MEASURE OF LOCATION

Mean is also known as average. For example the mean of 4, 5 6, 7 and 3 is 25/5 = 25. For a large data set, we normally make use of frequency distribution.

Example 1. Mean of a frequency distribution

Variable (x)	Frequency (f)	(xf)
4	2	8
5	1	5
6	3	18
7	4	28
8	1	8
9	1	9
	12	**76**

Mean = $\frac{\Sigma xf}{\Sigma f} = \frac{76}{12} = 6.33$

Mode = 7 because it is the number with the highest frequency

Types of mean

a) Arithmetic mean = *sum of all variable* = $\frac{\Sigma x}{n} = \frac{\Sigma x}{\Sigma f}$

number of variable

=

b) Geometric mean

The geometric mean of 2 numbers is the square root of their products e.g. geometric mean of 4 & 5 = $\sqrt{4}$ x $\sqrt{5}$ = $\sqrt{20}$

For three numbers 2, 3, 4, it is the cube root.

=

3

x^y

Geometric mean = $\sqrt[3]{2 \times 3 \times 4} = 2.884$

= $\sqrt[3]{24} = (24)^{1/3}$ = 24 shift 2.884

:. Geometric mean of 1, 2, 3, 4, 5 = $\sqrt[5]{120} = 2.605$

c) Harmonic mean

This is the reciprocal of arithmetic mean of sum of reciprocals.

Consider 5 numbers 1, 2, 3, 4, 5

Sum reciprocals $= \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}$

$= 1 + 0.5 + 0.333 + 0.25 + 0.2 = 2.28$

Arithmetic mean of sums (am)

= *1/5 of 2.28*

= *0.456*

Reciprocal of arithmetic mean $= \frac{1}{0.456} = 2.19$

Please note very well that harmonic mean is lower in value than the geometric mean.

Median: This is the number (or numbers) in the middle of a set of data after re-arranging the set of figures in ascending or descending order, or in the order of magnitude.

Example: 5, 6, 4, 7, 6, 6, 7, 8, 7, 7, 9, 4 can be re-arranged as 4, 4, 5, 6, 6, 6, 7, 7, 7, 7, 8, 9

Median = $= \frac{6+7}{2} = 6.5$

Median of a large population

When the number of measurements (N) is large i.e we have a large population, median of that population is taken as $\frac{1}{2}N^{th}$ measurement. i.e.

Median may be defined as the 50^{th} percentile, with half of the population above and the other half below it. Median divides the area of a histogram into two equal parts.

PERCENTILES

From the cumulative frequency graph (Ogive) given, 50^{th} percentile correlates with cumulative frequency of ½ x 108 = 54

The Ogive gives the number of plots per plant that is not more than 14

Therefore the 50^{th} percentile = 14 which is about the **median**

75^{th} percentile = 81 ≈ ¾ of 109 of Ogive

The number that corresponds to this is 17 if you trace the ogive vertically down

75^{th} percentile = ¾ $(N + 1)^{th}$ number

25^{th} percentile = ¼ $(N + 1)^{th}$ number

RELATIONSHIP BETWEEN MEAN, MEDIAN AND MODE

Mean - mode = 3 (mean – median) (1)

Mean – mode = 3mean – 3 medain (2)

- mode = 2 mean – 3 median (3)

:. Mode = 3 medain -2 mean (4)

Example 2

Weight (kg) (x)	Freq (f)	xf

73	1	73
72	2	144
71	3	213
70	2	140
69	4	276
Total	**12**	**876**
	∑f	∑xf

Mean $= \frac{\Sigma xf}{\Sigma f} = \frac{846}{12} = 70.5kg$

Median = (71 + 70)/2 = 71.5 kg because we have 69, 69, 69, 69, 70, 70. 71, 71, 71, 72, 72 and 73

Mode = 69 kg because it has the highest frequency

Example 3 Mean using a variable as origin

(x)	(deviation)	f	xf
73 – 70	3	1	3
72 – 70	2	2	4
71 – 70	1	3	3
70 – 70	0	2	0
70 – 69	-1	4	-4

Mean weight $= 70 + \frac{\Sigma fx}{\Sigma f}$

$= 70 + \frac{6}{12} = 70 + 0.5 = 70.5kg$ as in Example 2

Example 4

Number of insects	Freq	Deviation 15 as origin	(fx)
8	1	-7	-7
9	1	-6	-6
10	2	-5	-10
11	5	-4	-20
12	9	-3	-27
13	15	-2	-30
14	19	-1	-19
15	20	0	0
16	16	1	16
17	10	2	20
18	5	3	15
19	3	4	12
20	1	5	5
21	0	6	0
22	1	7	7
Total	**108**		**75**

