Introduction

This section on Basic Statistics is provided mainly as a handy reminder of the various formulae. A very useful and readable text book is "Facts from Figures" by M. J. Moroney, published by Penguin Books Ltd, Harmonsworth, Middlesex, England.

With the exception of the Arithmetic Mean, all of the basic formulae assume that the data to be analysed are drawn from a population that is Normally Distributed.

Arithmetic Mean

A mean or typical value of a set of test results can be derived in various ways. The most commonly used statistic is the Arithmetic Mean.

The Arithmetic Mean of a set of data is calculated as the sum of all values divided by the number of values. If n measurements are made, represented by X1 ... Xn, then:

Mean = (X1 + X2 + ... Xn) / n

In standard statistical notation, the symbol $\sum$ is used to denote "the sum of all such as", and the formula is written as

Mean = $\sum$ (Xi) / n

Coefficient of Variation

The percentage Coefficient of Variation, CV%, is another way of expressing scatter. It is the Standard Deviation expressed as a percentage of the Arithmetic Mean.

In general, the Standard Deviation will be greater the greater is the Mean. The CV allows the scatter to be compared between sets of data that have different mean values.

If the mean is M and the standard deviation is s, then

CV = 100 * s / M

Confidence Limits and Accuracy

A sample taken from a large population can provide only an approximation of the true behaviour of the population.

Measurements made on one sample of fabric cut from one fabric roll can give an estimate for the fabric properties of a multi-roll delivery, but the sample may not be representative of the whole delivery.

If more samples are taken from different rolls, then the observations can be averaged to give a better estimate.

The accuracy of these estimates in representing the true mean value of the whole delivery depends on:-

- the variability of the data, given by the Standard Deviation (s), or the Coefficient of Variation (CV).
- the Number of Observations (n).

The Confidence Limits (CL) and the Accuracy (A) of the estimate can be calculated, with a given degree of confidence, using the following formulae.

CL = t * s / $\sqrt{}$(n)

A = t * CV / $\sqrt{}$(n)

Where t is the Student's t statistic for the required confidence level (e.g. 95%).

Note:

*The term s / $\sqrt{}$(n) is the same as the Standard Error of the Mean, so that the Confidence Limit is t * SE (m).*

If the estimate of the mean is M, then there is a 95% chance that the true mean lies within M plus or minus CL, and the Accuracy of the estimate is plus or minus A%.

Number of Observations

For any single set of test data, there is no way of knowing in advance what will be the level of Accuracy, but there is a limit to the number of specimens that can be tested.

Not only is testing time-consuming, but it removes a certain amount of fabric from the delivery. A compromise must be found between the degree of accuracy and the level of sampling.

If the normal level of Standard Deviation of test data is reasonably well known, then it is possible to calculate how many samples must be tested in order to be reasonably sure of achieving a desired degree of Accuracy at a given level of Confidence.

All that is necessary is to rearrange the Accuracy formula.

$n = (t * CV / A)^2$

where t is the Student's t statistic, and CV is the Coefficient of Variation.

Note:

This formula only works well for data where the Mean Value is significantly (more than three standard deviations) different from zero.

There are two fabric parameters that do not fulfil this criterion. These are Spirality and shrinkage.

Both of these parameters can have positive or negative values, can have a mean of zero, and can have relatively large standard deviations. For such parameters, it is better to establish criteria for the Confidence Limits rather than the Coefficient of Variation.

Significance of Differences

Because of normal random variations in production, sampling and testing, it is rather unlikely that two separate determinations of a given parameter will be exactly the same.

There are occasions when it is necessary to decide whether a given determination of a given parameter is really different from the value that was expected.

For example, when checking a process calibration, new determinations for Reference Courses and Wales have to be compared with previous determinations made on previous batches of the same fabric quality.

Another example might be when a customer measures the weight per unit area of a delivery and complains that it is not the same as that which was specified.

A third example is when the Stitch Length in a batch of grey fabric is checked in the laboratory, to ensure that the electronic Course Length measurement device is working correctly.

There is a standard statistical procedure for the comparison of two sets of data, which makes use of the concepts of Standard Errors and Confidence Limits. The calculation is carried out in three steps.

1. Calculate the Standard Error of the difference between the two means, SE(d).
2. Calculate a value for Student's t from the difference between the means.
3. Compare the calculated value for t with that shown in the tables for the appropriate number of degrees of freedom.

