MEASUREMENT AND STATISTICAL TREATMENT OF EMPERICAL DATA

PRECISION, ACCURACY, and ERROR.

Precision refers to the variability among replicate measurements of the same quantity. Consider three determinations of the percentage energy loss in a conversion process, determined by one scientist, to be 2.63, 2.62, and 2.62 per cent, and three results obtained for the same energy loss, by a second scientist, to be 2.60, 2.75, and 2.81 per cent. The results of the first scientist exhibit much less variation among themselves than do those of the second, so the precision of the first set of results is better than that of the second.

Accuracy refers to the difference between a quantities' measured value and the *true* value of the quantity being measured. Strictly speaking, the true values are never known except in counting discrete objects ("there are exactly 22 students in this class") and in defined quantities. All other types of measurements, including mass, length, time, and charge, are actually comparisons to standards, and these comparisons must consist of measurements. So the term accuracy refers to the difference between a measured value and the value which is accepted as the true or correct value of the quantity measured.

The distinction between precision and accuracy may be likened to the result of shooting a series of arrows at an archery target -- *precision* refers to how close together the several arrows hit and accuracy refers to how close to the bull's-eye each lands. It is possible for a replicate series of measurements or determinations to be very precise and yet highly inaccurate. It is, however, quite meaningless to consider the accuracy of a series of values unless the precision is reasonably good. The scientist desires to achieve acceptable precision and accuracy in all of his work and to assess how accurate and precise his work and methods are.

Error. The scientist is continually interested in the cause and the magnitude of errors in his measurements. He examines the quantitative data he obtains not with the question as to whether error is present but rather with the question as to how much error and uncertainty exist. He recognizes that error is always present and that he will not completely eliminate error even though he does continually strive to recognize, to minimize, and to evaluate error in his measurements. Error may be arbitrarily divided into two categories, *systematic* and *random error.*

Systematic error are those one-sided errors which can be traced to a specific source, either in the strategic scheme of the experiment or in the apparatus used to perform it. Such errors can often be minimized by a modified plan of attack. Even when the errors cannot be completely suppressed in this way, an understanding of their origins often makes it possible to deduce a correction factor that can be applied to the final result, or at least to estimate the probable residual error in that result.

When one or more large errors appear to be present, it is frequently possible to discover their origins by a series of carefully controlled experiments in which the

experimental conditions and quantities are varied widely in a systematic way. The resultant error must follow one of three courses: (1) the error may remain relatively constant and independent of the experimental conditions, (2) the magnitude of the error may vary systematically with one or more of the experimental conditions, or (3) the error may persist as a random error.

If the error in a measurement proves to be constant in magnitude, such possibilities as instrument calibration must be considered. If a systematic variation of the error is evident, the parameter linked to his variation frequently indicates the cause. When an apparently random error is encountered it may be a systematic error linked to some experimental condition not yet investigated or controlled. For example, an apparently random error could ultimately prove to be associated with variations in atmospheric humidity, perhaps indicating that a chemical or material is absorbing water during the experiment.

Systematic errors tend to make the observed or calculated values consistently too high or too low. This means that systematic errors can make results highly inaccurate without affecting the precision of replicate results. Good precision does not necessarily mean good accuracy. Varying at least some experimental factors in replicate experiments can minimize the danger of retaining one-sided errors without recognizing their presence. In critical analyses, duplicate sets of samples should be analyzed by entirely different methods since it is unlikely that the same systematic errors would appear to the same extent in entirely different analytical procedures.

Random error. The cause of a random error may or may not be known. Some personal judgment is required in all measurements, such as in reading instrument dials or meters, noting just when a container is filled to a predetermined calibration mark, and so forth; and random inaccuracies are bound to occur. Some random errors arise within the method itself, such as impurity of a supposedly pure material, variations with stirring and with speed of mixing reagents, and so on. Random variations in room temperature and other environmental factors may introduce random error into analytical results.

