

Chapter 4

Elements of Scientific Measurement

4.1 Issues in measurement

In most practical situations, a scientist has to measure something. The quantity to be measured could be the average weight of adult male sparrows, the number of microbes in unit volume of a fluid, or the electrical charge of an electron.

In the first case the weight of each sparrow is really different from that of another, because of the variation inherent in that species. If one could somehow measure the weight of *all* adult male sparrows, one could obtain the true answer to the question. But that enterprise is practically impossible, due to constraints of time and money. So one has to obtain a smaller sample and has to obtain the mean. For the scientific question in hand, it could also be important to measure the variability within the species (because it is the variation that natural selection acts on). Even though the weight of a sparrow is a continuous variable, one always measures upto a definite accuracy: the least count of the instrument used.

In the second case, the microbes may not be uniformly distributed in the liquid, and for that reason one would have to take samples from different parts of the liquid. Counting such tiny

living organisms may also have its problems: each counting may not be accurate. One could miss or overcount. If one does not assume any particular propensity of the experimenter to miss or overcount, the error in counting may be assumed to be random.

In case of the measurement of the charge of an electron, the value to be measured is really a constant. But in the process of measurement, many errors would creep in—systematic as well as random. The scientist would have to remove all chances of systematic errors, but has to learn to live with random errors.

In all these cases there is some objective value of the quantity to be measured, and the scientist has to reach as close as possible to that value, subject to the constraints of time and money, by taking *samples*. Even in the case of measuring the charge of an electron, in each measurement he is really taking samples from a theoretically infinite number of possible readings.

4.2 Analyzing the sampled data

The first thing one does after recording the data is to obtain the mean given by

$$\text{Mean : } \bar{x} = \frac{1}{n}(x_1 + x_2 \cdots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i \quad (4.1)$$

One would also like to know how much the data points deviate from the mean value on an average. Thus one would like to obtain an average of $(x_i - \bar{x})$. But this quantity $(x_i - \bar{x})$ could be positive as well as negative, although the average of the quantities $(x_i - \bar{x})$ is zero. If we really want a measure of the deviation from the mean value, it should be a positive number. To overcome this problem, we take the square of $(x_i - \bar{x})$ and obtain its average. Thus, we obtain

$$\text{Variance : } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4.2)$$

The ‘standard deviation’ s is just the square root of the variance. This is thus a measure of how much the data points deviate from the mean value on an average.

It has been found that if the divisor in (4.2) is n , then s underestimates the population standard deviation σ . That is why the divisor in (4.2) is taken as $n - 1$ to compensate for it.

These are the values obtained from the samples. The ‘true’ mean of the population will be denoted by μ and the ‘true’ standard deviation of the population will be denoted by σ . The attempt is to obtain a value of \bar{x} as close as possible to μ and a value of s as close as possible to σ .

Example-1

You have conducted an experiment to measure a quantity x , and have obtained the following data.

5.23	4.97	4.78	5.05	5.34	4.78	4.92	4.89
5.10	5.22	4.80	4.94	5.06	4.96	5.03	5.15
5.26	4.92	4.78	4.98	5.01	5.19	5.08	5.15

Obtain the mean and standard deviation of x .

Solution: The mean of x is

$$\bar{x} = \frac{1}{24} \sum_{i=1}^n x_i = 5.018$$

and the standard deviation of x is

$$s = \sqrt{\frac{1}{23} \sum_{i=1}^n (x_i - \bar{x})^2} = 0.16 \quad \square$$

4.3 Distribution of the data

The next step is to obtain the frequency plots. In case of the weight of sparrows, one would have to divide the whole range of weights into discrete ‘bins’ and have to count how many data

points fell into each bin. The number of data points in each bin divided by the total number of data points give the normalized frequency of each bin. One then plots the graph of the weight versus the frequency of each weight. An example of such a graph is shown in Fig. 4.1. For different observations the graph may have different shapes.

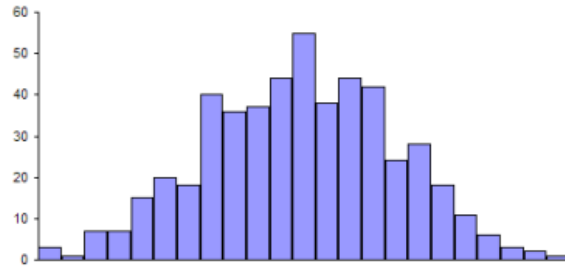


Figure 4.1: A typical histogram obtained from observational data

If the thing to be measured has an inherent variability (like the variation within a species), the frequency curve may have specific characteristics reflecting the nature of variability. But if the variation is purely due to random errors, one would expect a bell-shaped curve: the so-called ‘normal’ distribution. This curve is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad -\infty < x < \infty \quad (4.3)$$

where μ is the mean and σ is the standard deviation. The function is graphed in Fig. 4.2. The position of the standard deviation is shown on the graph. If a set of data has a smaller standard deviation, the graph is narrow and tall and if σ is large, it is broader and of shorter height.

