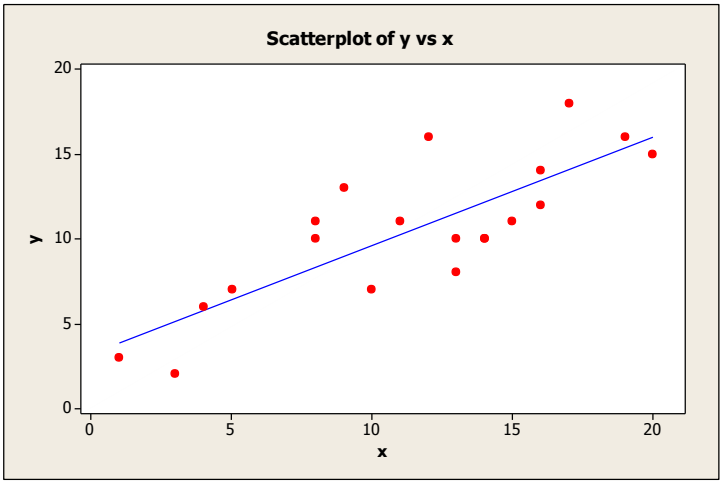
x- marks for English , y – marks for French

Answer

First find out values for below notations.

$\sum x^2 = 3142 \quad \sum y^2 = 2544 \quad \sum xy = 2711 \quad \bar{x} = 11.40 \quad \bar{y} = 10.5$

Figure 1.10: Scatter plot



$$r = \frac{\sum xy - n * \bar{x} * \bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - \bar{y}^2)}}$$

$$r = \frac{2711 - 20 * 11.4 * 10.5}{\sqrt{(3142 - (20 * 11.40^2))\,(2544 - (20 * 10.5^2)}}$$

$r = 0.809$

Correlation coefficient of two variables of marks for French and marks for English is equal to 0.809. It has a strong positive correlation, which indicates that there is a strong possibility of increasing English marks of a student when his/her French marks increase. Figure 1.10 also illustrates a strong positive correlation between English and French mark of students.

Example 1.3:

Given below is the price of 1 kg of 14 types of rice and the number of sales per day in a super market in Colombo. Store owner wants to identify if there is a relationship between price of the rice and amount of sales.

x- Price of 1 kg of rice in Sri Lankan rupees

y- Number of sales happened per day

x	71	71	72	76	77	75	74	73	74	71	72	70	69	68
y	10	9	8	7	4	4	5	4	6	6	5	9	12	14

In this scenario correlation can be used to identify the relationship. Equation of $r = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$ is used to calculate the correlation.

Table 1.2: Descriptive Statistics

	x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	71.00	10	-1.36	2.64	-3.58	1.84	6.97
2	71.00	9	-1.36	1.64	-2.23	2.69	2.69
3	72.00	8	-0.36	0.64	-0.23	0.41	0.41
4	76.00	7	3.64	-0.36	-1.31	0.13	0.13
5	77.00	4	4.64	-3.36	-15.59	11.29	11.29
6	75.00	4	2.64	-3.36	-8.87	11.29	11.29
7	74.00	5	1.64	-2.36	-3.87	5.57	5.57
8	73.00	4	0.64	-3.36	-2.15	11.29	11.29

9	74.00	6	1.64	-1.36	-2.23	1.85	1.85
10	71.00	6	-1.36	-1.36	1.85	1.85	1.85
11	72.00	5	-0.36	-2.36	0.85	5.57	5.57
12	70.00	9	-2.36	1.64	-3.87	2.69	2.69
13	69.00	12	-3.36	4.64	-15.59	21.53	21.53
14	68.00	14	-4.36	6.64	-28.95	44.09	44.09
Total	**1013.00**	**103**	**-0.03**	**-0.04**	**-85.77**	**122.09**	**127.21**

Calculate the averages of x and y.

$\bar{x}= 72.36 \quad \bar{y}=7.36$

$$r = \frac{\sum[(x - \bar{x})(y - \bar{y})]}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

$$r = \frac{-85.77}{\sqrt{122.09 * 127.21}}$$

$$r = -0.6862$$

There is a moderate negative correlation between price of rice and the number of sales. It can be concluded that when price of rice goes high sales start dropping moderately.

1.4 Spearman's Rank Correlation

Spearman's Rank Correlation Coefficient is used to measure relationship between two qualitative variables. In order to transfer qualitative variables in to measurable scale of data, method of "ranking the data according to a given phenomenon" can be used.

$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$ is an equation used to calculate Spearman's Rank Correlation.

Example 1.4
Below are the preferences on 10 restaurants of two random expats who moved to Colombo city. Restaurant names are A,B,C,D,E,F,G,H,I,J. 1 is the most favorable restaurant while 10 is the least favorable restaurant. Is there any relationship between the preferences of two expats?

