Lectures of Stat -106 (Biostatistics)

Text book

Biostatistics

Basic Concepts and Methodology for the Health Sciences

By

Wayne W. Daniel

Chapter 1

Introduction To Biostatistics

- ***Key words :***

  - *Statistics , data , Biostatistics,*
  - *Variable ,Population ,Sample*

Introduction
Some Basic concepts

Statistics is a field of study concerned with

1- collection, organization, summarization and analysis of data.

2- drawing of inferences about a body of data when only a part of the data is observed.

Statisticians try to interpret and

communicate the results to others.

Data:

- **The raw material of Statistics is data.**
- **We may define data as figures. Figures result from the process of counting or from taking a measurement.**

For example:

- **- When a hospital administrator counts the number of patients (counting).**
- **- When a nurse weighs a patient (measurement)**

* A variable:

It is a characteristic that takes on different values in different persons, places, or things.

For example:

- heart rate,

- the heights of adult males,

- the weights of preschool children,

- the ages of patients seen in a dental clinic.

* Biostatistics:

The tools of statistics are employed in many fields:

business, education, psychology, agriculture, economics, … etc.

When the data analyzed are derived from the biological science and medicine,

we use the term biostatistics to distinguish this particular application of statistical tools and concepts.

Types of variables

Quantitative

Qualitative

Quantitative Variables

It can be measured in the usual sense.

For example:

- **the heights of adult males,**
- **the weights of preschool children,**
- **the ages of patients seen in a**
- **dental clinic.**

Qualitative Variables

Many characteristics are not capable of being measured. Some of them can be ordered (called ordinal) and Some of them can't be ordered (called nominal).

For example:

- **classification of people into socio-economic groups**

-.**hair color**

Types of quantitative variables

Discrete

A discrete variable

is characterized by gaps or interruptions in the values that it can assume.

For example:

- **The number of daily admissions to a general hospital,**
- **The number of decayed, missing or filled teeth per child**
- **in an**
- **elementary**
- **school.**

Continuous

A continuous variable

can assume any value within a specified relevant interval of values assumed by the variable.

For example:

- **Height,**
- **weight,**
- **skull circumference.**

No matter how close together the observed heights of two people, we can find another person whose height falls somewhere in between.

Types of qualitative variables

Nominal

As the name implies it consist of "naming" or classifies into various mutually exclusive categories

For example:

- Male - female
- Sick - well
- Married – single - divorced

Ordinal

.Whenever qualitative observation

Can be ranked or ordered according to some criterion.

For example:

- Blood pressure (high-good-low)
- Grades (Excellent – V.good –good –fail)

* A population:

It is the largest collection of values of a random variable for which we have an interest at a particular time.

For example:

The weights of all the children enrolled in a certain elementary school.

Populations may be finite or infinite.

* A sample:

It is a part of a population.

For example:

The weights of only a fraction of these children.

Exercises

- Question (6) – Page 17
- Question (7) – Page 17

  “ Situation A , Situation B “

Exercises:

Q6: For each of the following variables indicate whether it is quantitative or qualitative variable:

(a)The blood type of some patient in the hospital. Qualitative Nominal

(b) Blood pressure level of a patient.

(Qualitative ordinal)

(c) Weights of babies born in a hospital

during a year. Quantitative continues

(d) Gender of babies born in a hospital during a year. Qualitative nominal

(e) The distance between the hospital to the house Quantitative continues

(f) Under-arm temperature of day-old infants born in a hospital. Quantitative continues

Q7: For each of the following situations,

answer questions a through d:

(a) What is the population?

(b) What is the sample in the study?

(c) What is the variable of interest?

(d) What is the type of the variable?

Situation A: A study of 300 households in a small southern town revealed that if she has school-age child present.

All households in a small southern town. (a) Population:

300 households in a small southern town. (b) Sample:

(c) Variable: Does households had school age child present.

(d) Variable is qualitative nominal.

- **Situation B: A study of 250 patients admitted to a hospital during the past year revealed that, on the average, the patients lived 15 miles from the**

hospital.

(a) Population: All patients admitted to a hospital during the past year.

(b) Sample: 250 patients admitted to a hospital during the past year.

(c) Variable: Distance the hospital live away from the hospital

Variable is Quantitative continuous. (d)

Choose the right answer: (For Students)

1-The variable is a

a. subset of the population.

b. parameter of the population.

c. relative frequency.

d. characteristic of the population to be measured.

e. class interval.

2-Which of the following is an example of discrete variable

a. the number of students taking statistics in this term at ksu.

b. the time to exercise daily.

c. whether or not someone has a disease

d. height of certain buildings

e. Level of education

3.Which of the following is not an example of discrete variable

a. the number of students at the class of statistics.

b. the number of times a child cry in a certain street.

c. the time to run a certain distance.

d. the number of buildings in a certain street.

e. number of educated persons in a family.

4.Which of the following is an example of qualitative variable

a. the blood pressure level.

b. the number of times a child brush his/her teeth.

c. whether or not someone fail in an exam.

d. Weight of babies at birth.

e. the time to run a certain distance.

5.The continuous variable is a

a. variable with a specific number of values.

b. variable which cant be measured.

c. variable takes on values within intervals.

d. variable with no mode.

e. qualitative variable.

6. which of the following is an example of continuous variable

a. The number of visitors of the clinic yesterday.

b. The time to finish the exam.

c. The number of patients suffering from certain disease.

d. Whether or not the answer is true.

7. The discrete variable is

a-qualitative variable.

b-variable takes on values within interval.

C-variable with a specific number of values.

d-variable with no mode.

8-Which of the following is an example of nominal variable :

a-age of visitors of a clinic.

b-The time to finish the exam.

c-Whether or not a person is infected by influenza.

d-Weight for a sample of girls .

9-The nominal variable is a

a-A variable with a specific number of values

b-Qualitative variable that can't be ordered.

c-variable takes on values within interval.

d-Quantitative variable .

10**-Which of the following is an example of nominal variable :**

a-The number of persons who are injured in accident.

b-The time to finish the exam.

c-Whether or not the medicine is effective.

d-Socio-economic level.

11-The ordinal variable is :

a-variable with a specific number of values.

b-variable takes on values within interval.

c-Qualitative variable that can be ordered.

d-Variable that has more than mode.

Chapter (2)

Strategies for understanding the meanings of Data

Pages(19 – 27)

- Key words

frequency table, bar chart ,range

width of interval

Histogram , Polygon

Descriptive Statistics

Frequency Distribution for Discrete Random Variables

Example:

Suppose that we take sample of size 16 from children in a primary school and get the following data about the number of their decayed teeth,

3,5,2,4,0,1,3,5,2,3,2,3,3,2,4,1

To construct a frequency table

We need three columns:

1.Variable name

2.Frequency (f):how manmber

3.Relative frequency(R.f)=

Frequency / n

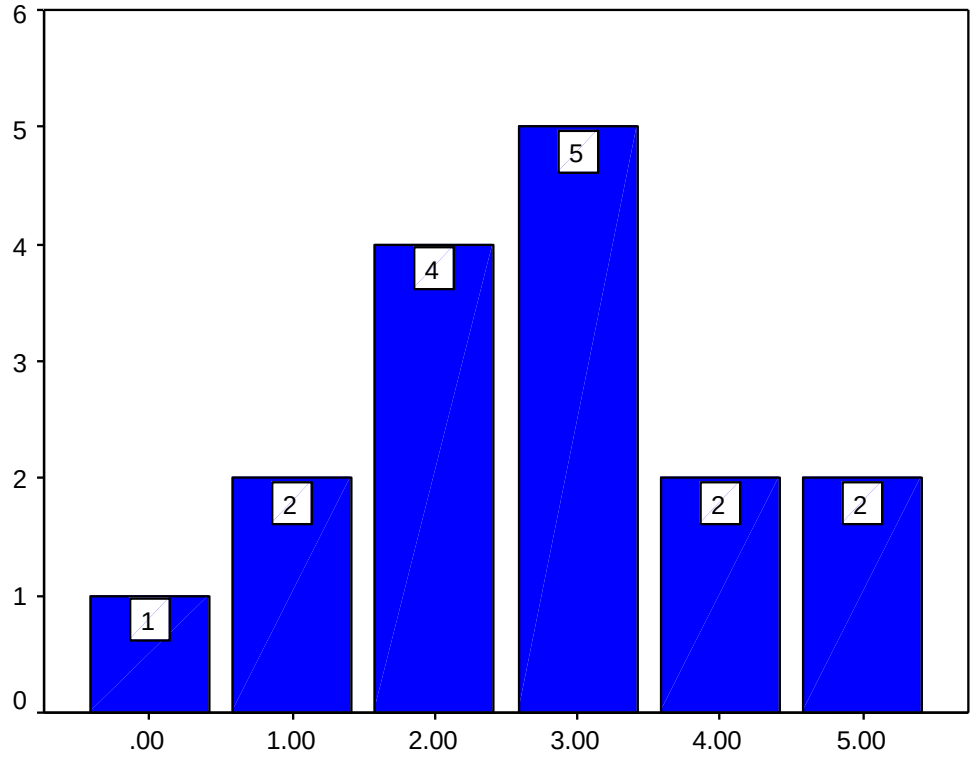
No. of decayed teeth	Frequency	Relative Frequency
0	1	0.0625
1	2	0.125
2	4	0.25
3	5	0.3125
4	2	0.125
5	2	0.125
Total	n = 16	1

Representing the simple frequency table using the bar chart

We can represent the above simple frequency table using the bar chart.

We can get :

1. **The sample size?**
2. **Number of children with decayed teeth=2?**
3. **Relative frequency of children with decayed teeth =4?**



2.3 Frequency Distribution for Continuous Random Variables

For large samples, we can't use the simple frequency table to represent the data.

We need to divide the data into groups or intervals or classes.

So, we need to determine:

The range (R).

It is the difference between the largest and the smallest observation in the data set.[R=Max –Min]

Class interval	Frequency (f)
30 – 39	11
40 – 49	46
50 – 59	70
60 – 69	45
70 – 79	16
80 – 89	1
Total	189

Sum of frequency =sample size=n

Example:
The following table gives the hemoglobin level (in g/dl) of a sample of 50 apparently (ظاهرياً) healthy men aged 20-24. Find the grouped frequency distribution for the data.

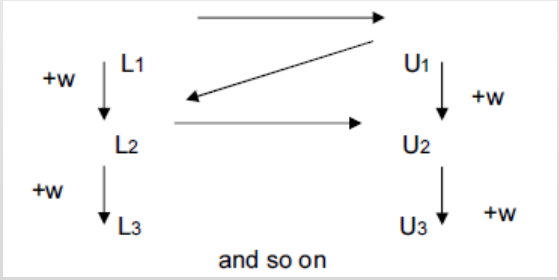
17	17.1	14.6	14	16.1	15.9	16.3	14.2	16.5
17.7	15.7	15.8	16.2	15.5	15.3	17.4	16.1	14.4
15.9	17.3	15.3	16.4	18.3	13.9	15	15.7	16.3
15.2	13.5	16.4	14.9	15.8	16.8	17.5	15.1	17.3
16.2	16.3	13.7	17.8	16.7	15.9	16.1	17.4	15.8

-What is the variable? The sample size?

- The max=18.8
-The min=13.5
-The range=max-min=18.8-13.5=4.8

In example to group the data we use a set of intervals, called class intervals.

The width (w) is the distance from the lower or upper limit of one class interval to the same limit of the next class interval.
Let we denote the lower limit and upper limit of the class interval by L and U, that is the first class is L1-U1, the second class is L2-U2 #
To find the class intervals we use the following relationship



The Cumulative Frequency:

It can be computed by adding successive frequencies.

The Cumulative Relative Frequency:

It can be computed by adding successive relative frequencies.

For the above example, the following table represents the cumulative frequency, the relative frequency, and the cumulative relative frequency.

R.f= freq/n

Class interval	Frequency Freq (f)	Cumulative Frequency	Relative Frequency R.f	Cumulative Relative Frequency
30 – 39	**11**	11	0.0582	0.0582
40 – 49	**46**	57	0.2434	-
50 – 59	**-**	127	-	0.6720
60 – 69	**45**	-	0.2381	0.9101
70 – 79	**16**	188	0.0847	0.9948
80 – 89	**1**	189	0.0053	1
Total	189		1	

Example:

- From the above frequency table, complete the table then answer the following questions:
- 1-The number of objects with age less than 50 years ?
- 2-The number of objects with age between 40-69 years ?
- 3-Relative frequency of objects with age between 70-79 years ?
- 4-Relative frequency of objects with age more than 69 years ?
- 5-The percentage of objects with age between 40-49 years ?

- 6- The percentage of objects with age less than 60 years ?

- 7-The Range (R) ?

- 8- Number of intervals (K)?

- 9- The width of the interval (W) ?

