Chapter 1

Measurements and Errors in measurement

This chapter introduces measurement errors and methods of describing them so that measured data is interpreted properly. Statistical principles involved in error analysis are discussed in sufficient detail. Concepts such as precision and accuracy are clearly explained. Different statistics useful in experimental studies are discussed. Tests of normality of error distribution and procedure for rejection of data are discussed. Results from sampling theory are discussed because of their use in the interpretation of sparse experimental data that is almost the rule.

1.1 Introduction

We recognize three reasons for making measurements as indicated in Figure 1.1. From the point of view of the present book measurements for commerce is outside its scope. Engineers design physical systems in the form of machines to serve some specified functions. The behavior of the parts of the machine during the operation of the machine needs to be examined or analyzed or designed such that it functions reliably. Such an activity needs data regarding the machine parts in terms of material properties. These are obtained by performing measurements in the laboratory.

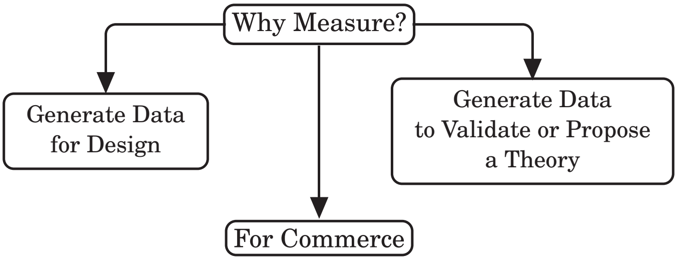


Figure 1.1: *Why make measurements?*

The scientific method consists in the study of nature to understand the way it works. Science proposes hypotheses or theories based on observations and these need to be validated with carefully performed experiments that use many measurements. When once a theory has been established it may be used to make predictions which may themselves be confirmed by further experiments.

1.1.1 Measurement categories

Broadly measurements may be categorized as given below.

- **Primary quantity**: It is possible that a single quantity that is directly measurable is of interest. An example is the measurement of the diameter of a cylindrical specimen. It is directly measured using an instrument such as vernier calipers. We shall refer to such a quantity as a primary quantity.
- **Derived quantity**: There are occasions when a quantity of interest is not directly measurable by a single measurement process. The quantity of interest needs to be estimated by using an appropriate relation involving *several* measured primary quantities. The measured quantity is thus a derived quantity. An example of a derived quantity is the determination of acceleration due to gravity (g) by finding the period (T) of a simple pendulum of length (L). T and L are the measured primary quantities while g is the derived quantity.
- **Probe or intrusive method**: It is common to place a *probe* inside a system to measure a physical quantity that is characteristic of the system. Since a probe invariably affects the measured quantity the measurement process is referred to as an intrusive type of measurement.

- **Non-intrusive method**: When the measurement process does not involve insertion of a probe into the system the method is referred to as being non-intrusive. Methods that use some naturally occurring process, like radiation emitted by a body to measure a desired quantity relating to the system, may be considered as non-intrusive. The measurement process may be assumed to be non-intrusive when the probe has negligible interaction with the system. A typical example for such a process is the use of Laser Doppler Velocimeter (LDV) to measure the velocity of a flowing fluid.

1.1.2 General measurement scheme

Figure 1.2 shows the schematic of a general measurement process. Not all the elements shown in Figure 1.2 may be present in a particular case.

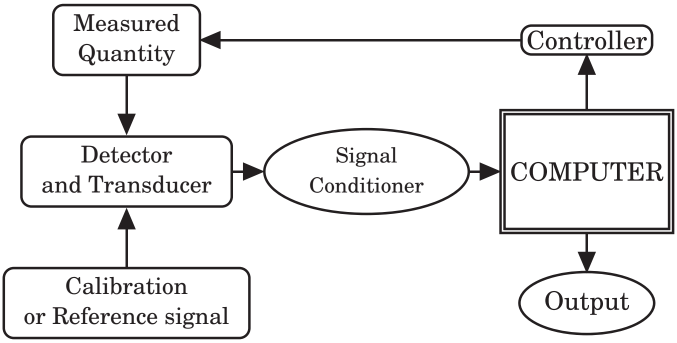


Figure 1.2: *Schematic of a general measurement system with controller*

The measurement process requires invariably a detector that responds to the measured quantity by producing a measurable change in some property of the detector. The change in the property of the detector is converted to a measurable output that may be either mechanical movement of a pointer over a scale or an electrical output that may be measured using an appropriate electrical circuit. This action of converting the measured quantity to a different form of output is done by a transducer. The output may be manipulated by a signal conditioner before it is recorded or stored in a computer. If the measurement process is part of a control application the computer can be used to drive the controller. The relationship that exists between the measured quantity and the output of the transducer may be obtained by calibration or by comparison with a reference value. The measurement system requires external power for its operation.

1.1.3 Some issues

- Errors - *Systematic or Random*
- Repeatability
- Calibration and Standards
- Linearity or Linearization

Any measurement, however carefully conducted, is subject to measurement errors. These errors make it difficult to ascertain the true value of the measured quantity. The nature of the error may be ascertained by repeating the measurement a number of times and looking at the spread of the values. If the spread in the data is small the measurement is repeatable and may be termed as being good. If we compare the measured quantity obtained by the use of any instrument and compare it with that obtained by a standardized instrument the two may show different performances as far as the repeatability is concerned. If we add or subtract a certain correction to make the two instruments give data with similar spread the correction is said to constitute a systematic error. Then the spread of data from each of the instruments will constitute random error.

The process of determining the systematic error is calibration. The response of a detector to the variation in the measured quantity may be linear or non-linear. In the past the tendency was to look for a linear response as the desired response. Even when the response of the detector was non-linear the practice was to make the response linear by suitable manipulation. With the advent of automatic recording of data using computers this practice is not necessary since software can take care of this aspect during the post-processing of the data.

1.2 Errors in measurement

Errors accompany any measurement, however well it has been conducted. The error may be inherent in the measurement process or it may be induced due to variations in the way the experiment is conducted. The errors may be classified as systematic errors and random errors.

1.2.1 Systematic errors *(Bias)*

Systematic error or bias is due to faulty or improperly calibrated instruments. These may be reduced or eliminated by careful choice and calibration of instruments. Sometimes bias may be linked to a specific cause and estimated by analysis. In such a case a correction may be applied to eliminate or reduce bias.

Bias is an indication of the accuracy of the measurement. Smaller the bias more accurate the data.

1.2.2 Random errors

Random errors are due to non-specific causes like natural disturbances that may occur during the measurement process. These cannot be eliminated. The magnitude of the spread in the data due to the presence of random errors is a measure of the precision of the data. Smaller the random error more precise is the data. Random errors are statistical in nature. These may be characterized by statistical analysis.

We shall explain these through the familiar example shown in Figure 1.3. Three different individuals with different skill levels are allowed to complete a round of target[1] practice. The outcome of the event is shown in the figure.

[1] Target shown in Figure 1.3 is non-standard and for purposes of illustration only. Standard targets are marked with 10 evenly spaced concentric rings.

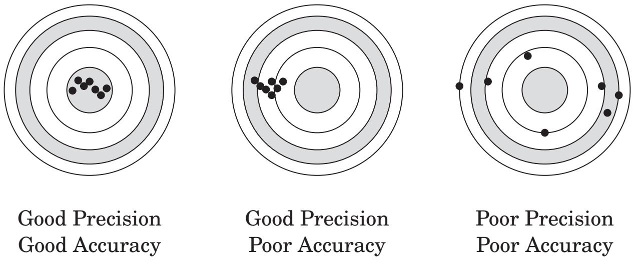


Figure 1.3: *Precision and accuracy explained through a familiar example*

It is evident that the target at the left belongs to a highly skilled shooter. This is characterized by all the shots in the inner most circle or the 'bull's eye'. The result indicates good accuracy as well as good precision. A measurement made well must be like this! The individual in the middle is precise but not accurate. Maybe it is due to a faulty bore of the gun. The individual at the right is an unskilled person who is behind on both counts. Most beginners fall into this category. The analogy is quite realistic since most students performing a measurement in the laboratory may be put into one of the three categories. A good experimentalist has to work hard to excel at it! The results shown in Figure 1.4 compare the response of an individual thermocouple (that measures temperature) and a standard thermocouple. The measurements are reported between room temperature (close to) $20°C$ and an upper limit of $500°C$. That there is a systematic variation between the two is clear from the figure that shows the trend of the measured temperatures indicated by the individual thermocouple. The systematic error appears to vary with the temperature. The data points indicated by the full symbols appear also to hug the trend line (we look at in detail at trend lines while discussing regression analysis of data in Section 2.2), which is a linear fit to the data. However the data points do not *lie* on it. This is due to random errors that are always present in any measurement. Actually the standard thermocouple would also have random errors that are *not* indicated in the figure. We have deliberately shown only the trend line for the standard thermocouple to avoid cluttering up the graph.

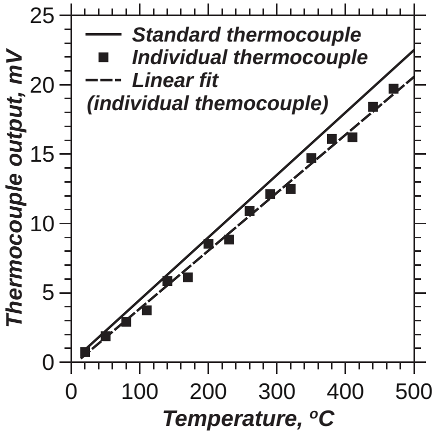


Figure 1.4: *Illustration of presence of systematic and random errors in data*

1.3 Statistical analysis of experimental data

1.3.1 Statistical analysis and best estimate from replicate data

Let a certain quantity x be measured repeatedly to get

$$x_i, \quad i = 1, n \tag{1.1}$$

Because of random errors these *will* all be different. How do we find the best estimate x_b for the true value of x? It is reasonable to assume that the best value be such that the measurements are *as precise* as they can be! In other words, the experimenter is confident that he has conducted the measurements with the best care and he is like the skilled shooter in the target practice example presented earlier! Thus, we minimize the variance with respect to the best estimate x_b of x. Thus we minimize:

$$S = \sum_{i=1}^{n} [x_i - x_b]^2 \tag{1.2}$$

This requires that

$$\frac{\partial S}{\partial x_b} = 2\sum_{i=1}^{n} [x_i - x_b](-1) = 0 \tag{1.3}$$

or

$$\boxed{x_b = \frac{\sum_{i=1}^{n} x_i}{n}} \tag{1.4}$$

The best estimate is thus nothing but the mean of all the individual measurements!

1.3.2 Error distribution

When a quantity is measured repeatedly it is expected that it will be *randomly* distributed around the best value. The random errors *may* be distributed as a normal distribution. If μ and σ are respectively the mean and the standard deviation,[2] then, the normal probability density is given by

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^2} \quad \text{or} \quad f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}}} \tag{1.5}$$

The distribution is also represented some times as $N(\mu, \sigma)$. Normal distribution is also referred to as Gaussian distribution [3] or "bell shaped curve". The probability that the error around the mean is $(x - \mu\ )$ is the area under the probability density

[2] The term standard deviation was first used by Karl Pearson, 1857-1936, an English mathematician, who has made outstanding contributions to the discipline of mathematical statistics

[3] Named after Johann Carl Friedrich Gauss, 1777-1855, a German mathematician and physical scientist

function between $(x-\mu)-\frac{dx}{2}$ and $(x-\mu)+\frac{dx}{2}$ represented by the product of the probability density and dx. The probability that the error is anywhere between $-\infty$ and x is thus given by the following integral:

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^2} dx \quad \text{or} \quad F(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{z^2}{2}} dz \tag{1.6}$$

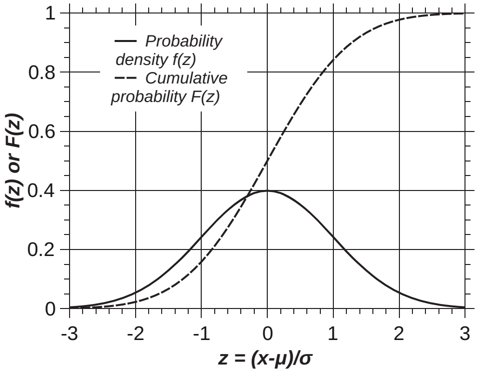


Figure 1.5: Normal probability density function and the cumulative probability

This is referred to as the cumulative probability. It is noted that if $x \to \infty$ the integral tends to 1. Thus the probability that the error is of all possible magnitudes (between $-\infty$ and $+\infty$) is unity! The integral is symmetric with respect to $z=0$ where $z=\frac{x-\mu}{\sigma}$, as is easily verified. The above integral is in fact the error integral that is a tabulated function. A plot showing $f(z)$ and $F(z)$ with respect to z is given in Figure 1.5.