Restaurant	A	B	C	D	E	F	G	H	I	J
1st Expat	2	3	5	6	9	10	1	7	4	8
2nd Expat	4	5	3	7	10	9	2	8	1	6

Answer
Table 1.3 : Ranking

Restaurant	1st Expat	2nd Expat	d	d^2
A	2	4	-2	4
B	3	5	-2	4
C	5	3	2	4
D	6	7	-1	1
E	9	10	-1	1
F	10	9	1	1
G	1	2	-1	1
H	7	8	-1	1
I	4	1	3	9
J	8	6	2	4
			Total	30

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6*30}{10*99}$$

$$\rho = 0.8182$$

Correlation coefficient is positive and strong. It can be concluded that there is a strong positive correlation between the preferences of the expats.

1.5 Significance of Correlation Coefficient

If the value of correlation coefficient is zero there is no relationship between variables. Hypothesis testing for population correlation coefficient can be done to determine the significance of the correlation coefficient or significance of the relationship between variables.

Given below are the hypothesis and equation for test statistics used to test significance of correlation coefficient.

Hypothesis

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Test statistic

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} *$$

Example 1.5

If correlation Coefficient of sample is 0.928 and sample size is 24, test the significance of the correlation coefficient of population with 95% of confidence

Answer

Hypothesis

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Significance level $= 0.05$

Test statistic

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} *$$

$$T = \frac{0.928\sqrt{22}}{\sqrt{1-0.928^2}} *$$

$T = 11.637$

Critical Value for t can be derived from t table (Refer to a T-table with $\alpha/2 = 0.05/2 = 0.025$)

Critical Value $t_{\alpha/2,n-2} = \pm 2.074$

Test statistic is in the rejection region. Therefore H_0 is rejected. It can be concluded with 95% of confidence that the correlation coefficient is

significantly different from zero. There is a significant relationship between given independent and dependent variables.

1.6 Correlation Matrix

Correlation Matrix is a table of correlation coefficients of dependent variable and two or more different independent variables. Table 1.4 shows the correlation matrix between dependent variable y and three other independent variables of x_1, x_2, x_3. (MINITAB COMMAND: Stat -> Basic Statistics -> Correlation)

Table 1.4: Correlation matrix

	y	x1	x2	x3
x1	-0.892			
	0.000			
x2	0.912	-0.816		
	0.000	0.002		
x3	-0.907	0.879	-0.866	
	0.000	0.000	0.001	
x4	-0.420	0.492	-0.156	0.425
	0.198	0.124	0.646	0.192

Cell Contents: Pearson correlation
P-Value

According to the table 1.4, Pearson correlation coefficient between y and x_1 is -0.892 which means there is a strong negative correlation between y and x_1. Below the coefficient value, the p-value or the probability value is mentioned. Significance level used for above correlation matrix is 5%. When the significance level of the test is 5%, p-value below 0.05 means the null hypothesis is rejected, if p value is greater than 0.05 then the null hypothesis is not rejected. Null Hypothesis used to test significance of correlation coefficient is "$H_0 : \rho = 0$" . According to table 1.4 the correlation coefficient of y and x_1 (-0.892) is significant with p-value of 0.000 which is less than 0.05. If a statistic is statistically significant then it means that the result is unlikely due to chance.

Correlation between y and x_4 is not significant as p-value (0.198) is greater than 0.05. Therefore it cannot be considered that there is an effective relationship between y and x_4. y has a significant positive correlation with x_2 (correlation = 0.912, p-value = 0.000) and significant negative relationship with x_3 (correlation = -0.907, p-value = 0.000). There are significant correlation between independent variables such as strong positive correlation between x_1 and x_3 (correlation = 0.807, p-value = 0.000); x_1 and x_2 (correlation = -0.816, p-value = 0.002) and x_2 and x_3 (correlation = -0.866, p-value = 0.001). Correlation among independent variables is called multicolinearity, which will be explained in details at chapter 5.

Book: Introduction to Hypothesis Testing

CHAPTER ONE: HYPOTHESIS TESTING

1.1 Introduction

Hypothesis is an assumption based on evidence. Hypothesis testing is the procedure of testing the hypothesis using a statistical experiment or a statistical method. Hypothesis are established according to scientific theories, observations from past or present studies, the experience of competitors, a relevant generally identified pattern and similarity between events.

Hypothesis is a statement on a population parameter, which is tested using the observations of a sample. The sample should be drawn out from the same population. Hypothesis also is a declaration made on an unknown parameter. In hypothesis testing, a hypothesis is tested using random samples and tries to generalize the hypothesis into its population.

A hypothesis should be concise, measurable, clear, reliable and easy to understand. These are the characteristics of a hypothesis. Hypothesis can be tested only if it is measurable. It should be specific and has a wide scope for further testing.

