

# Chapter 10

## Comparison and Rejection of Measurements

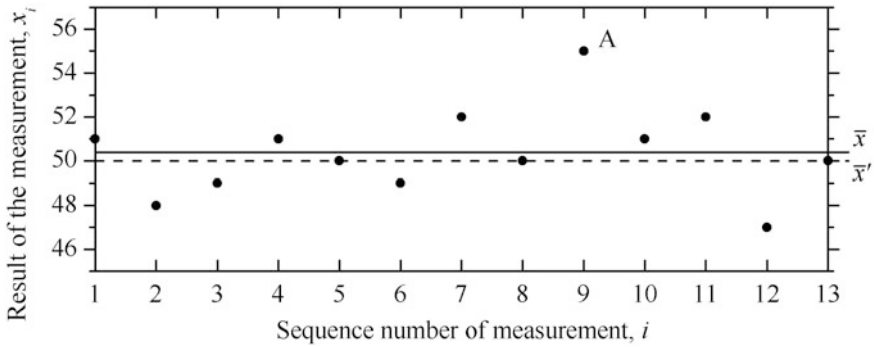
The question of whether to reject a measurement arises very often. The question is then whether the observed large deviation of a measurement from the mean of the series of measurements to which it belongs is due to expected and acceptable random errors, or is the result of a mistake during measurement. It must be borne in mind that many researchers are of the opinion that no measurement should be rejected, as a matter of principle, because this would alter the results based on subjective criteria. Nevertheless, the question frequently arises and we will present here the criteria by which a measurement, even if not finally rejected, is placed on our list as the result of some unusual and unknown sequence of events.

### 10.1 The Problem of the Rejection of Measurements

Assume that we have  $N$  results  $x_i$  ( $i = 1, 2, \dots, N$ ) of measurements of the magnitude  $x$ . These results have a mean value  $\bar{x}$  and sample standard deviation  $s_x$ . If the distribution of the errors in  $x$  is Gaussian, the results will be distributed about their mean with this standard deviation. If we had a large number of measurements, we would expect some values to differ by a large amount from the mean.

The question that arises is: given the number  $N$  of measurements, by how much must a result differ from the mean for us to conclude that the difference is improbable to be due to random errors, but is rather the result of a mistake during the experimental procedure and, therefore, this measurement must be rejected as unacceptable?

Figure 10.1 shows the results of 13 measurements of the magnitude  $x$ , in the order in which they were obtained. The results are also given in Table 10.1. We find that  $\sum x_i = 655$  and, therefore, the mean of the measurements is  $\bar{x} = 50.4$ . The mean value is marked in Fig. 10.1 by a horizontal (full) straight line at  $x = 50.4$ .



**Fig. 10.1** A series of 13 measurements, among which there is one considered for rejection (A)

**Table 10.1** The results of 13 measurements, one of which is a candidate for rejection

$i$	$x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	51	0.62	0.384
2	48	-2.38	5.664
3	50	-0.38	0.144
4	51	0.62	0.384
5	50	-0.38	0.144
6	48	-2.38	5.664
7	52	1.62	2.624
8	50	-0.38	0.144
9	55	4.62	21.344
10	51	0.62	0.384
11	52	1.62	2.624
12	47	-3.38	11.424
13	50	-0.38	0.144
$\Sigma =$	655		51.08

From the sum  $\sum (x_i - \bar{x})^2 = 51.08$ , the standard deviation of the measurements is calculated to be  $s_x = \sqrt{51.08/13} = 2.0$ .

The 9th measurement differs from the mean by  $d_9 = 55 - 50.4 = 4.6$ . The difference is equal to  $4.6/2.0 = 2.3$  standard deviations.

Assuming that the distribution of the measurements about the mean is normal, we may calculate the probability for a measurement to differ from the mean by at least 2.3 standard deviations either in the positive or the negative direction. From Table 4.2, we find that this probability is equal to  $1 - 2 \times 0.4893 = 0.021$  or 2.1%. We expect that, due to random errors, one measurement in 48 will differ from the mean by more than 2.3 standard deviations. In our measurements we find one such measurement in only 13. The expected number would be  $0.021 \times 13 = 0.27$ . The observed number, 1, is almost 4 times the expected. We therefore conclude that,

most probably, this measurement differs from the mean by as much as it does not because of random errors but due to other causes and we reject it.

Having rejected the 9th measurement, we recalculate the mean and the standard deviation of the remaining 12 values. We find  $\bar{x}' = 600/12 = 50$ . This new mean value is also marked in Fig. 10.1 with a horizontal dashed line. The sum  $\sum (x_i - \bar{x}')^2 = 28$ , gives a standard deviation of the 12 measurements  $s'_x = \sqrt{28/12} = 1.5$ .