The scientist can and should minimize random errors insofar as is feasible by careful work, by choice of schemes of analysis which have been or can be proven to be valid, and by keeping environmental factors as constant as possible. However, residual random errors will remain even when all reasonable efforts are made to ensure careful and accurate work.

A statistical probability analysis of random error provides two criteria for the recognition of random errors: (1) small deviations from the correct value are much more frequent than large ones; (2) positive and negative deviations of equal magnitude occur with about the same frequency. These two criteria are expressed graphically in the curve shown in Figure I below, which shows a normal distribution of the errors in an infinitely large number of experimental measurements, all of which are ideally perfect except for random errors. The characteristic distribution of errors, particularly as expressed in criterion, 2, suggests that, if a large number of determinations is made of the same quantity and if the measurement is affected only by random errors, the average of all the values should indicate directly the *correct* value. Even when relatively few measurements are made, the average provides a more reliable estimate of the correct value than does any one of the individual determinations, assuming that only random errors are present. The quantitative treatment of averages and of measures of precision and accuracy will be discussed in the following section with further reference to the normal distribution curve.

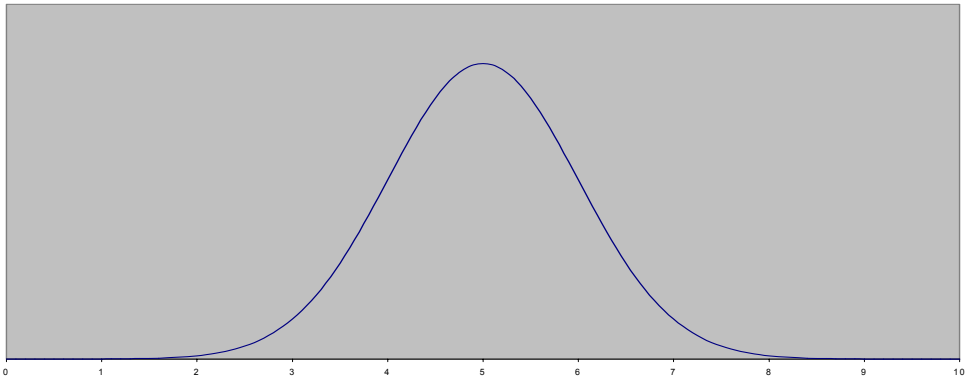

Figure I. Normal distribution curve showing frequency of measurement as a function of measured value. This curve has a mean of 5, and a standard deviation of 1. Note the measured value of maximum frequency (5) and the distance from that value to either inflection points (+1 or -1).

SIGNIFICANT FIGURES, MEASUREMENT, AND UNCERTAINTY.

Significant figures are a way of indicating uncertainty in a measurement. Significant figures are the digits necessary to express the results of a measurement to the precision with which it is made. **The number of significant figures is a count of the number of successively smaller powers of ten on the instrument (finer graduations), that the scientist was able to take advantage of in the measurement**. When using a scale, the usual practice is to estimate between the smallest marks on the scale to the next tenth smaller. This estimate is also considered a significant figure although somewhat uncertain. It is usually assumed that a scientist can estimate between the marks to an accuracy of $\pm$ 1/10 of the distance if the scale is reasonably constructed and the scientist is familiar with reading a scale.

Consider determining the mass of an object, first on a rough balance to the nearest tenth of a gram, and then on an analytical balance to the nearest ten thousandths of a gram. The results of the two being 11.2 g and 11.2169 g, respectively. Three digits are used in expressing the result of the first measurement and six for the second. Any fewer digits could not express the result of the measurements to the precision with which they were made, and no more digits could justifiably be used for either value; therefore, the first mass is expressed in three significant figures and the second in six.