Mean $= 15 + \left[\frac{\Sigma fx}{\Sigma f}\right] = 15 + \left[\frac{75-119}{108}\right] = 15\left[\frac{-44}{108}\right] = 15 - 0.4 = 14.6$

QUESTION 1 Try the above using 14 as origin

Table 1. A frequency distribution table

Number of pods/plant	Frequency

8	𝍸	1
9	𝍸	1
10	𝍸𝍸	2
11	𝍸𝍸𝍸	5
12	𝍸𝍸𝍸 𝍸𝍸𝍸	9
13	𝍸𝍸𝍸 𝍸𝍸𝍸 𝍸𝍸𝍸	15
14	𝍸𝍸𝍸 𝍸𝍸𝍸 𝍸𝍸𝍸 𝍸𝍸𝍸	19
15	𝍸𝍸𝍸 𝍸𝍸𝍸 𝍸𝍸𝍸 𝍸𝍸𝍸	20
16	𝍸𝍸𝍸 𝍸𝍸𝍸 𝍸𝍸𝍸 𝍸	16
17	𝍸𝍸𝍸 𝍸𝍸𝍸	10
18	𝍸𝍸𝍸	5
19	𝍸𝍸𝍸	3
20	𝍸	1
21	-	0
22	𝍸	1
Total number examined	108	

Frequency polygon

(a) Modal group - Equal grouping

Under 10		Frequency
10 – 11	𝍸	1
12 – 13	𝍸𝍸	1
14 – 15	𝍸𝍸𝍸𝍸	2
16 – 17	𝍸𝍸𝍸𝍸𝍸𝍸𝍸𝍸𝍸𝍸	5
18 – 19	⸺⸺⸺ 𝍸𝍸𝍸𝍸𝍸	9
20 – 22	𝍸𝍸	15

Unequal grouping	Frequency
Under 10	2
10 - 11	7
12 – 14	43
15 – 17	46
18 – 20	9
21 - 22	1
	108

Modal group = 15 - 17

HISTOGRAM OF WEIGHT DISTRIBUTION

Weight score (kg)	Frequency

8	1
9	1
10	2
11	5
12	9
13	15
14	19
15	20
16	16
17	10
18	5
19	3
20	1
21	0
22	1

CUMULATIVE FREQUENCY DISTRIBUTION OF NUMBER OF PODS PER PLANT

Number of pods per plant	Frequency	Cumulative frequency
8	1	1
9	1	2
10	2	4
11	5	9
12	9	18
13	15	33
14	19	52
15	20	72
16	16	88
17	10	98
18	5	103
19	3	106
20	1	107

21	0	107
22	1	108
	108	

THE CUMULATIVE FREQUENCY GRAPH

The cumulative frequency graph is also called “Ogive” It is a graphical way of representing the frequency distribution.

MEASURE OF DISPERSION

The various statistics used for the measurement of dispersion or spread are variance.

Standard deviation

Variance of the mean

Standard error of the mean

Coefficient of variability

(a) The sample variance (S^2) is given as $S^2 = \frac{\left[\Sigma x^2 - \frac{\Sigma x^2}{n}\right]}{n-1}$

Where

(Σx^2) = Sum of squares of variable x obtained by squaring each variable

$(\Sigma x)^2$ = Square of sum of variable x

n = number of variables

$\frac{(\Sigma x)^2}{n}$ is called the correlation term

Let us consider the following six variables 5.0, 5.5, 7.0, 8.0, 10.0, & 2.5

Where n = 6

Variable	x^2
5.0	25.0
5.5	30.25
7.0	49.0
8.0	64.0
10.0	100.0
2.5	6.25
Sum ∑38.0	**274.5**

Σx^2 = 274.5

Σx = 38.0

:. $(\Sigma x)^2$ = $(38.0)^2$ = 1444

$(\Sigma x)^2)/n$ = 1444/6 = 240.67

Variance (S^2) = $\frac{\Sigma x^2 - \left(\frac{\Sigma x)^2}{n}\right)}{n-1}$

= $(274.5 - 240.67)/5$

= $33.83/5 = 6.77$

(b) Standard deviation is the square root of variance i.e. the square root of variance gives standard deviation.

