NUMERICAL AND STATISTICAL COMPUTING (MCA-202-CR)

Autumn Session

UNIT 4

STANDARD DEVIATION:-

Standard deviation is defined as the square root of the mean of the square of the deviation from the arithmetic mean. The square root of the standard deviation σ^2 is called VARIANCE.

Standard deviation (SD) represented by the Greek letter sigma 'σ' or the Latin letter 's' is a measure that is used to quantify the amount of variation or dispersion of a set of data values. A low standard deviation indicates that the data points tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

The standard deviation of a random variable, statistical population, data set, or probability distribution is the square root of its variance. It is algebraically simpler, though in practice less robust, than the average absolute deviation. A useful property of the standard deviation is that, unlike the variance, it is expressed in the same units as the data. In addition to expressing the variability of a population, the standard deviation is commonly used to measure confidence in statistical conclusions.

Mathematically,

$$S.D. = \sigma = \sqrt{\frac{\sum f(x - x')^2}{\sum f}}$$

$$= \sqrt{\frac{\sum fd^2}{\sum f} - \left[\frac{\sum fd}{\sum f}\right]^2}$$

Where, x = data points; f = frequency; a = mean; d = x-a

For example, each of the three populations {0, 0, 14, 14}, {0, 6, 8, 14} and {6, 6, 8, 8} has a mean of 7. Their standard deviations are 7, 5, and 1, respectively. The third population has a much smaller standard deviation than the other two because its values are all close to 7. It will have the same units as the data points themselves. If, for instance, the data set {0, 6, 8,14} represents the ages of a population of four siblings in years, the standard deviation is 5 years. As another example, the population {1000, 1006, 1008,1014} may represent the distances traveled by four athletes, measured in meters. It has a mean of 1007 meters, and a standard deviation of 5 meters.

Example

Q: Calculate the mean and standard deviation for the following data:

Size of Item	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

Sol: Assumed mean= 9

x	**F**	**d=x-a**	**f*d**	**f*d^2**
6	3	-3	-9	27
7	6	-2	-12	24
8	9	-1	-9	9
9	13	0	0	0
10	8	+1	8	8
11	5	+2	10	20
12	4	+3	12	36
	$\sum f = 48$	$\sum fd = 0$	$\sum fd^2 = 124$	

$$\text{Mean} = a + \sum \frac{\sum fd}{\sum f}$$

$$S.D. = \sigma = \sqrt{\frac{\sum f(x - x')^2}{\sum f}}$$

$$= \sqrt{\frac{\sum fd^2}{\sum f} - \left[\frac{\sum fd}{\sum f}\right]^2}$$

$$= \sqrt{\frac{124}{48}} = 1.6$$

CORRELATION:

Correlation is any of a broad class of statistical relationships involving dependence, though in common usage it most often refers to the extent to which two variables have a linear relationship with each other. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling; however, correlation is not sufficient to demonstrate the presence of such a causal relationship (i.e., correlation does not imply causation).

Formally, dependence refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence. In loose usage, *correlation* can refer to any departure of two or more random variables from independence, but technically it refers to any of several more specialized types of relationship between mean values. There are several correlation coefficients, often denoted ρ or r, measuring the degree of correlation. The most common of these is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables (which may exist even if one is a nonlinear function of the other). Other correlation coefficients have been developed to be more robust than the Pearson correlation – that is, more sensitive to nonlinear relationships. Mutual information can also be applied to measure dependence between two variables.

Whenever two variables x and y are so related that an increase in the one is accompanied by an increase or decrease in the other, then the variables are said to be correlated.

e.g., the yield of crop varies with the amount of rainfall.

If an increase in one variable corresponds to an increase in the other, the correlation is said to be positive. If increase in one corresponds to the decrease in the other the correlation is said to be negative. If there is no relationship between the two variables, they are said to be independent.

KARL PEARSON'S COEFFICIENT OF CORRELATION:

Karl Pearson's Coefficient of Correlation(r) is defined by the relation

$$r = \frac{\sum XY}{\sqrt{(\sum X^2)(\sum Y^2)}}$$

$$= \frac{P}{\sigma_x \sigma_y}$$

Where X= X-X', Y=Y-Y'

i.e. X,Y are the deviations measured from their respective means,

$$P = \frac{\sum XY}{n}$$

= co-variance

and $\sigma_x \sigma_y$ being the standard derivations of these series.

Example

Q: Ten students got the following percentage of marks in Economics and Statistics.