One can integrate this curve in different ranges to find the fraction of the population that can be expected to lie in specific ranges:

- Approximately 68.3% of the population are within 1 standard

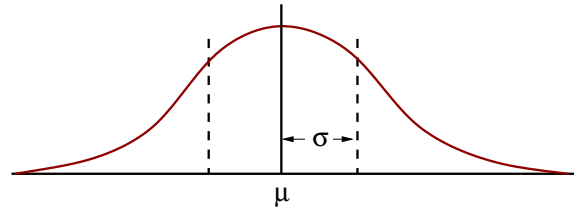


Figure 4.2: The normal distribution .

deviation of the mean (that is, between $\mu - \sigma$ and $\mu + \sigma$);

- Approximately 95.4% of the population are within 2 standard deviations of the mean (that is, between $\mu - 2\sigma$ and $\mu + 2\sigma$);
- Almost all the population – about 99.7% – are within 3 standard deviations of the mean (that is, between $\mu - 3\sigma$ and $\mu + 3\sigma$).

One is normally interested in finding the standard deviations within which certain specific percentages of the population lie. These are summarized as follows.

- Approximately 90% of the population are within 1.64 standard deviations of the mean;
- Approximately 95% of the population are within 1.96 standard deviations of the mean;
- Approximately 99% of the population are within 2.58 standard deviations of the mean.

It helps to remember these values.

In all observational or experimental situations, there is an objectively existing distribution for the quantity being measured, with ‘true’ values of the mean μ and the standard deviation σ . The challenge is to get as close as possible to these values through a finite number of samplings. Note that, after a scientist publishes his or her results, hundreds of scientists worldwide will repeat the

experiment or observation, and it is almost certain that others will not obtain exactly the same numbers obtained by that scientist. Given this reality, how can the scientist state the results in a way that would be acceptable to the scientific community?

4.4 Sampling

When a scientist conducts an experiment, often he has to measure the value of a parameter or a constant. Obviously the procedure of measurement of the mass of an electron will be quite different from the procedure of measuring the density of Earth's crust. The fundamental difference is that in the former case the quantity to be measured is really a constant (since all electrons have identical rest mass) while in the latter, the density varies from place to place and one is trying to measure the *average* density. In the class of measurement of the second category, one has to exercise great caution in choosing the samples from which the information about the average is to be extracted. Normally a random sampling is recommended. If variability is expected, one has to take great care to collect samples representing that variability.

Let us illustrate that with an example.

Suppose you are trying to find the character of soil in a given field. If you go to different locations within the field, or if you collect soil from different depths, you will find significant variability in the character of the soil. How, then, can one get a 'representative' sample which can be tested to get an idea about the character of soil in the field?

This is a typical problem of sampling, and the recommended procedure of soil collection gives a good idea about the sampling procedure to be adopted in other areas.

The recommended procedure is as follows.

We take a 10,000 sqm area in the field. We dig 15 cm holes at every 20 m distances, and collect some amount of soil (say, half a kg) from the top, bottom, and middle of the hole. Soil is collected

in a sack from similar holes 20 m apart. Then the collected soil is spread over a hard surface and is dried in the sun. After the soil is dry, it is mixed well and spread over a square area. The square is then divided into 4 equal parts. Two of them along the main diagonal are kept for further processing and the other two are discarded. The remaining part is again mixed well, spread over another square area, divided into four parts, two along the diagonal are kept and the rest are thrown away. The process is repeated until the remaining amount is around 2-3 kg. Then the bigger particles are ground, and the material is passed through a 2 mm sieve. The soil that goes through the sieve is considered to be the representative sample of the soil in the field, on which tests are done.

Similarly, standard procedures of sampling exist in almost all fields. One has to learn the procedures before proceeding to conduct any experiment. If such time-tested procedures are not available in a field of enquiry, one has to develop a procedure that ensures that the range of variability is aptly represented in the sample. The procedure adopted has to be clearly stated in the paper or scientific report.

After the samples are collected, one proceeds to measure the parameters in question.

4.5 Experimental Errors

All measurements are prone to errors, and a scientist has to be conscious of this fact when making measurements. Errors can be divided into two major categories:

Random errors: These are fluctuations in readings around the actual value being measured, caused by thermal and other sources of noise.

Systematic errors: These are consistent deviations of the measurement from the value being measured, caused by definite causative factors.

A researcher has to consciously try to avoid all possibilities of systematic errors. There are two general prescriptions of doing so. First, one has to check the calibration of each measuring instrument, because these may change with time and environmental conditions under which an experiment is conducted. Second, to do the measurement of a value in two or more different ways, because it is unlikely that the same systematic error would creep in two different sets of apparatus. Apart from these two prescriptions, there are no other general guidelines, because in different experiments different causative factors may be operative that influence the readings.

