MEASURES OF CENTRAL TENDENCY

Central tendency means average performance, while dispersion of a data is how it spreads from a central tendency. He measures of central tendency are: mean, mode and media.

The mean is the sum of observations divided by the number of observations

X1 X2 X3 X4 …………………………… Xn

N

X (pronounced as X bar) is the mean from sample where n is the mean from population. X is the estimate of n e.g.

0.9g, 1.2g, 1.4g, 1.6g, 1.8g

X   0.9 1.2 1.4 1.6 1.8   6.9

X = 6.9/5   1.38g/plot

N

The mean is a very good measure of central tendency where samples show what is called ***Normal distribution.***

Data is normally distributed when it has equal spread around the mean. It is not always the case that a set of data will be normally distributed. When the data is not normally distributed the mean is not a good measure of the central tendency

Mean / mode / median.

In the case of the skewed distribution the mean is located away from the central tendency

SAMPLE MEDIAN

Where data doesn't have a normal distribution mean doesn't represent the central tendency in such a situation, the median gives a better picture of central tendency which is defined as the central value when the data is arranged in ascending/descending numerical order.

e.g. 0.15g, 0.45g, 0.87g, 0.99g, 1.02g, 0.38g, 0.98g 1.00g Calculate the median of the data.

Solution: 0.15, 0.38, 0.45, 0.56, 0.87, 0.98, 0.99, 1.00, 1.02

0.87 = 0.71

Mode

This is very useful in determining central tendency when a very large data is involved. Given that the scores of 20 students in statistics examination are given below. Find the mode.

66.60.61.45.45.80.80.77.72.43.55.60.60.60. 98.40.80.62.41.42 he

Mode 60.

If a data is normally distributed the mode medium mean

MEASURES OF DISPERSION

The measures central tendency of a sample: Mean, median and mode do not tell us how value varies from the central tendency. This measurement of how values deviate from the mean is known as a measure of dispersion.

For instance, if the mean of ages of 1001 student in the University is 19 years. This could be 3 which means that the age range is 16 - 22 years in which the positive sign indicate the upper limit and the negative indicate the lower limit.

X - 19 years 3 i.e. Range 16 - 22

For every statistics that we measure there is a corresponding population parameter. Therefore the statistics/parameter for measuring dispersion are as foolows

Measures of Dispersion

Parameter/Statistics	*Population*	*Samples*
Variance		
Standard deviation		
Range		

Range provides the lower and upper values of a set of data e.g. if the dry matter yield of soyabean plant inoculated with strains of nitrogen fixing bacteria (Rhizobium) are given as follows:

Bacteria strains	*Dry matter yield (x)*	*x - x*
R_1	*10.2*	*10.28*
R_2	*11.1*	*1.18*
R_3	*9.8*	*-0.12*
R_4	*8.7*	*-1.22*
R_5	*12.2*	*2.28*
R_6	*13.1*	*3.18*
R_7	*9.4*	*-0.52*
R_8	*10.6*	*10.08*
R_9	*6.9*	*-3.02*
R_{10}	*7.2*	*-2.72*

The sum of deviations for the mean is 0

Calculate the mean, the range and the S.D.

$$Mean = x - \frac{99.2}{10} = 9.92$$

$\underline{x_1} = x$

n

x = mean

$x - x$ = deviation from the mean

$(x - x)2$ = square of deviation from the mean

Variance $S2 \ (x - x)$ while $S = (x - x)2$

$n - 1$ $n - 1$

Range x

Bacteria strains	*Dry matter (x)*	*X - x*	*(x – x)2*	*X2*
R_1	10.2	0.28	0.0784	104.04
R_2	11.1	1.18	1.3924	123.21
R_3	9.8	-0.12	0.0144	96.04
R_4	8.7	-1.22	5.1984	75.69
R_5	12.2	2.28	5.1984	148.84
R_6	13.1	3.18	10.1124	171.61
R_7	9.4	-0.52	0.2704	88.36
R_8	10.6	0.68	0.4624	112.36
R_9	6.9	-3.02	9.1204	17.61
R_{10}	7.2	-2.72	7.3984	51.84
	X = 9.92	= 0	35.5360	1019.6