If M1 and M2 are the two Means (M1 is the larger estimate), s1 and s2 are the corresponding Standard Deviations, and n1 and n2 are the corresponding numbers of observations, then

$SE(d) = \sqrt{[s1^2 / (n1 - 1) + s2^2 / (n2 - 1)]}$

$t = (M1 - M2) / SE(d)$

Degrees of Freedom = $n1 + n2 - 2$

If the calculated value for t is greater than the limit of t corresponding to the 95% Confidence Level, given in the Student's t table for the appropriate degrees of freedom, then the difference between the means is greater than would be expected as a result of normal random variation.

Significance of Differences: An Example

Student's t Test for Significance of Difference

The Stitch Length of each of ten rolls of knitted fabric is measured in the laboratory and compared with the log of the last ten records of stitch length given by the electronic Course Length device, used to set up the knitting machine. The mean values and standard deviations were found to be:

	Mean	s
Machine set-up	0.350	0.0030
Lab test	0.347	0.0025

Is there a real difference in the results, or can the discrepancy be accounted for by normal variation?

1. SE (d) = $\sqrt{[\,0.003 * 0.003 / 9 + 0.0025 * 0.0025 / 9\,]}$ = 0.00130
2. Calculated t = (0.350 - 0.347) / SE (d) = 2.308
3. Tabulated t for 18 degrees of freedom = 2.101

The limit of t at the 95% level of confidence is less than the calculated figure, so the difference is probably a real one. Further comparative data between the two methods must be obtained.

If the disagreement persists, then there is definitely a difference between the two methods and action should be taken to bring them into line.

Standard Deviation

The Standard Deviation, s, is a measure of the variation, or scatter, of the data about the mean value. It is calculated in six steps.

1. Calculate the arithmetic mean

   $M = \sum (X_i) / n$

2. Calculate the deviations by subtracting the Mean from each observation.

   Deviation $= (X_i - M)$

3. Square the deviations

   Squared Deviation $= (X_i - M)^2$

4. Calculate the sum of the squared deviations

   Sum of Squares $= \sum (X_i - \text{Mean})^2$

5. Divide the sum of squares by the number of observations minus 1

   Mean Squared Deviation $= \sum (X_i - M)^2 / (n-1)$

6. Take the square root

   $s = \sqrt{[\,\sum(X_i - M)^2 / (n - 1)\,]}$

Standard Deviation is also known as the Root-Mean-Square-Deviation.

The use of (n - 1) "degrees of freedom" rather than n, the number of observations, gives a more accurate result for small numbers of observations. This is known as Bessel's correction.

For a Normal Distribution approximately 68% of the data will lie within one standard deviation from the mean; 95% will be within plus or minus two standard deviations; 99% will be within plus or minus three standard deviations.

Standard Error of the Mean

For a given estimate of the Arithmetic Mean the accuracy of the estimate will depend on the number of observations. The greater the number of observations, the more accurate is the estimate of the Mean.

The Standard Error, SE, gives an estimate for the reliability of a particular estimate of the Mean (m). It is calculated as the Standard Deviation divided by the square root of the number of observations (n).

$SE(m) = s / \sqrt{(n)}$

Student's t Statistic

The Student's t table supplies a set of numbers that indicate the probability of finding a given value within a certain range of the Standard Deviation at certain confidence levels. It is found in standard statistical tables and is used in calculating Confidence Limits and Accuracy for test data. It is also used in testing the significance of the difference between two independent estimates of the same characteristic.

The value of t depends on the number of Degrees of Freedom, Df, and the level of confidence that is required in the result of a given calculation.

For a single set of data, Df is usually one less than the number of observations, (n - 1). If Df is reasonably large, say more than 25, then t is approximately equal to 2 at the 95% Confidence Level. This is approximately equivalent to saying that 95% of the observations are expected to lie within plus or minus two Standard Deviations from the Mean.

For routine decisions, a confidence level of 95% is generally used. For important decisions, the 99% confidence level might be chosen.

Some examples of the relationship between t and Df at the 95% and 99% Confidence Levels are given below:-

Df	t (95%)	t (99%)
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
6	2.447	3.707
10	2.228	3.169
15	2.131	2.947
25	2.060	2.787