But there are very definite prescriptions in dealing with random errors. The first prescription is that, the experiment should be so planned that it is possible to make a large number of measurements of the same quantity, under varying conditions. For example, if one is interested in measuring the electrical resistance of a sample, one should make arrangements for applying a variable voltage (which can be done with a potential divider) and to measure the current for each value of the voltage. When the measured values are tabulated, the resistance can be obtained by dividing the voltage across the sample by the current through it. We thus get a large number of measured values of the resistance, say, $x_1, x_2, x_3 \cdots x_n$ which typically will be slightly different from each other due to random error. One can then take the average of the measured values

$$\bar{x} = \frac{x_1 + x_2 + x_3 \cdots + x_n}{n}$$

and can hope that the positive errors will cancel out the negative ones, thus getting a mean value close to the actual value being measured.

But still many questions remain.

- How many observations need to be taken in order to obtain a confident estimate?
- Which value out of the large number of observations can be

stated as the “measured value”?

- How reliable will the measured mean be as an estimate of the actual value?
- How can one state the measured value so that it will be acceptable to the scientific community?
- Should he declare a range within which the actual value is likely to lie?
- With what level of confidence can the scientist state that the value to be measured actually lies within this range?

All these questions require statistical treatment.

4.6 Specifying the measured value

Modern science is heavily dependent on statistical methods. There is hardly any quantitative treatment of a problem in any branch of science where one can avoid this method. In most situations the use of statistical method becomes indispensable for a scientist.

Let us consider the first question. Suppose somehow you are able to take an *infinite number* of readings. These will give a distribution with a certain mean μ and a certain variance σ^2 . This mean would be a reliable representation of the value you are trying to measure. But we cannot physically take an infinite number of readings, and have to be constrained to a finite number, say n , i.e., you take *samples* from the theoretically infinite number of possible readings. The question is, how reliable will this estimate of mean be?

Let us consider a real-life situation. Suppose a biologist has discovered a new species of insect and wants to measure the average body-weight of these organisms¹. She will have to catch

¹If you are a geologist, you might think of the task of measuring the average density of the Earth's crust; if you are a physicist, you might think of measuring

a few of these insects and will have to weigh them. By the act of catching a few insects, she is actually 'sampling' from a population of insects, and typically the population will be much larger than the samples chosen. How can she make an objective estimate about the character of the species by taking a relatively small number of samples?

Note that the body-weight of the organisms in the insect species might have a distribution that is not a normal distribution. But if one could somehow capture all the organisms in that species and could measure them, she could obtain the 'true' mean μ and the 'true' standard deviation σ . But that is not physically possible. So she would actually measure these quantities from a finite sample. If she captures 10 individuals and measures them, the sample size is 10. From these values she could obtain the sample mean \bar{x} and sample standard deviation s .

Now, she could again capture another 10 individuals and measure them, i.e., she could again obtain another sample of size 10. Of course she will not get the same value of the sample mean \bar{x} and sample standard deviation s . Each time she repeats the experiment and obtains 10 samples, she will get different values. Now if she calculates the distribution of these *sample means*, what will the distribution be?

Similarly, when you are making any measurement (say, the charge of an electron), you are actually 'sampling' from an ideally infinite number of possible measurements. Suppose you have taken 10 measurements, and have obtained the mean value. Now if you repeat the experiment and take 10 more readings, will you get the same mean value? No. If you repeat the experiment a number of times (each time taking 10 readings), you will get a scatter of mean values. What will the distribution be?

the value of the gravitational constant G by a number of experimental runs; if you are a chemist, think of the task of measuring the specific gravity of a new compound that has been synthesized, etc. It helps to think of a problem from one's own field. Note that in all these cases, you are taking a small number of samples from a large 'population' of possible measurements.

4.6.1 The Central Limit Theorem

The answers to these questions come from the Central Limit Theorem:

For large sample sizes (at least 25), the sampling distribution of the mean for samples of size n from a population with mean μ and standard deviation σ may be approximated by a normal distribution with mean μ and standard deviation σ/\sqrt{n} .

Thus, the Central limit Theorem says that in all these cases the distribution of the sample means will be approximately a normal distribution. The more the sample size n , the better is the fit to a normal distribution curve. And this is independent of the actual distribution in the population.

Now, instead of taking 10 readings in each set, if you had taken 50 readings, the average values that come out in each experiment would be closer to the actual mean value μ , and so you would get a narrower Gaussian function. If you take 100 readings, it will be even narrower, i.e., with a smaller variance. The variance of the distribution of means, $\sigma_{\bar{x}}^2$, is thus inversely proportional to the number of readings, i.e.,

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

Therefore the standard deviation of the distribution of the sample means is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

That is what the Central Limit Theorem says: That the sample means will follow a normal distribution, with the same mean as that of the population (i.e., μ) and standard deviation σ/\sqrt{n} .