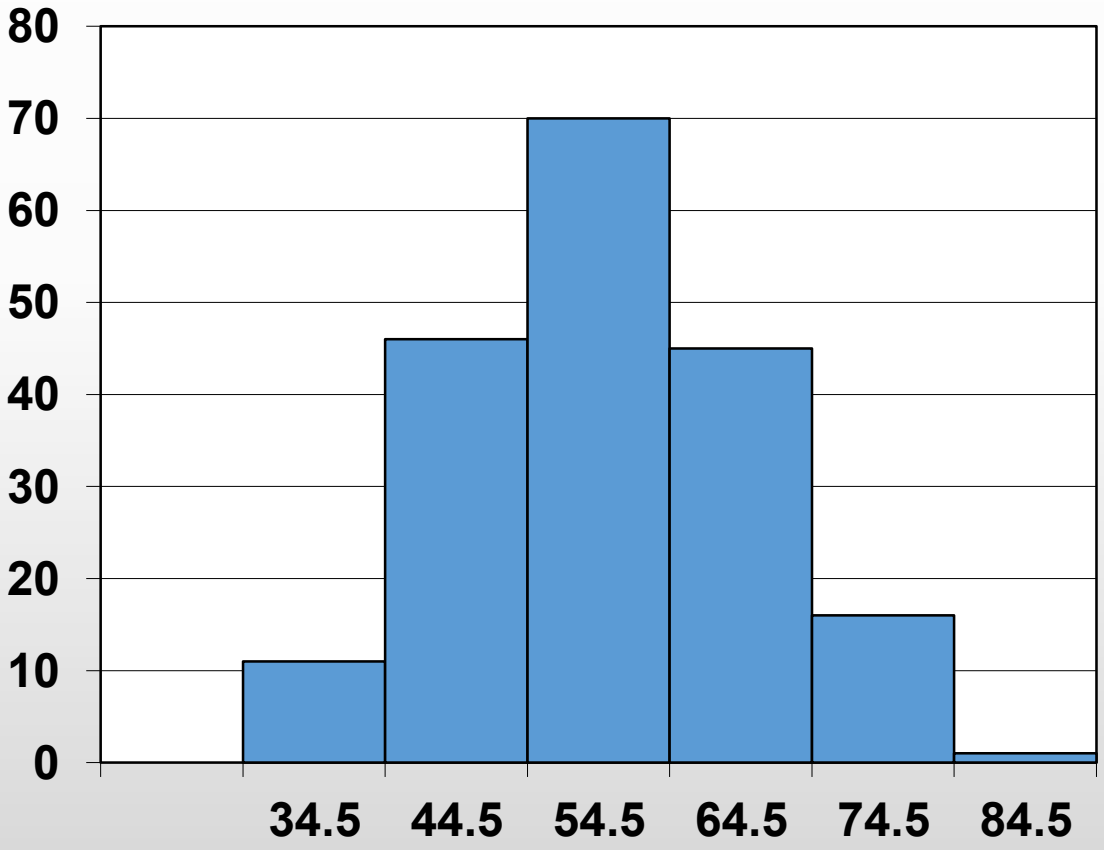
Displaying grouped frequency distributions

Grouped frequency distributions can be displayed by

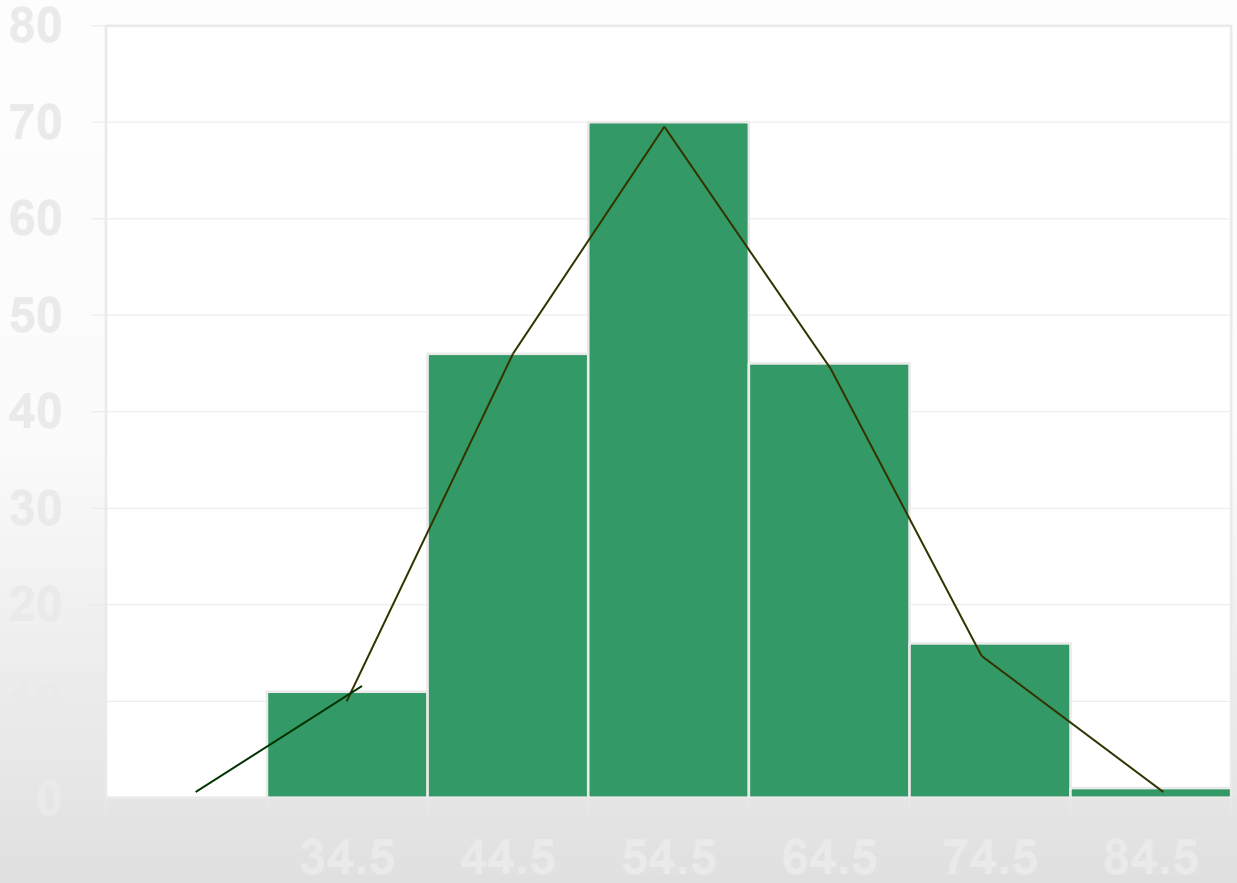
- Histogram
- Polygon
- Curves

(For frequency or relative frequency distributions)

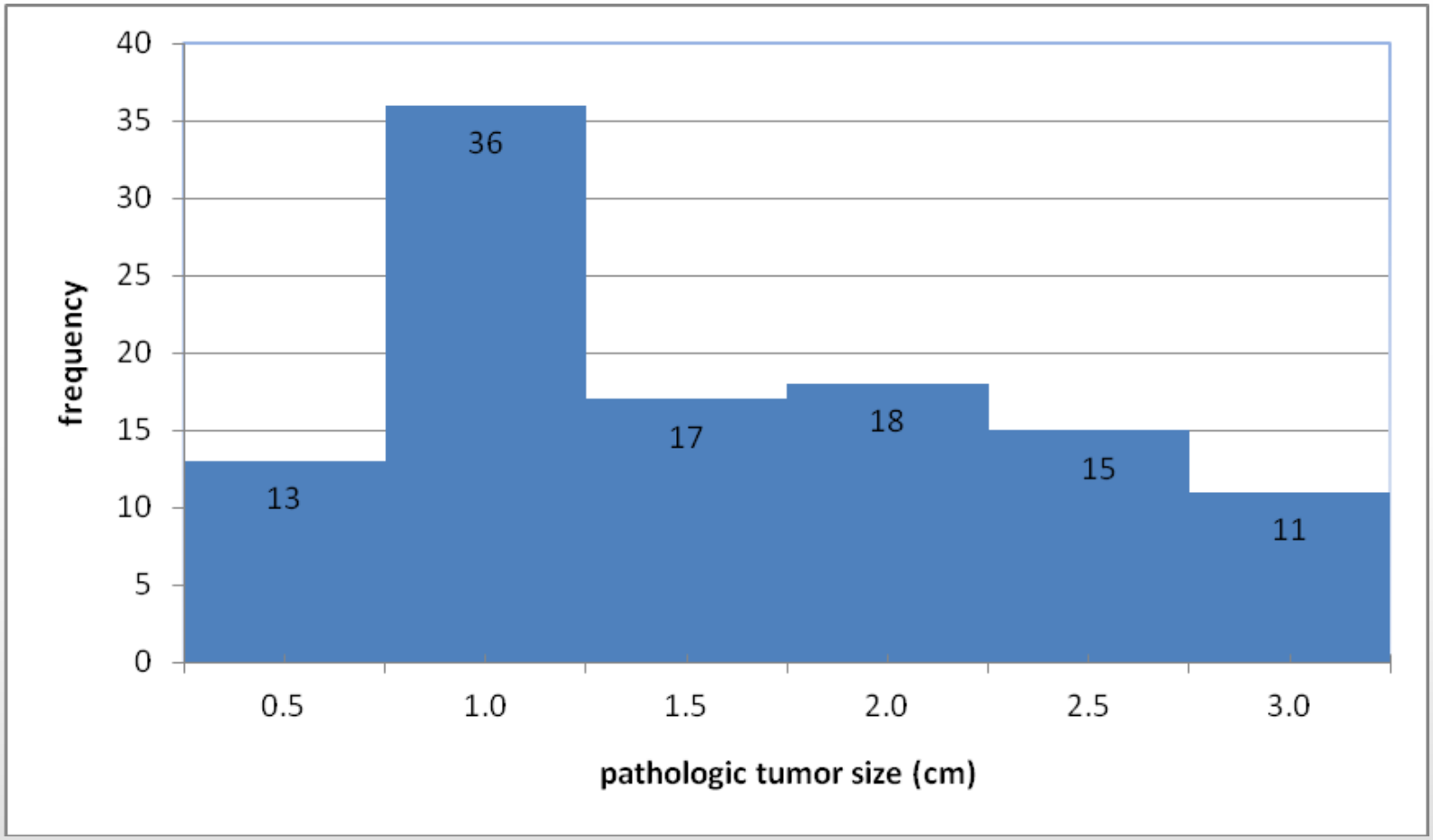
Representing the grouped frequency table using the Histogram



Representing the grouped frequency table using the Polygon



[H.W] the following histogram show the frequency distribution of pathologic tumor size (in cm) for a sample of 110 cancer patients:



1. The percent of cancer patients with approximate level of **pathologic tumor size** =2 cm is:

(a) 18% (b) 50% (c) 16.36% (d) 32.72% (e) 36% (f) 0%

2. The number of cancer patients with lowest **pathologic tumor size** is:

(a) 0.5 (b) 3 (c) 11 (d) 13 (e) 15 (f) 24

3. The approximate **size of pathologic tumor** with highest percentage of patients is:

(a) 3 (b) 1 (c) 36 (d) 110 (e) 32.72% (f) 16.36%

4. What the approximate value of the sample mean

(a) 18.33 (b) 36 (c) 1 (d) 1.75 (e) 1.586 (f) we can't find it

5. The mode equals

(a) 1 (b) 3 (c) 36 (d) 55 (e) 110 (f) we can't find it

6. The approximate value of the sample variance

(a) 3 (b) 0.6 (c) 0.774 (d) 1.586 (e) 110 (f) we can't find it

Exercises

- Pages : 31 – 34
- H.W.: 2.3.2, 2.3.6 , 2.3.7(a)

Exercises: (For Students)

Q2.3.5: The following table shows the number of hours 45 hospital patients slept following the administration of a certain anesthetic.

(a) From these data construct:

* A relative frequency distribution

Class Interval	Frequency
1-5	21
6-10	16
11-15	6
16-20	2
Total	45

(b) How many of the measurements are greater than 10? Ans: 8

(c) What percentage of the measurements are between 6-15 ?

Ans: 49%

(d) What proportion of the measurement is less than or equal 15? Ans: 0.96

Section (2.4) :
Descriptive Statistics
Measures of Central Tendency
Page 38 - 41

key words:

Descriptive Statistic, measure of central tendency ,statistic, parameter, mean (μ) ,median, mode.

The Statistic and The Parameter

- ***A Statistic:***

It is a descriptive measure computed from the data of a sample.

- ***A Parameter:***

It is a a descriptive measure computed from the data of a population.

Since it is difficult to measure a parameter from the population, a sample is drawn of size n, whose values are $\chi_1, \chi_2, \ldots, \chi_n$. From this data, we measure the statistic.

Measures of Central Tendency

A measure of central tendency is a measure which indicates where the middle of the data is.

The three most commonly used measures of central tendency are:

The Mean, the Median, and the Mode.

The Mean :

It is the average of the data.

The Population Mean:

Population mean: let $X_1, X_2, \ldots, X_N$ be the population values of the variable (usually unknown), then the population mean is $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$, then we use the sample mean to estimate or approximate it.

The Sample Mean:(or Average)

let $x_1, x_2, \ldots, x_N$ be the sample values of the variable, then the sample mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

The sample mean is an estimator of a population mean

Example:

Here is a random sample of size 10 of ages, where

$\chi_1 = 42$, $\chi_2 = 28$, $\chi_3 = 28$, $\chi_4 = 61$, $\chi_5 = 31$,

$\chi_6 = 23$, $\chi_7 = 50$, $\chi_8 = 34$, $\chi_9 = 32$, $\chi_{10} = 37$.

$$\bar{x} = \frac{42 + 28 + 28 + 61 + 31 + 23 + 50 + 34 + 32 + 37}{10} = 36.6$$

Properties of the Mean:

- **Uniqueness.** For a given set of data there is one and only one mean.
- **Simplicity.** It is easy to understand and to compute.
- **Affected by extreme values.** Since all values enter into the computation.