Consider next the measurement of an extremely small number, such as the number of moles of hydrogen ion in 1 liter of pure water at room temperature. This quantity can be measured, and the result could be written as 0.0000001 mole. Eight digits, including the zeros, have been used. However, the same number could be written as $1. \times 10^{-7}$ mole, in which case only one digit has been used exclusive of the exponential factor. Thus, the result of the measurement has only one significant figure no matter which way it is written, because only one digit is necessary to express the results of the measurement to the precision with which it was made. The zeros to the left of the 1 in 0.0000001 and the exponential factor of $\times 10^{-7}$ are used merely to locate the decimal point and do not fit into the definition of significant figures. Zero's leading the first non-zero digit are not significant, regardless of the position of the decimal point. A similar consideration is encountered in measurements of very large numbers. For example, the number of molecules in a mole of any compound can be written as 6.02×10^{23}, and this number contains three significant figures. The exponential factor again serves only to locate the decimal point.

It is important for each person making measurements to express the results of the measurements with the proper number of significant figures. Another scientist or engineer who reads and in any way uses or interprets the results of those measurements can usually tell (and will assume) at a glance how many significant figures are intended. Spreadsheets, such as EXCEL, do not understand significant figures! This job is left to the scientist.

There is a possibility a scientist could be confused about counting the number of significant figures when reading large numbers. For example, a recorded volume of 2000 ml might involve only one significant figure, meaning that the measured value was closer to 2000 than to 1000 or 3000. Alternatively, it could signify the measured quantity to be closer to 2000 than to 2001 or 1999, in which case four significant figures are indicated. Likewise, the number 2000 might intend only two or three figures to be significant. This possible uncertainty can be avoided very simply if the one who makes the measurements in the first place writes it in a exponential form (2×10^3, 2.0×10^3, 2.00×10^3) clearly showing whether he intends one, two, three, or four figures, respectively, to be significant. It is advisable to express the results of measurements in this exponential form whenever there can possibly be any confusion as to whether zeros to the left of the decimal point are significant or not. All measurements should include a decimal point, while a count requires no decimal point.

Absolute Uncertainty and Relative Uncertainty. Uncertainty in measured values may be considered from either of two distinct viewpoints. Absolute uncertainty is the uncertainty expressed directly in units of the measurement. A mass expressed as 10.2 g is presumably valid within a tenth of a gram, so the absolute uncertainty is one tenth of a gram. Similarly, a volume measurement written as 46.26 ml indicates an absolute uncertainty of one hundredth of a milliliter. Absolute uncertainties are expressed in the same units as the quantity being measured -- grams, liters, and so forth.

Relative uncertainty is the uncertainty expressed in terms of the magnitude of the quantity being measured. The mass 10.2 g is valid within one tenth of a gram and the entire quantity represents 102 tenths of a gram, so the relative uncertainty is about one part in 100 parts. The volume written, as 46.16 ml is correct to within one hundredth of a milliliter in 4626 hundredths of a milliliter, so the relative uncertainty is one part in 4626 parts, or about 0.2 part in a thousand. It is customary, but by no means necessary to express relative uncertainties as parts per hundred (per cent), as parts per thousand, or as parts per million. Relative uncertainties do not have dimensions of mass, volume, or the like because a relative uncertainty is simply a ratio between two numbers, both of which are in the same dimensional units.

To distinguish further between absolute and relative uncertainty, consider the results of mass determinations of two different objects on an analytical balance to be 0.0021 g and 0.5432 g. As written, the absolute uncertainty of each number is one ten-thousandth of a gram, yet the relative uncertainties differ widely -- one part in 20 for the first mass and one part in approximately 5000 for the other value.

Significant Figures in Mathematical Operations.

Very seldom is the result of an analytical determination based solely upon one measured value. For example, even the mass determination of a single sample normally requires two mass measurements, one before and one after removing a portion of the sample from a “weighing” bottle. The result of the second mass determination must be subtracted from the first to get the sample mass. Frequently, one measured value must be multiplied or divided by another. The scientist is concerned with significant figures not only in dealing with results of single measurements but also in conjunction with numbers computed mathematically from two or more measured quantities. The arithmetical operations of addition and subtraction may be considered together, as may multiplication and division.