This gives a way of finding out how good will be the measured value \bar{x} (using a small number of samples) as an estimate of the mean μ of the actual population. Let us illustrate that with an example.

Example 4.1: Suppose there is a ‘population’² with mean 2 (whatever the unit) and standard deviation 0.7. The distribution within that population is unknown. Now suppose you take a random sample of 20 individuals from that population. What is the probability that the sample mean that you get might be above 2.2?

Solution:

The Central Limit Theorem says that if you kept taking similar 20 samples again and again and plotted the distribution of the means, you would get a normal distribution with mean same as that of the population, i.e., 2, and standard deviation $0.7/\sqrt{20} = 0.156$. Now we need to find out the probability $P(\bar{x} \text{ lies above } 2.2)$

$$= P\left(\bar{x} \text{ lies } \frac{2.2 - 2}{0.156} = 1.28 \text{ standard deviations above the mean}\right)$$

The multiplier of the standard deviation is known as the z value, and the area under the normal distribution curve to the left of the z value are given in the z -tables (Tables 4.1 and 4.2). In this case we have to find out the area under the normal distribution curve that lies above 1.28 standard deviations. We see from Table 4.2 that the area to the left of 1.28 is 0.8997. Therefore that to the right is $1 - 0.8997 = 0.1003$ of the whole area under the normal curve (see Fig. 4.3). This implies that, if the scientist took 20 samples from the population, there will be 10% chance that she will get a mean value beyond 2.2, even though the actual mean is 2. Scientifically she should not rely on such an estimate.

Let us now check what will be the odds of getting such a bad estimate if she took 50 data points. According to the Central Limit Theorem, in this case the means will be distributed according to a standard distribution curve with mean 2 and standard deviation

²This could be the population of individuals in a species, or the ‘population’ of possible measurements of the mass of an electron, each measurement coming with a random error, etc.

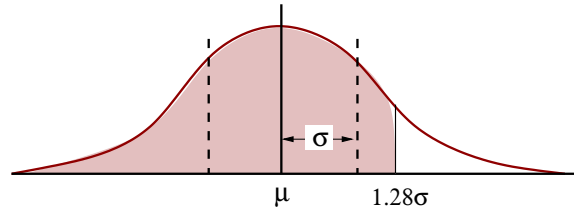


Figure 4.3: The distribution of the sample means in Example 1

$0.7/\sqrt{50} = 0.099$. Therefore we have to find

$$P\left(\bar{x} \text{ lies } \frac{2.2 - 2}{0.099} = 2.02 \text{ standard deviations above the mean}\right)$$

From Table 4.2 we find that 0.9783 fraction of the area lies to the left of this value, and so $1 - 0.9783 = 0.0217$ fraction lies to the right. This implies that the probability of getting a mean value beyond 2.2 goes down to 2.17% if she took 50 samples. \square

4.6.2 Standard error of the mean

The standard deviation of the distribution of the sample means, $\sigma_{\bar{x}}$, is called the “standard error of the mean”, and gives an estimate of the error in the mean obtained by taking a finite number of readings. Let us illustrate it with an example.

Example 4.2: Suppose you have measured a quantity 36 times and have obtained a sample mean $\bar{x} = 112.0$ and sample standard deviation $s = 40$. What is the probability that the actual mean μ lies in the range $[100, 124]$?

Solution:

Here we have a situation where we do not know the actual population mean μ and the population standard deviation σ , i.e., we do not know the actual distribution. But we want a good estimate of the population mean.

If we repeated the experiment of taking 36 samples again and again, we would get slightly different values each time which

will be distributed as a normal distribution, whose mean will be μ and standard deviation will be $\sigma/\sqrt{36} = \sigma/6$. Since we actually do not know the value of σ , the best we can do is to substitute it with what you know: the standard deviation s of the readings actually taken. Thus the sampling distribution will have a standard deviation $40/6=6.67$.

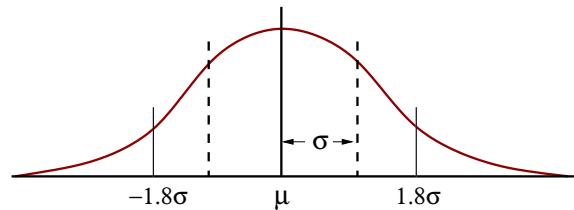


Figure 4.4: The distribution of the sample means in Example 2

Now we ask: what is the probability that the population mean μ is within 12 of $\bar{x} = 112$, i.e., $\mu \in [112 - 12, 112 + 12]$? This is the same as asking what is the probability that the quantity \bar{x} we have measured is within 12 of the population mean μ ? And the quantity 12 is actually $12/6.67=1.8$ standard deviations (see Fig. 4.4). To write it mathematically,

$$\begin{aligned} P(\mu \text{ is within } 12 \text{ of } \bar{x}) \\ &= P(\bar{x} \text{ is within } 12 \text{ of } \mu) \\ &= P(\bar{x} \text{ is within } 1.8 \text{ standard deviations of } \mu) \end{aligned}$$

From the z -table in Table 4.2 we see that the area under the normal curve below 1.8 standard deviations is 0.9641. The area under the curve from the mean to 1.8 standard deviations is $0.9641 - 0.5 = 0.4641$. Thus the area between -1.8 standard deviations to $+1.8$ standard deviations is $0.4641 \times 2 = 0.9282$.