Example:

Assume the values are 115, 110, 119, 117, 121 and 126. The mean = 118.

But assume that the values are 75, 75, 80, 80 and 280. The mean = 118, a value that is not representative of the set of data as a whole.

The Median:

When ordering the data, it is the observation that divide the set of observations into two equal parts such that half of the data are before it and the other are after it.

* If n is odd, the median will be the middle of observations. It will be the $(n+1)/2^{th}$ ordered observation.

When n = 11, then the median is the 6^{th} observation.

* If n is even, there are two middle observations. The median will be the mean of these two middle observations. It will be the mean of the [$(n/2)^{th}$, $(n/2 +1)^{th}$]ordered observation.

When n = 12, then the median is the 6.5^{th} observation, which is an observation halfway between the 6^{th} and 7^{th} ordered observation.

Example:

For the same random sample, the ordered observations will be as:

$$\underbrace{23, 28, 28, 31}, \underbrace{\boxed{32, 34}}_{5^{th}, 6^{th}} \underbrace{37, 42, 50, 61}.$$

Since n = 10, then the median is the 5.5th observation, i.e. = (32+34)/2 = 33.

Let the same random sample, the ordered observations will be as:

$$\underbrace{23, 28, 28, 31}, \underbrace{\boxed{32}}_{5^{th}} \underbrace{37, 42, 50, 61}.$$

Since n = 9, then the median is the 5th observation, i.e. = 32.

Properties of the Median:

- Uniqueness. For a given set of data there is one and only one median.
- Simplicity. It is easy to calculate.
- It is not affected by extreme values as is the mean.

The Mode:

It is the value which occurs most frequently.

If all values are different there is no mode.

Sometimes, there are more than one mode.

Example:

For the same random sample, the value 28 is repeated two times, so it is the mode.

Properties of the Mode:

- **Sometimes, it is not unique.**
- **It may be used for describing qualitative data.**
- **It is not affected by extreme values**

Examples

Find the mean and the mode for the following Relative Frequency?

Mode = 7
(has the higher frequency)

Age(x)	frequency (f)	X f
5 6 7 10	2 3 4 1	10 18 28 10
Total	10	66

Examples

Find the mode for the following grouped Frequency table?

Age	Frequency (f)	Midpoint (X)	X f
1 - 3 4 - 6 7 - 9 10 - 12	2 1 4 3	2 5 8 11	4 5 32 33
Total	10	_	74

Mode :interval(7 – 9)
(can't give exact number only the interval with higher Frequency)

Example

A sample of 80 families have been asked about the number of times to travel abroad. The computer results of the SPSS are given below

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	11	13.8	13.8	13.8
	1	7	8.8	8.8	22.5
	2	5	6.3	6.3	28.8
	3	6	7.5	7.5	36.3
	4	8	10.0	10.0	46.3
	5	6	7.5	7.5	53.8
	6	6	7.5	7.5	61.3
	7	10	12.5	12.5	73.8
	8	6	7.5	7.5	81.3
	9	6	7.5	7.5	88.8
	10	5	6.3	6.3	95.0
	11	2	2.5	2.5	97.5
	12	1	1.3	1.3	98.8
	13	1	1.3	1.3	100.0
	Total	80	100.0	100.0	
Total		**80**			

From above table:

A) The variable is

(a) Number of families

(b) Number of times to Travel abroad

(c) None of these

B) The type of the variable is

(a) Quantitative Discrete

(b) Qualitative

(c) Quantitative Continuous

(d) Normal

(e) Binomial

(f) None of these

C) Number of families that travelled abroad 7 times is

(a) 7.5 (b) 0 (c) 6 (d) 1 (e) 53.8 (f)10

D) The percentage of families that travelled abroad less than or equal to 10 times is

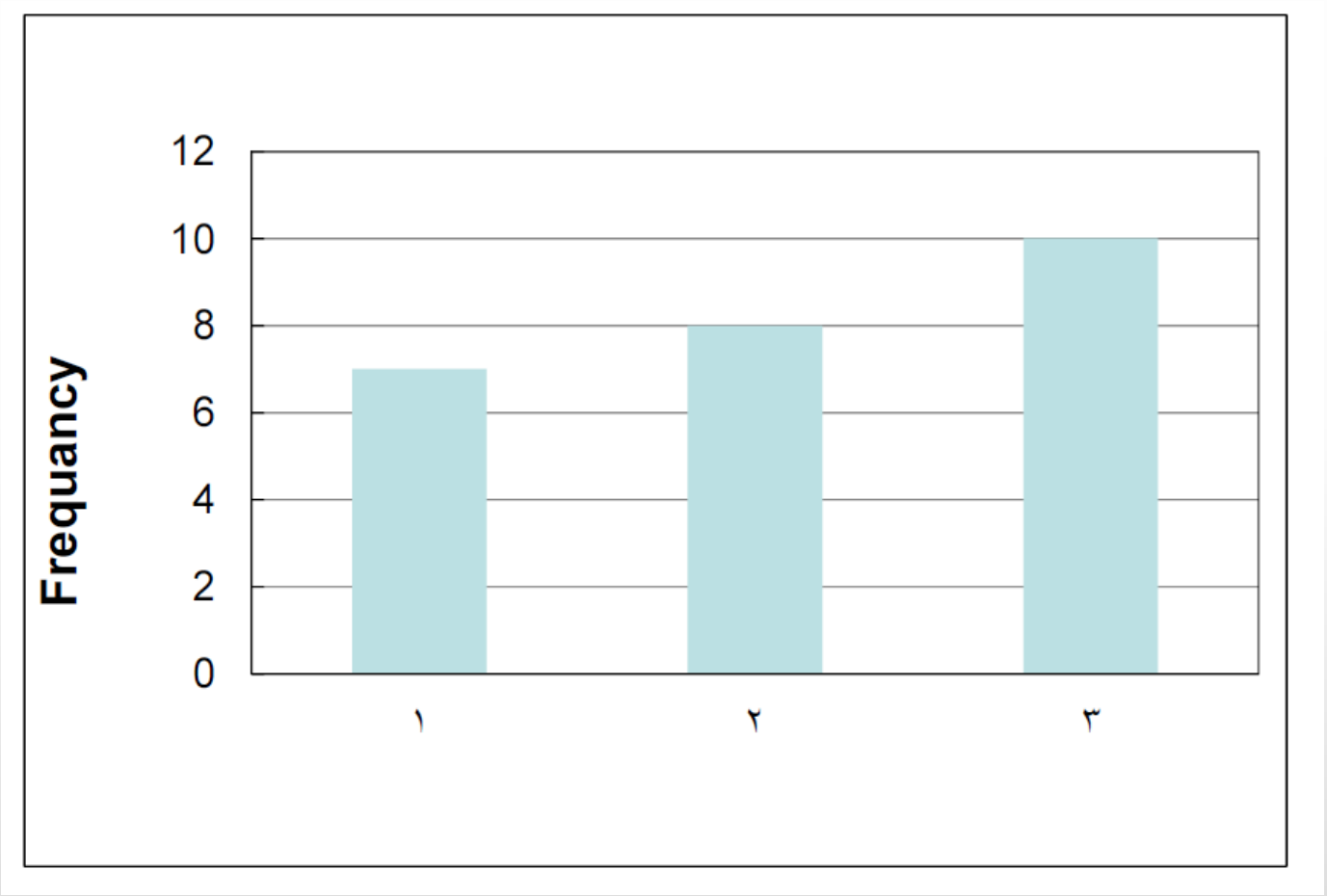
(a) 73.8% (b) 0 (c)100% (d) 88.8% (e) 5% (f) 95%

E) The mode of travel times is

(a) 13.8 (b) 0 (c) 6 (d) 11 (e) 80 (f)13

Example

By using the computer results of SPSS the plot of the number of courses in English that student takes in a year is obtained:



1) The type of the graph is:

(a) Bar chart (b) polygon (c) histogram (d) line (e) curve

2)The Variable is:

a)Number of students

b)Number of courses

c)English

d)Arabic

3) The total number of students who study in English is:

(a) 0 (b) 25 (c) 12 (d) 6 (e) 3

4) The number of students who study two courses in English is:

(a) 0 (b) 2 (c) 7 (d) 8 (e) 5

5)The number of students who study at least two courses in English is:

(a)7 (b) 8 (c) 15 (d) 18 (e) 25

6)The percent of students who study at most one course in English is:

(a)7% (b) 18% (c) 28% (d) 60% (e) 72%

7) The sample mean is

(a)2.55 (b) 255 (c) 3 (d) 3.1875 (e) 40

8)The sample mode is

(a)0 (b) 3 (c) 10 (d) 2 (e) no mode

2.5. Descriptive Statistics – Measures of Dispersion:

- A measure of dispersion conveys information regarding the amount of variability present in a set of data.
- Note:

1. If all the values are the same

→ There is no dispersion .

2. If all the values are different

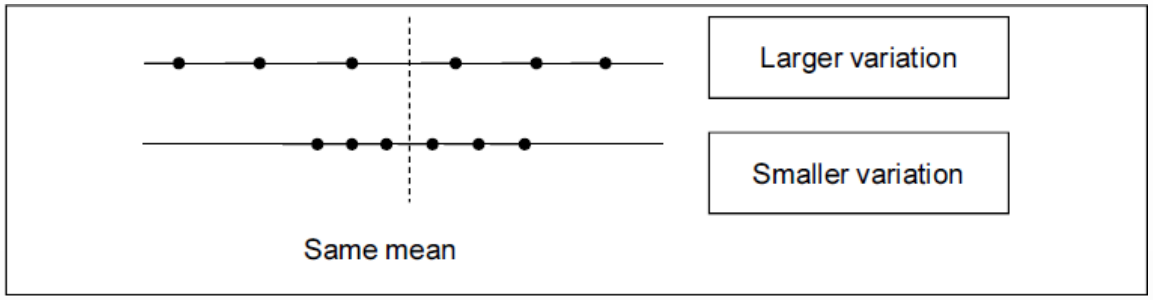
→ There is a dispersion:

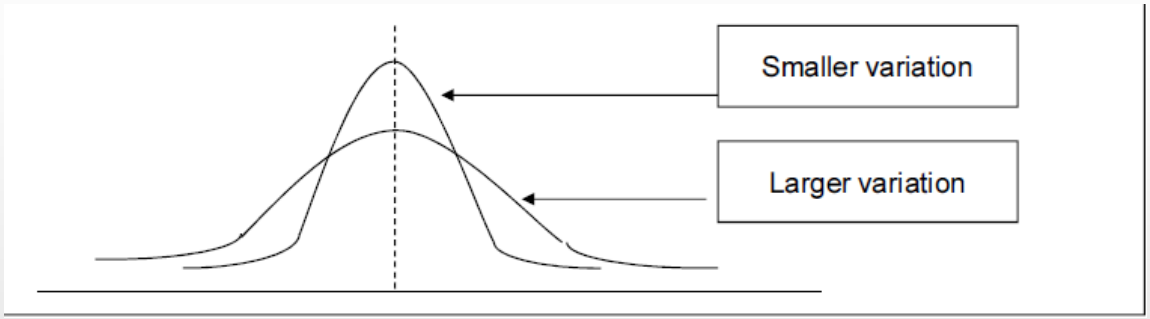
a). If the values close to each other

→ The amount of Dispersion small.

b) If the values are widely scattered

→ The Dispersion is greater.


Larger variation
Smaller variation
Same mean


Smaller variation
Larger variation

Ex. Figure 2.5.1 –Page 43

- ** **Measures of Dispersion are :**

1. Range (R).
2. Variance.
3. Standard deviation.
4. Coefficient of variation (C.V).

1.The Range (R):

- Range =Largest value- Smallest value = $X_L - X_S$
- **Note:**
- Range concern only onto two values
- **Example 2.5.1 Page 40:**
- Refer to Ex 2.4.2.Page 37
- Data:
- 43,66,61,64,65,38,59,57,57,50.
- Find Range?
- Range=66-38=28

2.The Variance:

- It measure dispersion relative to the scatter of the values a bout there mean.

a) Sample Variance (S^2) :

$S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$, **where $\bar{x}$ is sample mean**

- **Example 2.5.2 Page 40:**

Refer to Ex 2.4.2.Page 37

Find Sample Variance of ages , $\bar{x}$= 56

Solution:

S^2= [(43-56) ^{2} +(66-43) ^{2}+.....+(50-56) ^{2}]/ 10

= 900/10 = 90

b)Population Variance (σ^2) :

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$ where , μ is Population mean

3.The Standard Deviation:

is the square root of variance=$\sqrt{\text{variance}}$

a) Sample Standard Deviation =$S = \sqrt{S^2}$

b) Population Standard Deviation =$\sigma = \sqrt{\sigma^2}$

4.The Coefficient of Variation (C.V):

Is a measure use to compare the dispersion in two sets of data which is independent of the unit of the measurement .

$$C.V = \frac{S}{\bar{X}}(100)$$

where S: Sample standard deviation.