$X = 99.2$

$X2 = 1019.6$

$(x – x)2$

n i.e. if you are dealing with a population

$S2 = (x - x)2$ i.e if you are dealing with sample

$n - 1$

Variance $S2 = 35.5360 = 3.9484$

9

$S.D = S = 3.9484 = 1.987$

Mean $= 9.92 = 10$ Range $= 10.42$

Variance $= 1.987 \; 2$ or

Therefore the range is $9.92 \; 1.987$ i.e. $8 – 12$ or $7.92 - 11.92$

$S = (x – x)2$

$n – 1$

$S = x2 - x2$

$n - 1$

$x2$ = sum of square of x

$x2$ = sum of x squared

$n – 1$ = degree of freedom

$(x)2$ = correction factor which must be less than sum of square

N

$1019.16 – (99.2)2$

10

9

$S\ 1019.6 - 98.064$

9

$S = 35.536$

9

$S = 3.9484$

$S = 1.987$

$(x – x)2$ $\quad\quad$ $\dfrac{x2 - \dfrac{(x)2}{n}}{n - 1}$

COMPARISONS AMONG TREATMENT MEANS

In the analysis of variance, the null hypothesis (Ho) that is tested is always that all means are equal. If the F statistic is not significant then this Ho is not rejected and there is nothing more to do, except possibly try to make the experiment itself more precise/sensitive. If the Ho is rejected then at least one mean is significantly different from at aleast one other one. However, the limitation of the F-test is that it does not locate the specific difference(s) among the treatments.

FOUR BASIC ASSUMPTIONS UNDERLYING ANOVA

1. *Error terms are randomly, independently and normally distributed*
2. *Variances of different samples are homogenous i.e. they must not deviate from each other.*
3. *Variances and means of different samples are not correlated*
4. *The mean effects are additive (non-addivitivity results in heterogeneity of errors)*

Specific comparisons

1. *between pairs of treatments*
2. *between 2 groups of treatments*
3. *tend comparisons*

1. The first comparison is the simplest and most commonly used

2. The second comparison involves classifying the treatments into meaningful groups. A group may consist of one or more treatments and comparisons is made between aggregate means e.g. fertilized plots vs control; exotic varieties vs local varieties.

1 and 2 can be applied to any set of treatments.

3. The third comparison is limited to only treatments that are quantitative e.g fertilizer rates, distances of planting

The two basic procedures for testing the differences between treatment means are the least significant difference method (Lsd) and duncan's Multiple Range Test (DMRT).

LSD Method

This is the most commonly used and misused method of mean sesapartion. LSD is the most effective or valid test when there are only two means. The precision decreases as the number of means of treatments increases.

Rules for effective use of Lsd test

a. *Use the lsd test only when the F test in the ANOVA is significant.*
b. *Do not use the lsd when the possible pairs of means in the experiment exceeds five*
c. *Use the lad for pre-planned comparisons even if the treatments are more than 5*

To determine whether a difference between the treatments means is statistically significant, compare the observed difference with the computed lsd value. If the observed difference is larger *than than the lsd value, the two treatment means are significantly different at alpha level of significance. If the difference between any pair of means is smaller than the lsd value, the two treatments are not significantly different.*

Limitations of lsd test

a. Lsd assumes that the error is homogenous

b. It is often used to make many unplanned comparisons

c. It is not satisfying for all possible paired compatisons

Duncan's Multiple Range Test (DMRT)

DMRT was developed to alleviate the deficiencies of lsd. It is identical to lsd for adjacent means, but requires progressive higher values necessary for comparison. The use of DMRT should be limited to very specific situations where no prior knowledge is available about the performance of the treatments e.g selection of new varieties being compared

Advantages of DMRT

a. It takes into consideration the number of treatments

b. It permits decisions as to which differences are significant and which are not

c. F-test does not need to be significant before one can proceed

d. It uses a set of significant ranges, each range depending upon the number of means in that comparison

OUTLIER TEST

Outlier is the term used to denote the results of a replication which deviates greatly from the mean of all other replications of a specific treatment. They have an adverse and undesirable effect on trial results. They may be due to errors in plot measuring, tillage, application of the treatments, damage to plants during hoeing or by animals, robbery, mistakes in calculation e.t.c

N-fertilizer trial with 4 replicates

Yield: 31, 24, 8 and 25 kg/ha

SD of the 3 other means should be calculated and it must be determined if the SD of the presumed outlier is >4 times this SD

$$\frac{31+24+25}{3} = 26.7\ (\bar{x})$$

$$SD = \frac{(31-26.7)^2+(24-26.7)^2+(25-26.7)^2}{3-1} = 3.8$$

SD outlier = 8-26.7= 18.7

Since 18.7 is 4 times > 3.8, it is a true outlier

Coefficient of variation (CV)

When analyzing the trial, the magnitude of the experimental error permits us to assess the precision with which the trial has been conducted.

CV is the SD expressed as a percentage of the mean. It is a measure of the degree of repeatability of an experiment. It is said to be low of high in relation to a reference point.

$$CV = \frac{SD}{Mean} \times 100,\ \%$$

In field experiment, the CV should not exceed 10 -15%, otherwise it can be assumed that the trial has not been carried out satisfactorily. The results can only be used if reasons are stated in the Results and discussions Sections.