Addition and Subtraction Rule: *Decimal Places, not significant figures, control the precision of the results of the computation. The answer may only contain as many decimal places as is equal to the operand with the fewest number of decimal places.*

The concept is illustrated in the following example:

Mass of bottle plus sample		11.2169 g
Mass of bottle empty	-	10.8114 g
Mass of sample		.4055 g

Each of the quantities measured directly contains six significant figures and four decimal places, but the mass of the sample has only four significant figures and four decimal places.

Now, assume that one mass determination was made less precisely, so that the data are as follows:

Mass of bottle plus sample		11.2169 g
Mass of bottle alone	-	10.81 g
Mass of sample		.41g

The correct mass of the sample is not 0.4069 g but rather 0.41 g. With the decimal points aligned vertically, the computed result has no more decimal places than the number with the least number of decimal places. The mass of the sample has two decimal places, and two significant figures. Note that, with absolute uncertainties of 0.0001 and 0.01 g for the two numbers to be subtracted, the absolute uncertainty of the difference is 0.01 g.

Multiplication and Division Rule: *Significant figures, not decimal places, control the precision of the results of the computation. The answer may only contain as many significant figures as is equal to the operand with the fewest number of significant figures.*

The concept of significant figures in the operations of **multiplying and dividing** must be based upon relative uncertainties. A product or quotient should be expressed with sufficient significant figures to indicate a relative uncertainty comparable to that of the factor with the greatest relative uncertainty. Consider the problem :

$$\begin{array}{r} 9.678234\ \text{n} \\ \times\ 0.12\ \text{m} \\ \hline 1.2\ \text{nm.} \end{array}$$

Expressing this result as, for example, 1.1613 nm would be totally unjustifiable in view of the fact that the relative uncertainty of the second factor is one part in 12.

The rule that the relative error of a product or quotient is dependent upon the relative error of the least accurately known factor suggests the important generalization that, in measuring quantities which must be multiplied or divided to get a final result, it is advantageous to make all the measurements with approximately the same relative error. It is a waste of time to measure one quantity to one part in a hundred thousand if it must subsequently be multiplied by a number which cannot be measured any better than to within one part in a hundred. Similarly, it is advisable to measure quantities, which are to be combined by addition or subtraction to about the same absolute uncertainty. It would be foolish to take pains to measure one mass to a tenth of a milligram if it is to be added to a mass, which for some reason cannot be, measured any closer than to, say, 10 mg.

Again, it should be pointed out that computer programs and spreadsheets similar to EXCEL do not understand significant figures, nor the rules necessary in maintaining the correct number of decimal places and significant figures through arithmetic computations. If using a spreadsheet to perform data analysis, it is the responsibility of the scientist, using the rounding functions, to preserve the appropriate level of uncertainty.

STATISTICAL TREATMENT

Every scientist must develop a working familiarity with a few fundamental statistical concepts. In order to recognize errors and to minimize their effects upon the final result, the scientist must run each determination more than once, usually in triplicate or quadruplicate. Then he must combine the results of these replicate experiments to yield his answer for the determination. Statistical methods are employed in combining and in interpreting these replicate measurements.

Average. *The average is defined as a measure of central tendency of an event*. There are several methods of expressing the central tendency. Mean, median, and mode all estimate the central tendency of the data and can be called the average. Given an infinitely large normal distribution, the mean, median and mode would yield the same value, the *true* value. However, for a non infinite sample size, in the range of about 4 to 500 samples, one method of estimating the average is the simplest and at the same time about the best from a theoretical standpoint. This is the *arithmetic mean*, commonly called by the more general term average. It is obtained by adding the replicate results and dividing by the number of those results.

$$Xmean = \frac{1}{n}\sum_{i=1}^{n} X_i$$

EXCEL provides the arithmetic mean with the statistical function AVERAGE().

