BASIC STATISTICS

In this unit we learn the basic concepts of Statistics, need of Statistics, definition of data, raw data and grouped data, measures of central tendency consists mean, median mode and range and its calculations.

DEFINITION OF STATISTICS

"Statistics is the science of the measurement of social organism as a whole in all its manifestation" – A.L.Bowley

"Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data" – Croxton & Cowden

"Statistics is a body of methods for making decisions in the face of uncertainity" – Wallis and Roberts.

MEANING OF STATISTICS

Statistics is a branch of science methodology. It deals with the collection, classification, description and interpretation of data obtained by the conduct of surveys and experiments. Its essential purpose is to describe and draw inferences about the numerical properties of population

NEED AND IMPORTANCE OF STATISTICS IN RESEARCH

In general there are seven simple reasons why the students must take a course in statistics or develop some mastery on that subject.

1. To understand the literature

   One cannot read much of the literature in the physical education and sports without encountering statistical concepts method and techniques. Without proper understanding of the fundamentals of statistics it is very difficult to interpret and digest the results mentioned in the research articles, with a result very purpose of the article may not be clear to the students and he/she loses the interest in reading.

2. To fabricate the research problems

There is much difference between an abstract and rational thinking. A researcher needs to know in advance about his feasible hypothesis which he wishes to verify. He must be aware of proper statistical design and techniques which are possible to implement in his problem. Thus, it is not possible to fabricate the rationally correct research problems without proper appreciation of statistical concepts and techniques.

3. To develop scientific temper

   Statistics is essential part of any profession and training programme, so is true in case of physical education too, in order to facilitate the students to think rationally about any situation. During the match a player's decision for any move is purely based upon his scientific thinking. Infact a coach visualizes the training schedule of his trainee based upon his decision derived from his previous experiences; All this is possible only if one observes any fact justifiable as per the scientific arguments.

4. To asses the authenticity of research findings and to contradict the unjustifiable claims

   Several researchers publicly announce their findings based upon their research work. In order to assess the authenticity of their statement one can read their research report. But to prove the conformity between their statements and the actual fact one should e able to understand the statistical techniques used in the report for drawing conclusion. Many companies make a claim about their product by using research findings of the scientists. These claims may be tested by conducting an experiment under the required condition and analyzing the data. The knowledge of designing an experiment and various statistical techniques are essential to write off the unjustifiable claims.

5. To develop the indices on various characteristics and performances

   In order to assess an academic excellence of a student, his performance in the examination is used as an index. Similarly to assess the general fitness of a college student an index is required. Although several authors have suggested ways and means to find an index to measure various characteristics of an individual and for the measurement of performance on different events in sports, but it requires lot of statistical concepts to further improve the quality of such indices.

6. To develop norms on various traits

Performance on any trait like sit-ups, pull-ups and push-ups etc needs to be converted into a score by using a scale ranging from 0 to 100, to motivate a student. Conversion of such performance is known as norms. Such norms are easily understood by a common man and be used in the admission procedure of the students in schools and colleges.

7. To conduct research

Statistical concepts and techniques are important, in designing an experiment, in drawing a representative sample, in administering the test and in choosing the correct statistical test for conducting research and interpreting the results. Thus, it is extremely important to have an appropriate knowledge of statistical concepts, methods and techniques to conduct a research in an accurate manner for drawing the reliable conclusion.

Data

In almost all scientific researches, we interact with qualitative and quantitative scores. These scores need to be properly handled in a scientific way in order to draw inferences. The use of statistics has now spread throughout the field of science in general and thus incorporating the skill of data handling techniques is essential in the scientist's training.

Raw data

Any amount of qualitative or quantitative scores obtained as a result of measurement or due to an experiment are termed as raw scores.

GROUPED DATA

If the raw data is classified as class interval and frequency the data is called Grouped data.

FREQUENCY TABLE

One of the principal aims of statistics is to generalize population characteristics on the basis of sample observations. But if the sample observation is large it becomes unmanageable until and unless it is compiled. Quite a few information can be extracted from large amount of data when they are complied and presented by means of suitable tables.

MEASURES OF CENTRAL TENDENCY

Any statistical constant which describes an aggregate of a set of values is called a measure of central tendency. Broadly speaking measures of central tendency consists of three components viz., Mean, Median, Mode. Mean can further be subdivided into three types i.e., arithmetic mean, geometric mean and harmonic mean. Since we shall restrict our discussion to arithmetic mean only.

MEAN

The mean of a set of finite observations is the sum of all the observations divided by the total number of observations.