Many times we are interested in finding out the chances of error lying between two values in the form $\pm p\sigma$ around the mean or $\pm z$. This is referred to as the "confidence interval" and the corresponding cumulative probability specifies the chances of the error occurring within the confidence interval. Table 1.1 gives the confidence intervals that are useful in practice. A more complete table of confidence intervals is given in Appendix B as Table B.4.

Table 1.1: Confidence intervals according to normal distribution

p	0	±1	±2	±1.96	3	±2.58	±3.30
CP	0	0.6827	0.9545	**0.95**	0.9973	**0.99**	**0.999**

CP = Cumulative Probability

Example 1.1

A certain measurement gave the value of C, the specific heat of water, as 4200 J/kg^oC. The precision of measurement is specified by the standard deviation given by 25 J/kg^oC. If the measurement is repeated what is the probability that the value is within 4200 ± 35 J/kg^oC? You may assume that the error is normally distributed.

Solution :

The maximum and minimum values that are expected for the specific heat of water are given by

$$C_{min} = 4200 - 35 = 4165, \quad \text{and} \quad C_{max} = 4200 + 35 = 4235\ J/kg^oC$$

Scaling the values in terms of σ the spread of the readings should be $\pm\frac{35}{25}\sigma = \pm 1.4\sigma$ with respect to the mean or $z = \pm 1.4$. Cumulative probability required is nothing but the cumulative probability of $N(0,1)$ between -1.4 and $+1.4$. This is obtained from Table B.4 of Appendix B as $0.8385 \approx 0.84$. Thus roughly 84% of the time we should get the value of specific heat of water between 4165 and 4235 J/kg^oC.

Gaussian distribution: Consider an experiment where the outcome is either a success $s = 1$ or a failure $f = 0$. Assuming that probability of a success is p, we expect the value of successes in n trials to be $nps + n(1-p)f = np$. For example, if the number of experiments is 8, it is likely that 4 experiments will show success as the outcome, assuming that the chances of success or failure are the same i.e. $p = 0.5$. Since the outcome is countable in terms of number of successes or failures, the probability $B(k)$ of a certain number k of successes ($0 \le k \le n$), when n number of experiments have been performed, is given by the product of binomial coefficient $\frac{n!}{k!(n-k)!}$ (this represents the number of combinations that yield k successes) and $p^k(1-p)^{n-k}$. For example, if $n = 8$ and $p = 0.5$ we have $0.5^k(1-0.5)^{8-k} = 0.5^8 = \frac{1}{256}$ where the denominator 256 represents total number of possible ways of getting all possible combinations of outcomes when 8 experiments are performed. Thus we have the binomial distribution function

$$B(k;n,p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

where n and p are indicated as parameters, by placing them after a semicolon following k. The expression given above is one of the terms in the binomial $[p+(1-p)]^n$. In the special case $p = \frac{1}{2}$ of interest to us (we are expecting equal likelihood of positive and negative errors in measurements) this becomes

$$B(k;n,0.5) = \frac{n!}{2^n k!(n-k)!}$$

Binomial distribution is a function of a discrete variable k and satisfies the requirement that $\sum_{k=0}^{k=n} B(k;n,0.5) = 1$. Mean value of k can be obtained as

$$\bar{k} = \sum_{k=0}^{k=n} kB(k;n,0.5) = 0.5n$$

The variance may be shown to be given by

$$\sigma_k^2 = \sum_{k=0}^{k=n}(k-\bar{k})^2 B(k;n,0.5) = 0.25n = \frac{\bar{k}}{2}$$

For sufficiently large n the Binomial distribution is closely approximated by the Normal distribution or the Gaussian curve. Thus we have

$$B(k;n,0.5) \approx N(\mu,\sigma) = N\left(\bar{k}, \sqrt{\frac{\bar{k}}{2}}\right)$$

For example, with $n = 30$ we have $\mu = 15$, $\sigma = 2.739$, the Binomial distribution very closely resembles the Normal distribution as shown in Figure 1.6. Note that the Normal distribution function is a function of continuous variable $\frac{k-\bar{k}}{\sigma_k}$ when $n >> 1$.

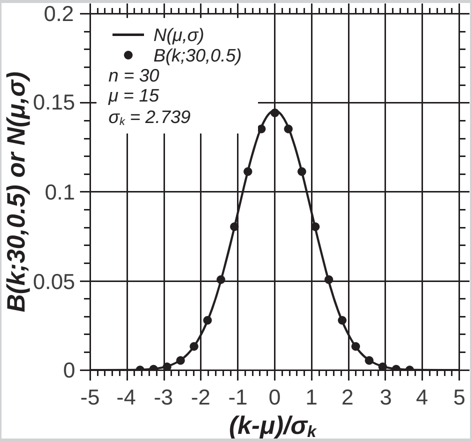


Figure 1.6: Comparison of Binomial and Normal distributions for $n = 30$

1.3.3 Principle of Least Squares

Earlier we have dealt with the method of obtaining the best estimate from replicate data based on minimization of variance. No mathematical proof was given as a basis for this. We shall now look at the above afresh, in the light of the fact that the errors are distributed normally, as has been made out above.

Consider a set of replicate data x_i. Let the best estimate for the measured quantity be x_b. The probability for a certain value x_i within the interval $x_i - \frac{dx_i}{2}$, $x_i + \frac{dx_i}{2}$ to occur in the measured data is given by the relation

$$p(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - x_b)^2}{2\sigma^2}} dx_i \tag{1.7}$$

The probability that the particular values of measured data are obtained in replicate measurements must be the compound probability given by the product of the individual probabilities. Thus

$$p = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i - x_b)^2}{2\sigma^2}} dx_i = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\Sigma_{i=1}^{n}\frac{(x_i - x_b)^2}{2\sigma^2}} \prod_{i=1}^{n} dx_i \tag{1.8}$$

The reason the set of data was obtained as replicate data is that it was the most probable! Since the intervals dx_i are arbitrary, the above will have to be maximized by the proper choice of x_b and σ such that the exponential factor is a maximum. Thus we have to choose x_b and σ such that

$$p' = \left(\frac{1}{\sigma}\right)^n e^{-\sum_{i=1}^{n}\frac{(x_i - x_b)^2}{2\sigma^2}} \tag{1.9}$$

has the largest possible value. As usual we set the derivatives $\frac{\partial p'}{\partial x_b} = \frac{\partial p'}{\partial \sigma} = 0$ to get the values of the two parameters x_b and σ . We have:

$$\frac{\partial p'}{\partial x_b} = \frac{1}{2}\left(\frac{1}{\sigma}\right)^{n+2} e^{-\sum_{i=1}^{n}\frac{(x_i - x_b)^2}{2\sigma^2}} \cdot \underline{\sum_{i=1}^{n} 2(x_i - x_b)\cdot(-1)} = 0 \tag{1.10}$$

The term shown with underline in Equation 1.10 should go to zero. Hence we should have

$$\sum_{i=1}^{n}(x_i - x_b) = 0 \quad \text{or} \quad x_b = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \tag{1.11}$$

It is thus clear that the best value is nothing but the mean value! We also have:

$$\frac{\partial p'}{\partial \sigma} = \underline{\left[-\frac{n}{\sigma^{n+1}} + \frac{1}{\sigma^{n+3}}\sum_{i=1}^{n}(x_i - x_b)^2\right]} e^{-\sum_{i=1}^{n}\frac{(x_i - x_b)^2}{2\sigma^2}} = 0 \tag{1.12}$$

or, again setting the underlined term to zero, we have

$$\boxed{\sigma^2 = \frac{\sum(x_i - x_b)^2}{n}} \tag{1.13}$$

Equation1.13 indicates that the parameter σ^2 is nothing but the variance of the data with respect to the mean! Thus the best value of the measured quantity and its spread are based on the minimization of the squares of errors with respect to the mean. This embodies what is referred to as the “Principle of Least Squares”. We shall be making use of this principle while considering regression.

1.3.4 Error estimation - single sample

In practice measurements are expensive in terms of cost and time. Hence it is seldom possible to record a large number of replicate data. In spite of this one would like to make generalizations regarding the parameters that describe the population from the parameters like mean and variance that characterize the sample. Population represents the totality of experiments that could have been conducted had we the resources to do it. For example, if we would like to study the characteristics of student performance under a schooling system, all the students in the schooling system would make up the population. In order to reduce the cost a statistician would draw a random sample of students, characterize this sample, and extrapolate by statistical theory to get parameters that describe the entire population. In engineering we may want to estimate a physical quantity based on a small number of experiments. The problem is not trivial and hence is discussed in some detail here.

Variance of the means

The problem is posed as given below:

- Replicate data is collected with n measurements in a set or a sample
- Several (possibly) such sets of data are collected
- Sample mean is m_s and sample variance is σ_s^2
- What is the mean and variance of the mean of all samples?

Population mean:

Let N be the total number of data in the entire population, if indeed, a large number of samples have been collected. Without loss of generality we assume that the population mean is zero. Hence we have

$$m = \sum_{i=1}^{N} \frac{x_i}{N} = 0 \tag{1.14}$$

Consider sample s whose members are identified as $x_{k,s}$ with $1 \le k \le n$. Mean of the sample then is

$$m_s = \sum_{k=1}^{N} \frac{x_{k,s}}{n} \tag{1.15}$$

The number of samples n_s each comprising n data, drawn out of the population N is given by $n_s =^N C_n$.[4] Mean of all sample means - $\bar{m}_s$ - is then given by

$$\bar{m}_s = \sum_{s=1}^{^NC_n} \frac{m_s}{^NC_n} \tag{1.16}$$

A particular data x_i will occur in $^{N-1}C_{n-1}$ samples as may be easily seen. Hence the summation in the above equation may be written as

$$\bar{m}_s = \sum_{i=1}^{N} \frac{^{N-1}C_{n-1} x_i}{n^N C_n} = \sum_{i=1}^{N} \frac{x_i}{N} = m = 0 \tag{1.17}$$

Thus the mean of all samples is also the population mean.

Population variance:

The population variance is given by

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - m)^2}{N} = \sum_{i=1}^{N} \frac{x_i^2}{N} \tag{1.18}$$

since the population mean is zero.

[4] This result is analogous to filling n bins with one object each, drawing objects from a container with N objects, without replacing them. Number of samples is just the number of ways the bins may be filled.