Therefore there is 92.82% chance that the actual population mean lies within ± 12 of the measured value. \square

Now let us consider the question: how many data points are necessary for a confident estimate of the population mean and

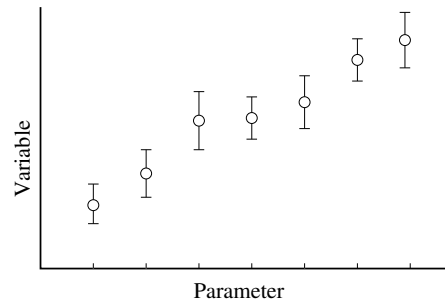


Figure 4.5: A typical graph showing the error bars.

standard deviation? The Central Limit Theorem says that the answer is: At least 25.

Thus, after you have obtained the mean from the data taken, the standard error of the mean is estimated to be

$$SE = \frac{\text{Standard deviation of the readings obtained}}{\sqrt{n}}$$

4.6.3 The error bar

Now let us consider the question: How can one state the measured value so that it will be acceptable to the scientific community? The standard procedure is to define a range within which the actual value is likely to lie. In many experimental situations, standard error of the mean is plotted as an ‘error-bar’ around the mean value.

A typical plot showing the error bars might look like Fig. 4.5. Notice that each error bar may have different length, because for each value of variable 1, the data points obtained for variable 2 may have different variance. The number of data points may also be different.

If one is measuring a length, one would be expected to express the measurement in the form

$$3.56 \text{ cm} \pm 0.03 \text{ cm}.$$

Here the error is expressed in absolute magnitude, and so it has a unit. The error can also be expressed as a percentage, i.e., a measurement of a length x can also be expressed as

$$\bar{x} \text{ cm} \pm \frac{\delta x}{\bar{x}} \times 100\%$$

where \bar{x} is the mean and δx is the standard error of the mean.

In this case the error is expressed as a fraction and will not have any unit.

Example 4.3: You have conducted an experiment to measure two values x and y , and have obtained the following data.

x	5.23	4.97	4.78	5.05	5.34	4.78	4.96	5.03
	5.15	5.26	4.92	4.78	4.98	5.01	5.19	5.08
	5.15	4.94	4.92	4.89	5.10	5.22	4.80	5.06
	4.87							
y	3.43	3.45	3.85	3.29	3.96	3.10	3.11	3.23
	3.43	3.24	3.29	3.24	3.16	3.45	3.23	3.19
	3.29	3.37	3.42	3.10	3.29	3.27	3.17	3.24
	3.58							

How will you describe the result scientifically?

Solution:

The mean values can be easily obtained as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 5.018$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = 3.335$$

The standard deviations are

$$\sigma_{sx} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = 0.16$$

$$\sigma_{sy} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = 0.211$$

Using the approximation $\sigma \approx s$, we get

$$\text{Standard error in } x = \frac{\sigma}{\sqrt{n}} = \frac{0.16}{5} = 0.032$$

$$\text{Standard error in } y = \frac{\sigma}{\sqrt{n}} = \frac{0.211}{5} = 0.042$$

Therefore, the results are to be specified as

$$x = 5.018 \pm 0.032$$

$$y = 3.335 \pm 0.042$$

□

4.7 Estimating with confidence

Thus, the usual experimental procedure is to obtain samples from a population, to obtain the mean and the standard deviation from the data, and to use these to estimate the characteristics of the population. The question is: How good is the sample mean as an estimate of the population mean? One normally approaches this question by finding an interval of values within which one can be fairly confident that the population mean lies.

We have seen earlier that the sampling distribution of the mean for samples of size n has mean μ and

$$\text{Standard error } SE = \frac{\sigma}{\sqrt{n}}.$$

Since this distribution is approximately normal, 95% of the samples will lie within $1.96 \times SE$ of the population mean. So, for approximately 95% of samples of size n , the difference between the sample mean \bar{x} and the population mean μ is less than $1.96 \sigma / \sqrt{n}$.

Thus, for approximately 95% of samples of size n , the population mean will lie between the two values

$$\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \bar{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}$$

The sample mean \bar{x} can be calculated from the data. Since we do not know the value of population standard deviation σ , we use the sample standard deviation s as an estimate of σ . With that, we can calculate the values of the two expressions above and hence can obtain an interval of values within which we could be fairly confident that the population mean μ lies. This leads to the conclusion that for 95% of samples of size n , the population mean μ lies between the two values

$$\bar{x} - 1.96 \times \frac{s}{\sqrt{n}} \quad \text{and} \quad \bar{x} + 1.96 \times \frac{s}{\sqrt{n}}$$

If the act of collecting samples is repeated many times, approximately 95% of the time we would find that the interval thus calculated did contain the true population mean, and in approximately 5% of the cases it would miss μ . That is why this interval is called *95% confidence interval* for the population mean.