Consider the following four results of the determination of the half life of a radio active sample:

22.64 sec
22.54 sec
22.61 sec
+22.53 sec
90.32 sec

The arithmetic mean or average, is 90.32 sec / 4 = 22.58 sec. Note the rules of significant figures have been applied to this computation.

Deviation. The average, as the measure of central tendency, is very important, but it does not in itself indicate all the information, which can be derived from a series of numerical results. The extent of the variations from this average is also of considerable interest. The variation of a single value from the average may be expressed simply as the difference between the two, and this difference is designated the deviation. Thus, if X_1, X_2, and X_3 represent the several numerical values and Xmean represents the arithmetic mean calculated as described in the preceding section, the several deviations (d_1, d_2, and d_3) are:

$d_1 = X_1 - Xmean$
$d_2 = X_2 - Xmean$
$d_3 = X_3 - Xmean$

It is conventional to subtract the arithmetic mean from the specific value, as indicated, and not vice versa. Thus, the deviation is positive if the one experimental value is greater than the arithmetic mean and negative if the arithmetic mean is greater. The algebraic sum of all the deviations in a set must equal zero, at least within the close limits set by rounding off numbers - a consequence of the definition of the arithmetic mean. The individual deviations may be expressed either in absolute units or in relative units. For example, the deviation of the mass 11. g from the mass 10. g is one in absolute units of grams, and it is one part in 10 relative units. The latter may also be expressed as 10 per cent or a 100 parts per thousand.

Standard deviation. The scientist is interested not just in averages and individual deviation values. He also needs a single number whereby he can represent the overall deviation within a series of replicate results. The standard deviation is a measure of the spread of the Normal Distribution curve as seen previously in figure 1. Given an infinite series of replicate results (as described by the normal distribution curve), the standard deviation would be the distance from the mean value to either inflection point value. Based on probability theory, it can be shown that 68% of these replicate results will lie within the bracket:

(Xmean - standard deviation) to (Xmean + standard deviation).

The estimate of the standard deviation of a sample, *s*, is computed as follows:

$$s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - X_{mean})^2}$$

EXCEL provides the standard deviation of a sample with the statistical function: STDEV().

Consider the following time data:

t (sec)

22.64
22.54
22.61
22.53
90.32

t_{mean} = 90.32/4 =22.58 sec

s = 0.05 sec.
Note again that the rules governing significant figures were employed in this calculation.

Confidence Limits. In order for us to recognize more fully the true significance of the arithmetic mean and the standard deviation, we must refer again to the curve of Figure 1. This curve, which may be derived mathematically, represents the normal distribution of the errors or deviations in an infinitely large sample size, which is ideally perfect except for random errors. It has already been pointed out that small deviations are much more frequent than large ones. This latter statement may be made quantitative with the use of the standard deviation. The mathematical treatment from which the normal distribution curve is derived reveals, for example, that 68 per cent of the individual deviations are less than the standard deviation, that 95 per cent are less than twice the standard deviation, and that 99 per cent are less than 2.5 times the standard deviation. In other words, 68 per cent of the X values fall within the range of Xmean $\pm$ *s*, 95 per cent within the range Xmean $\pm$ 2*s*, and 99 per cent within the range Xmean $\pm$ 2.5*s*.

Data, which can be interpreted strictly in terms of the normal distribution curve or on its mathematical origins, do not generally arise in most analytical situations. There are two reasons for this fact: the derivation specifies random errors only, whereas many analytical data are influenced by one-sided, systematic errors as well; the derivation specifies a large sample size (actually an infinite number) whereas only relatively small sample sizes are feasible in practical situations. Because of the first reason, one-sided, systematic errors must be eliminated before the concept of confidence limits (to be described shortly) can become applicable. As a consequence of the second reason, the scientist can never know with absolute certainty whether his arithmetic mean is the absolutely correct value unless he does run an extremely large number of determinations.