Roll no.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Marks in Economics	78	36	98	25	75	82	90	62	65	39
Marks in Statistics	84	51	91	60	68	62	86	58	53	47

Calculate the coefficient of correlation.

Solution: Let the marks of two subjects be denoted by x and y respectively.

Then the mean for x marks =650/10 = 65 and the mean of y marks =660/10=66.

If X and Y are the derivations of x's and y's from their respective means, then the data may be arranged in the following form:

X	Y	X= X- 65	Y=Y- 66	X^2	Y^2	XY
78	84	13	18	169	324	234
36	51	-29	-15	841	225	435
98	91	33	25	1089	625	825
25	60	-40	-6	1600	36	240
75	68	10	2	100	4	20
82	62	17	-4	289	16	-68
90	86	25	20	625	400	500
62	58	-3	-8	9	64	24
65	53	0	-13	0	169	0
39	47	-26	-19	676	361	494
650	660	0	0	5398	2224	2704

Here $\sum X^2 = 5398$; $\sum Y^2 = 2224$; $\sum XY = 2704$

$$r = \frac{\sum XY}{\sqrt{(\sum X^2)(\sum Y^2)}}$$

$$= \frac{2704}{\sqrt{5398 * 2224}}$$

$$= \frac{2704}{3457}$$

$$= 0.78$$

REGRESSION

Regression analysis is the method used for estimating the unknown values of one variable corresponding to the known value of another variable. If the scatter diagram indicates some relationship between two variables x and y, then the dots of the scatter diagram will be concentrated round a curve. This curve is called the 'curve of regression'.

LINE OF REGRESSION;

When the curve is a straight line, it is called a line of regression. A line of regression is the straight line which gives the best fit in the least square sense to the given frequency.

EQUATIONS TO THE LINES OF REGRESSION:

Regression of y on x

$$y\text{-}y' = r\,\frac{\sigma_y}{\sigma_x}(x - x')$$

Regression of x on y

$$x\text{-}x' = r\,\frac{\sigma_x}{\sigma_y}(y - y')$$

$r\,\frac{\sigma_y}{\sigma_x}\ and\ r\,\frac{\sigma_x}{\sigma_y}$ are known as the coefficients of regression.

Example

Q: Find the regression line of y on x for the following data:

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Estimate the value of y, when x=10.

SOLUTION:

S.No.	**X**	**Y**	**XY**	**X^2**
1	1	1	1	1
2	3	2	6	9
3	4	4	16	16
4	6	4	24	36
5	8	5	40	64
6	9	7	63	81
7	11	8	88	121
8	14	9	126	196
TOTAL	**56**	**40**	**364**	**524**

Let y=a+bx be the line of regression be the line of regression of y on x, where a and b are given by the following equations:

$\sum y = na + b\sum x$ => 40=8a+56b

$\sum xy = a\sum x + b\sum x^2$ => 364=56a+524b

Solving these two equations, we get

a= 6/11 and b=7/11

The equation of the required line is

y=6/11+7/11x

= 7x-11y+6=0

If x=10, y=6/11 +7/11(10) =76/11

HYPOTHESIS:

A hypothesis is a proposed explanation for a phenomenon. For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it. Scientists generally base scientific hypotheses on previous observations that cannot satisfactorily be explained with the available scientific theories. A working hypothesis is a provisionally accepted hypothesis proposed for further research[.]

A hypothesis requires more work by the researcher in order to either confirm or disprove it. In due course, a confirmed hypothesis may become part of a theory or occasionally may grow to become a theory itself.

In other words, a hypothesis is a prediction that can be tested.

E.g., imagine you have a test at school tomorrow. You stay out late and see a movie with friends. You know that when you study the night before, you get good grades. What do you think will happen on tomorrow's test?
When you answered this question, you formed a hypothesis. A hypothesis is a specific, testable prediction. It describes in concrete terms what you expect will happen in a certain circumstance. Your hypothesis may have been, 'If not studying lowers test performance and I do not study, then I will get a low grade on the test

Purpose of a Hypothesis

A hypothesis is used in an experiment to define the relationship between two variables. The purpose of a hypothesis is to find the answer to a question. A formalized hypothesis will force us to think about what results we should look for in an experiment.

The first variable is called the '*independent variable'*. This is the part of the experiment that can be changed and tested. The independent variable happens first and can be considered the cause of any changes in the outcome. The outcome is called the '*dependent variable*'. The independent variable in our previous example is not studying for a test. The dependent variable that you are using to measure outcome is your test score.