Variance of the means:

Let the variance of the means be σ_m^2. By definition we then have

$$ {}^{N}C_n \sigma_m^2 = \sum_{s=1}^{{}^{N}C_n} (m_s - \bar{m})^2 = \sum_{s=1}^{{}^{N}C_n} m_s^2 \tag{1.19}$$

Using Equation 1.15 we have, for the s^{th} sample

$$ m_s^2 = \left[\sum_{k=1}^{n} \frac{x_{k,s}}{n} \right]^2 = \frac{(x_1 + x_2 + ... + x_n)_s^2}{n^2} $$

On expanding squares, the right hand will contain terms such as x_l^2 and $2x_l x_m$ with $l \neq m$. Both l and m are bounded between 1 and n. Hence the summation over s indicated in Equation 1.19 will have x_l^2 appearing ${}^{N-1}C_{n-1}$ times and $2x_l x_m$ appearing ${}^{N-2}C_{n-2}$ times. Hence we have

$$ {}^{N}C_n \sigma_m^2 = \frac{1}{n^2} \left[\sum_{i=1}^{N} {}^{N-1}C_{n-1} x_i^2 + 2 \underbrace{\sum_{i=1}^{N} \sum_{j=1}^{N} {}^{N-2}C_{n-2} x_i x_j}_{i \neq j} \right] = 0 \tag{1.20}$$

However, since the population mean is zero, we have

$$ \left(\sum_{i=1}^{N} x_i \right)^2 = \sum_{i=1}^{N} x_i^2 + 2 \underbrace{\sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j}_{i \neq j} = 0 $$

Hence we have $2 \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j = - \sum_{i=1}^{N} x_i^2$. Substitute this in Equation 1.20 and simplify to get

$$ \sigma_m^2 = \frac{1}{n^2} \underbrace{\left(\frac{{}^{N-1}C_{n-1}}{{}^{N}C_n} \right)}_{=\frac{n}{N}} \underbrace{\left[1 - \left(\frac{{}^{N-2}C_{n-2}}{{}^{N-1}C_{n-1}} \right) \right]}_{=1-\frac{n-1}{N-1}} \sum_{i=1}^{N} x_i^2 = \frac{(N-n)}{n(N-1)} \sigma^2 \tag{1.21}$$

If $n << N$ and N is large, the above relation may be approximated as

$$ \boxed{\sigma_m^2 = \frac{1 - \frac{n}{N}}{n\left(1 - \frac{1}{N}\right)} \sigma^2 \approx \frac{\sigma^2}{n}} \tag{1.22}$$

Estimate of variance:

- Sample variance - how is it related to the population variance?
- Let the sample variance from its own mean m_s be σ_s^2
- i.e. $\sigma_s^2 = \frac{1}{n} \sum_{i=1}^{n} x_{i,s}^2 - m_s^2$

The mean of all the sample variances may be calculated by summing all the sample variances and dividing it by the number of samples. Since the total number of samples is ${}^{N}C_n$

$$\bar{\sigma}_s^2 = \frac{1}{{}^{N}C_n}\left[\frac{{}^{N-1}C_{n-1}}{n}\sum_{i=1}^{N} x_i^2 - \sum_{j=1}^{{}^{N}C_n} m_{j,s}^2\right] = \sigma^2 - \sigma_m^2 \tag{1.23}$$

Combine this with Equation 1.21 and simplify to get

$$\sigma_s^2 = \frac{N(n-1)}{n(N-1)}\sigma^2 \tag{1.24}$$

If $n << N$ the above relation will be approximated as

$$\boxed{\sigma_s^2 \approx \sigma^2\left(1 - \frac{1}{n}\right)} \tag{1.25}$$

Error estimator σ_e:

The last expression may be written down in the more explicit form

$$\boxed{\sigma_e^2 = \sum_1^n \frac{(x_i - m_s)^2}{n-1}} \tag{1.26}$$

Essentially the experimenter has only one sample and the above formula tells him how the variance of the single sample is related to the variance of the population!

Physical interpretation

Equation 1.26 may be interpreted using physical arguments. Since the mean (the best value) is obtained by one use of all the available data, the degrees of freedom available (units of information available) is one less than before. Hence the error estimator *should* use the factor $(n - 1)$ rather than n in the denominator! The estimator thus obtained is referred to as unbiased variance.

Example 1.2

Resistance of a certain resistor is measured repeatedly to obtain the following data.

Expt. No.	1	2	3	4	5	6	7	8	9
$R, k\Omega$	1.22	1.23	1.26	1.21	1.22	1.22	1.22	1.24	1.19

What is the best estimate for the resistance? What is the error with 95% confidence?

Solution :

Best estimate is the mean of the data.

$$\bar{R} = \frac{4 \times 1.22 + 1.23 + 1.26 + 1.21 + 1.24 + 1.19}{9} = 1.223\,k\Omega \approx 1.22\,k\Omega$$

Standard deviation of the error σ_e:

$$\text{Unbiased Variance } \sigma_e^2 = \frac{\begin{bmatrix} 4 \times 1.22^2 + 1.23^2 + 1.26^2 \\ +1.21^2 + 1.24^2 + 1.19^2 \end{bmatrix}}{8} - 1.223^2 = 3.75 \times 10^{-4}\, k\Omega^2$$

Hence

$$\sigma_e = \sqrt{3.75 \times 10^{-5}} = 0.0194\, k\Omega$$

The corresponding error estimate based on 95% confidence interval is

$$\text{Error} = 1.96 \times \sigma_e = 1.96 \times 0.0194 = 0.038\, k\Omega$$

Example 1.3

Thickness of a metal sheet (in *mm*) is measured repeatedly to obtain the following replicate data. What is the best estimate for the sheet thickness? What is the variance of the distribution of errors with respect to the best value? Specify an error estimate to the mean value based on *99%* confidence.

No.	1	2	3	4	5	6
t, mm	0.202	0.198	0.197	0.215	0.199	0.194
No.	7	8	9	10	11	12
t, mm	0.204	0.198	0.194	0.195	0.201	0.202

Solution :

The best estimate for the metal sheet thickness is $\bar{t}$, the mean of the 12 measured values. This is given by

$$\bar{t} = \frac{\begin{bmatrix} 2 \times 0.202 + 2 \times 0.198 + 0.197^2 + 0.215 \\ +0.199 + 2 \times 0.194 + 0.204 + 0.195 + 0.201 \end{bmatrix}}{12} = 0.200\, mm$$

The variance with respect to the mean or the best value is then given by

$$\begin{aligned} \sigma_e^2 &= \frac{\begin{bmatrix} 2 \times 0.202^2 + 2 \times 0.198^2 + 0.197^2 + 0.215^2 \\ +0.199^2 + 2 \times 0.194^2 + 0.204^2 + 0.195^2 + 0.201^2 \end{bmatrix}}{11} - 0.2^2 \\ &= 3.3174 \times 10^{-5}\, mm^2 \end{aligned}$$

The corresponding standard deviation is given by

$$\sigma_e = \sqrt{3.3174 \times 10^{-5}} = 0.0058\, mm$$

The corresponding error estimate based on 95% confidence is

$$\text{Error} = 2.58 \times \sigma_e = 2.58 \times 0.0058 = 0.0149\, mm$$

1.3.5 Student t distribution

We have seen above that, an estimate for the variance of the population, based on a single sample of n values is given by $\sigma_e^2 = \frac{n(N-1)}{N(n-1)}\sigma_s^2$. We also know that the variance of the sample means is given by the expression $\sigma_m^2 = \frac{\sigma^2}{n}$. In practice only one sample of n values may have been obtained experimentally. The population variance σ^2 is, in fact, not known and hence we use σ_e^2 for σ^2 and hence we have $\sigma_m^2 = \frac{\sigma_e^2}{n}$. Thus the standard deviation of the means is given by $\sigma_m = \frac{\sigma_e}{\sqrt{n}}$. The advantage of doing this is that the standard deviation of the population may be calculated based on this expression even though the population variance is not known. Now we consider the following function given by

$$T_n = \frac{m_s - m}{\frac{\sigma_e}{\sqrt{n}}} \tag{1.27}$$

Note that σ_e is a random variable and hence the function T given by Equation 1.27 is not standard normal. The distribution is referred to as the "Student t distribution" and is defined as $T_n = t_{n-1}$. This distribution depends on n and is given by the following expression

$$\boxed{t(y,d) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi d}\,\Gamma\left(\frac{d}{2}\right)}\left[1+\frac{y^2}{d}\right]^{-\frac{(d+1)}{2}}} \tag{1.28}$$

Here d is the degrees of freedom given by $d = n-1$ and Γ is the Gamma function also referred to as the Generalized Factorial function. The argument $y = m_s - m$ represents the difference between the sample mean and the population mean. This distribution was discovered by a British mathematician who published his work under the pseudonym "Student".[5] For large n (or d) the t - distribution approaches the normal distribution with 0 mean and unit variance given by $N(0,1)$. For small n the distribution is wider than the Normal distribution with larger areas in the tails of the distribution. Plots in Figure 1.7 show these trends. We see that for $d = 30$,

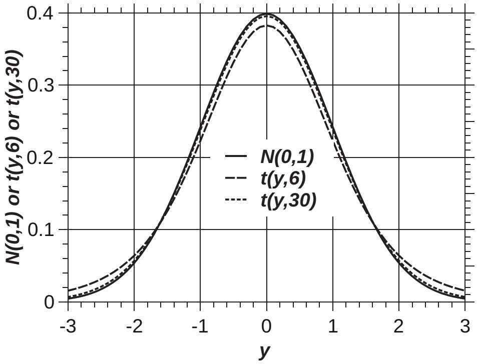


Figure 1.7: Comparison of t - distributions with the Normal distribution

[5] real name William Sealy Gosset, 1876-1937, well known statistician

the t -distribution is quite close to the normal distribution. If the number of samples available is more than about 30 one may simply use the N distribution.

In statistical analysis of data what are more important are the confidence intervals that are appropriate with the t - distributions. These are, in deed, larger than the corresponding values for the Normal distribution. A short table useful for analysis is given as Table B.5. Notice that the 95% confidence interval tends asymptotically to $\pm 1.96\sigma$, characteristic also of the Normal distribution.

Example 1.4

The temperature of a controlled space was measured at random intervals and the spot values are given by the following *10* values:

Trial	1	2	3	4	5
Temperature $^\circ C$	45.3	44.2	45.5	43.5	46.2
Trial	6	7	8	9	10
Temperature $^\circ C$	46.4	43.8	47	45.5	44.4

The control was expected to maintain the temperature at $44.5^\circ C$. How would you describe the above observations?

Solution :

The number of data in the sample is $n = 10$.
The number of degrees of freedom is $d = n - 1 = 9$.
Tabulation of data helps in pursuing the statistical analysis of data. Sample mean is the arithmetic mean of all the spot values of temperature while the estimated variance is based on d.

Trial No.	1	2	3	4	5
Temperature,$^\circ C$	45.3	44.2	45.5	43.5	46.2
Square of error with respect to mean	0.0144	0.9604	0.1024	2.8224	1.0404
Trial No.	6	7	8	9	10
Temperature,$^\circ C$	46.4	43.8	47	45.5	44.4
Square of error with respect to mean	1.4884	1.9044	3.3124	0.1024	0.6084

Sample mean m_s= $45.2^\circ C$
Estimated variance σ_e^2=1.373
Estimated standard deviation σ_e=$1.172^\circ C$
The population mean is specified to be $m = 44.5^\circ C$.
Hence the t - value based on the data may now be calculated as

$$t = \frac{m - m_s}{\sigma_e}\sqrt{n} = \frac{45.2 - 44.5}{1.172}\sqrt{10} = 1.835$$

The 95% Confidence Interval for the t - distribution with $d = 9$ is read from Table B.5 in Appendix B as 2.262. Since the t - value is less than the 95% Confidence Interval we conclude that the sample indicates satisfactory functioning of the controller.

Example 1.5

A sample of 6 resistors is picked up from a lot during a manufacturing process. The resistances are measured in the laboratory and the values are found to be 1020, 1040, 995, 1066, 970 and 992 Ω. The manufacturer will label all the resistors as being equal to a mean value of 1000 Ω. Is this justified? Also specify a tolerance for the resistors from this lot.

Solution :

It is convenient to make a spreadsheet as in Table 1.2.

The table shows the sample mean, the population mean, estimated standard error and finally the value of t calculated for the sample of 6 measured resistance values with degrees of freedom of $d = 6 - 1 = 5$. The 95% confidence interval for t with $d = 5$ is 2.571 (Table B.5). Since the t calculated from the sample is less than this the manufacturer is justified in labeling the resistors as having a mean value of 1000 Ω. We may now use the 95% confidence interval to specify the tolerance. We thus have

$$\text{Tolerance} = \pm 2.571 \times 35.2 = \pm 90.5\,\Omega$$

The resistors from this lot may be labeled as 1000 Ω nominal with 10% tolerance.