Hypothesis is tested using a confidence interval or the level of significance of the hypothesis. If a hypothesis is tested using 5% of significance level, it means that this hypothesis can be claimed with 95% of confidence level.

Null and alternative hypothesis are the two hypothesis established in hypothesis testing.

Null Hypothesis

First estimated hypothesis is called null hypothesis which is denoted by H_0. This null hypothesis normally states opposite of the assumption on the test. It is a statement that is formed to reject when the test is significant.

Example : If the assumption is to check if there is a difference between the statistic and the parameter when creating null hypothesis we uses negativity. Therefore null hypothesis will be "there is no difference between the statistic and the parameter"

Alternative Hypothesis

Alternative hypothesis is the hypothesis established against the null hypothesis. It is denoted by H_1.

In other words, the hypothesis which is true when null hypothesis is rejected is alternative hypothesis. Conclusion of the test is normally included in alternative hypothesis.

Example:

It is assumed that the average IQ level of a student in a school is more than 82. In order to testify this assumption sample of 100 students are taken and found that this group has an average IQ level of 79 with standard deviation of 3. Check if the hypothesis is acceptable at 95% of confidence level.

In above question it is assumed that IQ level of a student in a school is more than 82. Therefore the null hypothesis here is mean of the population is not greater than 82 ($\mu \leq 82$). Alternative hypothesis is $\mu > 82$.

Then in order to test these sample 100 students were selected. Therefore the sample size is n and n is 100. In this sample average is 79 ($\bar{x} = 79$) and the standard deviation is 3. Using these data , the null hypothesis can be tested using 5% of significance level.

This hypothesis testing procedure will be discussed in coming chapters.

Steps of Hypothesis testing

1. Establishing hypothesis (Null and Alternative)
2. Choose the alpha value (significance level 5% and 1% are common)
3. Calculate the test statistics using a suitable distribution
4. Find the decision rule (critical value helps to take a decision on the hypothesis)
5. Make the decision
6. Conclusion and recommendation

In below picture it demonstrates how a hypothesis can be tested using a normal distribution.

Significance level - α

Level of significance is the probability level that is used to reject or accept the hypothesis. α is the significance level and it is also the type one error.

Test of significance allows us to take decisions based on sample results.

Critical Value

Critical value is the point of value, which the test statistic is compared with. Critical value defines the rejection region. Rejection region is the area where the null hypothesis is rejected if the test statistic lies in that region.

Test statistic is a value calculated using the statistics drawn from the sample. Direction of the hypothesis is important in deciding the rejection region. (Rejection region will be discussed later)

Decision rule

Decision rule defines when to reject the null hypothesis. Critical value helps in establishing the decision rule. Direction of the test is important in decision rule.

1.2 Directional and Non-directional Hypothesis

Non-directional Hypothesis

Figure 1.1 : Two tailed test

Non-directional hypothesis doesn't predict a direction for the behavior of the variable and it is widely used when an empirical theory is not involved. There are two directions of interests in non directional hypothesis; therefore the critical area is located at two ends of the distribution as the above distribution illustrates. When critical area is spread at two ends of the distributions, it is also called two tailed tests.

Below non directional null hypothesis indicates that there is no significance difference between θ_1 and θ_2. $\theta_1 \neq \theta_2$ means θ_1 can be lesser than θ_2 or θ_1 can be greater than θ_2. Therefore there are two directions in below hypothesis. This hypothesis is illustrated in figure 1.1: two tailed test. Theta (θ) means any parameter.

In two tailed test or non directional hypothesis testing alpha (α) value is divided into 2. If 5% significance level (95% confidence) is used for two tail test one rejection region gets only 2.5% of significance level (0.05/2 = 0.025)

Non directional hypothesis

H_0: $\theta_1 = \theta_2$ equal

H_1: $\theta_1 \neq \theta_2$ not equal

If the test statistics falls at either side of rejection region then the null hypothesis will be rejected. If it falls on acceptance region, it can be concluded that there are no significance evidence to reject the null hypothesis.

Directional Hypothesis

When a direction of the assumption or hypothesis is specified clearly, it is called a directional hypothesis. Direction can be specified using statements "greater than" or "less than". Directional hypothesis are more powerful and accurate than non directional hypothesis as it is more specific in its direction.

Critical value indicated by the directional hypothesis is concentrated at only one side of the distribution (negative or positive). As the critical value concentrates at only one side , it is also called one tail test.

Directional Hypothesis -Type 1

H_0: $\theta_1 \geq \theta_2$ Greater than or equal

H_1: $\theta_1 < \theta_2$ Less than

Above directional null hypothesis indicates that θ_1 is greater than or equal to θ_2. Alternative hypothesis means θ_1 is less than θ_2. There is only one direction for this hypothesis and the critical value will be negative side in normal and t distribution. α value is not divided into to 2.