$\bar{X}$: Sample mean.

Example 2.5.3 Page 46:

- Suppose two samples of human males yield the following data:

	Sampe1	Sample2
Age	25-year-olds	11year-olds
Mean weight	145 pound	80 pound
Standard deviation	10 pound	10 pound

We wish to know which is more variable.

Solution:

c.v (Sample1)= (10/145)*100= 6.9

c.v (Sample2)= (10/80)*100= 12.5

Then age of 11-years old(sample2) is more variation.

Exercises

Pages : 52 – 53

Questions: 2.5.1 , 2.5.2 ,2.5.3

H.W.: 2.5.4 , 2.5.5, 2.5.6, 2.5.14

* Also you can solve in the review questions page 57:

Q: 12,13,14,15,16, 19

Exercises:

For each of the data sets in the following exercises Use calculator if possible)(compute:

(a) The mean

(b) The median

(c) The mode

(d) The range

(e) The variance

(f) The standard deviation

(g) The coefficient of variatio

Q2.5.3:

Butz et al. (A-10) evaluated the duration of benefit derived from the use of noninvasive positive-pressure ventilation by patients with amyotrophic lateral sclerosis on symptoms, quality of life, and survival. One of the variables of interest is partial pressure of arterial carbon dioxide (PaCO2). The values below (mm of Hg) reflect the result of baseline testing on 30 subjects as established by arterial blood gas analyses.

seven rat pups from the experiment involving the carotid artery.

500 570 560 570 450 560 570

(a) The mean (b) The median

Ans: 540 Ans: 560

(c) The mode (d) The range

Ans: 570 Ans: 120

(e) The variance (f) The standard deviation

Ans: 2200 Ans: 46.90

(g) The coefficient of variation Ans: 8.69%

H.W :Q2.5.1:

Porcellini et al. (A-8) studied 13 HIV-positive patients who were treated with highly active antiretroviral therapy (HAART) for at least 6 months. The CD4 T cell counts ($\times 10^6$/L) at baseline for the 13 subjects are listed below.

230 205 313 207 227 245 173

58 103 181 105 301 169

Q2.5.2:H.W Shrair and Jasper (A-9) investigated whether decreasing the venous return in young rats would affect ultrasonic vocalizations (USVs). Their research showed no significant change in the number of ultrasonic vocalizations when blood was removed from either the superior vena cava or the carotid artery. Another important variable measured was the heart rate (bmp) during the withdrawal of blood. The data below presents the heart rate of

40.0 47.0 34.0 42.0 54.0 48.0 53.6 56.9 58.0
45.0 54.5 54.0 43.0 44.3 53.9 41.8 33.0 43.1
52.4 37.9 34.5 40.1 33.0 59.9 62.6 54.1 45.7
40.6 56.6 59.0

(a) The mean

Ans: 47.72

(b) The median

Ans: 46.35

(c) The mode

Ans: 33, 54

(d) The range

Ans: 29.6

(e) The variance

Ans: 84.135

(f) The standard deviation

Ans: 9.17251

(g) The coefficient of variation

(H.W)Q2.5.4:

According to Starch et al. (A-11), hamstring tendon grafts have been the "weak link" in anterior cruciate ligament reconstruction. In a controlled laboratory study, they compared two techniques for reconstruction : either an interference screw or a central sleeve and screw on the tibial side. For eight cadaveric knees, the measurements below represent the required force (in Newtones) at which initial failure of graft strands occurred for the central sleeve and screw technique.

172.5 216.63 212.62 98.97 66.95 239.76 19.57 195.72

(a) The mean (b) The median

Ans: 152.84 Ans: 184.11

(c) The mode (d) The range

Ans: no mode Ans: 220.19

(e) The variance (f) The standard deviation

Ans: 6494.732 Ans: 80.5899

(g) The coefficient of variation Ans: 52.73%

Q2.5.5: Cardosi et al. (A-12) performed a 4 years retrospective review of 102 women undergoing radical hysterectomy for cervical or endometrial cancer. Catheter-associated urinary tract infection was observed in 12 of the subjects. Below are the numbers of postoperative days until diagnosis of the infection for each subject experiencing an infection.

16 10 49 15 6 15 8 19 11 22 13 17

(a) The mean (b) The median

Ans: 16.75 Ans: 15

(c) The mode (d) The range

Ans: 15 Ans: 43

(e) The variance (f) The standard deviation

Ans: 124.0227 Ans: 11.1365

(g) The coefficient of variation Ans: 66.49%

Q2.5.6: The purpose of a study by Nozama et al. (A-13) was to evaluate the outcome of surgical repair of pars interarticularis defect by segmental wire fixation in young adults with lumbar spondylolysis. The authors found that segmental wire fixation historically has been successful in the treatment of nonathletes with spondylolysis, but no information existed on the results of this type of surgery in athletes. In a retrospective study, the authors found 20 subjects who had the surgery between 1993 and 2000. For these subjects, the data below represent the duration in months of follow-up care after the operation.

103 68 62 60 60 54 49 44 42 41 38 36 34 30 19 19 19 19 19 17 16

(a) The mean

Ans: 41.5

(b) The median

Ans: 39.5

(c) The mode

Ans: 19

(d) The range

Ans: 87

(e) The variance

Ans: 490.264

(f) The standard deviation

Ans: 22.1419

(g) The coefficient of variation Ans: 53.35%

Q2.5.14: In a pilot study, Huizinga et al. (A-14) wanted to gain more insight into the psychosocial consequences for children of a parent with cancer. For the study, 14 families participated in semistructured interviews and completed standardized questionnaires. Below is the age of the sick parent with cancer (in years) for the 14 families.

37 48 53 46 42 49 44 38 32 32 51 51 48 41

(a) The mean
Ans: 43.7143

(b) The median
Ans: 45

(c) The mode
Ans: 32, 51

(d) The range
Ans: 21

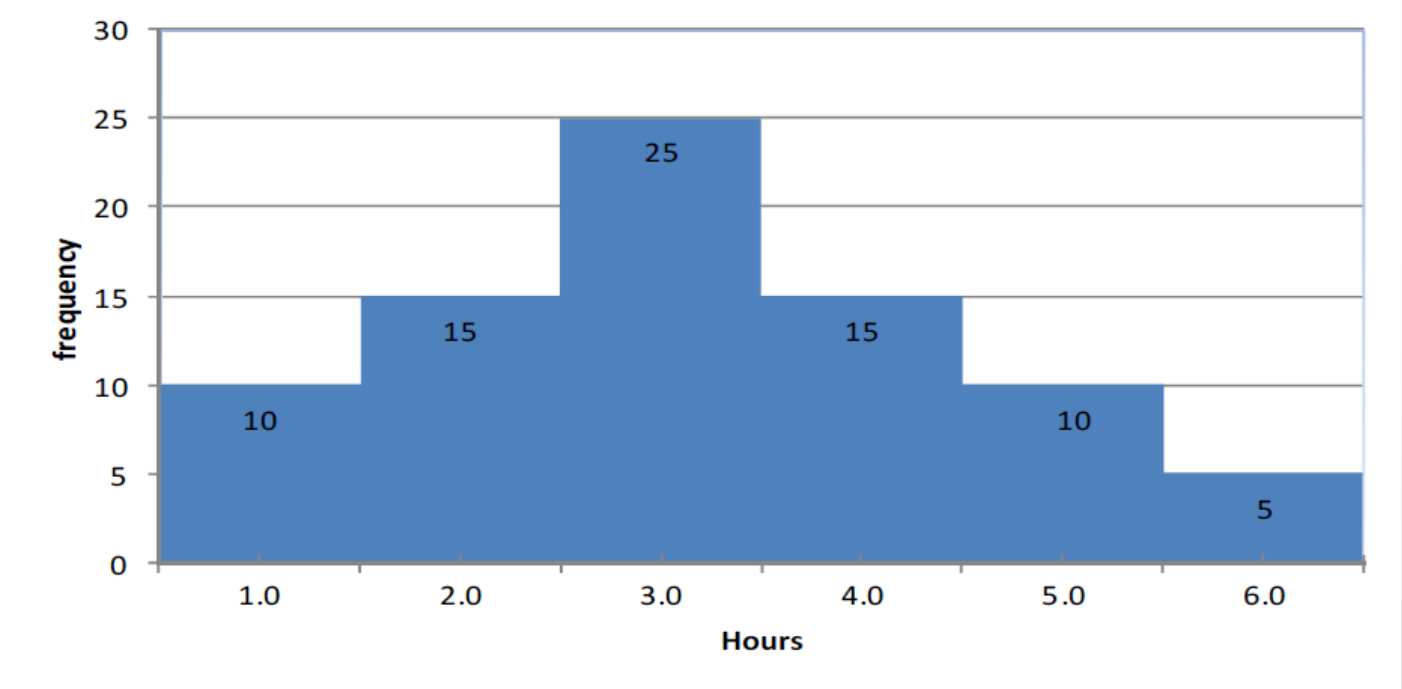
(e) The variance (f) The standard deviation

Ans: 48.0659 Ans: 6.93296

(g) The coefficient of variation Ans: 15.8597%

Example :

For a sample of patients, we obtain the following graph for approximated hours spent without pain after a certain surgery



1) The type of the graph is:

a) Bar chart (b) polygon (c) histogram (d) line (e) curve

2) The number of patients stayed the longest time without pain is:

a) 10 (b) 15 (c) 6 (d) 5 (e) 80

3)The percent of patients spent 3.5 hours or more without pain is:

a)37.5% (b) 68.75% (c) 18.75% (d) 50% (e) 25%

4)The lowest number of hours spent without pain is:

a)10 (b) 1 (c) 0.5 (d) 5 (e) 25 (f) 6.5

5)What the approximate value of the sample mean

a)2.55 (b) 255 (c) 3 (d) 3.1875 (e) 40 (f) we can't find it

6)The sample mode equals

a)80 (b) 3 (c) 15 (d) 2,4 (e) 6 (f) we can't find it

The SPSS computer results of the age of patients in one of the Riyadh hospitals are given below

Find :

a) Variable name

b) The type of the variable

c) The mode

d) The mean age of the patients

e) The median age of the patients

f) The variance

g) Sample size

h) The coefficient of variation

Statistics

AGE

N	Valid	20
	Missing	0
Mean		4.6000
Median		4.5000
Mode		5.00
Std. Deviation		2.23371
Percentiles	25	3.0000
	50	4.5000
	75	6.0000

screw on the tibial side. For eight cadaveric knees, the measurements below represent the required force (in Newtones) at which initial failure of graft strands occurred for the central sleeve and screw technique.

172.5 216.63 212.62 98.97 66.95 239.76 19.57 195.72

(a) The mean (b) The median

Ans: 152.84 Ans: 184.11

(c) The mode (d) The range

Ans: no mode Ans: 220.19

(e) The variance (f) The standard deviation

Ans: 6494.732 Ans: 80.5899

(g) The coefficient of variation Ans: 52.73%

Solution:

1. Data:, $n_C=10$, $n_{SCI}=10$, $S_C=21.8$, $S_{SCI}=133.1$,$\alpha=0.05$.

- $\bar{X}_C = 126.1$ $\bar{X}_{SCI} = 133.1$
- (calculated from data)

2.Assumption: Two population are normal, σ^2_1 , σ^2_2 are unknown but equal

3. Hypotheses:

H_0: $\mu_C = \mu_{SCI} \rightarrow \mu_C - \mu_{SCI} = 0$

H_A: $\mu_C < \mu_{SCI} \rightarrow \mu_C - \mu_{SCI} < 0$

- **4.Test Statistic:**

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(126.1 - 133.1) - 0}{\sqrt{756.04}\sqrt{\frac{1}{10} + \frac{1}{10}}} = -0.569$$

- Where,

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{9(21.8)^2 + 9(32.3)^2}{10 + 10 - 2} = 756.04$$

5. Decision Rule:

Reject H_0 if $T < -T_{1-\alpha,(n_1+n_2-2)}$

$T_{1-\alpha,(n_1+n_2-2)} = T_{0.95,18} = 1.7341$

(from table E)

6-**Conclusion:** Fail to reject H_0

since $-0.569 < -1.7341$

Or

Fail to reject H_0 since $p = -1.33 > \alpha = 0.05$

7.4 Paired Comparisons:

- In this section, we are interested in comparing the means of two related (non-independent/dependent) normal populations.
- In other words, we wish to make statistical inference for the difference between the means of two related normal populations.
- Paired t-Test concerns about testing the equality of the means of two related normal populations.

Examples of related populations are:

1. Height of the father and height of his son.
2. Mark of the student in MATH and his mark in STAT.
3. Pulse rate of the patient before and after the medical treatment.
4. Hemoglobin level of the patient before and after the medical treatment.

Example: (effectiveness of a diet program)
Suppose that we are interested in studying the effectiveness of a certain diet program. Let the random variables X and Y are as follows:

X = the weight of the individual before the diet program

Y= the weight of the same individual after the diet program

We assume that the distributions of these random variables are normal with means μ_1 and μ_2, respectively.

These two variables are related (dependent/non-independent) because they are measured on the same individual.