Among the remaining 12 measurements, the value with initial order number 12 is now the candidate for rejection. This measurement differs from the new mean by  $|47 - 50|/1.5 = 2$  new standard deviations. From Table 4.2 we find that the probability of finding a measurement which differs from the mean by at least 2 standard deviations, either in the positive or the negative direction, is equal to  $1 - 2 \times 0.4773 = 0.045$  or 4.5%. The expected number of such results in our 12 measurements is 0.54. Obviously this measurement must not be rejected, as the ratio of the observed to the expected number is near unity.

There seems to be a need for a criterion which, although not entirely objective, would at least be commonly agreed to, thus reducing to a certain degree the subjective factor. One such criterion we will examine below.

## 10.2 Chauvenet's Criterion

Chauvenet proposed the following criterion for the rejection of measurements:

*A measurement belonging to a group of  $N$  measurements is rejected if its difference from the mean of the measurements is such that the probability of observation of such a difference or greater is less than  $1/(2N)$ .*

In other words, Chauvenet's criterion rejects a measurement if the expected number of such measurements with a differences from the mean equal to or larger than its deviation is less than  $\frac{1}{2}$ . Obviously, the number  $\frac{1}{2}$  is arbitrary and this is one of the objections against this criterion or any other criterion.

For use with Chauvenet's criterion, Table 10.2 gives the probability that the absolute difference of a value of  $x$  from the mean  $\bar{x}$  is equal to or greater than  $v$  times the standard deviation  $s_x$  of the measurements, as a function of  $v$ . The difference of  $x$  from the mean  $\bar{x}$  is expressed, in units of  $s_x$ , as  $|x - \bar{x}|/s_x$  and the probability for this difference to be equal to or greater than  $v$  is denoted by  $\Pr\left\{\frac{|x - \bar{x}|}{s_x} \geq v\right\}$ .

According to Eq. (4.65), the probability for a value of  $x$  to differ from the mean  $\bar{x}$  by more than  $v$  times the standard deviation  $s_x$ , is given by the relation

$$\Pr\{x \leq \bar{x} - vs_x \text{ or } x \geq \bar{x} + vs_x\} = 1 - \operatorname{erf}\left(\frac{v}{\sqrt{2}}\right) \equiv \operatorname{erfc}\left(\frac{v}{\sqrt{2}}\right) = 1 - 2\Phi(v). \quad (10.1)$$

**Table 10.2** Probability for the absolute difference of a value of  $x$  from the mean  $\bar{x}$  being equal to or greater than  $v$  times the standard deviation of the measurements,  $s_x$ 

$v$	$\Pr\left\{\frac{ x-\bar{x} }{s_x} \geq v\right\}$	$v$	$\Pr\left\{\frac{ x-\bar{x} }{s_x} \geq v\right\}$	$v$	$\Pr\left\{\frac{ x-\bar{x} }{s_x} \geq v\right\}$	$v$	$\Pr\left\{\frac{ x-\bar{x} }{s_x} \geq v\right\}$
1.6	0.1096	2.1	0.0357	2.6	0.00932	3.1	0.001936
1.7	0.0891	2.2	0.0278	2.7	0.00694	3.2	0.001374
1.8	0.0719	2.3	0.0214	2.8	0.00512	3.3	0.000966
1.9	0.0574	2.4	0.01640	2.9	0.00374	3.4	0.000674
2	0.0455	2.5	0.01242	3	0.00270	3.5	0.000466

We wish to find the limit for rejection of values,  $v_C \equiv \frac{|x-\bar{x}|}{s_x}$ , according to Chauvenet's criterion. From Eq. (10.1) we have  $\Pr\{x \leq \bar{x} - v_C s_x \text{ or } x \geq \bar{x} + v_C s_x\} = \frac{1}{2N}$  or

$$\operatorname{erf}\left(\frac{v_C}{\sqrt{2}}\right) = 1 - \frac{1}{2N} \quad \text{and} \quad \Phi(v_C) = \frac{1}{2} - \frac{1}{4N} \quad (10.2)$$

from which, for a given  $N$ , we may find the corresponding value of  $v_C$ .

Given in the Table 10.3, for various values of the number  $N$  of the measurements, is the limit for rejection,  $v_C \equiv |x - \bar{x}|/s_x$ , according to Chauvenet's criterion. A value, out of a total of  $N$  values, is rejected if the absolute value of its difference from the sample mean,  $|x - \bar{x}|$ , is larger than  $v_C s_x$ , where the value of  $v_C$  is found from the table for the corresponding  $N$ . For example, one measurement in a series of 10 measurements is rejected if it differs from the sample mean by more than  $1.96 s_x$ .