This interval is our answer to the question: “how good is the sample mean \bar{x} as an estimate of the population mean μ ?”

Now, we have seen in Chapter 4 that 68.3% of the data points lie within the range $[\mu - \text{SE}, \mu + \text{SE}]$. Therefore stating this error bar amounts to saying that the ‘true’ population mean would lie within this range with 68.3% confidence level. Some experiments demand a higher confidence level, typically 95%, and for that error bar will have to be $\pm 1.96 \text{ SE}$.

Some physics experiments, especially the ones that test the correctness of a theory like the existence of Higgs’ boson or gravitational waves, demand a much higher level of confidence before such an announcement is made. Typically announcements are made stating that ‘five-sigma’ confidence level is achieved. This means that the range is taken as five times the standard error.

Only one in 3.5 million data points lie outside this range, i.e., if the results were due to chance (not caused by the phenomenon in question) then the obtained result can occur to most once in 3.5 million repetitions of the experiment.

Still a question may remain in your mind: Was it a good idea to replace σ by s ? Doesn't it introduce error in our estimate of the 95% confidence interval? There is actually an intuitive justification for it. You have seen that the sample mean varies from sample to sample less for large sample sizes than for small ones. In a similar way, it can be shown that the sample standard deviation also varies from sample to sample less for large sample sizes than for small ones. Thus, the larger the sample size, the better s is as an estimate of σ . Moreover, due to the division by \sqrt{n} , for large sample sizes s/\sqrt{n} would not be much different from σ/\sqrt{n} , and the confidence interval thus calculated would really contain the population mean μ in 95% of the cases.

Now notice a few things.

- Since the 95% confidence interval is proportional to $1/\sqrt{n}$, you would need to take samples four times as large in order to halve the widths of confidence intervals.
- The calculation of a 95% confidence interval does not depend on the size of the population. The only assumption is that the population size is much larger than the sample size. The interval will remain the same if you have drawn a sample of 100 from a population of 10,000 or 10^7 .
- The calculation of this interval does depend on the size of the sample, because we have assumed that the sampling distribution follows a normal curve. This is true only if the sample size is at least 25. Moreover, for small n the replacement of σ by s becomes questionable. But still, a scientist may encounter situations where it is impossible (or expensive) to draw a large number of samples ($n < 25$). In that case the distribution of the sample means cannot be assumed to be normal, and

one has to fit it to some other distribution (generally the t -distribution). We shall come to this issue later.

- Following a similar line of argument, any other confidence interval can also be calculated. For example, a 99% confidence interval would be given by

$$\left[\bar{x} - 2.58 \times \frac{s}{\sqrt{n}}, \quad \bar{x} + 2.58 \times \frac{s}{\sqrt{n}} \right]$$

because 99% of the area under the normal distribution curve lies between -2.58 to $+2.58$ standard deviations.

Example 4.4: For the data in Example 3, define a range of x in which the “true” value of x must lie with 99% probability.

Solution:

From the data, we get the standard error of x as

$$SE_x = \frac{\sigma}{\sqrt{n}} = \frac{0.16}{5} = 0.032$$

Therefore the true value of x will lie in the range

$$\begin{aligned} & [\bar{x} - 2.58 SE_x, \bar{x} + 2.58 SE_x] \\ &= [5.018 - 2.58 \times 0.032, 5.018 + 2.58 \times 0.032] \\ &= [4.937, 5.102] \end{aligned}$$

□

4.8 When the data size is small

The above procedures are applicable to situations where the data size is reasonably large (at least 25), without which the Central Limit Theorem would not be applicable. But there are situations in which it is difficult (or very expensive) to obtain many data points. What to do in such cases?

We have seen earlier that, if the number of samples is sufficiently large, the sampling distribution of the mean follows a normal distribution. In that case we defined a quantity

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

which followed a normal distribution. Then we used the z -table to obtain the probability of getting a z value at least that large (or that small).

Where the data size n is small, the sampling distribution of the mean would not follow a normal distribution. But it follows a different distribution, called the t -distribution, whose characteristics can then be used to derive meaningful results. In this case the quantity t is defined the same way:

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Then one can use the t -table to derive similar conclusions.

Let us illustrate this with an example.

Example 4.5: A scientist could take only 9 measurements on the mass of a particle, and the measured values were 16.2, 19.7, 21.8, 15.6, 19.0, 18.7, 16.9, 21.7, 20.2 (in suitable units). Do the data provide sufficient evidence to say that the mass of the particle is less than 21? Here “sufficient evidence” implies that the probability that the statement is wrong is less than 0.01 or 1%.