Populations:

1-st population (X): weights before a diet program

mean = μ_1

2-nd population (Y): weights after the diet program

mean = μ_2

where:

$$\mu_D = \mu_1 - \mu_2$$

- We calculate the following quantities:
  - The differences (D-observations):

$$D_i = X_i - Y_i \quad (i=1, 2, \ldots, n)$$

  - Sample mean of the D-observations (differences):

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{D_1 + D_2 + \ldots + D_n}{n}$$

  - Sample variance of the D-observations (differences):

$$S_D^2 = \frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1} = \frac{(D_1 - \overline{D})^2 + (D_2 - \overline{D})^2 + \cdots + (D_n - \overline{D})^2}{n-1}$$

  - Sample standard deviation of the D-observations:

$$S_D = \sqrt{S_D^2}$$

Confidence Interval for the Difference between the Means of Two Related Normal Populations ($\mu_D = \mu_1 - \mu_2$):

In this section, we consider constructing a confidence interval for the difference between the means of two related (non-independent) normal populations. As before, let us define the difference between the two means as follows:

$$\mu_D = \mu_1 - \mu_2$$

where μ_1 is the mean of the first population and μ_2 is the mean of the second population. We assume that the two normal populations are not independent.

Result:

A $(1-\alpha)100\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is:

- We select a random sample of n individuals. At the beginning of the study, we record the individuals' weights before the diet program (X). At the end of the diet program, we record the individuals' weights after the program (Y). We end up with the following information and calculations:

Individual i	Weight before X_i	Weight after Y_i	Difference $D_i = X_i - Y_i$
1	X_1	Y_1	$D_1 = X_1 - Y_1$
2	X_2	Y_2	$D_2 = X_2 - Y_2$
.	.	.	
.	.	.	

$$\bar{D} \pm t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$$

$$\bar{D} - t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}} < \mu_D < \bar{D} + t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$$

where:

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n} \quad , \quad S_D = \sqrt{S_D^2} \, , \quad S_D^2 = \frac{\sum_{i=1}^{n} \left(D_i - \bar{D} \right)^2}{n-1} \, , \quad \mathrm{df} = \nu = \mathrm{n} - 1.$$

Example:

Consider the data given in the previous numerical example:

Individual (i)	1	2	3	4	5	6	7	8	9	10
Weight before (X_i)	86.6	80.2	91.5	80.6	82.3	81.9	88.4	85.3	83.1	82.1
Weight after (Y_i)	79.7	85.9	81.7	82.5	77.9	85.8	81.3	74.7	68.3	69.7

Find a 95% confidence interval for the difference between the mean of weights before the diet program (μ_1) and the mean of weights after the diet program (μ_2).

Solution:

Calculations:

i	X_i	Y_i	$D_i = X_i - Y_i$
1	86.6	79.7	6.9
2	80.2	85.9	-5.7
3	91.5	81.7	9.8
4	80.6	82.5	-1.9
5	82.3	77.9	4.4
6	81.9	85.8	-3.9
7	88.4	81.3	7.1
8	85.3	74.7	10.6
9	83.1	68.3	14.8
10	82.1	69.7	12.4
sum	$\sum X$=842	$\sum Y$=787.5	$\sum D$=54.5

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{54.5}{10} = 5.45$$

$$S_D^2 = \frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1} = \frac{(6.9-5.45)^2 + \ldots + (12.4-5.45)^2}{10-1} = 50.3283$$

$$S_D = \sqrt{S_D^2} = \sqrt{50.3283} = 7.09$$

We need to find a 95% confidence interval for $\mu_D = \mu_1 - \mu_2$:

$$\overline{D} \pm t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$$

We have found:

$$\overline{D} = 5.45 \quad , \quad S_D^2 = 50.3283 \quad , \quad S_D = \sqrt{S_D^2} = 7.09$$

The value of the reliability coefficient $t_{1-\frac{\alpha}{2}}$ $(df = \nu = n-1 = 9)$ is

$t_{1-\frac{\alpha}{2}} = t_{0.975} = 2.262$.

Therefore, a 95% confidence interval for $\mu_D = \mu_1 - \mu_2$ is

$$5.45 \pm (2.262) \frac{7.09}{\sqrt{10}}$$

$$5.45 \pm 5.0715$$

$$0.38 < \mu_D < 10.52$$

$$0.38 < \mu_1 - \mu_2 < 10.52$$

H_o: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_2$

Equivalently,

H_o: $\mu_1 - \mu_2 = 0$

H_A: $\mu_1 - \mu_2 \neq 0$

Equivalently,

H_o: $\mu_D = 0$

H_A: $\mu_D \neq 0$

where:

$$\mu_D = \mu_1 - \mu_2$$

- We calculate the following quantities:
  - The differences (D-observations):

    $$D_i = X_i - Y_i \quad (i=1, 2, \ldots, n)$$

  - Sample mean of the D-observations (differences):

    $$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{D_1 + D_2 + \ldots + D_n}{n}$$

  - Sample variance of the D-observations (differences):

    $$S_D^2 = \frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1} = \frac{(D_1 - \overline{D})^2 + (D_2 - \overline{D})^2 + \cdots + (D_n - \overline{D})^2}{n-1}$$

  - Sample standard deviation of the D-observations:

    $$S_D = \sqrt{S_D^2}$$

Test statistic is

$$t = \frac{\overline{D}}{S_D / \sqrt{n}} \sim t(n-1)$$

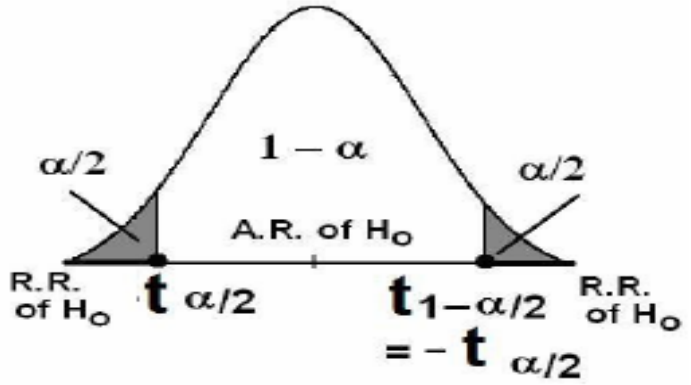
This statistic has a t-distribution with df = $\nu = n-1$.

- Rejection Region of H_o:

  Critical values are: $t_{\alpha/2}$ and $t_{1-\alpha/2} = -t_{\alpha/2}$.

  The rejection region (critical region) at the significance level α is:

$$t < t_{\alpha/2} \quad \text{or} \quad t > t_{1-\alpha/2} = -t_{\alpha/2}$$



- Decision:

We reject H_o and accept H_A at the significance level α if $T \in$ R.R. ,i.e., if:

$$t < t_{\alpha/2} \quad \text{or} \quad t > t_{1-\alpha/2} = -t_{\alpha/2}$$

Numerical Example:

In the previous example, suppose that the sample size was 10 and the data were as follows:

Individual (i)	1	2	3	4	5	6	7	8	9	10
Weight before (X_i)	86.6	80.2	91.5	80.6	82.3	81.9	88.4	85.3	83.1	82.1
Weight after (Y_i)	79.7	85.9	81.7	82.5	77.9	85.8	81.3	74.7	68.3	69.7

Does these data provide sufficient evidence to allow us to conclude that the diet program is effective? Use α=0.05 and assume that the populations are normal.

Solution:

μ_1 = the mean of weights before the diet program

μ_2 = the mean of weights after the diet program

Hypotheses:

H_o: $\mu_1 = \mu_2$ (H_o: the diet program is not effective)

$H_A: \mu_1 \neq \mu_2$ (H_A: the diet program is effective)

Equivalently,

$H_o: \mu_D = 0$

$H_A: \mu_D \neq 0$ (where: $\mu_D = \mu_1 - \mu_2$)

We have found:

$$\overline{D} = 5.45 \quad , \quad S_D^2 = 50.3283 \quad , \quad S_D = \sqrt{S_D^2} = 7.09$$

Degrees of freedom:

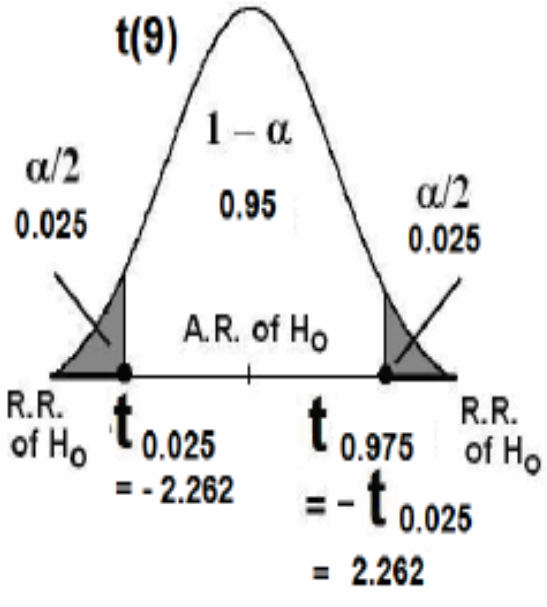
$$df = \nu = n-1 = 10-1 = 9$$

Significance level: $\alpha = 0.05$

Rejection Region of H_o:

Critical values: $t_{0.025} = -2.262$ and $t_{0.975} = -t_{0.025} = 2.262$

Critical Region: $t < -2.262$ or $t > 2.262$



Decision:

Since t= 2.43 ∈R.R., i.e., t=2.43 > $t_{0.975}$=–$t_{0.025}$=2.262, we reject:

H_o: $\mu_1 = \mu_2$ (the diet program is not effective)

and we accept:

H_1: $\mu_1 \neq \mu_2$ (the diet program is effective)

Consequently, we conclude that the diet program is effective at α=0.05.

7.5 Hypothesis Testing A single population proportion:

• Testing hypothesis about population proportion (P) is carried out in much the same way as for mean when condition is necessary for using normal curve are met

• We have the following steps:

1.Data: sample size (n), sample proportion($\hat{p}$) , P_0

$$\hat{p} = \frac{\text{no.of element in thesample with some charachtaristic}}{\text{Total no.of element in thesample}} = \frac{a}{n}$$

2. Assumptions :normal distribution **,**

- 3.Hypotheses:
- we have three cases
- Case I : H_0: P = P_0

  H_A: P ≠ P_0
- Case II : H_0: P = P_0

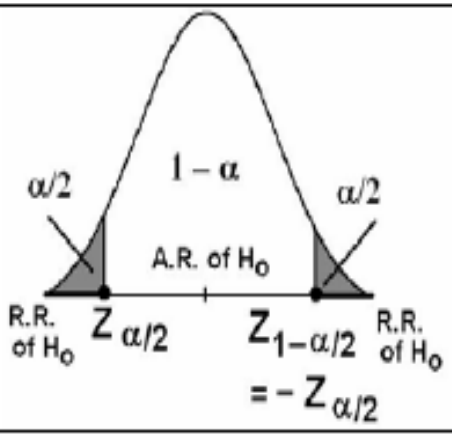
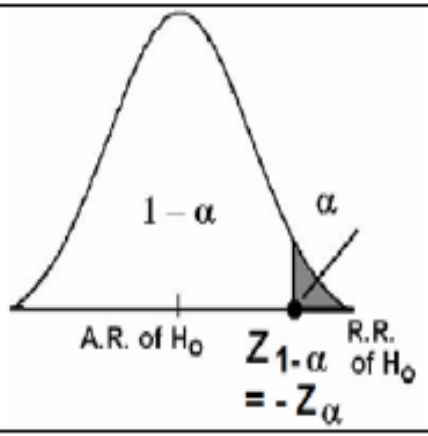
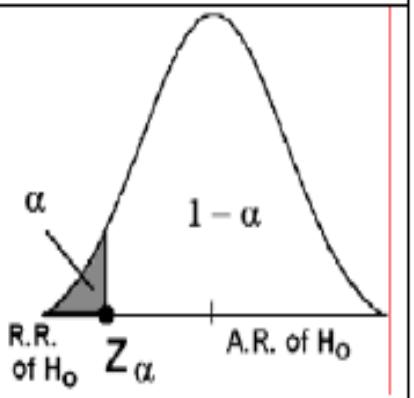
  H_A: P > P_0
- Case III : H_0: P = P_0

  H_A: P < P_0

4.Test Statistic:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

Where H_0 is true ,is distributed approximately as the standard normal

Hypotheses	$H_o: p = p_o$ $H_A: p \neq p_o$	$H_o: p \leq p_o$ $H_A: p > p_o$	$H_o: p \geq p_o$ $H_A: p < p_o$
Test Statistic (T.S.)	$Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \sim N(0,1)$		
R.R. & A.R. of H_o			
Decision:	Reject H_o (and accept H_A) at the significance level α if:		
	$Z < Z_{\alpha/2}$ or $Z > Z_{1-\alpha/2} = -Z_{\alpha/2}$ Two-Sided Test	$Z > Z_{1-\alpha} = -Z_{\alpha}$ One-Sided Test	$Z < Z_{\alpha}$ One-Sided Test

Example7.5.1 page 259

Wagen collected data on a sample of 301 Hispanic women Living in Texas .One variable of interest was the percentage of subjects with impaired fasting glucose (IFG). In the study,24 women were classified in the (IFG) stage .The article cites population estimates for (IFG) among Hispanic women in Texas as 6.3 percent .Is there sufficient evidence to indicate that the population Hispanic women in Texas has a prevalence of IFG higher than 6.3 percent ,let α=0.05

$\hat{p} = \frac{a}{n} = \frac{24}{301} = 0.08$

Solution:

1.Data: n = 301, p_0 = 6.3/100=0.063 ,a=24,

q_0 =1- p_0 = 1- 0.063 =0.937, α=0.05

2. Assumptions : $\hat{p}$is approximately normaly distributed

3.Hypotheses:

- we have three cases
- H_0: P = 0.063

  H_A: P > 0.063
- 4.Test Statistic :

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.08 - 0.063}{\sqrt{\frac{0.063(0.937)}{301}}} = 1.21$$

5.Decision Rule: Reject H_0 if $Z > Z_{1-\alpha}$

Where $Z_{1-\alpha} = Z_{1-0.05} = Z_{0.95} = 1.645$

6. Conclusion: Fail to reject H_0

Since

$$Z = 1.21 > Z_{1-\alpha} = 1.645$$

Or ,

If P-value = 0.1131,

fail to reject $H_0 \rightarrow P > \alpha$

- Exercises:
- Questions : Page 234 -237
- 7.2.1,7.8.2 ,7.3.1,7.3.6 ,7.5.2 ,,7.6.1

- H.W:
- 7.2.8,7.2.9, 7.2.11, 7.2.15,7.3.7,7.3.8,7.3.10
- 7.5.3,7.6.4

Exercises

Q7.5.2:

In an article in the journal Health and Place, found that among 2428 boys aged from 7 to 12 years, 461 were over weight or obese. On the basis of this study ,can we conclude that more than 15 percent of boys aged from 7 to 12 years in the sampled population are over weight or obese?