Square root of S^2 $= \sqrt{S^2} = S$

∴ *Standard deviation S* = $\sqrt{S^2}$

From the example above S = $\sqrt{6.77} = 2.60$

(c) Mean of the samples (x) is

given as $(x) = \frac{\Sigma x}{n} = \frac{38}{6} = 6.33$

Coefficient of variability (CV) = Standard deviation = 2.60/6.33 = 0.410

6.33

Expressed as a percentage CV = 0.41 x 100 = 41.0%

(d) Variance of the mean (S^2 x) is given as variance divided by the number of variables

n = 6

S^2 x = S^2/n

= 6.77/6 = 1.128

(e) Standard error of the mean simply called standard error is the square root of the variance of the mean

Standard error = √variance of the mean

$$= \sqrt{S^2} = \frac{\sqrt{S^2}}{n} = \sqrt{1.128} = 1.062$$

Note that Standard error = $\sqrt{\frac{S^2}{n}} = \frac{S}{\sqrt{n}} = Sx$

Where S = Standard deviation

TEST OF HYPOTHESIS

An hypothesis is an assumption about a parameter or population which may or may not be true.

Type of Hypothesis

1. *Null hypothesis (H_O)*
2. *Alternative hypothesis (Ha)*

Ho is the hypothesis to be tested for acceptance or rejection depending on the result of an experiment. It usually contains an equality statement or sign so that a confidence interval can be constructed around the parameter e.g.

Mean value of the two soil samples (μ_1 & μ_2) are the same or similar

Ho : $\mu_1 = \mu_2$ means are equal soils are similar.

Ha is the hypothesis taken as true when the Ho is false.

Ha : $\mu_1 \neq \mu_2$ means are not equal soils are not similar

Where there are more than two means we say

Ho : $\mu 1 = \mu 2 = \mu 3 = \mu 4$ …….

In the alternative.

Ha : $\mu 1, \neq \mu 2 \neq \mu 3 \neq \mu 4$

Test Statistics

- *The basis for any scientific experimentation is to set up a hypothesis – the Null hypothesis and when the null hypothesis is disputed, we accept the alternative hypothesis.*
- *However, the rejection or acceptance of any hypothesis must be done on an objective and rational basis and not what we think about the outcome of our investigation.*
- *Therefore, the use of appropriate test statistic will provide the rational and objective acceptance or rejection of hypothesis under various circumstances. It is very important to note that each test statistic is based on certain assumptions rather than the mathematic calculations.*
- *The assumptions must be known and fulfilled because any conclusions based on the test results will be meaningless if the assumptions are not fulfilled.*

The most useful and commonly used test statistics for testing hypothesis are

1. *The t test, which compares the means of two samples and tests the null hypothesis that the two means are the same*
2. *The (Chi square) X^2 test, which compares how well some data fit a model or an idea situation you already have.*
3. *The F-test, which compares the variances or standard deviations of two samples to see if they come from the same population.*
4. *Correlation coefficient, which measures the association between two variables.*

Before we go into specific examples of the above, let us see what we understand by the following term.

Significant Levels

- *Each of any of the test statistics above has the percentage point (probability) at which the null hypothesis can be accepted or rejected i.e. 20%, 10%, 5%, 1%, and so on.*
- *The use of the probability level will allow us to know how likely it is that we would have obtained our results, if the null hypothesis is true.*
- *The point at which we reject the null hypothesis has to be decided on the basis of the penalties for being wrong. The question is "What is the penalty for rejecting the null hypothesis when we should have accepted it?, or the penalty of accepting when we should have rejected? This decision on the percentage point or the significant level is very important.*

In agricultural and biological experiments, 5% is normally taken as the significant level.

If the chance (probability) of getting results that are different from those predicted by null hypothesis is greater than 5% (P> 0.05), then we say that there is no significant difference between the observed and the predicted and thus, we accept the null hypothesis.

On the other hand, if the chance of getting results is less than 5% (P< 0.05), we say that the observed and predicted are significantly different at 5% level and thus, we reject the null hypothesis and accept the alternative hypothesis.