Table 1.2: Spreadsheet for Example 1.5

Resistor Number	Resistance Value Ω	Square of Error
1	1020	38.44
2	1040	686.44
3	995	353.44
4	1066	2724.84
5	970	1918.44
6	992	475.24
$m_s =$	1013.8	
$\sigma_e^2 =$	1239.368	
$\sigma_e =$	35.2	
$t =$	0.963	

1.3.6 Test for normality

Distribution of random errors in measurements are obtained by repeated measurements of a physical quantity. These are done by keeping the conditions under which the experiments are conducted invariant. For example, in most measurements, pressure and temperature may affect the outcome and hence these need to be kept fixed during the experiment. Once replicate data is collected we should like to ascertain the error distribution so that we may draw conclusions on the quality of the measurement based on the distribution of errors. Specifically we would like to ascertain whether random errors are distributed normally.

Box and whisker plot

Many methods are available for test of normality of a sample distribution. Shape of the histogram may indicate whether the distribution is close to being normal. Alternately, we look for symmetry, lower and upper quartile values and the minimum and maximum values to make a "box and whisker plot" to check for normality as shown in Figure 1.8. The box and whisker plot shown here is for a sample of data that follows closely a normal distribution. If the sample size is large the values on the box and whisker plot are like those indicated here: $Min = -2.33$, $Q1 = -0.67$, $Median = 0$, $Q3 = 0.67$, $Max = 2.33$. The values shown are the z values for $N(0,1)$.

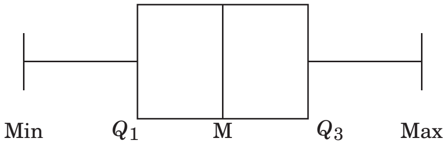


Figure 1.8: Box and whisker plot: M- Median, Q_1 - Lower quartile, Q_3 - Upper quartile, Min - Lower extreme, Max - Upper extreme

Box and whiskers plot is a construct introduced by Tukey[6] and gives a summary of the distribution. Obvious asymmetry, outliers and sharpness of the distribution may be gleaned by looking at the plot.

Example 1.6

A sample data consists of 15 values shown in the table. i is the serial number of data and v_i is the corresponding value. Make a box and whisker plot and comment on the nature of the distribution.

i	1	2	3	4	5	6	7	8
v_i	-1.829	-1.259	-1.187	-0.884	-0.854	-0.745	-0.343	-0.130
i	9	10	11	12	13	14	15	
v_i	0.097	0.224	0.844	1.098	1.112	1.806	2.050	

Solution :

[6] J. W. Tukey, "Box-and-Whisker Plots", Section 2C in Exploratory Data Analysis, Reading, Mass, Addison-Wesley, pp. 39-43, 1977.

We note that the data is already arranged in the ascending order. Hence the minimum and maximum values are the first and last entries in the table. Thus $Min = -1.829$ and $Max = 2.050$.

Since there are an odd number of data values, the median is the value corresponding to the eighth data point i.e. $Median = -0.130$. Lower quartile Q_1 may be calculated as the median of first eight data points and hence we have $Q_1 = \frac{-0.884 - 0.854}{2} = -0.869$. Similarly the upper quartile Q_3 is obtained as the median of data points 8 through 15. Thus $Q_3 = \frac{0.844 + 1.098}{2} = 0.971$. The standard deviation may easily be calculated as $\sigma_e = 1.171$. It may be verified that the mean of all data is 0. With these values we make a box and whisker plot as shown in Figure 1.9. It is seen that the mean is slightly larger than the median and hence the distribution is heavier to the right and marginally skewed.

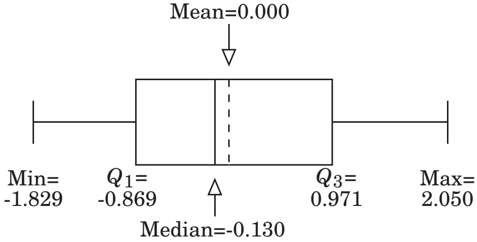


Figure 1.9: Box and whisker plot for Example 1.6

QtiPlot software has an option to make a menu driven Box and Whisker plot. To demonstrate its use we solve a typical problem below.

Example 1.7

Repeated measurement of a certain quantity gives replicate data. The deviate from mean is calculated to yield deviate data which has zero mean. Scaling the deviates with respect to the standard deviation gives the data shown rank ordered in the table. Use QtiPlot to make a Box and Whisker plot and comment on the quality of the data.

Rank Order	$\frac{d_i}{\sigma}$	Rank Order	$\frac{d_i}{\sigma}$	Rank Order	$\frac{d_i}{\sigma}$	Rank Order	$\frac{d_i}{\sigma}$
1	-1.946	6	-0.442	11	0.275	16	0.520
2	-1.057	7	-0.193	12	0.305	17	0.585
3	-0.829	8	-0.164	13	0.316	18	1.102
4	-0.675	9	-0.091	14	0.398	19	1.919
5	-0.660	10	-0.028	15	0.489	20	2.520

Solution :

Key in the given data in two columns of QtiPlot table. Invoke the option “Box plot” under “Statistical graphs” menu to obtain the Box and Whisker plot shown below (Figure 1.10).

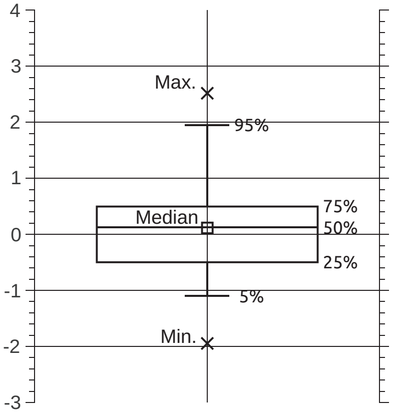


Figure 1.10: Box and Whisker plot for Example 1.7 obtained using QtiPlot

QtiPlot also gives the following statistics:

28/04/14 9:12 PM	Statistics for Table 1:
Min = −1.95	$D1$ (1st decile) = −0.85
$Q1$ (1st quartile) = −0.50	Median = 0.12
$Q3$ (3rd quartile) = 0.50	$D9$ (9th decile) = 1.18
Max = 2.52	Size = 20

All the calculations required for the plot are automatically done by QtiPlot and the box plot is the output. In this plot the minimum and maximum are shown by ‘×’ and the median by □. Other percentiles are as explained earlier.

In the present example there is a slight amount of skewness in the distribution. However there is no indication that the distribution is not normal.

Q-Q plot

Another useful graphical method, that helps in identifying match or mismatch with a normal distribution, is a Quantile-Quantile plot or a Q-Q plot. This plot is made with z values for a normal distribution against the z values for the sample. The sample data is arranged in ascending order and the deviates with respect to the sample mean or divided by the unbiased estimate of the sample standard deviation. The rank order of the samples are used for calculating the probability as $p(i) = \frac{i - 0.5}{n}$ where i is the rank order and n is the number of data in the sample. The corresponding z values for a normal distribution are calculated consulting a table.

If the sample is close to being normal the data points will lie close to a 45° line (called the parity line) assuming that the same scale is used along the two axes. Departures from linearity will be clear indication of non normal behavior of the sample. We consider the sample in Example 1.6 and demonstrate how a Q-Q plot is made and what conclusions we may draw from it.

Example 1.8

Consider the sample of data of Example 1.6 and make a Q-Q plot. Draw conclusions regarding the distribution underlying the sample.

Solution :

Since the data is already presented in ascending order rank order is the same as the number in column 1 (see table below). With the mean being zero and estimated standard deviation being $\sigma_e = 1.171$ the data is recalculated as $z(i) = \dfrac{v_i - \bar{v}}{\sigma_e}$ and is in the third column of the table. The probabilities are calculated based on the rank order and are in column 4. $z_N(i)$ is calculated using built in function NORMSINV($p(i)$) in the spreadsheet program.[7]

Q-Q plot is obtained by plotting z_N along the abscissa and z along the ordinate. The parity line is obtained by plotting a 45° line passing through the origin. In the present case the Q-Q plot is as shown in Figure 1.11.

The data points in the Q-Q plot lie close to the parity line. Also the data points are distributed evenly around the parity line and do not present any systematic variation. Hence it is safe to conclude that the sample of data is from a normal distribution.

Median value of $v(i)$ is calculated as -0.130. With the mean being zero, skewnwss in the sample distribution is represented by skewness = $\dfrac{3(mean - median)}{\sigma_e} = \dfrac{3(0-0.130)}{1.171} = 0.333$ (Note that skewness is bounded between -3 and +3). This is less than critical value of 0.863 and hence is considered to be not significant.[8] Skewness in the sample is purely due to chance and not due to non normality of the distribution.

i	$v(i)$	$z(i)$	$p(i)$	$z_N(i)$
1	-1.829	-1.561	0.033	-1.834
2	-1.259	-1.074	0.100	-1.282
3	-1.187	-1.013	0.167	-0.967
4	-0.884	-0.755	0.233	-0.728
5	-0.854	-0.729	0.300	-0.524

··· Continued on next page

[7] Most spreadsheet programs have built in functions useful for statistical analysis. The reader may familiarize herself/himself with these.

[8] D.P. Doane and L.E. Seward, Measuring Skewness: A Forgotten Statistic?, Journal of Statistics Education, Vol. 19, No.2, 2011. Critical values are taken from this reference.

continued from previous page···

i	$v(i)$	$z(i)$	$p(i)$	$z_N(i)$
6	-0.745	-0.636	0.367	-0.341
7	-0.343	-0.293	0.433	-0.168
8	-0.130	-0.111	0.500	0.000
9	0.097	0.083	0.567	0.168
10	0.224	0.192	0.633	0.341
11	0.844	0.721	0.700	0.524
12	1.098	0.937	0.767	0.728
13	1.112	0.949	0.833	0.967
14	1.806	1.541	0.900	1.282
15	2.050	1.750	0.967	1.834
$\bar{v}$ =	0.000			
σ_e =	1.171			

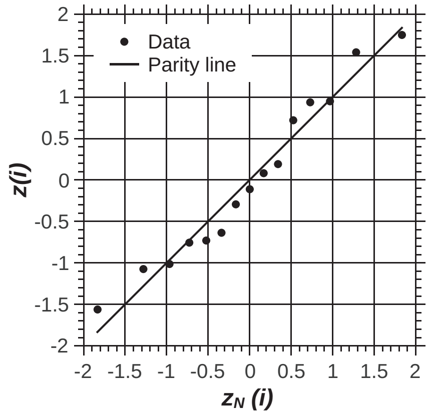


Figure 1.11: *Q-Q plot for sample data in Example 1.8*

Jarque-Bera test for normality

It is well known that the normal distribution is symmetric with respect to the mean. In other words the distribution is not skewed and hence the third moment defined as

$$\boxed{g_1 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma}\right)^3 \quad \text{or} \quad g_1 = \underbrace{\frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma_e}\right)^3}_{\text{Unbiased estimator, small sample}}} \tag{1.29}$$

is zero. The fourth moment, known as the Kurtosis, is defined by the relation

$$\boxed{g_2=\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^4-3 \text{ or } g_2=\underbrace{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma_e}\right)^4-\frac{3(n-1)^2}{(n-2)(n-3)}}_{\text{Unbiased estimator, small sample}}} \tag{1.30}$$

also has a value of zero. Any departure of the sample data from normality would mean that these two quantities are non-zero. Note that g_1 and g_2 are available as functions in spreadsheet programs invoked by SKEW(Sample Data) and KURT(Sample Data) where the sample data is a column of numbers invoked as with the other functions. In case of Jarque-Bera test the statistic JB is defined as

$$\boxed{JB=\frac{n}{6}\left[g_1^2+\frac{g_2^2}{4}\right]} \tag{1.31}$$

The critical values for JB are obtained by simulations and useful critical value tables are available from references.[9]

Example 1.9

Fifteen deviates arranged in ascending order forms a sample (second column of the spreadsheet). The sample is expected to follow a normal distribution. Test it using JB test.