Let $\alpha = 0.1$

Solution :

1.Data :

2. Assumption :

3. Hypothesis :

4.Test statistic :

5.Decision Rule

6. Decision :

7.6 Hypothesis Testing :The Difference between two population proportion:

- Testing hypothesis about two population proportion ($P_{1,,}\ P_2$) is carried out in much the same way as for difference between two means when condition is necessary for using normal curve are met
- We have the following steps:

1.Data: sample size (n_1 ,n_2), sample proportions($\hat{P}_1, \hat{P}_2$), Characteristic in two samples (x_1 , x_2),

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

2- Assumption : Two populations are independent .

- 3.Hypotheses:
- we have three cases
- Case I : H_0: $P_1 = P_2 \rightarrow P_1 - P_2 = 0$

  H_A: $P_1 \neq P_2 \rightarrow P_1 - P_2 \neq 0$
- Case II : H_0: $P_1 = P_2 \rightarrow P_1 - P_2 = 0$

  H_A: $P_1 > P_2 \rightarrow P_1 - P_2 > 0$
- Case III : H_0: $P_1 = P_2 \rightarrow P_1 - P_2 = 0$

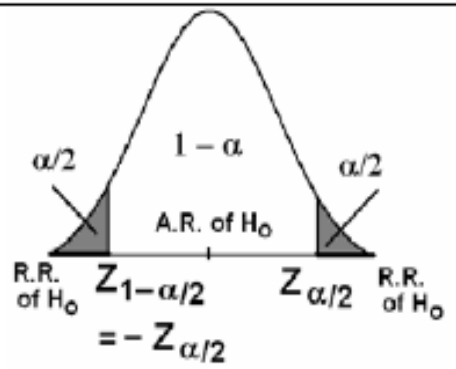
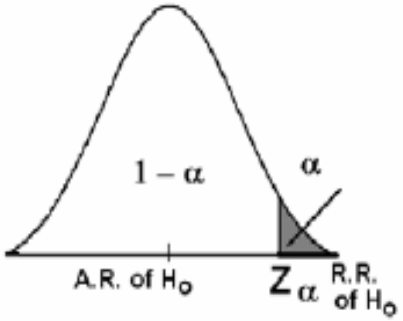
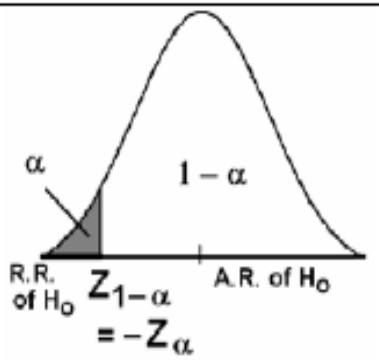
  H_A: $P_1 < P_2 \rightarrow P_1 - P_2 < 0$

4.Test Statistic:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}$$

Where H_0 is true ,is distributed approximately as the standard normal

Testing Procedure:

Hypotheses	$H_o: p_1 - p_2 = 0$ $H_A: p_1 - p_2 \neq 0$	$H_o: p_1 - p_2 \leq 0$ $H_A: p_1 - p_2 > 0$	$H_o: p_1 - p_2 \geq 0$ $H_A: p_1 - p_2 < 0$
Test Statistic (T.S.)	$Z = \dfrac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n_1} + \dfrac{\bar{p}(1-\bar{p})}{n_2}}} \sim N(0,1)$		
R.R. and A.R. of H_o			
Decision:	Reject H_o (and accept H_1) at the significance level α if $Z \in$ R.R.:		
Critical Values	$Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$ Two-Sided Test	$Z > Z_{\alpha}$ One-Sided Test	$Z < -Z_{\alpha}$ One-Sided Test

Example7.6.1 page 262

Noonan is a genetic condition that can affect the heart growth, blood clotting and mental and physical development. Noonan examined the stature of men and women with Noonan. The study contained 29 Male and 44 female adults. One of the cut-off values used to assess stature was the third percentile of adult height .Eleven of the males fell below the third percentile of adult male height ,while 24 of the female fell below the third percentile of female adult height .Does this study provide sufficient evidence for us to conclude that among subjects with Noonan ,females are more likely than males to fall below the respective of adult height? Let α=0.05

Solution:

1.Data: $n_M = 29$, $n_F = 44$, $x_M = 11$, $x_F = 24$, $\alpha = 0.05$

$$\bar{p} = \frac{x_M + x_F}{n_M + n_F} = \frac{11+24}{29+44} = 0.479$$

$$\hat{p}_M = \frac{x_m}{n_M} = \frac{11}{29} = 0.379,\ \hat{p}_F = \frac{x_F}{n_F} = \frac{24}{44} = 0.545$$

2- Assumption : Two populations are independent .

3.Hypotheses:

- Case II : $H_0: P_F = P_M \rightarrow P_F - P_M = 0$

  $H_A: P_F > P_M \rightarrow P_F - P_M > 0$

- 4.Test Statistic:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{(0.545 - 0.379) - 0}{\sqrt{\frac{(0.479)(0.521)}{44} + \frac{(0.479)(0.521)}{29}}} = 1.39$$

5.Decision Rule:

Reject H_0 if $Z > Z_{1-\alpha}$, Where $Z_{1-\alpha} = Z_{1-0.05} = Z_{0.95} = 1.645$

6. Conclusion: Fail to reject H_0

Since $Z = 1.39 > Z_{1-\alpha=} 1.645$

Or , If P-value = 0.0823 → fail to reject H_0 → $P > \alpha$

- Exercises:
- Questions : Page 234 -237
- 7.2.1,7.8.2 ,7.3.1,7.3.6 ,7.5.2 ,,7.6.1

- H.W:
- 7.2.8,7.2.9, 7.2.11, 7.2.15,7.3.7,7.3.8,7.3.10
- 7.5.3,7.6.4

Chapter 9
Statistical Inference and The Relationship between two variables

Prepared By : Dr. Shuhrat Khan

REGRESSION
CORRELATION
ANALYSIS OF VARIANCE

EQUATION OF REGRESSION

- Regression, Correlation and Analysis of Covariance are all statistical techniques that use the idea that one variable say, may be related to one or more variables through an equation. Here we consider the relationship of two variables only in a linear form, which is called linear regression and linear correlation; or simple regression and correlation. The relationships between more than two variables, called multiple regression and correlation will be considered later.
- Simple regression uses the relationship between the two variables to obtain information about one variable by knowing the values of the other. The equation showing this type of relationship is called simple linear regression equation. The related method of correlation is used to measure how strong the relationship is between the two variables is.

119

Line of Regression

DEPENDENT VARIABLE

INDEPENDENT VARIABLE

TWO RANDOM VARIABLE

OR

BIVARIATE

RANDOM

VARIABLE

- **Simple Linear Regression:**
- Suppose that we are interested in a variable Y, but we want to know about its relationship to another variable X or we want to use X to predict (or estimate) the value of Y that might be obtained without actually measuring it, provided the relationship between the two can be expressed by a line.' X' is usually called the **independent variable** and 'Y' is called the **dependent variable.**
-
- We assume that the values of variable X are either fixed or random. By fixed, we mean that the values are chosen by researcher--- either an experimental unit (patient) is given this value of X (such as the dosage of drug or a unit (patient) is chosen which is known to have this value of X.
- By random, we mean that units (patients) are chosen at random from all the possible units,, and both variables X and Y are measured.
- We also assume that for each value of x of X, there is a whole range or population of possible Y values and that the mean of the Y population at X = x, denoted by $\mu_{y/x}$, is a linear function of x. That is,
-
- $\mu_{y/x} = \alpha + \beta x$

ESTIMATION

We select a sample of n observations (x_i, y_i) from the population, WITH the goals

- Estimate α and β.
- Predict the value of Y at a given value x of X.
- Make tests to draw conclusions about the model and its usefulness.

- We estimate the parameters α and β by ‘a’ and ‘b’ respectively by using sample regression line:
- $\hat{Y} = a + bx$
- Where we calculate
-

ESTIMATION AND CALCULATION OF CONSTANTS , "a" AND "b"

$$a = \bar{y} - b\bar{x}$$

$$B = \frac{(\sum xiyi - n\bar{x}\,\bar{y})}{(\sum x_i{}^2 - n\bar{x}^2)}$$

EXAMPLE

- investigators at a sports health centre are interested in the relationship between oxygen consumption and exercise time in athletes recovering from injury. Appropriate mechanics for exercising and measuring oxygen consumption are set up, and the results are presented below:
  - x variable

exercise time (min)	y variable oxygen consumption
0.5	620
1.0	630
1.5	800
2.0	840
2.5	840
3.0	870
3.5	1010
4.0	940
4.5	950
5.0	1130

calculations

- $\overline{x} = 2.75 \qquad \bar{y} = 863 \qquad N = 10$

  $\sum x = 27.5 \qquad \sum y = 8630 \qquad (\sum x)^2 = 756.25 \qquad (\sum y)^2 = 74476900 \qquad \sum xy = 25750$

  $\sum x^2 = 96.25 \qquad \sum y^2 = 7672500$

$$b = \frac{(25750 - 10\text{x}2.75\text{x}863)}{(96.25 - 10\text{x}2.75^2)} = 97.82$$

$a = \bar{y} - b\bar{x}$ or $a = 863 - (97.82\,X2.75) = 594$

$\hat{y}\ for\ given\ x = 2.8 = 594 + (97.82 \text{ x } 2.8) = 868$ units

Pearson's Correlation Coefficient

- With the aid of Pearson's correlation coefficient (*r*), we can determine the strength and the direction of the relationship between *X* and *Y* variables,
- both of which have been measured and they must be quantitative.
- For example, we might be interested in examining the association between height and weight for the following sample of eight children:

Height and weights of 8 children

Child	Height(inches)X	Weight(pounds)Y
A	49	81
B	50	88
C	53	87
D	55	99
E	60	91
F	55	89
G	60	95
H	50	90
Average	(= 54 inches)	(= 90 pounds)

Scatter plot for 8 babies

height	weight
49	81
50	88
53	
55	
60	
55	
60	
50	

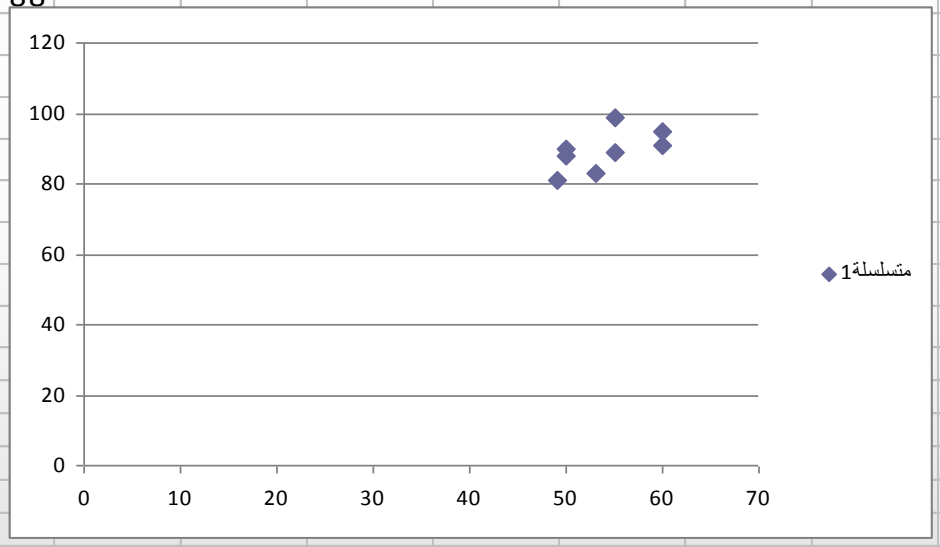


Table : The Strength of a Correlation

Value of r (positive or negative)	Meaning
0.00 to 0.19	A very weak correlation
0.20 to 0.39	A weak correlation
0.40 to 0.69	A modest correlation
0.70 to 0.89	A strong correlation
0.90 to 1.00	A very strong correlation