Even with a few determinations, however, he can specify a range of values centered upon his arithmetic mean and then state that there is a 50-50 chance, 95 chances out of 100, 99 chances out of 100, or any other desired probability that the true value does lie within that range. That is, he can know and specify the probability that the true answer lies within a given range, and he can indicate that range using the arithmetic mean and the standard deviation. That range is designated as the confidence limit, and the likelihood that the true value lies within that range is designated the probability. The probability is conveniently expressed in percentage units.

When a sample size is less than infinite, the normal distribution curve is not used. The appropriate distribution curve is the Students' – t distribution. Rather than attempting to explain the use of the t-distribution, Table I has been prepared so the student may select the appropriate "fudge factor" for which to multiply his standard deviation in order to adjust it for a sample size of "n" at a selected percent probability. The sample size is represented by n, and f_{50}, f_{95}, and f_{99} are the *factors* by which the standard deviation of an individual result must be multiplied to yield the confidence limits for 50, 95, and 99 per cent probability, respectively, in the form Xmean $\pm$ f *s*. Thus, the scientist may conclude that the true value lies within the range Xmean $\pm$ f_{50} *s* and he has a 50-50 change of being correct, or he may state the true value lies within the range Xmean $\pm$ f_{95}*s* and be 95 per cent certain of being correct.

TABLE I. FACTORS FOR CALCULATING CONFIDENCE LIMITS

n	f_{50}	f_{95}	f_{99}
2	0.7071	8.9846	45.0115
3	0.4714	2.4841	5.7302
4	0.3824	1.5912	2.9204
5	0.3312	1.2417	2.0590
6	0.2967	1.0494	1.6461
7	0.2712	0.9248	1.4013
8	0.2514	0.8360	1.2373
9	0.2355	0.7687	1.1185
10	0.2222	0.7154	1.0277
11	0.2110	0.6718	0.9556
12	0.2013	0.6354	0.8966
13	0.1929	0.6043	0.8472
14	0.1854	0.5774	0.8051
15	0.1788	0.5538	0.7686
16	0.1728	0.5329	0.7367
17	0.1674	0.5142	0.7084
18	0.1624	0.4973	0.6831
19	0.1579	0.4820	0.6604
20	0.1538	0.4680	0.6397

{ f factors above computed from 2 tail t distribution: $f = t / (n)^{1/2}$ }

Note that the "f" values in Table I , 95% probability, may be computed using EXCEL as follows:

$$f_{95} = \text{TINV}(1-.95,N-1)/\text{SQRT}(N)$$

Given n individual X values for calculating the average, Xmean, the true value may be expected to lie within the range Xmean $\pm f_\alpha s$ with a % probability as indicated by the f subscript, α.

The use of Table I may be illustrated by the following example. Four results of a coefficient of friction determination yielded an arithmetic mean (Xmean) of .231 with a standard deviation (*s*) of 0.050 . From Table I, for n = 4, f_{50} is 0.3824; so there is a 50-50 likelihood that the true value lies within the range .231 $\pm$ (0.3824 x 0.050), or .231 $\pm$ 0.019 . Similarly, there is a 95 per cent probability that the true value lies within the range .231 $\pm$ .080, and a 99 per cent probability that it is .231 $\pm$ .15 . It is clear from this example, and from the table, that the limits must be widened as the required probability of being correct is increased. It is also evident from the table that the importance of each additional trial beyond three or four diminishes as the total number n increases. These factors are in keeping with common sense - statistical concepts should be considered as a means of putting common sense on a quantitative foundation, but not as a substitute for common sense itself.

The probability value used in expressing the results of an analytical determination is quite arbitrary. In any case, the probability chosen should be stated or otherwise indicated. Probabilities of 95 and 99 per cent are most commonly employed in analytical work, whereas a 50 per cent probability is also

useful in student work. Therefore, 50, 95, and 99 per cent data are included in Table I, although other probabilities could be used and occasionally are.

Rejection of an Observation.