Solution :

Extract of spreadsheet used is shown below as a table. The first two columns show the deviates in ascending order. The appropriate statistical parameters have been calculated using the functions available in the spreadsheet program and presented in the last column. The JB value that characterizes the sample and the critical value taken from the cited reference is also shown in the last column.

Data		Data		Parameters	
No.	Deviate	No.	Deviate	$\mu =$	0.000
1	-1.829	9	0.097	$\sigma_e =$	1.171
2	-1.259	10	0.224	$g_1 =$	0.321
3	-1.187	11	0.844	$g_2 =$	-0.916
4	-0.884	12	1.098	$JB =$	0.781
5	-0.854	13	1.112	$\alpha =$	0.05
6	-0.745	14	1.806	$JB_{crit} =$	3.768
7	-0.343	15	2.050		
8	-0.130				

Since $JB < JB_{crit}$ the hypothesis that the deviates are normally distributed is valid.

[9] Thadewald T., and Büning H, Working Paper - "Jarque-Bera test and its competitors for testing normality: A power comparison", accessed at www.econstor.eu/bitstream/10419/49919/1/668828234.pdf

χ^2 test for normality

Another useful test for normality is the χ^2 test that is recommended to be used for large samples with $n \geq 50$. We work with the frequencies of occurrence rather than the magnitude of the data values. Essentially, the test compares the observed frequencies with the expected frequencies according to normal distribution, to test for normality. We work with "binned" data such that each bin contains at least five values.

Let the data consist of n values v_i arranged in ascending order. We create bins by defining lower and upper bound values for the data to group the data. Each group fills a bin of certain width. The number of data values which are within a particular bin is the observed frequency f_O for that particular bin. Let there be k bins. The sum of frequencies in all the bins is thus equal to n i.e. $\sum\limits_{i=1}^{k} f_{O,i} = n$.

Arrange the data now by calculating the z values defined, as usual as, $z_i = \dfrac{v_i - \bar{v}}{\sigma_e}$. Compute the z values that bound a particular bin and calculate the cumulative probability that the chosen values are within the particular bin. Multiply this probability with the number of data n to get the expected frequency f_E. We see that $\sum\limits_{i=1}^{k} f_{E,i} = n$.

The statistic χ^2 (refer Chapter 2 for a more detailed discussion about the χ^2 distribution) defined as

$$\boxed{\chi^2 = \sum_{i=1}^{k} \frac{\left(f_{O,i} - f_{E,i}\right)^2}{f_{E,i}}} \tag{1.32}$$

follows the χ^2 distribution which is one sided and has a range of $0 \leq \chi^2 \leq \infty$. The critical χ^2 value is calculated based on the chosen α for a particular number of degrees of freedom ν. In the case of normality test of a sample there are c columns and n rows of data. Hence $\nu = (n-1)\times(c-1)$. Tables of χ^2 is available in the Appendix as Table B.6 in Appendix B.

Example 1.10

A set of 50 deviates arranged in ascending order has been obtained and shown in Table 1.4. Perform χ^2 test of normality on this sample.

Solution :

Step 1 We calculate the mean and variance of the sample and hence order the data according the z values. The bounds are rephrased in terms of z values. We create bins using the lower and upper bounds as shown in Table 1.5. The corresponding observed frequencies are also shown.

Step 2 The cumulative probabilities are obtained by subtracting the cumulative probabilities between $z = -\infty$ to $z =$ left bound from the cumulative probability between $z = -\infty$ to $z =$ right bound for each bin. In the spreadsheet

Table 1.4: Deviates data for Example 1.10

i	d_i	i	d_i	i	d_i	i	d_i
1	-2.803	14	-0.972	27	-0.084	40	0.660
2	-1.781	15	-0.928	28	-0.020	41	0.677
3	-1.571	16	-0.815	29	0.070	42	0.749
4	-1.555	17	-0.701	30	0.132	43	0.753
5	-1.512	18	-0.605	31	0.151	44	0.868
6	-1.422	19	-0.576	32	0.156	45	0.967
7	-1.388	20	-0.562	33	0.236	46	1.070
8	-1.213	21	-0.556	34	0.375	47	1.255
9	-1.204	22	-0.393	35	0.398	48	1.335
10	-1.128	23	-0.358	36	0.405	49	1.552
11	-1.049	24	-0.346	37	0.543	50	1.841
12	-0.987	25	-0.277	38	0.547		
13	-0.982	26	-0.249	39	0.654		

we use the function NORMSDIST(z) for this purpose. The expected frequencies are then obtained by multiplying the probability of occurrence within each bin and the total number of data n. The expected frequencies are indicated in the 5th column of the table.

Step 3 We calculate the value of χ^2 as $\chi^2 = 5.877$ from the observed and expected frequencies from the binned data as shown by the sum of entries in the last column of the table. Calculations have been performed using a spreadsheet program. The critical χ^2 for $\nu = (9-1) \times (2-1) = 8$ and $\alpha = 0.1$ is calculated using the function CHISQINV(0.9,8)[10] as 13.362. Since the calculated χ^2 is less than the critical value there is no reason to doubt the normality of the sample data.

Table 1.5: Expected frequencies and evaluation of χ^2 for data in Example 1.10

Bin No. i	Bin bounds	$f_{O,i}$	$f_{E,i}$	$\dfrac{(f_{O,i}-f_{E,i})^2}{f_{E,i}}$
1	$-\infty$ to -1.2	5	5.753	0.099
2	-1.2 to -1	4	2.179	1.521
3	-1 to -0.7	5	4.165	0.167
4	-0.7 to -0.2	7	8.939	0.421
5	-0.2 to 0.2	6	7.926	0.468
6	0.2 to 0.4	7	3.808	2.675
7	0.4 to 0.7	4	5.131	0.249
8	0.7 to 1	5	4.165	0.167
9	1 to ∞	7	7.933	0.110
Total number of data =		50	$\chi^2 =$	5.877

[10] Argument 0.9 is the probability which is nothing but $1-\alpha$ with $\alpha = 0.1$.

Example 1.11

5 machines are used to produce identical parts in a factory setting. Number of parts made by each machine and the number rejected during inspection are shown in the table. Test the hypothesis that all machines are of equal quality.

Machine No. i	1	2	3	4	5
No. of parts made n_i	300	400	250	150	200
No. of parts rejected, O_i	24	13	10	16	8

Solution :

The hypothesis that all the machines are of similar quality requires that the number of parts rejected be the same proportion of the number of parts made by each machine. Thus the expected number of parts rejected are given by $E_i = \frac{n_i \times \sum_{i=1}^{5} O_i}{\sum_{i=1}^{5} n_i}$.

We round the E_i to whole numbers and have the following:

i	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
1	24	16	4.0000
2	13	22	3.6818
3	10	14	1.1429
4	16	8	8.0000
5	8	11	0.8182
$\chi^2 =$			17.643
$\nu =$			4
$\chi^2(\alpha = 0.05, \nu) =$			9.4877

Since $\chi^2 > \chi^2_{critical}$ the hypothesis is not sustained. The machines are not all of the same quality.

1.3.7 Nonparametric tests

In experimental studies, we often need to compare samples of data and decide whether they follow the same distribution. What the distribution itself is, as long as it is continuous, may be secondary and hence the parameters that characterize the distributions are not important for the proposed test. We essentially devise a nonparametric test for comparing two samples of data. For example, if two samples of data have been collected at different times, we would like to know whether there are significant changes between them.

A commonly used nonparametric test is the Kolmogorov Smirnov[11] or KS two sample test.

[11] Andrey Nikolaevich Kolmogorov 1903 - 1987, Russian mathematician; Nikolai Vasilyevich Smirnov 1900 - 1966, Russian mathematician

Kolmogorov Smirnov two sample test

Let the first sample consist of n_1 data and the second sample n_2 data. Order both the data in ascending order. We assume that each data is independent and identically distributed. Assume that an empirical distribution function (EDF) is assumed such that the probability is defined by a uniform distribution. Hence the cumulative empirical probability distribution (CEDF) jumps by $\frac{1}{n_1}$ at each data point for the first sample and by $\frac{1}{n_2}$ for the second sample. We make a plot of CEDF vs data value for the two samples on the same graph. The supremum of the difference between the two cumulative probabilities D_{max} is the statistic of interest to us. Critical value of $D_{crit}(n_1, n_2, \alpha)$ depends on the number of data and the significance level α.[12] The null hypothesis H_0 is that the two samples follow the same distribution. If $D_{max} < D_{crit}$ the hypothesis is accepted. Otherwise it is concluded that the two samples do not follow the same distribution.

An example is presented below to demonstrate the KS two sample test.

Example 1.12

Two samples of data were collected by two batches of students taking part in a laboratory class. Each batch had 20 students (i.e. $n_1 = n_2 = 20$) and the task consisted of measuring a physical quantity d using the same method. The data has been arranged in ascending order and presented in the table below. Perform a KS test to determine whether the two samples follow the same distribution.

Data No.i	Batch 1 d_i	Batch 2 d_i	Data No.i	Batch 1 d_i	Batch 2 d_i
1	4.541	4.543	11	5.040	4.930
2	4.611	4.586	12	5.053	4.940
3	4.708	4.781	13	5.099	4.965
4	4.737	4.808	14	5.122	5.031
5	4.739	4.817	15	5.164	5.033
6	4.777	4.835	16	5.171	5.096
7	4.912	4.836	17	5.215	5.181
8	4.939	4.865	18	5.249	5.345
9	4.939	4.882	19	5.266	5.431
10	5.031	4.905	20	5.311	5.489

Solution :

Step 1 Since both samples have the same number of data the CEDF is the same for the two samples. As we pass each data point the CEDF increases by $\frac{1}{20} = 0.05$. We make a plot of CEDF versus the data value as shown in Figure 1.12.

[12] Table of critical values may be downloaded from "www.soest.hawaii.edu/wessel/courses/gg313/Critical_KS.pdf"

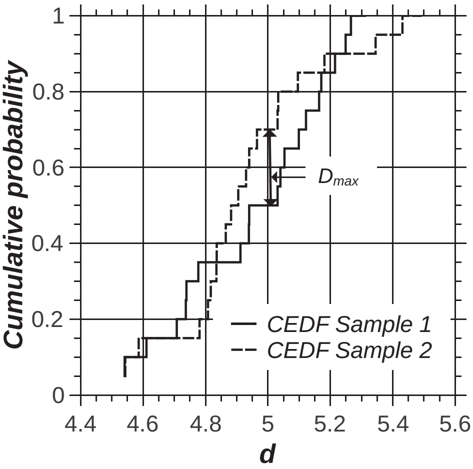


Figure 1.12: Cumulative probability plot for two samples in Example 1.12

Step 2 The maximum value of D is equal to 0.2 as shown in Figure 1.12.

Step 3 Critical value (based on reference cited) is obtained for $n_1 = n_2 = 20$ and $\alpha = 0.05$ as $D_{crit} = 1.36\sqrt{\dfrac{20+20}{20\times 20}} = 0.430$.

Step 4 Since $D_{max} < D_{crit}$ the hypothesis that the two samples are from the same distribution holds. This means that the observed differences between the two samples are solely due to chance.

Kolmogorov-Smirnov test for normality

The KS two sample test has to be modified to use it as a test for normality of a single sample. In that case the test is also known as Kolmogorov-Smirnov *goodness of fit* test. We assume that the mean of the sample μ is known and calculate the standard deviation with respect to the mean σ_e, the usual way, based on an unbiased estimator. We compare the ECDF with the normal distribution function based on $z = \dfrac{x-\mu}{\sigma_e}$. The cumulative normal probability is then calculated, after transforming the measured values to corresponding z values. The maximum difference between the two cumulative probabilities is the statistic that is used to test H_0.

Since the standard deviation has been calculated by using the sample data the critical values need a correction as given by Lilliefors[13].

Example 1.13

Consider the first sample data in Example 1.12. Test whether it is distributed normally using KS test for normality.