FORMULA FOR CORRELATION COEFFECIENT (r)

$$r = (\sum(X - \bar{x})(Y - \bar{y}) / \sqrt{(\sum(X - \bar{x})^2 \sum(Y - \bar{y})^2)}$$

- With Pearson's *r*, $\sum(X - \bar{X})(Y - \bar{Y})$
- means that we add the products of the deviations to see if the positive products or negative products are more abundant and sizable. Positive products indicate cases in which the variables go in the same direction (that is, both taller or heavier than average or both shorter and lighter than average);
- negative products indicate cases in which the variables go in opposite directions (that is, taller but lighter than average or shorter but heavier than average).
-

Computational Formula for Pearsons's Correlation Coefficient r

Where SP (sum of the product), SSx (Sum of the squares for x) and SSy (sum of the **squares for y) can be computed as follows:**

$$SP = \sum XY - N\bar{X}\bar{Y} = (\sum(X - \bar{x})(Y - \bar{y})$$

$$\text{SSx} = \sum X^2 - N\bar{X}^2 = (\sum(X - \bar{x})^2$$

$$\text{SSy} = \sum Y^2 - \text{N}\,\bar{Y}^2 = \sum(Y - \bar{y})^2$$

$$r = \frac{SP}{\sqrt{SSx \times SSy}} = \frac{(\sum(X - \bar{x})(Y - \bar{y})}{\sqrt{SSx \times SSy}}$$

$$= \frac{\sum XY - N\bar{X}\bar{Y}}{\sqrt{(\sum X^2 - N\bar{X}^2)(\sum Y^2 - N\,\bar{Y}^2)}}$$

$$r = \frac{\sum XY - N\bar{X}\bar{Y}}{\sqrt{(\sum X^2 - N\bar{X}^2)(\sum Y^2 - N\bar{Y}^2)}} = \frac{981 - 8(10.5)(11.5)}{\sqrt{[946 - (10.5)^2][1118 - (11.5)^2]}}$$

$$\frac{15}{\sqrt{(64)(60)}} = (15)/(61.97) = +.24$$

	XY	Y^2	X^2	Y	X	Child
	144	144	12	144	A	12
	80	64	100	8	10	B
	72	144	36	12	6	C
	176	121	256	11	16	D
	80	64	100	10	8	E
	72	64	81	8	9	F
	192	256	144	16	12	G
	165	225	121	15	11	H
	∑	**84**	**92**	**946**	**1118**	**981**

Table 2 : Chest circumference and Birth Weight of 10 babies

X(cm)	y(kg)	x^2	y^2	xy
22.4	2.00	501.76	4.00	44.8
27.5	2.25	756.25	5.06	61.88
28.5	2.10	812.25	4.41	59.85
28.5	2.35	812.25	5.52	66.98
29.4	2.45	864.36	6.00	72.03
29.4	2.50	864.36	6.25	73.5
30.5	2.80	930.25	7.84	85.4
32.0	2.80	1024.0	7.84	89.6
31.4	2.55	985.96	6.50	80.07
32.5	3.00	1056.25	9.00	97.5
TOTAL				
292.1	24.8	8607.69	62.42	731.61

Checking for significance

$$r = \frac{71.92}{\sqrt{754.45 \times 9.16}} = 0.86$$

- There appears to be a strong between chest circumference and birth weight in babies.
- We need to check that such a correlation is unlikely to have arisen by in a sample of ten babies.
- Tables are available that gives the significant values of this correlation ratio at two probability levels.
- First we need to work out degrees of freedom. They are the number of pair of observations less two, that is (n – 2)= 8.
- Looking at the table we find that our calculated value of 0.86 exceeds the tabulated value at 8 df of 0.765 at p= 0.01. Our correlation is therefore statistically highly significant.

Chapter 12

Analysis of Frequency Data

An Introduction to the Chi-Square Distribution

Prepared By : Dr. Shuhrat Khan

TESTS OF INDEPENDENCE

- To test whether two criteria of classification are independent . For example socioeconomic status and area of residence of people in a city are independent.
- We divide our sample according to status, low, medium and high incomes etc. and the same samples is categorized according to urban, rural or suburban and slums etc.
- Put the first criterion in columns equal in number to classification of 1^{st} criteria (Socioeconomic status) and the 2^{nd} in rows, where the no. of rows equal to the no. of categories of 2^{nd} criteria (areas of cities).

The Contingency Table

- Table Two-Way Classification of sample

First Criterion of Classification →

Second Criterion ↓	1	2	3	…..	c	Total
1	N_{11}	N_{12}	N_{13}	……	N_{1c}	$N_{1.}$
2	N_{21}	N_{22}	N_{23}	……	N_{2c}	$N_{2.}$
3	N_{31}	N_{32}	N_{33}	……	N_{3c}	$N_{3.}$
.	.	.	.	……	.	.
.	.	.	.		.	.
r	N_{r1}	N_{r2}	N_{r3}		N_{rc}	$N_{r.}$
Total	$N_{.1}$	$N_{.2}$	$N_{.3}$	……	$N_{.c}$	N

Observed versus Expected Frequencies

- O_{ij} : The frequencies in ith row and jth column given in any contingency table are called observed frequencies that result form the cross classification according to the two classifications.
- e_{ij} :Expected frequencies on the assumption of independence of two criterion are calculated by multiplying the marginal totals of any cell and then dividing by total frequency
- Formula:

$$e_{ij} = \frac{(N_{i\bullet}(N_{\bullet j})}{N}$$

Chi-square Test

- After the calculations of expected frequency,

  Prepare a table for expected frequencies and use Chi-square

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(o_i - e_i)^2}{e_i} \right]$$

  Where summation is for all values of r xc = k cells.

- D.F.: the degrees of freedom for using the table are (r-1)(c-1) for α level of significance
- Note that the test is always one-sided.

Example 12.401(page 613)

The researcher are interested to determine that preconception use of folic acid and race are independent. The data is:

Observed Frequencies Table

	Use of Folic Acid		total
	Yes	No	
White	260	299	559
Black	15	41	56
Other	7	14	21
Total	282	354	636

Text Book

Expected frequencies Table

	Yes	no	Total
White	(282)(559)/636 =247.86	(354)(559)/636 =311.14	559
Black	(282)(56)/636 =24.83	(354)(559) = 31.17	56
Others	(282)((21) =9.31	21x354/636 =11.69	21

Calculations and Testing

- Data: See the given table
- Assumption: Simple random sample
- Hypothesis: H_0: race and use of folic acid are independent

  HA: the two variables are not independent. Let $\alpha = 0.05$
- The test statistic is Chi Square given earlier
- Distribution when H_0 is true chi-square is valid with (r-1)(c-1) = (3-1)(2-1)= 2 d.f.
- Decision Rule: Reject H0 if value of χ^2 is greater than

$$\chi^2_{\alpha,(r-1)(c-1)} = 5.991$$

- Calculations

$$\chi^2 = (260-247.86)^2/247.86 + (299-311.14)^2/311.14 + \ldots\ldots + (14-11.69)^2/11.69 = \underline{9.091}$$

Conclusion

- Statistical decision. We reject H_0 since 9.08960> 5.991

- Conclusion: we conclude that H_0 is false, and that there is a relationship between race and preconception use of folic acid.
- P value. Since 7.378< 9.08960< 9.210, 0.01<p <0.025
- We also reject the hypothesis at 0.025 level of significance but do not reject it at 0.01 level.
- Solve Ex12.4.1 and 12.4.5 (p 620 & P 622)

ODDS RATIO

- In a retrospective study, samples are selected from those who have the disease called '*cases'* and those who do not have the disease called *'controls'* . The investigator looks back (have a *retrospective look)* at the subjects and determines which one have (or had) and which one do not have (or did not have) the risk factor.
- The data is classified into 2x2 table, for comparing cases and controls for risk factor *ODDS RATIO* IS CALCULATED
- ODDS are defined to be the ratio of probability of success to the probability of failure.
- The estimate of population odds ratio is

$$\therefore \hat{OR} = \frac{a/b}{cld} = \frac{ad}{bc}$$

ODDS RATIO

- Where a, b, c and d are the numbers given in the following table:

Risk Factor ↓	Sample		Total
	Cases	Control	
Present	a	b	a + b
Absent	c	d	c + d
Total	a + c	b + d	

- We may construct 100(1-α)%CI for OR by formula:
-

$$R^{1\pm(Z_{\alpha/2}/\sqrt{X^2})}$$

Example 12.7.2 for Odds Ratio

- Example 12.5.7.2 page 640: Data relates to the obesity status of children aged 5-6 and the smoking status of their mothers during pregnancy
- Hence OR for table
- is :

Obesity status

$$OR = \frac{(64)(3496)}{(342)(68)} = 9.62$$

Smoking status(during Pregnancy)	cases	Non-cases	Total
Smoked throughout	64	342	406
Never smoked	68	3496	3564
Total	132	3838	3970

Confidence Interval for Odds Ratio

The (1-α) 100% Confidence Interval for Odds Ratio is:

$$\hat{OR}^{1\pm(z_\alpha/\sqrt{X^2})}$$

Where

$$X^2 = \frac{n(ad-bc)^2}{(a+c)(a+d)(b+c)(b+d)}$$

For Example 12.5.7.2 we have: a=64, b=342, c=68, d=3496 , therefore:

$$X^2 = 3970(64\times 3496 - 342\times 68)^2 \Big/ (132)(3838)(406)(3564) = 217.68$$

Its 95% CI is:

$$\hat{OR}^{1\pm(z_\alpha/\sqrt{X^2})} = 9.62^{1\pm(1.96/\sqrt{217.6831})}$$

- or (7.12, 13.00)

Interpretation of Example 12.7.2 Data

- The 95% confidence interval (7.12, 13.00)

  mean that we are 95% confident that the population odds ratio is somewhere between 7.12 and 13.00
- Since the interval does not contain 1, in fact contains values larger than one, we conclude that, in Pop. Obese children (cases) are more likely than non-obese children (non-cases) to have had a mother who smoked throughout the pregnancy.
- Solve Ex 12.7.4 (page 646)

Interpretation of ODDS RATIO

- The sample odds ratio provides an estimate of the relative risk of population in the case of a rare disease.
- The odds ratio can assume values between 0 to ∞.
- A value of 1 indicate no association between risk factor and disease status.
- A value greater than one indicates increased odds of having the disease among subjects in whom the risk factor is present.

Chapter 13
Special Techniques for use when population parameters and/or population distributions are unknoen

pages 683-689

Prepared By : Dr. Shuhrat Khan

NON-PARAMETRIC STATISTICS

- The t-test, z-test etc. were all parametric tests as they were based n the assumptions of normality or known variances.

- When we make no assumptions about the sample population or about the population parameters the tests are called non-parametric and *distribution-free*.

ADVANTAGES OF NON-PARAMETRIC STATISTICS

- Testing hypothesis about simple statements (not involving parametric values) e.g.

  The two criteria are independent (test for independence)

  The data fits well to a given distribution (goodness of fit test)
- Distribution Free: Non-parametric tests may be used when the form of the sampled population is unknown.
- Computationally easy
- Analysis possible for ranking or categorical data (data which is not based on measurement scale)

The Sign Test

- This test is used as an alternative to t-test, when normality assumption is not met
- The only assumption is that the distribution of the underlying variable (data) is continuous.
- Test focuses on median rather than mean.
- The test is based on signs, plus and minuses
- Test is used for one sample as well as for two samples

Example
(One Sample Sign Test)

Score of 10 mentally retarded girls

We wish to know
if Median of population is
different from 5.

Solution:

Data: is about scores of 10
mentally retarded girls

Girl	Score	Girl	Score
1	4	6	6
2	5	7	10
3	8	8	7
4	8	9	6
5	9	10	6

Assumption: The measurements are continuous variable.

Continued…….

- **Hypotheses:** H_0: The population median is 5

  H_A: The population median is not 5

  Let $\alpha = 0.05$

- **Test Statistic:** The test statistic for the sign test is either the observed number of plus signs or the observed number of minus signs. The nature of the alternative hypothesis determines which of these test statistics is appropriate. In a given test, any one of the following alternative hypotheses is possible:

  H_A: $P(+) > P(-)$ one-sided alternative

  H_A: $P(+) < P(-)$ one-sided alternative

  H_A: $P(+) \neq P(-)$ two-sided alternative

Continued…….

- If the alternative hypothesis is H_A: $P(+) > P(-)$ a sufficiently small number of minus signs causes rejection of H_0. The test statistic is the number of minus signs.
- If the alternative hypothesis is H_A: $P(+) < P(-)$ a sufficiently small number of plus signs causes rejection of H_0. The test statistic is the number of plus signs.
- If the alternative hypothesis is H_A: $P(+) \neq P(-)$ either a sufficiently small number of plus signs or a sufficiently small number of minus signs causes rejection of the null hypothesis. We may take as the test statistic the less frequently occurring sign.

Continued…….

- **Distribution of test statistic:** If we assign a plus sign to those scores that lie above the hypothesized median and a minus to those that fall below.