### Example 10.1

Apply Chauvenet's criterion to the measurements of Table 10.1.

The  $N = 13$  measurements of the table have a mean equal to  $\bar{x} = 50.4$  and a standard deviation  $s_x = 2.0$ . The 9th measurement differs from the mean by

$$\frac{|x_9 - \bar{x}|}{s_x} = \frac{|55 - 50.4|}{2.0} = \frac{4.6}{2.0} = 2.3 \text{ standard deviations.}$$

From Table 10.2 we find that the probability for a difference from the mean greater than or equal to  $2.3s_x$  is  $\Pr\left\{\frac{|x-\bar{x}|}{s_x} \geq 2.3\right\} = 0.0214$ . For  $N = 13$ , it is  $1/(2N) = 0.0385$ .

Since it is  $\Pr\left\{\frac{|x-\bar{x}|}{s_x} \geq 2.3\right\} < 1/(2N)$ , the 9th measurement is rejected.

**Table 10.3** The rejection limit of a value,  $v_C \equiv |x - \bar{x}|/s_x$ , according to Chauvenet's criterion, as a function of the number  $N$  of measurements

$N$	$v_C$	$N$	$v_C$	$N$	$v_C$	$N$	$v_C$
–	–	21	2.26	41	2.51	110	2.84
–	–	22	2.28	42	2.51	120	2.87
–	–	23	2.29	43	2.52	130	2.89
4	1.53	24	2.31	44	2.53	140	2.92
5	1.65	25	2.33	45	2.54	150	2.93
6	1.73	26	2.34	46	2.55	160	2.94
7	1.80	27	2.35	47	2.55	170	2.96
8	1.86	28	2.37	48	2.56	180	2.98
9	1.91	29	2.38	49	2.57	190	3.00
10	1.96	30	2.39	50	2.58	200	3.02
11	2.00	31	2.41	55	2.61	218	3.05
12	2.04	32	2.42	60	2.64	258	3.10
13	2.07	33	2.43	65	2.67	306	3.15
14	2.10	34	2.44	70	2.69	364	3.20
15	2.13	35	2.45	75	2.72	433	3.25
16	2.16	36	2.46	80	2.74	517	3.30
17	2.18	37	2.47	85	2.76	619	3.35
18	2.20	38	2.48	90	2.77	742	3.40
19	2.22	39	2.49	95	2.79	892	3.45
20	2.24	40	2.50	100	2.81	1075	3.50

Alternatively, from Table 10.3 we find that, for 13 measurements, a measurement that differs from the mean by more than 2.07  $s_x$  is rejected. For a difference of 2.3  $s_x$ , the rejection is justified.

The remaining 12 measurements now have a mean  $\bar{x}' = 50$  and a standard deviation of  $s'_x = 1.5$ .

The measurement with number 12 differs from the new mean by

$$\frac{|x_{12} - \bar{x}'|}{s'_x} = \frac{|47 - 50|}{1.5} = \frac{3}{1.5} = 2.0 \text{ standard deviations.}$$

From Table 10.2 we find that the probability for a difference from the mean greater or equal to  $2.0s_x$  is  $\Pr\left\{\frac{|x - \bar{x}|}{s_x} \geq 2.0\right\} = 0.0455$ . For  $N = 12$ , it is  $1/(2N) = 0.0417$ .

Since it is  $\Pr\left\{\frac{|x - \bar{x}|}{s_x} \geq 2.0\right\} > 1/(2N)$ , the 12th measurement is not rejected.

Alternatively, from Table 10.3 we find that, for  $N = 12$ , a measurement that differs from the mean by up to 2.04  $s_x$  is not rejected. With a difference of  $2.0s_x$ , the 12th measurement should not be rejected.

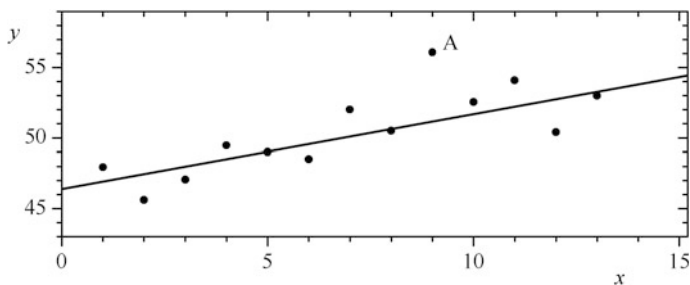
### 10.3 Comments Concerning the Rejection of Measurements

