Standard Errors and Confidence Intervals

Introduction

In the document 'Data Description, Populations and the Normal Distribution' a sample had been obtained from the population of heights of 5-year-old boys. If we assume that this variable has a Normal distribution (an assumption that is, in fact, entirely reasonable) then it will have a population mean, μ, whose value is very likely to be of interest. As this is a population parameter we will never know its true value, because we will never have a complete enumeration of the population. Consequently we will have to content ourselves with knowing what we can about μ on the basis of random samples drawn from the population.

The natural way to estimate μ is to compute the mean, *m*, of the sample and say that this value is our *estimate* of μ. The mean of the sample of 99 heights in the sample given in 'Data Description, Populations and the Normal Distribution' is 108.34 cm. Had we only measured the heights of the first ten boys on this sample the value obtained would have been 107.77 cm. If 20 boys had been measured then the value would have been 107.68 cm. For a sample of 1000 heights, the mean would have been 108.01 cm. (this is a hypothetical sample generated in the way described in 'Data Description, Populations and the Normal Distribution' assuming a population mean of 108 cm and a population SD of 4.7 cm). These results can be summarised in the following table:

Sample size	Sample mean (cm.)
10	107.77
20	107.68
99	108.34
1000	108.01

Each of these sample means is a legitimate estimate of μ - indeed, a single height measurement, such as the first measurement, 117.9 cm, is a legitimate estimate of μ. Faced with this choice, which mean should be chosen and, more importantly, why?

If issues of resources are ignored (see later), then most investigators would intuitively say that the mean of the sample of size 1000 was the best one. This intuition would be based on the notion that by using data from 1000 boys, the sample mean was based on more information and so must be better. A slightly more precise way of saying this is to say that a large sample will provide a more representative cross-section of the population and therefore the mean of this sample would be closer to the mean of the whole population than a mean based on a smaller sample.

Principles behind the Standard Error

The observation that the mean of a large sample is, in some sense, going to be closer to the mean of the population is the key to a more precise understanding of why the mean of a larger sample is 'better' than the mean of a smaller sample. In order to demonstrate this, and to formulate more precisely what is meant by 'better', it is useful to generate not just one sample of data from a population, but many samples of the same size from that population. If we postulate particular parameter values (such as μ = 108 cm, σ=4.7 cm, as was done above) then the computer can generate any

number of samples of any given size. Of course, this is not at all how things are in practice, where parameters are unknown. Moreover, drawing even a single sample from a population involves a great deal of work and drawing many samples is out of the question. However, the following is merely to demonstrate the important ideas: how the repeated drawing of samples is circumvented in practice is explained the next section.

Suppose that 500 samples, each of size 10 are drawn a Normal population with mean 108 and standard deviation 4.7 (values that accord closely with those for the heights of five-year-old boys). The mean of each of these samples can be computed and a histogram of the resulting 500 means can be plotted: see figure 1.