- **Decision R**

  For H_A: $P(+$ s signs is les

Girl	1	2	3	4	5	6	7	8	9	10
Score relative to median = 5	-	0	+	+	+	+	+	+	+	+

Continued…….

- For H_A: $P(+) > P(-)$ reject H_0 if, when H_0 if true, the probability of observing k or fewer minus signs is less than or equal to α.
- For H_A: $P(+) < P(-)$, reject H_0 if the probability of observing, when H_0 is true, k or fewer plus signs is equal to or less than α.
- For H_A: $P(+) \neq P(-)$, reject H_0 if (given that H_0 is true) the probability of obtaining a value of *k* as extreme as or more extreme than was actually computed is equal to or less than α/2.
- **Calculation of test statistic:** The probability of observing k or fewer minus signs when given a sample of size n and parameter *p* by evaluating the following expression:

$$P(X \leq k \mid n, p) = \sum_{x=0}^{k} C_x^n \; p^x \; q^{n-x}$$

Continued…….

For our example we would compute

$$C_0^9(0.5)^0(0.5)^{9-0} + C_1^9(0.5)^1(0.5)^{9-1}$$
$$= 0.00195 + 0.01758 = 0.0195$$

- **Statistical decision:** In Appendix Table B we find

  P (k ≤ 1 | 9, 0.5) = 0.0195

- **Conclusion:** Since 0.0195 is less than 0.025, we reject the null hypothesis and conclude that the median score is not 5.
- ***p* value:** The *p* value for this test is 2(0.0195) = 0.0390, because it is two-sided test.

SIGN TEST----Paired Data

This is used an alternative to t-test for paired observations, when the underlying assumptions of t test are not met.

Null Hypothesis to be tested the median difference is zero.

OR

$$P(X_i > Y_i) = P(Y_i > X_i)$$

Subtract Yi from Xi , if Yi is less than Xi , the sign of the difference is (+), if Yi is greater than Xi , the sign of the difference is (-), so that

$$H_0 : P(+) = P(-) = 0.5$$

TEST STATISTIC: As before is k, the no of least occurring of Plus or minus signs.

SIGN TEST----Example 13.3.2

A dental research team matched 12 pairs of 24 patients in age, sex, intelligence. Six months later random evaluation showed the following score (low score score is higher level of hygiene)

pair no.	1	2	3	4	5	6	7	8	9	10	11	12	
instructed	1.5	2.0	3.5	3.0	3.5	2.5	2.0	1.5	1.5	2.0	3.0	2.0	
Not instructed	2.0	2.0	4.0	2.5	4.0	3.0	3.5	3.0	2.5	2.5	2.5	2.5	
Difference	-	0	-	+	-	-	-	-	-	-	+	-	

1. **Data: scores of dental hygiene, one member instructed how to brush and other remained uninstructed.**
2. **Assumption: the variable of dist is continues**
3. **H_o : The median of the difference is zero** [P(+) =P(-)]

   H_A : The median of the difference is negative

   [P(+) <P(-)]

Continued…….

Let α be 0.05

4. Test Statistic: The test statistic is the number of plus signs which occurs less frequent. i.e. k = 2

5. Distribution of k is binomial with n= 11 (as one observation is discarded) and p= 0.5

6. Decision Rule: Reject H_0 if P(k≤2| 11,0.5) ≤ 0.05.

7. Calculations:

P(k≤2/11,0.5)=

$$\sum_{k=0}^{2} \binom{11}{k} (0.5)^{k} (0.5)^{11-k}$$

Table B or calculations show the probability is equal to 0.0327 which is less than 0.05, we

must reject H_0 .

8. Conclusion: median difference is negative and instructions are beneficial

9. p value: Since it is one sided test the p-value is p= .0327

NON-PARAMETRIC STATISTICS

- The t-test, z-test etc. were all parametric tests as they were based n the assumptions of normality or known variances.

- When we make no assumptions about the sample population or about the population parameters the tests are called non-parametric and *distribution-free*.

EXAMPLE 1

Cardiac output (liters/minute) was measured by thermodilution in a simple random sample of 15 postcardiac surgical patients in the left lateral position. The results were as follows:

4.91	**4.10**	**6.74**	**7.27**	**7.42**	**7.50**	**6.56**	**4.64**
5.98	**3.14**	**3.23**	**5.80**	**6.17**	**5.39**	**5.77**	

We wish to know if we can conclude on the basis of these data that the population mean is different from 5.05.

Solution:

1. **Data.** As given above

2. **Assumptions.** We assume that the requirements for the application of the Wilcoxon signed-ranks test are met.

3. **Hypothesis.**

$H_0: \mu = 5.05$

$H_A: \mu \neq 5.05$

Let $\alpha = 0.05$.

EXAMPLE 1

4. **Test Statistic.** The test statistic will be $T+$ or $T-$, whichever is smaller, called the test statistic T.

5. **Distribution of test statistic.** Critical values of the test statistic are given in Table K of the Appendix.

6. **Decision rule**. We will reject H_0 if the computed value of T is less than or equal to 25, the critical value $n = 15$, and $\alpha/2 = 0.0240$, the closest value to 0.0250 in Table K.

7. **Calculation of test statistic.** The calculation of the test statistic is shown in Table.

8. **Statistical decision.** Since 34 is greater than 25, we are unable to reject H_0.

Cardiac output	$d_i = x_i - 5.05$	Rank of $\lvert d_i \rvert$	Signed Rank of $\lvert d_i \rvert$
4.91	-0.14	1	-1
4.10	-0.95	7	-7
6.74	+1.69	10	+10
7.27	+2.22	13	+13
7.42	+2.37	14	+14
7.50	+2.45	15	+15
6.56	+1.51	9	+9
4.64	-0.41	3	-3
5.98	+0.93	6	+6
3.14	-1.91	12	-12
3.23	-1.82	11	-11
5.80	+0.75	5	+5
6.17	+1.12	8	+8
5.39	+0.34	2	+2
5.77	+0.72	4	+4
			$T+ = 86$, $T- = 34$, $T = 34$

EXAMPLE 1

8. **Statistical decision.** Since 34 is greater than 25, we are unable to reject H_0.

9. **Conclusion.** We conclude that the population mean may be 5.05

10. ***p* value.** From Table K we see that the p value is p = 2(0.0757) = 0.1514

EXAMPLE 2

A researcher designed an experiment to assess the effects of prolonged inhalation of cadmium oxide. Fifteen laboratory animals served as experimental subjects, while 10 similar animals served as controls. The variable of interest was hemoglobin level following the experiment. The results are shown in Table 2.

We wish to know if we can conclude that prolonged inhalation of cadmium oxide reduces hemoglobin level.

EXAMPLE 2

TABLE 2. HEMOGLOBIN DETERMINATIONS (GRAMS) FOR 25 LABORATORY ANIMALS

EXPOSED ANIMALS (X)	UNEXPOSED ANIMALS (Y)
14.4	17.4
14.2	16.2
13.8	17.1
16.5	17.5
14.1	15.0
16.6	16.0
15.9	16.9
15.6	15.0
14.1	16.3
15.3	16.8
15.7	
16.7	
13.7	
15.3	

EXAMPLE 2

Solution:

1. **Data.** See table above

2. **Assumptions.** We presume that the assumptions of the Mann-Whitney test are met.

3. **Hypothesis.**

$$H_0: M_x \geq M_y$$

$$H_A: M_x < M_y$$

where M_x is the median of a population of animals exposed to cadmium oxide and M_y is the median of a population of animals not exposed to the substance. Suppose we let $\alpha = 0.05$.

EXAMPLE 2

4. **Test Statistic.** The test statistic is

$$T = S - \frac{n(n+1)}{2}$$

where n is the number of sample X observations and S is the sum of the ranks assigned to the sample observations from the population of X values. The choice of which sample's values we label as X is arbitrary.

X	13.7	13.8	14.0	14.1	14.1	14.2	14.4			15.3	15.3	15.6
Rank	1	2	3	4.5	4.5	6	7			10.5	10.5	12
Y								15.0	15.0			
Rank								8.5	8.5			

X	15.7	15.9				16.5	16.6	16.7					
Rank	13	14				18.	19	20					
Y			16.0	16.2	16.3				16.8	16.9	17.1	17.4	17.5
Rank			15	16	17				21	22	23	24	25

Sum of the *Y* ranks = *S* = 145

TABLE 2. ORIGINAL DATA AND RANKS

EXAMPLE 2

5. **Distribution of test statistic.** The critical values are given in Table K.

6. **Decision Rule.** Reject H_0: $M_x \geq M_y$, if the computed *T* is less than w_α with n, the number of X observations; m the number of Y observations and α, the chosen level of significance.

If the null hypothesis were of the types

$$H_0: M_x \leq M_y$$

$$H_A: M_x > M_y$$

Reject H_0: $M_x \leq M_y$ if the computed *T* is greater than $w_{1-\alpha}$, where $W_{1-\alpha} = nm - W_\alpha$.

EXAMPLE 2

For the two-sided test situation with

$$H_0: M_x = M_y$$

$$H_A: M_x \neq M_y$$

Reject $H_0: M_x = M_y$ if the computed value of *T* is either less than $w_{\alpha/2}$ or greater than $w_{1-\alpha/2}$, where $w_{\alpha/2}$ is the critical value of *T* for *n, m* and *α/2* given in Appendix II Table K and $w_{1-\alpha/2} = nm - w_{\alpha/2}$.

For this example the decision rule of *T* is smaller than 45, the critical value of the test statistic for $n = 15$, $m = 10$, and $\alpha = 0.05$ found in Table K.

EXAMPLE 2

7. **Calculation of test statistic.** We have $S = 145$, so that

$$T = 145 - \frac{15(15+1)}{2} = 25$$

8. **Statistical Decision**. When we enter Table K with $n = 15$, $m = 10$, and $\alpha = 0.05$, we find the critical value of $w_{1-\alpha}$ to be 45. Since 25 is less than 45, we reject H_0.

9. **Conclusion**. We conclude that M_X is smaller than M_Y. This leads us to the conclusion that prolonged inhalation of cadmium oxide does reduce the hemoglobin level.

Since $22 < 25 < 30$, we have for this test

$0.005 > p > 0.001$.

EXAMPLE 2

When either *n* or *m* is greater than 20 we cannot use Appendix Table K to obtain critical values for the Mann-Whitney test. When this is the case we may compute

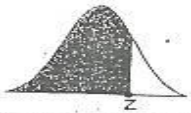
$$z = \frac{T - mn/2}{\sqrt{nm(n+m+1)/12}}$$

And compare the result, for significance, with critical values of the standard normal distribution.

Table B: Quantiles of Student's t-distributions with given degrees of freedom for selected cumulative probabilities.

df	Cumulative probability				
	0.90	0.95	0.975	0.99	0.995
1	3.078	6.3138	12.706	31.821	63.657
2	1.886	2.9200	4.3027	6.965	9.9248
3	1.638	2.3534	3.1825	4.541	5.8409
4	1.533	2.1318	2.7764	3.747	4.6041
5	1.476	2.0150	2.5706	3.365	4.0321
6	1.440	1.9432	2.4469	3.143	3.7074
7	1.415	1.8946	2.3646	2.998	3.4995
8	1.397	1.8595	2.3060	2.896	3.3554
9	1.383	1.8331	2.2622	2.821	3.2498
10	1.372	1.8125	2.2281	2.764	3.1693
11	1.363	1.7959	2.2010	2.718	3.1058
12	1.356	1.7823	2.1788	2.681	3.0545
13	1.350	1.7709	2.1604	2.650	3.0123
14	1.345	1.7613	2.1448	2.624	2.9768
15	1.341	1.7530	2.1315	2.602	2.9467
16	1.337	1.7459	2.1199	2.583	2.9208
17	1.333	1.7396	2.1098	2.567	2.8982
18	1.330	1.7341	2.1009	2.552	2.8784
19	1.328	1.7291	2.0930	2.539	2.8609
20	1.325	1.7247	2.0860	2.528	2.8453
21	1.323	1.7207	2.0796	2.518	2.8314
22	1.321	1.7171	2.0739	2.508	2.8188
23	1.319	1.7139	2.0687	2.500	2.8073
24	1.318	1.7109	2.0639	2.492	2.7969
25	1.316	1.7081	2.0595	2.485	2.7874
26	1.315	1.7056	2.0555	2.479	2.7787
27	1.314	1.7033	2.0518	2.473	2.7707
28	1.313	1.7011	2.0484	2.467	2.7633
29	1.311	1.6991	2.0452	2.462	2.7564
30	1.310	1.6973	2.0423	2.457	2.7500
35	1.3062	1.6896	2.0301	2.438	2.7239
40	1.3031	1.6839	2.0211	2.423	2.7045
45	1.3007	1.6794	2.0141	2.412	2.6896
50	1.2987	1.6759	2.0086	2.403	2.6778
60	1.2959	1.6707	2.0003	2.390	2.6603
70	1.2938	1.6669	1.9945	2.381	2.6480
90	1.2910	1.6620	1.9867	2.368	2.6316
120	1.2887	1.6577	1.9799	2.358	2.6175
160	1.2869	1.6545	1.9749	2.350	2.6070
200	1.2858	1.6525	1.9719	2.345	2.6006
∞	1.282	1.645	1.96	2.326	2.576

Areas Under The Standard Normal Curve



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7996	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9462	0.9453	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9554	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898*	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998