Before we take a stand on the subject of the ‘rejection’ of measurements, we must clarify exactly what we mean by this term. We examine the two main different possibilities below:

In the case we have examined (Fig. 10.1), we had a series of measurements of the same quantity, which were taken under identical experimental conditions, as much as this was possible. Of course, it is not possible to maintain the conditions unchanged. It is good practice, in all experimental work, to keep detailed notes for everything happening and resort to these in an effort to find what had possibly changed during the taking of the ‘suspect’ measurement and which led to the difference observed. It is, however, rather improbable that the causes have been recorded, given that, if changes in the experimental conditions had been noted, they would have been corrected before the execution of the measurement. The opinion that the measurement must be repeated is not the solution. In a series of  $N$  measurements, the measurement has been repeated  $N - 1$  times! If the rejection of results alters significantly the final result, then it might be necessary, if it is still possible, for more measurements to be made. Of course, the danger exists here that we keep making measurements until we have a result we like. This would have much more serious consequences than the rejection of a measurement. Summarizing, however, we would say that, in cases such as this, the use of Chauvenet’s criterion is justified. In no case, however, should the criterion be used for the rejection of two or more measurements, even when the number of measurements is large, in which case the rejection of a measurement would not alter the final result significantly. In cases when the criterion suggests the rejection of two or more measurements, the possibility should be seriously examined that the distribution of the parent population is not normal, the deviations being more important at its tails. Naturally, when a measurement has been rejected, this should be stated.

Someone using systematically Chauvenet’s criterion in his work, sooner or later is bound to reject measurements which he should not reject. Large deviations are improbable but not impossible!

The second case is that in which we are dealing with measurements which are not expected to give the same result, and one differs significantly from its expected value. For example, if we measure the values of variable  $y$  as a function of another,  $x$ , and by the use of some method (such as the method of least squares, to be developed in the next chapter) we find the best mathematical relation between the two magnitudes (Fig. 10.2), then some point may deviate by so much from the expected value, that it is probable that this value is the result of a mistake (point A in Fig. 10.2). If the method used in the fitting of a curve to the results also gives the



**Fig. 10.2** A series of 13 measurements of  $y$  as a function of  $x$ , among which there is one candidate for rejection (A)

expected error in  $y$  for every value of  $x$ , it is possible to apply criteria for the rejection of some value.

In this case, greater attention is needed to the rejection of a measurement. The reason we might wish to reject ‘erroneous’ results is so that we then apply the curve-fitting method again to the remaining results and find more accurate values for the parameters of the function relating  $x$  and  $y$ . In contrast to the last case, however, we do not have other measurements taken under the same experimental conditions, with which we would compare the result under investigation. The right way to face the problem is to return to the laboratory (assuming that, quite wrongly, we have left it before detecting the problem!) and perform more measurements in the region of the suspect point. In the example of Fig. 10.2 this would mean the region between  $x = 8$  and  $x = 10$ . Only then will we be in a position to decide whether the measurement should be rejected or that the relation  $y(x)$  does not behave as assumed in this region and, perhaps, that the deviation is due to a hitherto unknown phenomenon.

In cases that the problem cannot be resolved in the way described above, the criterion will indicate whether something unusual is happening in that region of values, by giving an estimate of how probable it is for the deviation to be due to random errors. We must not forget, however, that for  $|x - \bar{x}|/s_x > 2.5$  our confidence that the normal distribution accurately describes the deviations is low.

We will conclude our discussion by mentioning one of the many cases in the history of science when not rejecting a value that appeared wrong lead to important discoveries. Rayleigh and Ramsay, in 1894, noticed that nitrogen produced in the laboratory was lighter than atmospheric nitrogen by 0.5%. The fact that they did not interpret the difference as being the result of random errors but considered the deviation to be real, lead them to conclude that an unknown gas was present in the sample of what was thought to be pure atmospheric nitrogen. Thus argon was discovered. Of course, in support of the statistical analysis of experimental results, it must be said that it was the knowledge of the possible errors in the measurements that lead the two scientists to suspect that the deviation was statistically significant.

## 10.4 Comparison of the Means of Two Series of Measurements of the Same Quantity

The need frequently arises for the comparison of the results of two series of measurements of the same quantity. The two series of measurements were, possibly, performed by the same experimenter at different times or were performed by different researchers. It is also possible that the same quantity was measured using two different experimental methods. A classical example is the measurement, in an educational laboratory, of a universal constant or the property of a material and the comparison of the results with the generally accepted values for these magnitudes, found in tables.