Every scientist is occasionally confronted with a series of results of replicate determinations, one of which appears to be far out of line with the others. Even experienced scientists encounter the same situation. Consider the series of results:

22.64 sec.
22.54 sec.
22.22 sec.
22.69 sec.

The third value appears to be out of line. If this third determination were subject to an obvious large one-sided error the result could immediately be rejected prior to computing the arithmetic mean and confidence limits. However, in a small series of data such as this, all four values could be valid for ascertaining the arithmetic mean. The beginning student is perhaps too greatly inclined to discard a datum which does not seem to agree with the body of his measurements, so it is apparent that some standard criterion for such rejection is necessary.

Any value may be rejected if a particular reason for its inaccuracy is known. If it is known that part of a material was spilled or that a container leaked, that result may be discarded at once. This is systematic error. Other times, the scientist may suspect that a systematic error may have arisen in one sample but he may not be certain. If such is the case, he should include that sample and then discard the result if it appears particularly erroneous in the proper direction. If no experimental reason for rejection is known but a value still appears out of line, some statistical test must be employed before deciding whether to reject an observation. One such test developed by Dean and Dixon is explored here.

Dean and Dixon's rejection test is based upon the differences between the highest and lowest values as calculated both with and without the suspicious value. Let R1 be the difference between the highest and lowest values (Range) with all values included, and let R2 be the difference between the highest and lowest values excluding the suspicious one***. If the ratio R1/R2 exceeds the critical value listed for the appropriate n number in Table II, the suspected observation should be rejected;*** otherwise, it should be retained. For each n value, there are two critical R1/R2 ratios listed in Table II: one for the 95 per cent probability level and one for the 99 per cent probability level. When the 95 per cent column is used, the chance of an extreme value being rejected when it should have been retained is 5 per cent, whereas there is only a 1 per cent chance that a value rejected on the basis of the 99 per cent column should really have been retained.

TABLE II. FACTORS FOR RETENTION OF REJECTION OF EXTREME VALUES

Critical Values of R1/R2

N	Ratio at 95 % probability	Ratio at 99 % probability
3	16.9	83.3
4	4.3	9.0
5	2.8	4.6
6	2.3	3.3
8	1.9	2.4
10	1.7	2.1

{R.B. Dean and W.J. Dixon: Simplified statistics for small numbers of observations. Anal. Chem. 23,636 (1951)}

Consider again these four values:

22.64 sec
22.54 sec
22.22 sec
22.69 sec

R1 is (22.69 - 22.22) or 0.47
R2 is (22.69 - 22.54) or 0.15
The ratio R1/R2 is 0.47/0.15 or 3.1
The computed ratio (3.1) is less than the critical values from Table II (4.3 @95% and 9.0 @99%, for n=4), so the value should be retained.

Consider next these four values: 22.64, 22.69, 22.65, and 22.22. Here, R1 is 0.47 and R2 is 0.05 excluding the 22.22 value.
The computed R1/R2 is 9.4, which is greater than the critical value from Table II at 99% , so the 22.22 value is rejected.

It is suggested that the 99 per cent probability column of Table II be employed in student work in deciding whether to reject an observation, unless your professor instructs you differently. In any case, you should recognize, even quantitatively, what the residual chances are that a rejected value should have been retained.

If more than one value is doubtful, this test can be repeated after the most extreme value has been rejected. This is not ordinarily recommended, however. If two values were doubtful in a series of only four or so, it would be much better to repeat the whole experiment to obtain more values. It should be noted that the effect upon the arithmetic mean of one or even two discordant values is relatively less significant when there are many values than when there are only a few.

Comparing Averages.