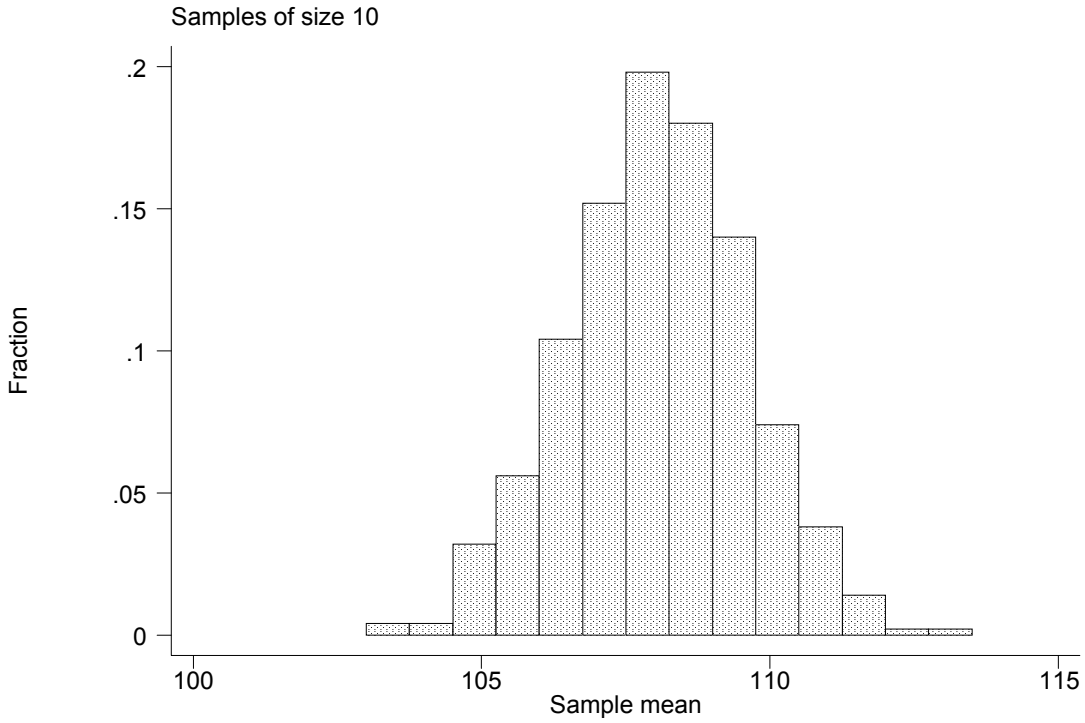


Figure 1

The principal thing to note about figure 1 is that the histogram is much less dispersed than the corresponding histogram of individual heights seen in figure 3 of 'Data Description, Populations and the Normal Distribution'. The data in the latter occupied the range from 95 to 120 cm whereas the figure above is largely confined to the interval 105 to 112 cm. Two other features of figure 1 should also be noted: first the distribution of the sample means does appear to be centred on $\mu = 108$ and second the distribution appears to have the shape of a Normal distribution.

Repeating this exercise but with 500 samples each of size 1000 gives the histogram in figure 2. This histogram is also centred on 108 cm but is even less dispersed than that in figure 1, with all the sample means being between 107 and 109 cm and virtually all of them being between 107.5 and 108.5 cm.

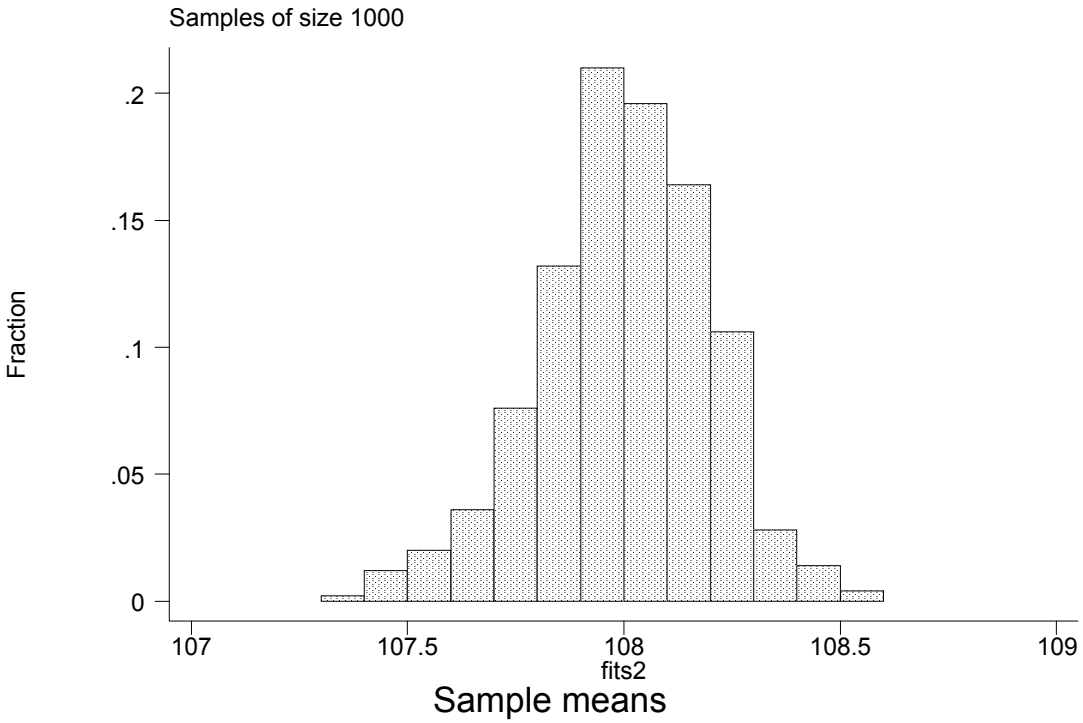


Figure 2

The difference between figures 1 and 2 clearly shows the benefit of taking larger samples and provides the basis for preferring one mean over another. Means based on larger samples provide more precise estimates of the underlying population mean because the distribution of the sample mean becomes more concentrated about μ as the sample size increases. Indeed, it is possible to be quantitative about this, as the standard deviations of the distributions in figures 1 and 2 measure how precisely the sample mean estimates the population mean. The standard deviation of the means of samples of size 10 (figure 1) is 1.54 cm and the corresponding figure for samples of size 1000 (figure 2) is 0.20 cm. This figure measures how precisely the sample mean estimates the population mean and is called the *standard error of the mean* (SEM) or, more simply, the *standard error* (SE).