We suppose that we have two means,  $\bar{x}_1$  and  $\bar{x}_2$ , and their corresponding standard deviations,  $\sigma_{\bar{x}_1}$  and  $\sigma_{\bar{x}_2}$ . The two series must be considered to have been taken from the parent population of the infinite possible values that may result in the measurement of the magnitude  $x$ . Being finite samples from the same parent population, they are not expected to agree completely. The question of whether the two samples originated from the same parent population is answered in Statistics by Student's  $t$ -test. Criteria which are simpler to apply are used for experimental results in the laboratory, of which we will describe only two:

1. The two results are considered to be in agreement with each other (or, better, that there are no serious indications for the presence of systematic errors), if the difference of their two means is smaller than or equal to the sum of the standard deviations of the two values:

$$|\bar{x}_2 - \bar{x}_1| \leq \sigma_{\bar{x}_1} + \sigma_{\bar{x}_2}. \quad (10.3)$$

2. The two results are considered to be in agreement with each other if the difference of their two means is smaller than or equal to the standard deviation of the difference of the two values:

$$|\bar{x}_2 - \bar{x}_1| \leq \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2}. \quad (10.4)$$

We will consider the second criterion as somewhat more correct and we will use it in the example that follows.

### Example 10.2

In an educational laboratory, two students determined experimentally, by two different methods, the absolute value  $e$  of the charge of the electron and found the following values:



$$e_1 = (1.62 \pm 0.02) \times 10^{-19} \text{ C} \quad \text{and} \quad e_2 = (1.59 \pm 0.03) \times 10^{-19} \text{ C}$$

Check whether the two results are consistent with each other and if they agree with the value of  $e$  generally accepted today.

The absolute value of the difference of the two results is  $|e_2 - e_1| = 0.03 \times 10^{-19} \text{ C}$ .

The standard deviation of the difference of the two mean values is

$$\sigma_{e_2 - e_1} = \sqrt{(0.02)^2 + (0.03)^2} \times 10^{-19} = 0.036 \times 10^{-19} = 0.04 \times 10^{-19} \text{ C}.$$

Since it is  $|e_2 - e_1| < \sigma_{e_2 - e_1}$ , the two results are considered to agree with each other.

The absolute value of the electronic charge is given in tables as  $e = 1.602\,176\,565(35) \times 10^{-19} \text{ C}$ .

The standard error in this value is relatively negligible compared to the errors in the values obtained by the students. Therefore, as standard deviation of the difference of  $e$  and  $e_1$  will be taken to be the standard deviation of  $e_1$ , i.e.  $0.02 \times 10^{-19} \text{ C}$ . The difference between the values of  $e$  and  $e_1$  is somewhat smaller than the standard deviation of their difference. We therefore conclude that the value  $e_1$  is in agreement with the generally accepted today value of  $e$ , within the limits of the errors of the measurements. The absolute difference of  $e_2$  from  $e$  is  $0.01 \times 10^{-19} \text{ C}$ , which is quite smaller than the standard deviation of the difference of  $e_2$  and  $e$ , which is equal to  $0.03 \times 10^{-19} \text{ C}$ . Thus,  $e_2$  is considered to be in agreement with the generally accepted today value of  $e$ , within the limits of the errors of the measurements.

## Problems

10.1 The results of 13 measurements of the quantity  $x$  are:

$$9 \quad 6 \quad 5 \quad 9 \quad 7 \quad 9 \quad 6 \quad 10 \quad 4 \quad 7 \quad 8 \quad 5 \quad 13.$$

- Find the mean  $\bar{x}$  and the standard deviation  $s_x$  of the measurements.
- Should the 13th measurement,  $x = 13$ , be rejected according to Chauvenet's criterion?

10.2 The results of 10 measurements are:

$$126 \quad 72 \quad 162 \quad 144 \quad 252 \quad 162 \quad 135 \quad 135 \quad 153 \quad 117.$$

Is there a result that should be rejected according to Chauvenet's criterion?

10.3 A series of 37 measurements resulted in the values  $x_r$  with frequencies  $n_r$

$x_r$	31.9	32.0	32.2	32.3	32.4	32.5	32.6	33.0
$n_r$	1	3	7	12	6	6	1	1

Use Chauvenet's criterion in order to decide whether the last measurement should be rejected.

10.4 Two series of measurements of the same quantity gave the results  $x_1 = 1.518 \pm 0.012$  and  $x_2 = 1.535 \pm 0.015$ . Are the two results mutually compatible?

10.5 Two series of measurements of the same quantity gave the results  $x_1 = 163 \pm 6$  and  $x_2 = 180 \pm 4$ . Are the two results mutually compatible?