Above directional hypothesis –type 1 is illustrated using a normal distribution assuming θ_1 and θ_2 are normally distributed.
If the test statistics falls at rejection region then the null hypothesis will be rejected. If it falls on acceptance region, it is concluded that there are no significance evidence to reject the null hypothesis.

Directional Hypothesis –Type 2
H_0: $\theta_1 \leq \theta_2$ Less than or equal
H_1: $\theta_1 > \theta_2$ Greater than

Above directional null hypothesis indicates that θ_1 is less than or equal to θ_2. Alternative hypothesis means θ_1 is greater than θ_2. There is only one direction for this hypothesis and the critical value will be positive in normal and t distribution. α value is not divided into to 2.

Above directional hypothesis is illustrated using a normal distribution assuming θ_1 and θ_2 are normally distributed.

If the test statistics falls at rejection region then the null hypothesis will be rejected. If it falls on acceptance region, it can be concluded that there are no significance evidence to reject the null hypothesis.

One Tail test

Directional hypothesis undergoes one tail tests. In one tail test, there is only one direction of interest $(\theta_1 < \theta_2)$ or $(\theta_1 > \theta_2)$, therefore critical region falls only at one end of the distribution. Therefore α value is not divided

Two Tail Test

In two tail test critical area of a distribution is located at two side of the distribution. Therefore α value is divided into two sides. Hypothesis shows two directions of interests. $(\theta_1 = \theta_2)$ when a hypothesis indicates equality, then it can be either greater than or less than.

Whether the statistical test is one tailed or two tailed will influence immensely to the interpretation of the outcome of the test.

1.3 Type I and Type II Error

Two types of errors are occurred, if hypothesis testing is not done correctly. These errors are called type I and type II error. Below table 1.1 explains how these errors occurs.

Table 1.1: Errors in hypothesis testing

	H_0 true	H_1 false
H_0 Reject	Type I error α	No error (1-β)
H_1 not reject	No error (1-α)	Type II error β

P(rejecting H_0 when H_0 is true) = α
Probability of rejecting null hypothesis when it is true is called type I error and it is similar to alpha value.

P(not rejecting H_1 when H_1 is not true) = β
Probability of not rejecting alternative hypothesis when it is not true is called type II error and it is similar to beta value.

These errors should be reduced for a competent hypothesis testing. "Decreasing the significance level of the test" (α) helps reducing type I error in a test. Increasing the size of sample (selecting a large sample) helps decreasing type II error.

Power of the test indicates how good the test performs. Power of the test is calculated reducing the probability of type II error by 1 (1-β).

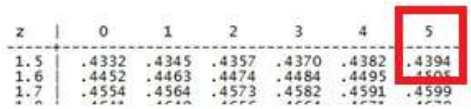
Z- test, t test , chi test and F test are four main tests which are used in hypothesis testing. In coming chapters application of these testing methods will be explicitly described.

1.4 How to calculate p-value

P value is a measure of probability in the critical region. When the p value is close to zero (smaller p value) then the significance of the statistical test is greater. When there is a smaller p value then it indicates that there is effective evidence which supports alternative hypothesis.

Let's calculate the p-value for z=1.55 , when the test is two tail.

Table 1.2: z- table for half of the normal distribution



z	0	1	2	3	4	5
1.5	.4332	.4345	.4357	.4370	.4382	.4394
1.6	.4452	.4463	.4474	.4484	.4495	.4505
1.7	.4554	.4564	.4573	.4582	.4591	.4599

When z is equal to 1.55, probability value of the table (for one side) is 0.4394 (Look at above 1.2 z-table). Then it should be subtracted by 0.5 (0.5-0.4394 = 0.0606). This is the p-value for one side of the two tailed normal distribution.
In order to calculate the p-value for whole distribution as it is two tailed 0.0606 should be multiplied by 2. Therefore the p value is 0.0606*2 = 0.1212.

When the significance level (α) of the test is 5% (confidence level is 95%) , p-value is greater than 0.05 (0.1212 > 0.05). Therefore null hypothesis (H_0) cannot be rejected.

Visit http://www.anushabooks.com/ for more details

Author's other books:

1) Probability Questions and Answers
2) Probability and Distributions
3) Introduction to Hypothesis Testing
4) Introduction to Regression Analysis
5) Introduction to Time Series Analysis
6) Introduction to Categorical Data Analysis
7) Introduction to Multivariate Data Analysis

Visit below link for Author's other book purchases https://www.amazon.com/dp/B08FJ4JHC6?binding=kindle_edition&ref=dbs_dp_rwt_sb_pc_tukn

"If you wish more statistics books for affordable price, please leave a review on Amazon stating the feedback and also the topics you wish to have books on."