Solution :

[13] Lilliefors, H., "On the Kolmogorov-Smirnov test for normality with mean and variance unknown", Journal of the American Statistical Association, Vol. 62. pp. 399-402, 1967

Step 1 Calculations are best performed using a spreadsheet as shown below. The mean μ and standard deviation σ_e of the sample are calculated using the data d_i for $1 \le i \le 20$. Based on these the z values are obtained. The corresponding cumulative normal probabilities P_i are then calculated. The D_i and D'_i are obtained as indicated. The maximum of the last two columns is the D_{max} required to perform the KS normality test (shown as bold entry).

i	d_i	$z_i = \frac{d_i - \mu}{\sigma_e}$	P_i	$p_i = \frac{i-1}{n}$	$p'_i = \frac{i}{n}$	$D_i = \lvert P_i - p_i\rvert$	$D'_i = \lvert P_i - p'_i\rvert$
1	4.541	-1.914	0.028	0	0.05	0.028	0.022
2	4.611	-1.611	0.054	0.05	0.1	0.004	0.046
3	4.708	-1.188	0.117	0.1	0.15	0.017	0.033
4	4.737	-1.063	0.144	0.15	0.2	0.006	0.056
5	4.739	-1.055	0.146	0.2	0.25	0.054	0.104
6	4.777	-0.890	0.187	0.25	0.3	0.063	0.113
7	4.912	-0.301	0.382	0.3	0.35	0.082	0.032
8	4.939	-0.184	0.427	0.35	0.4	0.077	0.027
9	4.940	-0.180	0.429	0.4	0.45	0.029	0.021
10	5.031	0.218	0.586	0.45	0.5	**0.136**	0.086
11	5.040	0.257	0.602	0.5	0.55	0.102	0.052
12	5.053	0.313	0.623	0.55	0.6	0.073	0.023
13	5.099	0.510	0.695	0.6	0.65	0.095	0.045
14	5.122	0.613	0.730	0.65	0.7	0.080	0.030
15	5.164	0.795	0.787	0.7	0.75	0.087	0.037
16	5.171	0.825	0.795	0.75	0.8	0.045	0.005
17	5.215	1.016	0.845	0.8	0.85	0.045	0.005
18	5.249	1.166	0.878	0.85	0.9	0.028	0.022
19	5.266	1.239	0.892	0.9	0.95	0.008	0.058
20	5.311	1.432	0.924	0.95	1	0.026	0.076
$\mu =$	4.981						
$\sigma_e =$	0.230						

Step 2 A plot is made, with sample values along the abscissa and the cumulative probabilities along the ordinate, as shown in Figure 1.13. The maximum D_{max} is identified on the figure.

Step 3 We note that the maximum deviate is 0.136. The critical value for comparison is obtained from a recent paper[14] as 0.192 for $\alpha = 0.05$.

Step 4 Since $D_{max} < D_{crit}$ the hypothesis that the sample data is from a normal distribution is valid.

[14] Hervé Abdi ansd Paul Molin, Lilliefors/Van Soest's test of normality, In: Neil Salkind (Ed.), Encyclopedia of Measurement and Statistics, Thousand Oaks (CA): Sage, 2007

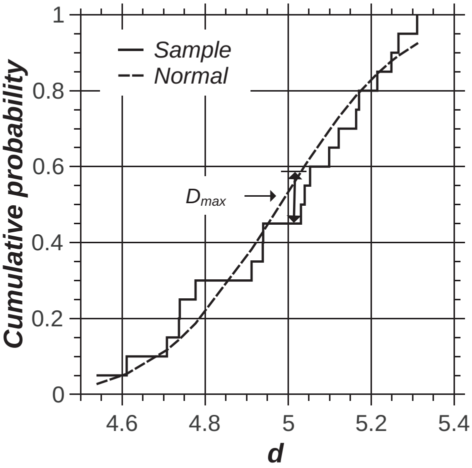


Figure 1.13: Cumulative probability plot for sample data in Example 1.13

1.3.8 Outliers and their rejection

We have seen that most data should lie within the bracket $\pm 3\sigma$ (if a large number of data has been collected) around the mean granting that the error distribution is normal. Those data that lie outside this range are called outliers. Even though there may be special cases where an outlier may be physically meaningful it is unusual to get such errors in normal practice. Since the outlier will change the mean and standard deviation it is best to reject such outliers unless there is a reason to believe them to be important. There are a large number of statistical tests that may be used to determine or reject outliers. From the point of view of the present book we shall discuss a few of the more useful ones.

Chauvenet's criterion for discarding outliers

Chauvenet's [15] criterion states that outliers (one or more than one) may be discarded as spurious or suspicious data if the data is outside a range on either side of the mean with probability less than $\frac{1}{2n}$ where n is the number of data. For example, if the number of data is 20, the probability we are looking for is $\frac{1}{2 \times 20} = 0.025$ and this corresponds to the region outside the cumulative probability interval 0.025 to 0.975 of the standard normal distribution $N(0,1)$. The corresponding critical value of the confidence interval is ± 2.24. Similarly we may evaluate critical values for different number of data as shown in Table 1.7. This table is useful in applying the Chauvenet criterion for data sets with different number of data points. Chauvenet's test is recommended to be used only once. After rejection of data the statistical parameters calculated using the rest of the sample is accepted.

An example is given below to show the effect of outliers and also the improvement in the deduced results when outliers are discarded.

Example 1.14

A certain experiment has been conducted by 12 students using the same

[15] after William Chauvenet, 1820-1870, American mathematician

Table 1.7: *Critical value for outlier according to Chauvenet's criterion*

n	$\left\lvert\frac{d_{max}}{\sigma}\right\rvert$	n	$\left\lvert\frac{d_{max}}{\sigma}\right\rvert$	n	$\left\lvert\frac{d_{max}}{\sigma}\right\rvert$
3	1.38	14	2.10	80	2.74
4	1.54	16	2.15	100	2.81
5	1.65	18	2.20	150	2.93
6	1.73	20	2.24	200	3.02
7	1.81	25	2.33	300	3.14
8	1.86	30	2.39	400	3.23
9	1.91	40	2.49	500	3.29
10	1.96	50	2.57	1000	3.48
12	2.04	60	2.64		

Note: d_{max} = maximum deviation in the data set

experimental set up, to determine the acceleration due to gravity g in m/s^2. The value of g estimated by various students is given in the following table. What is the best estimate for the value of the measured quantity? Specify a suitable error bar. Would you like to discard any data? If so which ones and why? What are the mean and error bar when you discard spurious data?

Student No.	1	2	3	4	5	6
Value g	9.628	9.813	9.729	9.81	9.836	9.718
Student No.	7	8	9	10	11	12
Value g	9.666	9.725	9.615	9.999	9.701	8.245

Solution :

Step 1 Calculate first with all data included Calculation has been made, using a spreadsheet program which calculated the mean and the variance using in-built functions, AVERAGE(number 1,number 2,...) or AVERAGE(B1:B12) to calculate the mean, VAR(number 1,number2,...) or VAR(B1:B12) to calculate the variance or STDEV(number 1,number2,...) or STDEV(B1:B12) to calculate the standard deviation [16] assuming that the data is entered in column B and occupies rows 1 to 12. Error with respect to the mean may be calculated for each entry as shown in column C. Absolute value of the error divided by the standard deviation is shown in column D. It is seen from Table 1.7 that this ratio is more than the critical value 2.04 for the data collected by student number 12. It represents an outlier which is suspect data. We may delete this data and recalculate the statistical parameters.

Step 2 Calculate next after discarding data number 12

The calculations after dropping the data point 12 are shown in the columns E - H. The entries are self explanatory. We see that the mean has changed

[16] Spreadsheet uses the unbiased estimate for the variance and the standard deviation

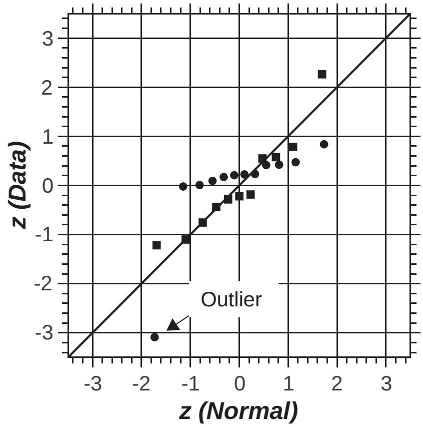
significantly and is more representative of the data. The standard deviation or the spread with respect to the mean has also changed significantly.

The revised calculations paint a much better picture of the data collected by the students in the class.

Step 3 It is interesting to make Q-Q plots for the two samples - (a) before rejecting outlier and (b) after rejecting outlier.

The single outlier has the effect of introducing a pattern to the departure from the parity line. The single outlier indicated by the arrow in Figure 1.14, of course, is far away from the parity line. However when the outlier is removed all the data points move close to the parity line and do not show any specific pattern of variation. Hence one may conclude that errors in the experimental data are normally distributed.

Column identifier							
A	B	C	D	E	F	G	H
No.	Value g	Error $g-\bar{g}$	$\left\lvert\frac{g-\bar{g}}{\sigma}\right\rvert$	No.	Value g	Error $g-\bar{g}$	$\left\lvert\frac{g-\bar{g}}{\sigma}\right\rvert$
1	9.628	0.004	0.010	1	9.628	-0.121	1.098
2	9.813	0.189	0.424	2	9.813	0.064	0.580
3	9.729	0.105	0.236	3	9.729	-0.020	0.182
4	9.81	0.186	0.417	4	9.81	0.061	0.552
5	9.836	0.212	0.475	5	9.836	0.087	0.788
6	9.718	0.094	0.211	6	9.718	-0.031	0.282
7	9.666	0.042	0.095	7	9.666	-0.083	0.753
8	9.725	0.101	0.227	8	9.725	-0.024	0.218
9	9.615	-0.009	0.020	9	9.615	-0.134	1.216
10	9.999	0.375	0.840	10	9.999	0.250	2.266
11	9.701	0.077	0.173	11	9.701	-0.048	0.436
12	8.245	-1.379	3.086				
$\bar{g}$	9.624			$\bar{g}$	9.749		
σ_e	0.447			σ_e	0.110		



● *z - before data rejection*
■ *z - after data rejection*
— *parity line*

Figure 1.14: *Q-Q plots before and after rejection of one outlier in Example 1.14*

Pierce's criterion for discarding outliers

Pierce's[17] criterion is useful if we have to discard more than one outlier. Rejection of data is based on the principle - to quote the author -

> "that the proposed observations should be rejected when the probability of the system of errors obtained by retaining them is less than that of the system of errors obtained by their rejection multiplied by the probability of making so many, and no more, abnormal observations"

In a recent publication Ross[18] has discussed in detail the use of Pierce's criterion for rejection of abnormal data. He has also provided a table of critical values that correspond to the probabilities of obtaining "so many, and no more, abnormal observations". The detailed method given in this paper (the reader should read this paper) is summarized below:

1. Number of data = n. Calculate m_s and σ_s
2. Assume one data is suspect
3. Read off R from table in paper by Ross
4. If any deviate (absolute value) $> R\sigma_s$ discard the corresponding data.
5. Assume a second data may be suspect.
6. Read off R from table in paper by Ross. No change in n for now.
7. If any deviate (absolute value) $> R\sigma_s$ discard the corresponding data.
8. Continue till no more data needs to be rejected.
9. Use the reduced number of data after rejection of all suspect data. Let number of data equal to n_1.
10. Calculate m_s and σ_s with n_1.
11. Repeat 2 - 9 with new parameters.
12. Break the loop whenever there is no scope for rejecting any more data.

An example is worked out now to demonstrate the above method.

Example 1.15

In a laboratory class students were asked to measure a certain physical quantity and came up with the readings given in the table. Use Pierce's test to discard abnormal data. How would you summarize the data after rejecting the abnormal data points?

Is it reasonable to assume that the deviates are normally distributed? Base your decision making a box and whisker plot.

[17] after Benjamin Peirce, 1809 - 1880, an American mathematician who authored the paper "Criterion for rejection of doubtful observations", The Astronomical Journal, Vol. II, No.21, pp. 161-163, 1852.