Let's use the previous example again to illustrate these ideas. The hypothesis is testable because you will receive a score on your test performance. It is measurable because you can compare test scores received from when you did study and test scores received from when you did not study.

A hypothesis should always:

Explain what you expect to happen
Be clear and understandable
Be testable
Be measurable

And contain an independent and dependent variable

Working hypothesis

A working hypothesis is a hypothesis that is provisionally accepted as a basis for further research in the hope that a tenable theory will be produced, even if the hypothesis ultimately fails. Like all hypotheses, a working hypothesis is constructed as a statement of expectations, which can be linked to the exploratory research purpose in empirical investigation. Working hypotheses are often used as a conceptual framework in qualitative research.

Null hypothesis

Null hypothesis is based for analyzing the problem. Null hypothesis is the hypothesis of no difference. Thus, we presume that there is no significant difference between the observed value and the expected value. Then, we shall test whether this hypothesis is satisfied by the data or not. If the hypothesis is not approved the difference is considered to be significant. If hypothesis is approved then the difference would be described as due to sampling fluctuations. Null hypothesis is denoted by H_0.

Errors

We decide to accept or to reject the lot after examining a sample from it. As such we are liable to commit the following two types of errors:

TYPE I ERROR: If H_0 is rejected while it should have been accepted.

TYPE II ERROR: If H_0 is accepted while it should have been rejected.

Testing a hypothesis

On the basis of sample information, we make certain decisions and based on these decisions we make certain assumptions. These assumptions are known as '*statistical hypothesis'*. There h. There hypothesis are tested, Assuming the hypothesis correct we calculate the probability of getting the observed sample. If this probability is less than a certain assigned value, the hypothesis is to be rejected.

TEST OF SIGNIFICANCE:

The tests which enable us to decide whether to accept or reject the null hypothesis is called the tests of significance. If the difference between the sample values and the population values are so large (i.e. lies in critical area), it is to be rejected.

THE CHI- SQUARE DISTRIBUTION:

Chi- square is a measure of actual divergence of the observed and expected frequencies. If f_0 is the observed frequency and f_e the expected frequency of a class interval, then x^2 is defined as

$$x^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

DEGREE OF FREEDOM

The degree of freedom refers to the number of "independent constraints" in a set of data.

Degree of freedom= (r-1) (c-1)

Where, r= number of rows and c= number of columns

GOODNESS OF FIT:

The value of x^2 is used to find the divergence of the observed frequency from the expected frequency.

If the value of P is high, the fit is said to be good. It means that there is no significant divergence between the observed and expected data.

If the curve of the expected frequency is superimposed on the curve of observed frequencies there would not be much divergence between the two. The fit would be good. If the value of P is small, the fit is said to be poor.

STEPS FOR TESTING:

1. Calculate the value of x^2.
2. From the table, read the value of x^2 for a given degree of freedom.
3. Find out the probability P corresponding to the value of x^2
4. If P > 0.05, the value is not significant and it is a good fit.
5. If P< 0.05, the deviations are significant.

Example

Q: From the following table, showing the number of plants having certain characters, test the hypothesis that the flower color is independent of flatness of leaf.
(Note: x^2 =0.0158 at 0.1level of significance)

	Flat Leaves	Curled Leaves	Total
White Flowers	99	36	135
Red Flowers	20	5	25
Total	119	41	160

Solution;
Null hypothesis: the flower color is dependent of flatness of leaf. The following table shows the theoretical frequencies:

	Flat Leaves	Curled Leaves	Total
White Flowers	$\frac{135 * 119}{160} = 100$	$\frac{135 * 41}{160} = 35$	135

Red Flowers	$\frac{25 * 119}{160} = 19$	$\frac{25 * 41}{160} = 6$	25
Total	119	41	160

$$x^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

$$x^2 = \frac{(99-100)^2}{100} + \frac{(36-35)^2}{35} + \frac{(20-19)^2}{19} + \frac{(5-6)^2}{6}$$

$$x^2 = \frac{1}{100} + \frac{1}{35} + \frac{1}{19} + \frac{1}{6}$$

$$x^2 = 0.2579$$

Degree of freedom = (r-1) (c-1) = (2-1)(2-1) = 1

We have $x^2 = 0.0158$ at 0.1 level of significance.

0.2579 > 0.0158

This leads to the conclusion that the hypothesis is wrong and the flower color is independent of flatness of leaf at the 0.1 level of significance.