Consider the example of some determinations of the density of nitrogen, which were performed in the laboratory of Lord Rayleigh in 1894. Batches of nitrogen were prepared by various means from the chemical compounds NO, N_2O, and NH_4NO_2 and also from dry, carbon dioxide-free air by several methods of removing oxygen. Measurements of the mass of nitrogen required to fill a certain flask under specified conditions revealed for that 10 batches of "chemical nitrogen" an arithmetic mean of 2.29971 g, and for nine batches of "atmospheric nitrogen" an arithmetic mean of 2.31022 g. The overall standard deviation within each group can be considered to be about 0.00030. A question arose: “Was there a significant difference between the two averages?” That is, was the density of the "chemical nitrogen" the same as that of the "atmospheric nitrogen" or, more basically, was nitrogen from both sources the same?

We can answer this question on the basis of the t-test for comparing averages. This test will be presented empirically here along with recognition of its statistical validity. The quantity t is defined as

$$t = \frac{|X_1\text{mean} - X_2\text{mean}|}{S_p}\sqrt{\frac{n_1 * n_2}{n_1 + n_2}}$$

in which $\mathbf{X_1mean}$ and $\mathbf{X_2mean}$ are the two averages, $\mathbf{n_1}$ and $\mathbf{n_2}$ are the number of individual values averaged to obtain $\mathbf{X_1mean}$ and $\mathbf{X_2mean}$ respectively, and $\boldsymbol{Sp}$ is the common (or pooled) standard deviation.

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Critical t values at 95 and 99 per cent probability levels are listed in Table III. The phrase degrees of freedom is a common statistical term which is simply $n_1 + n_2 - 2$ in this application. If an observed or calculated t exceeds the indicated critical t value, the chances are 95 out of 100 or 99 out of 100 (depending upon which critical t value of Table III is used) that the averages are significantly different.

TABLE III. CRITICAL t VALUES FOR COMPARISON OF AVERAGES

Critical t Value at 95% and 99 % probability level (2 tail)

D.F.	95%	99%
1	12.706	63.656
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
7	2.365	3.499
8	2.306	3.355
9	2.262	3.250
10	2.228	3.169
11	2.201	3.106
12	2.179	3.055
13	2.160	3.012
14	2.145	2.977
15	2.131	2.947
16	2.120	2.921
17	2.110	2.898
18	2.101	2.878
19	2.093	2.861
20	2.086	2.845

{D.F., degrees of freedom, is $n_1 + n_2 - 2$.}

The above table values can be computed for any probability and degrees of freedom using EXCEL. The EXCEL function: TINV(1.0-.95,10) , will yield 2.228, the "t" value for 10 degrees of freedom at 95% probability.

The t value for the nitrogen data may be calculated as follows:

$$t = \frac{|X_1\text{mean} - X_2\text{mean}|}{S_p}\sqrt{\frac{n_1 * n_2}{n_1 + n_2}}$$

$$t = \frac{(2.31022 - 2.29971)*[10*9/(10+9)]^{1/2}}{.00030}$$

$$t = 76$$

From Table III, $t_{computed} > t_{critical}$, thus the averages are different.

$$76 > t_{(99\%,df=17)} = 2.898$$

So the "chemical nitrogen" and the "atmospheric nitrogen" are almost certainly different. Lord Rayleigh, employing a somewhat different but comparable statistical test, recognized this difference; and this fact led directly to the discovery shortly thereafter of the so-called inert gases in the atmosphere!

Consider four gravimetric determinations of chloride in a particular sample yielding the arithmetic mean, 20.44% Cl, and four volumetric determinations of chloride in the same sample yielding the arithmetic mean 20.54 per cent Cl, both with standard deviations of about 0.08. Are the results of the gavimetric and volumetric methods significantly different?

The t test provides the answer.

$$t = \frac{|X_1\text{mean} - X_2\text{mean}|}{S_p}\sqrt{\frac{n_1 * n_2}{n_1 + n_2}}$$

$$t = \frac{(20.54-20.44)}{.08} * [4*4/(4+4)]^{1/2}$$

$$t = 1.77$$

The critical t values for D.F. = 6 ($n_1 + n_2 - 2$) are 2.5 and 3.7 at the two listed probability levels, we may not conclude that the two averages are significantly different.