[18] Stephen M. Ross, Peirce's criterion for the elimination of suspect experimental data, Journal of Engineering Technology, Fall 2003, pp. 38-41

i	v_i	i	v_i	i	v
1	4.590	8	4.952	15	5.062
2	4.764	9	4.78	16	5.129
3	4.854	10	5.106	17	4.452
4	5.254	11	5.039	18	4.959
5	4.894	12	5.143	19	4.903
6	4.998	13	5.442		
7	5.114	14	4.813		

Solution :

Step 1 Consider all the data in the sample i.e. $n = 19$. We calculate the mean and standard deviation of the sample as $m_s = 4.960$ and $\sigma_e = 0.230$.

Step 2 We would like to check if a single data is abnormal. From table critical value is $R = 2.185$ for $n = 19$ and one abnormal data. We calculate the maximum possible deviate as $R \times \sigma_e = 2.185 \times 0.230 = 0.502$. We calculate absolute values of all the 19 deviates and pick the maximum value. The maximum value is found to correspond to student No. 17 and is 0.508. Since this value is greater than the critical value of 0.502 we discard this data.

Step 3 Now assume that there may be a second data that is abnormal. From table critical value is $R = 1.890$ for $n = 19$ and two abnormal data. We calculate the maximum possible deviate as $R \times \sigma_e = 1.890 \times 0.230 = 0.434$. The second largest value is found to correspond to student No. 13 and is 0.482. Since this value is greater than the critical value of 0.434 we discard this data also.

Step 4 Now assume that there may be a third data that is abnormal. From table critical value is $R = 1.707$ for $n = 19$ and three abnormal data. We calculate the maximum possible deviate as $R \times \sigma_s = 1.707 \times 0.230 = 0.392$. The third largest value is found to correspond to student No. 1 and is 0.370. Since this value is less than the critical value of 0.392 we conclude that there are no more abnormal data.

Step 5 We discard the two abnormal data found above and end up with a sample containing 17 data points. The mean and standard deviation of this data set are given by $m_s = 4.962$ and $\sigma_e = 0.170$.

Step 6 We would like to check if a single data is abnormal in the reduced set obtained by discarding the outliers found previously. From table critical value is $R = 2.134$ for $n = 17$ and one abnormal data. We calculate the maximum possible deviate as $R \times \sigma_e = 2.134 \times 0.170 = 0.362$. We calculate absolute values of all the 17 deviates and pick the maximum value. The maximum value is found to correspond to student No. 1 and is 0.372. Since this value is greater than the critical value of 0.362 we discard this data.

Step 7 We would like to check if a second data is abnormal in the reduced set obtained by discarding the outliers found previously. From table critical value is $R = 1.836$ for $n = 17$ and one abnormal data. We calculate the maximum possible deviate as $R \times \sigma_s = 1.836 \times 0.170 = 0.311$. The second largest deviate is found to correspond to student No. 4 and is 0.292. Since this value is less than the critical value of 0.311 we conclude that there are no more abnormal data.

Step 8 We break the loop and accept three data as abnormal and hence have a sample containing 16 values. A repeat of the Pierce's test for these data shows that no more abnormal data are present in the sample. The pruned sample is shown in the following table.

i	v_i	i	v_i	i	v
2	4.764	8	4.952	15	5.062
3	4.854	9	4.78	16	5.129
4	5.254	10	5.106	18	4.959
5	4.894	11	5.039	19	4.903
6	4.998	12	5.143		
7	5.114	14	4.813		

The mean and standard deviation values characterizing the pruned sample are $m_s = 4.985\,cm$ and $\sigma_e = 0.144\,cm$.

Step 9 We now calculate the parameters required for nmaking a Box and Whisker plot.

Mean =	4.985	Median =	4.979
Sigma =	0.144	Quartile 3 =	5.108
Minimum =	4.764	Maximum =	5.254
Quartile 1 =	4.884		

We make a Box and Whisker plot as Figure 1.15. We have also included the three discarded data points in this plot. The plot shows good symmetry and spread close to a normal distribution. Hence it is reasonable to assume that the data is distributed normally. Also, it seems the three outliers that were discarded, had some calculation errors!

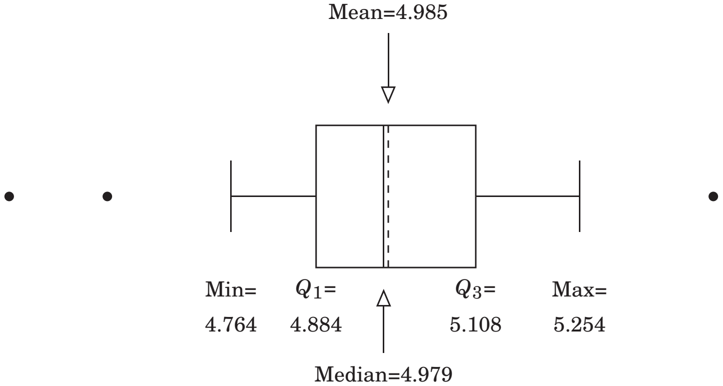


Figure 1.15: Box and whisker plot for pruned data of Example 1.15. Discarded outliers are shown by • in the plot.

Thompson τ for discarding outliers

Another test useful for discarding abnormal data from a sample is the Thompson τ test. Consider a data of n samples in which we would like to discard abnormal data. Calculate the mean and the standard deviation of the sample, the usual way. Calculate the absolute value of the difference between data and the mean i.e. calculate $d_i = |v_i - \bar{v}|$. Determine the largest, $d_{i,max}$ among these. The Thomson statistic τ is given by

$$\boxed{\tau(\alpha,n) = \frac{t_{\alpha/2}(n-1)}{\sqrt{n}\sqrt{n-2+t^2_{\alpha/2}}}} \tag{1.33}$$

The Student t value is calculated based on a chosen α and the number of data n. For example, if $\alpha = 0.1$, we calculate t as TINV(0.05,n), a function available in spreadsheet programs. Corresponding critical τ value may be calculated based on the definition given above. If $d_{i,max} > \sigma_e \tau$ discard the data.

If a data has been discarded as abnormal, redo the above steps with $n-1$ data to discard a second outlier, if it exists.

Continue the process till no more outliers are found.

Example 1.16

The period of a simple pendulum was repeatedly measured and the replicate data is tabulated below.

i	T_i	i	T_i	i	T_i	i	T_i
1	2.013	6	1.987	11	1.992	16	1.998
2	2.000	7	2.012	12	**1.786**	17	2.008
3	**2.225**	8	1.997	13	2.003	18	1.981
4	2.000	9	1.994	14	1.983	19	1.989
5	1.991	10	1.998	15	2.016	20	2.004

Bold entries represent likely outliers

Discard outliers using Thompson τ test. Calculate the mean and standard deviation of pruned data.

Solution :

Step 1 We calculate a few critical values of τ making use of function available in spreadsheet. We take $n = 20$ to $n = 17$ expecting no more than 4 abnormal data. The critical values are tabulated below.

n	17	18	19	20
$\tau(0.1,n)$	1.871	1.876	1.881	1.885

We note in passing that as $n \to \infty$ the critical value of τ tends to 1.96. Note also that all calculations have been rounded to 3 digits after decimals.

Step 2 The mean and standard deviation of the given sample consisting of 20 replicate data is calculated as $\bar{T} = 1.999$ and $\sigma_e = 0.072$. The biggest absolute deviate is identified as data No. 3 of $T_3 = 2.225$ and $d_3 = |2.225 - 1.999| = 0.226$. With $n = 20$, we have, $\tau \times \sigma_e = 1.885 \times 0.072 = 0.136$. Since $d_3 > 0.136$ we discard this data as an outlier.

Step 3 After discarding T_3 perform the above calculations with the rest of the 19 data points. The mean and standard deviation are given by $\bar{T} = 1.987$ and $\sigma_e = 0.050$. Identify the maximum deviation (absolute value) as that corresponding to T_{12} in the original sample (or T_{11} in the pruned sample). The maximum deviate of 0.201 is compared with critical value of $\tau(0.1, 19) \times \sigma_e = 0.093$. Discard T_{12} in the original sample.

Step 4 Repeat the calculations with 18 data and show that no more data is to be rejected.

Step 5 The statistical parameters that represent the pruned data are calculated with $n = 18$ and are given by $\bar{T} = 1.998$ and $\sigma_e = 0.010$.

Dixon's Q test

Ordered sample data is characterized by the difference between the extreme values, also known as the range. If abnormal data is present, it may be at either end of the table. We may identify such an outlier by comparing the difference between the suspected outlier and its nearest neighbor to the range and decide whether to retain or reject data. In the Dixon's Q (Rejection Quotient) test we calculate the ratio of the difference alluded to above with the range and compare this ratio with a critical Dixon Q value to make a decision. Table of critical Q values are presented in a paper by Rorabacher.[19]

Several Q values are defined as given in the cited reference. Correspondingly different Q's are used to make a decision regarding retention or rejection of data. We shall demonstrate the method by taking a simple example.

Example 1.17

A certain experiment has been conducted by 12 students using the same experimental set up. The value of a measured quantity estimated by various students is given in the following table. Would you like to discard any data? If so which ones and why? Base your decision on Dixon's test. What are the mean and error estimator for the pruned data.

i	v_i	i	v_i
1	7.962	7	8.000
2	8.147	8	8.059
3	8.063	9	7.949
4	8.144	10	8.664
5	8.170	11	8.035
6	8.052	12	6.579

[19] D. B. Rorabacher, Statistical Treatment for Rejection of Deviant Values: Critical Values of Dixon's "Q" Parameter and Related Subrange Ratios at the 95% Confidence Level, Analytical Chemistry, Vol. 63, NO. 2, pp. 139-146, 1991

Solution :

Step 1 Arrnage the data in ascending order as shown below.

i	v_i	i	v_i
1	**6.579**	7	8.059
2	7.949	8	8.063
3	7.962	9	8.144
4	8.000	10	8.147
5	8.035	11	8.170
6	8.052	12	**8.664**

Supected outliers are shown **bold**. The difference between adjacent readings in ordered data is the basis for this suspicion. Thus we suspect v_1 or v_{12}.

Step 2 We use Dixon's ratio defined as $r_{10} = \dfrac{v_2 - v_1}{v_{12} - v_1}$ or $r_{10} = \dfrac{v_{12} - v_{11}}{v_{12} - v_1}$, following the cited reference.

$$r_{10} = \frac{7.949 - 6.579}{8.664 - 6.579} = 0.657 \quad \text{or} \quad r_{10} = \frac{8.664 - 8.170}{8.664 - 6.579} = 0.237$$

Step 3 Critical value for r_{10} with $n = 12$ is read from the table as $r_{10,\text{critical}} = 0.426$. At once we see that v_1 is to be discarded.

Step 4 The pruned data at this stage will consist of 11 values from $i = 2$ to $i = 12$ from the original set. The data may be renumbered so that i spans from 1 to 11. The last data point is the next suspect data.

Step 5 The reader may redo the test with the 11 data and show that the last data point is also to be discarded.

Step 6 Finally the following data set with 10 data points is obtained.

i	v_i	i	v_i
1	7.949	6	8.059
2	7.962	7	8.063
3	8.000	8	8.144
4	8.035	9	8.147
5	8.052	10	8.170

Step 7 We report the mean and standard deviation for the pruned data as $\bar{v} = 8.058$ and $\sigma_e = 0.077$.

1.4 Propagation of errors

Replicate data collected by measuring a single quantity, enables us to calculate the best value and characterize the spread by the variance with respect to the best value, using the principle of least squares. Now we look at the case of a derived quantity that is estimated from the measurement of several primary quantities. The question that needs to be answered is the following:

"A derived quantity D is estimated using a formula that involves the primary quantities $a_1, a_2 ... a_m$. Each one of these is available in terms of the respective best values $\bar{a}_1, \bar{a}_2 ... \bar{a}_m$ and the respective variances $\sigma_1^2, \sigma_2^2 ... \sigma_m^2$. What is the best estimate for D and what is the corresponding variance σ_D?"

