Introduction To Applied Statistics.

Reshmi dey
16/10/2021

Statistics defined.

- Statistics as numerical data.

"Statistics are classified facts representing the conditions of the people in a state, specially those facts which can be stated in number or in tables of numbers or in any tabular or classified arrangement." —Webster

"Statistics are numerical statement of facts in any department of enquity placed in relation to each other."
—Bowley

- Statistics as statistical methods.

"Statistics is the science of estimates and probabilities."
—Boddington

"Statistics is the branch of scientific method which deals with the data obtained by counting or measuring the properties of populations of population of natural phenomena." —Kendall

"Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data." —Croxton and Cowden

Types of Datas

Categorical Data or Qualitative Data types.

It means that this type of data can't be counted or measured easily using numbers and therefore divided into categories. The gender of a person (male, female, or others) is a good example of this data type.
There are two subcategories under this:
Nominal or unordered - Set of values that don't possess a natural ordering. The colour of a phone can be considered as a nominal data type as we can't compare one colour with others.The gender of a person is another one where we can't differentiate between male, female, or others.
Ordinal or Ordered - Set of values that possess a natural ordering. For example the size of clothes produced by certain clothing brand like small, medium, large, Xl, etc.

Types of Datas

Quantitative Data Type

Data types that can be counted or calculated using numbers. For examples number of students in a class, number of planets in the universe. There can be infinite numbers of values, for examples number of stars in the galaxy.

The two subcategories which describe them clearly are:

Discrete - The numerical values which fall under integers or whole numbers. For example shoe size, number of bones in a human body, etc

Continuous - The fractional numbers are considered as continuous values. For example height, weight, temperature, length, etc.

Averages or Measures of central Tendency

Central tendency is a single value in a data set which is representative of the entire distribution. It helps us to calculate an accurate description of the entire data.

The following are the five measures of central tendency :-

(i) Mean or Arithmetic mean,
(ii) Median,
(iii) Mode,
(iv) Geometric mean, and
(V) Harmonic mean.

Mean

Arithmetic mean of a set of observations is defined as the sum of observations divided by the number of observations.
The mean of n observations x1, x2,…, xn is given by

$\bar{X} = (x1 + x2 + \ldots + xn) / n$

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

In case of grouped distribution, where f is the frequency of the variable x, we use the formula

Mean of Grouped Data:

$$\bar{x} = \frac{\sum fx}{n}$$

where: $\bar{x}$ = *mean*

f = *frequency of each class*

x = *mid-interval value of each class*

n = *total frequency*

$\sum fx$ = *sum of the producst of mid − interval values and their corresponding frequency*

Caption

For example we have a data set {1,4,2,6,7,4,6,2} of random numbers, and we need to find the mean

We first will find the sum of the numbers in the set, 1 + 4 + 2 + 6 + 7 + 4 + 6 + 2 = 32

Now, number of observations here is 8

Therefore our mean is, $\bar{X} = 32 / 8 = 4$

Another example.

Q. Find the arithmetic mean of following frequency distribution :

x:	1	2	3	4	5	6	7
f:	5	9	12	17	14	10	6

Solution:-

X	f	fx
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	9	42
Total	73	299

$$\bar{X} = 299/\ 73 = 4.09$$

Q.

Marks :-	0-10	10-20	20-30	30-40	40-50	50-60
No. of students:-	12	18	27	20	17	6

Solution:-

Marks	No. of students (f)	Mid point(x)	fx
0-10	12	(0+10)/2 = 5	60
10-20	18	(10+20)/2 = 15	270
20-30	27	(20+30)/2 = 25	675
30-40	20	(30+40)/2 = 35	700
40-50	17	(40+50)/2 = 45	765
50-60	6	(50+60)/2 = 55	330
Total	100		2800

Mean, $\bar{X} = 2800/100 = 28.$

Median

The value of a variable that divides it into two equal parts. A median is a positional average, because the number of observations preceding and succeeding the median are equal.

A very important point for calculating median is that the data should be sorted either in ascending order or descending.

For even number of observations,

Median = [(n/2)th + (n+2/2)th]/2

For odd number of observations,

Median = [(n+1)/2]th observation

Q. Obtain the median of the following distribution

(i).	X :	1	2	3	5	6	8	9	10
(ii).	Y :	4	3	7	23	44	45	22	
(iii).	R:	102	288	345	55	56	67	45	66

Solution:-

(i). Here, n = 8 (even)

Median = [(8/2)th+ (8+2/2)th] /2= [(4th + 5th)]/2 = (5 + 6)/2 = 11/2 = 5.2

(ii). Here, n = 7(odd)

Sorted observation is 3, 4, 7, 22, 23, 44, 45

Median = (7+1)/2 th = 8/2 = 4th observation = 22

(iii). Here, n= 8(even)

Sorted observation is 45, 55, 56, 66, 67, 102, 288, 345

Median = [(8/2)th+ (8+2/2)th] /2= [(4th + 5th)]/2 = 66+67/2 = 66.5

Mode

Most frequently occurring value in a set , in other words mode is the value of the variable which is predominant in the series.

The mode of the data set {2, 3, 5, 7, 8, 2, 5, 6, 2} is 2 since it is occurring 3 times.

The mode of the data set {11, 56, 11, 44, 56, 33, 76, 67, 23,} is 11 and 56.

Mode for continuous frequency distribution

$$\text{Mode} = l + \frac{h(f1 - f0)}{(f1 - f0) - (f2 - f1)}$$

Here , l = lower limit of the modal class, h is the magnitude , f1 is the frequency of modal class, and f0 and f2 are the frequencies of class preceding and succeeding the modal class respectively.

Q. Find the mode for the following distribution

Class interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	5	8	7	12	28	20	10	10

Solution:-

Here the maximum frequency is 28. Hence the modal class is 40-50.

And l = 40 and h = 10

Mode = 40 + {10(28 - 12) / [(28 - 12) -(20-28)] }

= 40 + {160/24}

= 40 + 6.667 (approx)

= 46.667

Measures of positions

Values which divide the series into equal parts.

Quartiles - The three points that divide the series into four equal parts are called quartiles. The first or the lower quartile, Q1, is the value below 25% and the second, Q2, also the median which has 50% observations before and after it. And the third or upper quartile,Q3, has 75% of data before it .

Deciles - The nine points which divide the series into ten equal parts.

Percentiles - The ninety nine points which divide the series into 100 equal parts.

Location of a percentile = (n+1)(p/100), where is p is the value of desired percentile.

Interquartile Range

IQR = (Q3 - Q1)

Q. Find the interquartile range.
23, 24, 24, 25, 26, 26, 27, 28, 29, 31

Solution:- 23 24 24 25 26 | 26 27 28 29 31

Q1 = 24
Q3 = 28

Interquartile range here will be = Q3 - Q1
= 28 - 24 = 4

N = 45
28th
Location of a percentile = (45+1) * (p/100) = 46 * 0.28 = 12.88