Generally the smaller the probability or chance (say 0.01 or 0.001), the more likely it is that our results are due to some realistic biological phenomena and not due to random chance. The corollary is that the larger the value, the less likely it is that our results are due to a real biological effects and the more likely they are due to chance

P Value	**0.20**	**0.10**	**0.05**	**0.01**	**0.001**
Accepting the Null hypothesis			Rejecting the null hypothesis		

P > 0.05	P < 0.05	P < 0.01	P< 0.001
Not quite significant (Lack of Confidence)	Significant	Highly significant	Very highly significant
Repeat of experiment may be necessary	Fairly confident	Very confident	Almost certain

---

The Student's t test

The test is not particularly reserved for students. If was first developed by William Gosset in 1908 who at that time was not allowed by his employer (Guinness) to publish under his own name. He therefore referred to himself as "Student"

The t test is used to compare the means of two samples or sets of data to see whether they come from the same population or not. The null hypothesis is that the two means come from the population and that any difference between them is due to sampling error or chance.

The assumptions of a t test are

1. *The populations from which samples are taken are normally distributed.*
2. *The samples have similar standard deviations.*
3. *The samples were collected independently.*

The test statistic is given as t $= \frac{(x_1 - x_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(x_1 - x_2)}{Sx}$

x_1 and x_2 are the two means

S is standard deviation of the population

Sx is called standard error.

n_1 and n_2 are the number of observations

thus, $SX = \frac{\sqrt{S^2}}{n} = \frac{S}{\sqrt{n}}$

The t test is very robust because the conclusions based on it has some level of validity even if the assumptions on which it is based are not completely fulfilled.

t test can be used in a number of ways

1. *To compare two sample means as above*
2. *To compare a sample means with a standard when the mean and variance of the population are known.*

Example 1

The mean yield of maize in Nigeria is 5 ton\ha (a standard). However, a new maize variety with a yield of 7.6 tons\ha is introduced.

Test if the two means are similar or not.

$H_0 : \mu_1 = \mu_2$ i.e. $5.0 = 7.6$

$Ha : \mu_1 \neq \mu_2$ i.e. $5.0 \neq 7.6$

Assuming that the mean yield of the improved variety is obtained from 64 samples i.e n = 64 with 6.77 as variance

$$t = \frac{x - \mu}{Sx}$$

s = Standard deviation

n = Number if observation

x = Mean of new (improved) variety = 7,6 kg/ha

μ = Standard mean = 5.0 kg/ha

Sx = Standard error = $\frac{\sqrt{S^2}}{n} = \sqrt{\frac{6.77}{64}} = \sqrt{0.1058}$

t = (7.6 – 5.0)/0.325 = 7.69

Thus, t_{cal} = 7.69

Tabulate t value at 63 df = 2.0

Ho : μ = μ makes the test a 2-tail test

α- level = 0.05/2 = 0.025 and from the above, $t_{cal} > t_{tab}$

We can then conclude that the two means are significantly difference at 5% probability i.e.

Ho $7.60 \neq 5.0$ $(P < 0.05)$

We reject the null hypothesis and conclude that the two means differed significantly.

Example 2

Sample vs Standard - When the means and sample variance are not known.

When the samples are given, one can calculate the mean of samples and then the variance of the sample. Then the sample means can be compared with the Standard

X yield of cassava = 10 ton/ha (Standard)

Another eight samples 8, 10, 12, 11, 9, 14, 8, 10 are taken

Sample mean = 10.25

Sample variance = 4.21 and n = 8

$$\therefore \quad Sx = \sqrt{\frac{4.21}{8}} = \sqrt{0.527} = 0.725$$

$$t = \frac{x-\mu}{Sx} = \frac{10.25-10.00}{0.725} = 0.344$$

Thus, $t_{cal} = 0.344$

Number of samples (n) = 8 and α – level is 0.05/2 = 0.025

Therefore, $t_{0.025}\ df_7 = 2.365$

Also, $t_{0.05}\ df_7 = 3.97$

$t_{cal} < t_{tabulated}$

Therefore, we accept the null hypothesis and conclude that there is no significant difference between the two means i.e. $P > 0.05$ ($\mu_1 = \mu_2$)

Example 3

Expected or standard mean = 10.29

Samples are (x) 8, 10, 7, 6, 9, 10, 8 :. $\Sigma x = 58$

Squares (x^2) 64, 100, 49, 36, 81, 100, 64 :. $\Sigma x^2 = 49.4$

Observed Mean (x) = $\frac{\Sigma x}{n} = 8.29$

Variance = $\frac{\left[\Sigma x^2 - \frac{(\Sigma x)^2}{7}\right]}{6} = \frac{494 - 480.6}{6} = 2.24$

6

Standard error = $\sqrt{\frac{\delta^2}{n}} = \frac{\delta}{\sqrt{n}} = (2.24)^{\frac{1}{2}} = 0.566$

t = (10.29 – 8.29)/0.566

= 3.54 = t_{cal}

$t_{tabulated}$ = 4.39

Thus at 5% probability level

$T_{cal} < t_{tab}$

We accept the null hypothesis (P> 0.05) and conclude that there is no significant difference between 10.29 and 8.29 i.e $\mu_1 = \mu_2$

The F tests

F-tests are tests of variances that make use of the ratio of two variances particularly in the analysis of variance (ANOVA) table to determine whether or not two samples come from the same population.