Solution: From the data, we find that the sample mean is $\bar{x} = 18.87$ and sample standard deviation is $s = 2.2583$. The prediction we have to test is that the population mean $\mu < 21$. This is the same as checking the odds of getting $\bar{x} = 18.87$ if the value of μ were 21. So our approach will be to assume $\mu = 21$ and to check the probability of getting $\bar{x} = 18.87$. If the probability is less than 0.01, there will be less than 1% chance of making an error.

Using the data, we get

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{18.87 - 21}{\frac{2.2583}{\sqrt{9}}} = -2.83$$

Now, we need to look at the t -table in Table 4.3 to locate the threshold value of t that has a significance level of 1%. The columns are arranged according to the “significance level” (which is the area under the t -distribution curve beyond that value of t). In this case we are looking for 1% significance level. The rows are arranged according to the degree of freedom, which is one less than the number of data points, i.e., $n - 1$. Here the number of data points is 9. Therefore the degree of freedom is $n - 1 = 8$. For the above degree of freedom and significance level, we find $t = 3.355$. Therefore the probability of getting a t value higher than 3.355 is 1%. Since the t -distribution is symmetrical about zero, the probability of getting a t value below -3.355 is also 1%.

The value of t we got in our case is -2.83 , which is above -3.355 . This implies that if the mean is $\mu = 21$, the probability of getting $\bar{x} = 18.87$ is more than 1%. Thus, from the data if we state that the population mean $\mu < 21$, there will be more than 1% chance of committing an error. \square

4.9 Box and whisker plots

It may be noticed that a plot of the experimental results showing the error bars (like Fig. 4.5) does not give the information about the spread of the data obtained. That is why in some applications where such information are important, a different way of presenting the results may be preferred. This is called a ‘box-and-whisker’ plot, a typical representation of which is shown in Fig. 4.6.

In producing such a plot, the data are first arranged in ascending order. The minimum value and the maximum value thus obtained gives the extremities of the ‘whiskers’ of the plot. Then

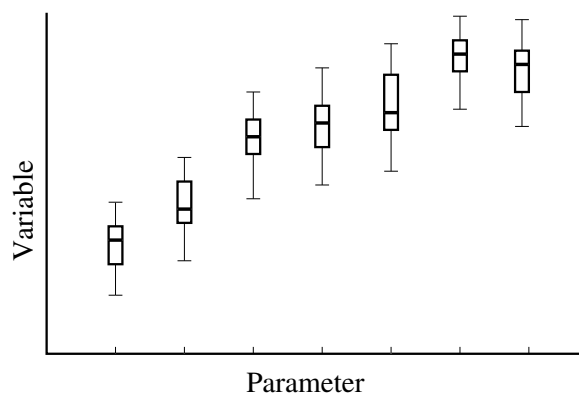


Figure 4.6: A typical box and whisker plot

one has to obtain the median, which is nothing but the middle value. If the number of data points is odd, the middle number is easy to identify. If there are an even number of data points, two numbers will appear at the middle and one has to take the mean of these two numbers. This median gives the mid-point of the plot, called second quartile, or Q2 (see Fig. 4.7).

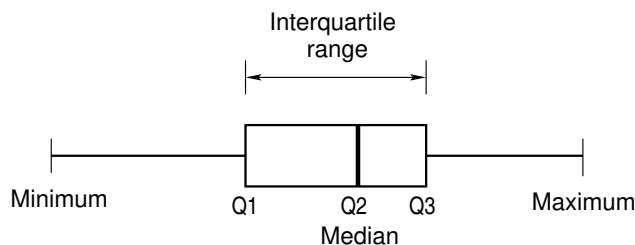


Figure 4.7: The ranges in a box and whisker plot

Then one has to obtain the median of the data points below Q2. That gives another value, called first quartile or Q1. Similarly one obtains the median of the data points above Q2, which gives the third quartile, or Q3. The range between Q1 and Q3 is called the interquartile range (IQR), which is plotted as a box. The range between the minimum and Q1, and that between Q3 and the

maximum are plotted as 'whiskers'. Thus the representation of a typical data set would look like Fig. 4.7. One characteristic feature of such a plot is that 25% of the data lie in each of the four ranges shown in the plot.

Sometimes one gets some data points that lie way outside the natural range of the data. These are called the 'outliers'. The box plot also enables one to identify and present the outliers. The usual method is that the data points lying outside 1.5 times the interquartile range outside the box are called outliers. Thus the 'minimum' of the whisker may be placed at the data point lying above $(Q1 - 1.5 \times IQR)$ and the 'maximum' may be placed at the data point lying below $(Q3 + 1.5 \times IQR)$, and any data point falling outside this range may be shown as 'outlier'.