Further, if the mean is calculated for a finite population it is said to be population mean and if it is for a finite sample it is termed as sample mean.

Calculation of Mean from Raw Data

Mean = $\overline{X} = \frac{\sum X}{N}$, Where N is the number of observation

Example

Following are the chest circumference of eight athletes in cm: 31,36,34,33,38,37,40,39. Calculate mean chest circumference.

$$\overline{X} = \frac{\sum X}{N}$$

$$= \frac{31+36+34+33+38+37+40+39}{8}$$

$$= \frac{288}{8}$$

$$= 36$$

Hence the mean chest circumference of eight athletes is 36 cm.

Calculation of Mean from Grouped Data

Mean = $\overline{X} = \frac{\sum fx}{N}$, Here N= $\sum f$

Example

Consider the scores are arranged in class intervals, calculate mean

CI	Frequency (f)
51-55	3
46-50	6
41-45	4
36-40	5
31-35	1
26-30	3
21-25	2

Solution

CI	X	Frequency (f)	fx
51-56	53	3	159
46-51	48	6	288
41-46	43	4	172
36-41	38	5	190
31-36	33	1	33
26-31	28	3	84
21-25	23	2	46
	N= 24		$\sum fx = 972$

Mean = $\overline{X} = \frac{\sum fx}{N}$

$= \frac{972}{24}$

$= 40.5$

Deviation Method for calculation of Mean

In using the deviation method, computation of mean becomes easy. Here X is transformed into a new variable d such that

$d = \frac{x - A}{h}$, where A= Assumed mean

h = width of the consecutive class intervals.

Mean = $A + \frac{\sum fd}{h} \times h$

Example

Consider the above example

CI	X	Frequency (f)	$d = \frac{x - A}{h}$	fd
51-57	53	3	3	9
46-52	48	6	2	12
41-47	43	4	1	4
36-42	38	5	0	0
31-37	33	1	-1	-1
26-32	28	3	-2	-6
21-25	23	2	-3	-6
		N= 24		$\sum fd = 12$

Mean = $A + \frac{\sum fd}{h} \times h$

Assumed mean A=38, h = 5

= $38 + \frac{12}{24} \times 5$

=40.5

Median

Another useful measure of central tendency is median. The median is an average used in a situation when scores are in the form of ranks. It is a positional value above and below which 50% of the scores lies. It is denoted as M_d .

Definition

Median is a middle most core in a distribution such that half the observations fall above it and half below it.

Median = $M_d = \left(\frac{N+1}{2}\right)^{th} score$

Calculation with Raw data

Case 1: N is odd

Consider the following values of x

X: 6, 38, 33, 28, 21, 15, 23.

Ascending order : 6, 15, 21, 23, 28, 33, 38

Here N=7 Median = $M_d = \left(\frac{N+1}{2}\right)^{th} score$

$= \left(\frac{8}{2}\right)^{th} score$

$= 4^{th}$ score

= 23

Case 2 : N is even

X : 6, 33, 28, 21, 15, 23

Ascending order : 6, 15, 21, 23, 28, 33

Here N= 6 Median = $M_d = \left(\frac{N+1}{2}\right)^{th} score$

$= \left(\frac{7}{2}\right)^{th} score$

$= \frac{3rd\ score + 4th\ score}{2}$

$= \frac{21+23}{2}$

=22

Calculation with Grouped data

Median = $M_d = L + \frac{\frac{N}{2} - cf}{f} \times h$

Where L= Lower limit of the median class

cf= cumulative frequency of the class interval just above the median class

h= width of the class interval

Example

Consider the frequency distribution of fitness test scores

CI	Frequency (f)
45-49	3

40-44	6
35-39	4
30-34	2
25-29	5
20-24	3
15-19	7
10-14	6

Solution

CI	Frequency (f)	cf	
9.5-14.5	6	6	
14.5-19.5	7	13	
19.5-24.5	3	16	
24.5-29.5	5	21	Median class
29.5-34.5	2	23	
34.5-39.5	4	27	
39.5-44.5	6	33	
44.5-49.5	3	36	

Median = $M_d = L + \frac{\frac{N}{2} - cf}{f} \times h$

L= 24.5, N/2 = 18, cf = 16 f= 5

$$M_d = 24.5 + \frac{18-16}{5} \times 5$$

= 26.5

MODE

Mode is defined as a score which occurs maximum number of times in a distribution. It is represented by M_o.

Calculation with Raw data

A score which is repeated maximum number of times in the distribution.