We have, by definition

$$D = D(a_1, a_2 ... a_m) \tag{1.34}$$

It is *obvious* that the best value of D should correspond to that obtained by using the best values for the a's. Thus, the best estimate for D is given by $\bar{D}$ as

$$\bar{D} = D(\bar{a}_1, \bar{a}_2 ... \bar{a}_m) \tag{1.35}$$

Again, by definition, we should have:

$$\sigma_D^2 = \sum_{i=1}^{n} \frac{[D_i - \bar{D}]^2}{n-1} \tag{1.36}$$

In the above expression n represents the number of measurements that have been made and subscript i stands for the experiment number. The i^{th} estimate of D is given by

$$D_i = D(a_{1i}, a_{2i} ... a_{mi}) \tag{1.37}$$

If we assume that the spread in values are small compared to the mean or the best values (this is what one would expect from a good experiment), the difference between the i^{th} estimate and the best value may be written using a Taylor expansion around the best value as

$$\sigma_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{\partial D}{\partial a_1} \Delta a_{1i} + \frac{\partial D}{\partial a_2} \Delta a_{2i} + \ldots + \frac{\partial D}{\partial a_m} \Delta a_{mi} \right]^2 \tag{1.38}$$

where the partial derivatives are all evaluated at the best values for the a_i. The partial derivatives evaluated at the best values of a_i are also known as influence coefficients, usually represented as I_{a_i}. Note that only the first partial derivatives are retained in the above expansion. If a_i are all independent of one another then the errors in these are unrelated to one another and hence the cross terms $\sum_{i=1}^{N} \Delta a_{mi} \cdot \Delta a_{ki} = 0$ for $m \neq k$. Thus Equation 1.38 may be rewritten as

$$\sigma_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left[\left(\frac{\partial D}{\partial a_1} \Delta a_{1i} \right)^2 + \left(\frac{\partial D}{\partial a_2} \Delta a_{2i} \right)^2 + \ldots + \left(\frac{\partial D}{\partial a_m} \Delta a_{mi} \right)^2 \right] \tag{1.39}$$

Noting that $\sum_{i=1}^{n} \frac{\Delta a_{ji}^2}{n-1} = \sigma_j^2$ we may recast Equation 1.39in the form

$$\sigma_D^2 = \left[\left(\frac{\partial D}{\partial a_1} \sigma_1 \right)^2 + \left(\frac{\partial D}{\partial a_2} \sigma_2 \right)^2 + \ldots + \left(\frac{\partial D}{\partial a_m} \sigma_m \right)^2 \right] \tag{1.40}$$

Equation 1.40 is the error propagation formula. It may also be recast in the form

$$\boxed{\sigma_D = \pm\sqrt{\left(\frac{\partial D}{\partial a_1}\sigma_1\right)^2 + \left(\frac{\partial D}{\partial a_2}\sigma_2\right)^2 + \ldots + \left(\frac{\partial D}{\partial a_m}\sigma_m\right)^2}} \tag{1.41}$$

or

$$\boxed{\sigma_D = \pm\sqrt{(I_{a_1}\sigma_1)^2 + (I_{a_2}\sigma_2)^2 + \cdots + (I_{a_m}\sigma_m)^2}} \tag{1.42}$$

Example 1.18

A derived quantity D follows the relation $D = 0.023a_1^{0.8}a_2^{0.3}$ where a_1 and a_2 are measured quantities. In a certain case it has been determined that $a_1 = 20000 \pm 125$ and $a_2 = 5.5 \pm 0.2$. Determine the nominal value of D and specify a suitable uncertainty for the same. Which of the two quantities a_1 or a_2 has a bigger influence on the uncertainty in D?

Solution :

Step 1 Using the mean values we first estimate the best value for D.

$$D = 0.023 \times 20000^{0.8} \times 5.5^{0.3} = 105.8$$

Step 2 In this case it is possible to perform logarithmic differentiation to get the error propagation formula. Taking logarithms on both sides of the defining relation between D and the a's we have

$$\ln(D) = \ln(0.023) + 0.8\ln(a_1) + 0.3\ln(a_2)$$

Differentiating the above we get

$$\frac{dD}{D} = 0.8\frac{da_1}{a_1} + 0.3\frac{da_2}{a_2} \quad \text{or} \quad dD = 0.8\frac{Dda_1}{a_1} + 0.3\frac{Dda_2}{a_2}$$

We recognize the influence coefficients as $I_{a_1} = \frac{0.8D}{a_1}$ and $I_{a_2} = \frac{0.8D}{a_2}$. These may be used in Equation 1.42 to obtain the desired result.

Step 3 Calculate the influence coefficients now.

$$I_{a_1} = \frac{0.8 \times 105.8}{20000} = 4.234 \times 10^{-5}; \quad I_{a_2} = \frac{0.3 \times 105.8}{5.5} = 5.773$$

Step 4 Set $\sigma_{a_1} = 125$ and $\sigma_{a_2} = 0.2$, use Equation 1.42 to get

$$\sigma_D = \sqrt{(4.234 \times 10^{-5} \times 125)^2 + (5.773 \times 0.2)^2} = 1.270$$

Step 5 We also have the following:

$$I_{a_1}\sigma_{a_1} = 4.234 \times 10^{-5} \times 125 = 0.529; \quad I_{a_2}\sigma_{a_2} = 5.773 \times 0.2 = 1.155$$

The uncertainty in measured quantity a_2 has a bigger influence on the uncertainty of D.

Role of variances: In Example 1.18 we have seen that uncertainty in the measured quantity a_2 had a larger influence on the uncertainty in the derived quantity D. The error propagation formula simply states that the variance in the derived quantity is weighted sum of variances of the measured quantities. The weights are the respective squares of influence coefficients. The fractional contribution of the respective weighted variances to the variance of the derived quantity gives us the relative influences of the variances in the measured quantities. In Example 1.18 the fractional contributions to variance in D from a_1 is $\frac{(I_{a_1}\sigma_{a_1})^2}{\sigma_D^2} = \frac{0.529^2}{1.270^2} = \frac{0.280}{1.613} = 0.174$ while that from a_2 is $\frac{(I_{a_2}\sigma_{a_2})^2}{\sigma_D^2} = \frac{1.155^2}{1.270^2} = \frac{1.333}{1.613} = 0.826$.

Example 1.19

Two resistances R_1 and R_2 are given as *1000±25* Ω and *500±10* Ω. Determine the equivalent resistance when these two are connected in a) parallel and b) series. Also determine the uncertainties in these two cases.

Solution :

Given Data: $R_1 = 1000\,\Omega$, $R_2 = 500\,\Omega$, $\sigma_1 = 25\,\Omega$, $\sigma_2 = 10\,\Omega$,

Case a) Resistances connected in parallel:

Equivalent resistance is

$$R_p = \frac{R_1 \cdot R_2}{R_1 + R_2} = \frac{1000 \times 500}{1000 + 500} = 333.3\,\Omega$$

The influence coefficients are

$$I_1 = \frac{\partial R_p}{\partial R_1} = \frac{R_2}{R_1 + R2} - \frac{R_1 \times R_2}{(R_1 + R_2)^2} = \frac{500}{1000 + 500} - \frac{1000 \times 500}{(1000 + 500)^2} = 0.111$$

$$I_2 = \frac{\partial R_p}{\partial R_2} = \frac{R_1}{R_1 + R2} - \frac{R_1 \times R_2}{(R_1 + R_2)^2} = \frac{1000}{1000 + 500} - \frac{1000 \times 500}{(1000 + 500)^2} = 0.444$$

Hence the uncertainty in the equivalent resistance is

$$\begin{aligned} \sigma_s &= \pm\sqrt{(I_1\sigma_1)^2 + (I_2\sigma_2)^2} \\ &= \pm\sqrt{(0.111 \times 25)^2 + (0.444 \times 10)^2} = \pm 5.24\,\Omega \end{aligned}$$

Case b) Resistances connected in series:

Equivalent resistance is

$$R_s = R_1 + R_2 = 1000 + 500 = 1500\,\Omega$$

The influence coefficients are

$$I_1 = \frac{\partial R_s}{\partial R_1} = 1;\ I_2 = \frac{\partial R_s}{\partial R_2} = 1$$

Hence the uncertainty in the equivalent resistance is

$$\sigma_s = \pm\sqrt{(I_1\sigma_1)^2 + (I_2\sigma_2)^2} = \pm\sqrt{(1 \times 25)^2 + (1 \times 10)^2} = \pm 26.93\,\Omega$$

1.5 Specifications of instruments and their performance

In this section we look at the limitations introduced by the instruments used for making measurements. In the past measuring instruments were mostly of the analog type with the reading displayed by a pointer moving past a scale. Resolution of such instruments were basically limited to the smallest scale division. Of course, in addition, the manufacturer would also specify the accuracy as a percentage of the full scale reading, based on calibration with reference to a standard.

In recent times most measuring instruments are digital in nature and the performance figures are specified somewhat differently. Take the example of meter that displays $4\frac{1}{2}$ digits[20]. The reading of the instrument may be anywhere between 0.0000 and 1.9999. The number of counts[21] is 20000. Accuracy specification is usually represented in the form $\pm$(%of reading + counts). For example, typical specification of a DMM (Digital Multi Meter)) is of form $\pm$(0.5 %of reading + 5 counts) when DC voltage is being measured. In a typical example, we may be measuring the voltage of a DC source whose nominal value is 1.5 V. This DMM would give a reading in between 1.492 and 1.508 V. We take an example below to show how the instrument specification affects the measurement.

Example 1.20

A resistor is picked up from a lot labeled 150 Ω with a precision of 1%. Its value is measured using a DMM which has a range of 0 − 600.0 Ω, accuracy of $\pm$(0.9 %of reading + 2 counts). What would be the expected outcome of the measurement?

Solution :

Step 1 The nominal value of the resistor is $R = 150\,\Omega$. Precision of 1% would mean that it may have a minimum value of $R_l = 148.5\,\Omega$ and a maximum value of $R_m = 151.5\,\Omega$.

Step 2 Let us assume that the actual value of the resistor is R_l. The meter will then give either of the readings shown in the last row below:

R_l =	148.5	Ω
0.9% of R_l =	1.3	Ω
Error due to count =	0.2	Ω
Reading is either 147.0 Ω or 150.0 Ω		

Step 3 Let us assume that the actual value of the resistor is R_m. The meter will then give either of the readings shown in the last row below:

[20] Most significant digit can be either 0 or 1 while all other digits may have any value between 0 and 9.

[21] Number of levels equals the number of counts.

R_m =	151.5	Ω
0.9% of R_m =	1.4	Ω
Error due to count =	0.2	Ω

Reading is either 149.9 Ω or 153.1Ω

Step 4 Thus the actual value of the resistance indicated by the instrument may be anywhere between 147.0 and 153.1 Ω.

Example 1.21

In a wind tunnel air flow is maintained steady at a nominal speed of 15 m/s. A vane anemometer is used to measure the wind speed. Specify an error bar for the measurement if the resolution (4 digit display) of the anemometer is 0.01 m/s and the accuracy specification by the manufacturer of the anemometer is ±(3% reading + 0.20) m/s.

Solution :

The nominal value of the wind speed is $V = 15\ m/s$. We assume that the reading of the anemometer is this value i.e. 15.00 m/s. Using the accuracy specification the uncertainty is calculated as follows:

1. 3%of reading $= \dfrac{3 \times 15.00}{100} = 0.45\ m/s$
2. Total uncertainty is equal to $0.45 + 0.2 = 0.65\ m/s$.

Hence the measured velocity is specified as $V = 15 \pm 0.65\ m/s$. In terms of percentages the uncertainty in the wind speed is ±4.33%.

Concluding remarks

This chapter has set the tone for the rest of the book by presenting statistical principles that play important role in analysis and interpretation of experimental results. Properties of normal distribution are relevant in most measurements and hence have been discussed in detail. Test for normality and data rejection based on sound statistical principles have been presented. Other topics considered include sampling theory and error propagation.