In some instances, standard deviations, are not necessarily the variances, of the samples are used. Thus, F tests can be used to test if two populations have equal variance. i.e they are used to compare variances.

Usually, the larger the F ratio, the greater is the difference between the means that connect the variances.

Assuming that we want to consider an experiment with seven samples using a t test at 0.05 significant level. There will be just 1 in 20 chances of concluding that the samples come from populations with different means when in fact they do not. It will also mean that we shall have 21 ($7C_2$) different comparisons to be able to decide that the two means from the population differ significantly even though in reality they come from the same population.

Therefore the use of F-tests in the analysis of variance is both safer and more efficient. They are used to test the significance of mean squares of different source of variation in the ANOVA tables. In fact, it is good to make use of t-tests only when the F-tests are significant.

F-tests are based on one-tail tests and the test statistic is given as

$$F = \frac{\delta_1^2}{\delta_2^2}$$

Where δ_1^2 and δ_2^2 are variances or mean squares.

The table below is a typical ANOVA table involving four genotypes of cowpea in four replicates in a given environment.

Source	df	Mean square	Observed F-ratio	Expected F-ratio
Replication	3	2299.25	1.43	3.86
Genotype	3	11024.18	6.87	3.86
Error	9	1605.34		

Total	15

Note that the mean square (MS) is the ratio of Sum of Square (SS) to the degree of freedom (df).

Thus SS = df x MS

Note also that the expected F-ratio was taken at 5% probability level taking cognizance of the df for genotype in the horizontal part of the F table and df of error in the vertical column.

From the ANOVA table one can conclude that there is significant difference among the four cowpea genotypes because the observed F-ratio is larger than the expected. However, there was no effect of replication suggesting that replications were virtually similar.

The chi-square (X^2) test

(X^2) is used in agriculture to test whether the observed values in an experiment agree with the expected values in a set of quantitative data.

It is also used to test whether the effect of a set of treatment depends on another effect or factor. When used to test treatment effects, there is need to construct a 2 – way contingency table.

Thus, $X^2 = \frac{\sum (\text{Observed} - \text{Expected})^2}{\text{Expected}}$

Example 1

A breeder postulates that the phenotypes of the progeny of a certain di hybrid ratio are 9:3:3:1 in the F_2 generations.

Examination of 800 members of the F_2 generation revealed that 439, 168, 133 and 60 values were observed for the four phenotypes. He would suspect linkage if the ratio does not agree with 9:3:3:1.

Ho: Observed value = expected value

Ha : Observed value ≠ expected value

Note that n = 4 and thus, number of phenotypes degree of freedom (df) = 3

$X^2_{0.05}$ df 3 = 7.81 from the table

Phenotype	Expected ratio	Observed (O)	Number	Expected number (E)	0 - E	$(0 – E)^2$	$(0 –E)^2/E$
1	9/16	439		450	-11	121	0.27
2	3/16	168		150	18	324	2.16
3	3/16	133		150	-17	289	1.93
4	1/16	60		50	-10	100	2.00
∑ (sum)	**1**	**800**		**800**			**6.36**

$X^2_{cal} = \sum (0 – E)^2 /E = 6.36$

$X^2_{cal} < X^2_{tab}$

$X^2_{tab} = 7.81$

Since the calculated (observed) chi-square value is less than the tabulated or expected value, one can conclude that there is no significant difference between the observed and expected ratio i.e.

H_O: Observed = expected ratio is true

There is no linkage as there has been no deviation from the expected ratio.

Example 2

	Observe	Expected	O – E	$(O – E)^2$	$(O – E)^2/E$	
1	10	15	5	25	5/3	= 1.666
2	35	30	5	25	5/6	= 0.833

1	15	15	0	0	0	
4	**60**	**60**				**2.499**

$X^2ob = 2.5$

$X^2_{tab} = 5.991$

$X^2ob < X^2_{tab}$

There is no significant difference between the observed X^2 and expected X^2

We conclude that null hypothesis (H_o) is rejected