Steps in calculating the mode

1. Select the value of x which is repeated max number of times in a distribution.
2. If two adjacent values of x are repeated equal number of times and are greater than the frequency of occurrence of the remaining values, then the mode shall be the average of the two such adjacent scores.
3. If two non adjacent values of x are such that the frequencies of both are greater than the frequencies of all other values, then each value of x may be taken as a mode and the distribution shall be known as bimodal.

Example

Calculate mode for the given raw data

10, 30, 80, 60, 20, 10

Mode =10

Calculation with Grouped data

Mode = $M_o = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$

L= the lower limit of the modal class

f_m= the frequency of the modal class

f_1= the frequency of the class just above the modal class

f_2= the frequency of the class just below the modal class

h= width of the class interval

Example

Consider the following frequency table

CI	f	
44-48	3	
39-43	6	Modal class
34-38	2	
29-33	5	
24-28	4	

$L = 38.5 \qquad f_m = 6 \quad f_1 = 2,\ f_2 = 3\ h = 5$

Mode = $M_o = L + \dfrac{f_m - f_1}{2f_m - f_1 - f_2} \times h$

Mode = $M_o = 38.5 + \dfrac{6-2}{12-2-3} \times 5$

= 41.36

RANGE

The range is the crudest measure of variability.

Range = Max. score – Min. score

The range gives us an idea about the maximum variation of scores in the distribution and thus it is used in constructing the class interval from the raw scores. Since it depends upon two scores, it is the most unstable measures of variability.

Example

Consider the following scores:

7,12,16,19,36,42,47,96.

Range =96-7=90.

Merits and Demerits of Range

1. It is simple to understand
2. It is easy to calculate
3. In certain types of problems like quality control, weather forecasts, share price analysis etc., range is most widely used.

Demerits

1. It is very much affected by the extreme items
2. It cannot be calculated from open end class intervals
3. It is based on only two extreme observations
4. It is not suitable for mathematical treatment
5. It is a very rarely used measure.

MEAN DEVIATION

Mean deviation is another measure of variability which depends upon the absolute deviation of scores from any average value. It is represented by MD. Mean deviation is the average of the absolute deviation of scores about any measure of central tendency or any constant value. If mean deviation is calculated about mean, it is called as mean deviation about mean and mean deviation can also calculated using mode and median.

For Raw data

$$MD = \frac{1}{N}\sum|x - \bar{x}|$$

Example

Calculate mean deviation for the following data

X: 36,28,32,22,16,24,21,10,12,9

Solution

x	$x-\bar{x}$	$\lvert x-\bar{x}\rvert$
36	15	15
28	7	7
32	11	11
22	1	1
16	-5	5
24	3	3
21	0	0
10	-11	11
12	-9	9
9	-12	12
$\sum x = 210$		$\sum \lvert x-\bar{x}\rvert = 74$

$$\bar{x} = \frac{210}{10} = 21$$

MD= $\frac{1}{N}\sum \lvert x-\bar{x}\rvert$

= $\frac{1}{10} \times 71 = 7.4$

For Grouped Data

MD = $\frac{1}{N}\sum f\lvert x-\bar{x}\rvert$

Example

Calculate mean deviation for the distribution of sit-ups scores

CI	f
46-50	3
41-45	2
36-40	6
31-35	4
26-30	9
21-25	3
16-20	6
11-15	2
6-10	4
1-5	3

Solution

CI	x	f	d	fd	$x-\bar{x}$	$f\lvert x-\bar{x}\rvert$
46-51	48	3	4	12	21	63
41-46	43	2	3	6	16	32
36-41	38	6	2	12	11	66
31-36	33	4	1	4	6	24
26-31	28	9	0	0	1	9
21-26	23	3	-1	-3	4	12
16-21	18	6	-2	-12	9	54
11-16	13	2	-3	-6	14	28
6-11	8	4	-4	-16	19	76
1-5	3	3	-5	-5	24	24
		N=40		$\sum fd=-8$		$\sum f\lvert x-\bar{x}\rvert=388$

Mean = $A + \frac{\sum fd}{h} \times h$

Assumed mean A=28 h = 5

$= 28 + \frac{-8}{40} \times 5$

=27

MD = $\frac{1}{N} \sum f|x - \bar{x}|$

$= \frac{1}{40} \times 388$

=9.7

Quartile Deviation

Quartile deviation measures the variability of middle 50% of scores about median. Quartile deviation would be an appropriate measure in a situation where median is the best measure of central tendency. It is represented by QD

$$QD = \frac{Q_3 - Q_1}{2}$$

Where Q_1 and Q_3 are first and third quartiles. Q_1 is the point in the distribution which has 25 percent of the sores below it and similarly $Q_3$is the point which has 75 percent of the scores below it.