Such outliers may result from experimental or observational errors, but may also result from some phenomenon not yet discovered. That is why one cannot simply ignore an outlier or delete it from a data set. Outliers have to be faithfully presented in the paper, though these may be ignored in further analysis of the data.

Example 4.6:

Consider the following data set:

17.2, 15.9, 16.7, 18.3, 15.0, 19.3, 20.2, 16.3, 17.9, 15.3, 10.1, 19.1, 18.2

Obtain the box and whisker plot.

Solution:

Arranging the data in ascending order, we get

10.1, 15.0, 15.3, 15.9, 16.3, 16.7, 17.2, 17.9, 18.2, 18.3, 19.1, 19.3, 20.2

It has 13 data points, which is an odd number. So the 7th data point, 17.2, is the median.

There are 6 data points below and above the median, which is an even number. So we get $Q1$ by taking the mean of the 3rd and 4th entries and get $Q1=15.6$. Similarly we get $Q3$ as the mean of the 10th and 11th entries, and get $Q3=18.7$.

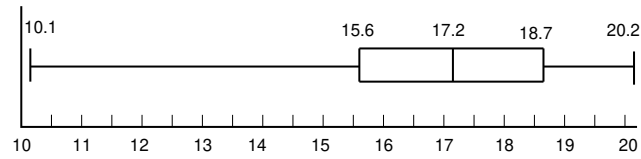


Figure 4.8: The box and whisker plot for the whole data set given in the Example

Thus the box and whisker plot becomes as shown in Fig. 4.8

Now let us see if any data point can be identified as an outlier. The IQR is $18.7 - 15.6 = 3.1$. Going below lowest point of the box by $1.5 \times \text{IQR}$ gives 10.95. We see that there is one data point below that value. Therefore we can declare this point as an outlier, and set the ends of the whisker at the last data point above 10.95. This value is 15.0. Going above Q3 by $1.5 \times \text{IQR}$ gives 23.35. This is above the highest point of the data set. Thus there is no outlier in the higher side. The resulting plot, excluding the outlier, is shown in Fig. 4.9.

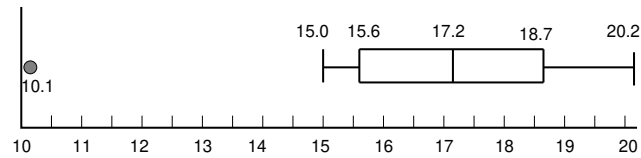


Figure 4.9: The box and whisker plot excluding the outlier.

Exercise

1. Suppose you have measured the concentration of a substance in a solution by taking 25 samples from different parts of the liquid and have obtained the following data (the figures are in ppm).

4.87	5.23	4.97	4.78	5.05	5.34	4.78	4.96	5.03
	5.15	5.26	4.92	4.78	4.98	5.01	5.19	5.08
	5.15	4.94	4.92	4.89	5.10	5.22	4.80	5.06

If you are to report the result in a paper, how will you specify the measured value? With what confidence level can you state that the actual value μ lies in the range 5 ppm to 5.036 ppm?

2. A scientist could take only 10 measurements on the molecular weight of a new compound, and the measured values were 57.6, 55.4, 58.9, 56.3, 55.3, 58.9, 54.9, 57.3, 58.1, 54.8. Do the data provide sufficient evidence to say that the molecular weight of the compound is less than 59? Here “sufficient evidence” implies that the probability that the statement is wrong is less than 0.01 or 1%.
3. Suppose that 15% of the 1750 students at a college have experienced extreme level of stress during the past month. A newspaper doesn't know the figure, but they are curious what it is, so they decide to ask a random sample of 160 students if they have experienced extreme levels of stress during the past month. 16 students replied “yes” to the question.

In case the assumption is true, what is the probability of getting the ‘yes’ answer from 16 or less number of students out of 160? What conclusion regarding the assumption can you scientifically draw based on the observation?

4. Find the median (Q2), lower quartile (Q1) and upper quartile (Q3) for the following data obtained in an experiment. Identify if there is any outlier and draw a box-and-whisker plot.

{48, 56, 75, 50, 46, 5, 52, 49, 53, 42, 55, 50, 58, 40, 102}

What is the mean value that you can report?

Standard Normal Probabilities

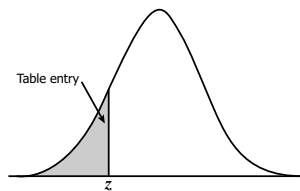


Table entry for z is the area under the standard normal curve to the left of z .

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0006	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Table 4.1: The z -table for negative values of z

Standard Normal Probabilities

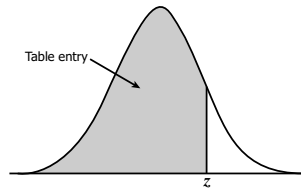


Table entry for z is the area under the standard normal curve to the left of z .

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table 4.2: The z -table for positive values of z

Degrees of freedom	Significance level					
	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.043	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.158	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Table 4.3: The t -table