Quartile deviation is a measure of variability taken about median.

Calculation of QD from Raw data

To obtain the $Q_1$and Q_3 we shall arrange the scores in ascending order.

$$Q_1 = \left(\frac{N+1}{4}\right)^{th} score$$

$$Q_3 = \left(\frac{3(N+1)}{4}\right)^{th} score$$

Consider the of measurements on X: 1,6,3,12,16,18,9,20.

After arranging the scores in ascending order

X: 1,3,6,9,12,16,18,20

Here N=8

$$Q_1 = \left(\frac{N+1}{4}\right)^{th} score$$

$$= \left(\frac{9}{4}\right)^{th} score = 2.25^{th} score$$

= 2nd score + ¼ (3rd score -2nd score)

=3.75

$$Q_3 = \left(\frac{3(N+1)}{4}\right)^{th} score$$

$$= \left(\frac{3(9)}{4}\right)^{th} score$$

$$= \left(\frac{27}{4}\right)^{th} score$$

= 6th score + 3/4(7th score -6th score)

= 17.5

Thus QD= $\frac{Q_3 - Q_1}{2} = \frac{17.5 - 3.75}{2} = 6.88$

Calculation of QD from Grouped Data

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times h$$

$$Q_3 = L + \frac{\frac{3N}{4} - cf}{f} \times h$$

Example

Calculate the QD for the distribution of scores in a cricket tournament

CI	f	cf	
91-100	1	41	
81-90	3	40	
71-80	4	37	
61-70	2	33	
51-60	6	31	Q_3 class
41-50	5	25	
31-40	8	20	
21-30	6	12	Q_1 class
11-20	4	6	
1-10	2	2	

Since N/4 = 41/4=10.25

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times h$$

$$= 20.5 + \frac{1025 - 6}{6} \times 10$$

= 27.58

$$Q_3 = L + \frac{\frac{3N}{4} - cf}{f} \times h$$

$$= 50.5 + \frac{30.75. - 25}{6} \times 10$$

= 60.08

Thus QD= $\frac{Q_3 - Q_1}{2} = \frac{60.08 - 27.58}{2} = 16.25$

STANDARD DEVIATION

Among all the measures of variability standard deviation is the most stable measure and thus it is widely used in many statistical operations. Standard deviation measures the variability of scores about the mean value in the distribution. Standard deviation is the positive square root of the average of the squared deviations of all the scores from their mean. It is represented by a Greek letter. σ is a symbol used for population standard deviation and 's' is used for sample standard deviation.

Computation of Standard Deviation from Raw Data

$$S = \sqrt{\frac{1}{N}\sum (x - \bar{x})^2}$$

Example

Calculate SD for the following data

X: 19,17,16,12,5,3,6,2

Solution

X	$(x-\bar{x})$	$(x-\bar{x})^2$
19	9	81
17	7	49
16	6	36
12	2	4
5	-5	25
3	-7	49
6	-4	16
2	-8	64
$\sum x = 80$		$\sum (x-\bar{x})^2 = 324$

Therefore N=8, $\bar{x} = \frac{\sum x}{N} = 10$

S= $\sqrt{\frac{1}{N}\sum (x-\bar{x})^2}$

= $\sqrt{\frac{1}{8} \times 324}$

= 6.36

Computation for SD from Grouped Data

S= $\sqrt{\frac{1}{N}\sum f(x-\bar{x})^2}$ OR S = $h\sqrt{\frac{1}{N}\sum fd^2 - \left(\frac{\sum fd}{N}\right)^2}$

Example

Compute SD for the scores on psychological test

CI	x	f	d	fd	fd^2
76-80	78	3	5	15	75
71-75	73	6	4	24	96
66-70	68	4	3	12	36
61-65	63	5	2	10	20
56-60	58	2	1	2	2
51-55	53	5	0	0	0
46-50	48	4	-1	-4	4
41-45	43	5	-2	-10	20
36-40	38	3	-3	-9	27
31-35	33	2	-4	-8	32
26-30	28	1	-5	-5	25
		N=40		$\sum fd = 27$	$\sum fd^2 = 337$

$$S = h\sqrt{\frac{1}{N}\sum fd^2 - \left(\frac{\sum fd}{N}\right)^2}$$

$$= 5\sqrt{\frac{1}{40} \times 337 - \left(\frac{27}{40}\right)^2}$$

=14.12