Calculating the Standard Error from a single sample

Figures 1 and 2 demonstrate how the standard error gives a useful measure of how far a sample estimate can be expected to depart from the underlying parameter. However, in order to compute a value for the SE not one but many samples had to be drawn from the population. This was achieved by asking the computer to generate the samples. While this is a useful device for illustrating important ideas, it is plainly artificial and does not provide a method that can be applied in practice.

The solution to the problem is to use a theoretical result of great importance in statistics. This result states that:

> *If a variable has population standard deviation σ then the standard error of the mean of a sample of size n is* $\frac{\sigma}{\sqrt{n}}$.

So, the SE of the mean of a sample of size 10 is $\sigma/\sqrt{10} \approx 0.316\sigma$, and that of a sample of size 1000 is ten times less than this, namely 0.0316σ. For a standard deviation of 4.7 these figures are 1.49 and 0.149 cm respectively. These are broadly similar to the values obtained for the standard deviations of the distributions pictured in figures 1 and 2, namely 1.54 and 0.20 cm respectively.

The value of this formula is that we can estimate the quantity $\frac{\sigma}{\sqrt{n}}$ from a single sample. The population standard deviation σ can be estimated by the standard deviation, *s*, of the single sample that is to hand, so the estimated SE is computed as $\frac{s}{\sqrt{n}}$.

Nomenclature

It might be asked, when the SE is simply the standard deviation of the distribution of sample means, why a term other than standard deviation is necessary. In fact a reasonably cogent argument can be made for abandoning the term *standard error*. However, there are at least three reasons why standard error is used.

i) While the standard error is a form of standard deviation, it is a very special form. It is the standard deviation of a hypothetical distribution that is never actually observed. Values for the standard error usually need to be computed using a theoretically derived formula.

ii) Standard errors and standard deviations are put to different uses. Standard deviations are descriptive tools that indicate the dispersion in a sample. A standard error is an inferential tool, which measures the precision of estimates of population parameters.

iii) The fact that a standard error is a form of standard deviation can readily give rise to confusion. *Standard error* is the term that has been widely used for the standard deviation of the distribution of sample means and to change nomenclature now may cause even greater confusion.

Using the Standard Error

The mean of the sample of 99 heights given in 'Data Description, Populations and the Normal Distribution' is 108.34 cm and its standard error is 0.52 cm. While the foregoing discussion shows that the SE is a useful measure of how well the mean of this sample estimates the population mean μ, it is not clear precisely how this information should be used.

It is commonplace to indicate the 'error' in a quantity by quoting the value plus or minus the estimated error. Perhaps we could quote the mean as 108.34 ± 0.52 cm.? Certainly, this is often the impression given in the medical literature, where it is common to find table headings of 'Mean (± SE)' and graphs with bars that are a standard error in length stretching out above and below some mean. However, this is a misleading practice that should be discouraged. This is because it invites the reader to suppose that the μ must lie between 108.34 – 0.52 = 107.82 and 108.34 + 0.52 = 108.86 cm., which is false.

The problem can be understood by recalling that for a Normal distribution most values (in fact about 95% of them) lie between $\mu - 2\sigma$ and $\mu + 2\sigma$. As the distribution of sample means has mean μ and standard deviation equal to the SE, there is a 95% chance that the sample mean, *m*, is between $\mu \pm 2\text{SE}$, which amounts to saying that there is a 95% chance that μ lies between $m \pm 2\text{SE}$ [§]. Consequently it is much more accurate to assert that the population mean lies in the interval $m \pm 2\text{SE}$ than that implied by headings such as mean ± SE. For the sample of heights of 99 boys, this interval is (107.3, 109.38), which is wider than the interval in the previous paragraph.

Confidence Intervals

For the reasons that have just been outlined, the interval $\left(m - 2\frac{\sigma}{\sqrt{n}}, m + 2\frac{\sigma}{\sqrt{n}} \right)$ is, approximately, the *95% confidence interval* for μ. A more exact definition is available and is explained in the Appendix.

An alternative name, widely used by methodological statisticians but not often encountered in the applied literature, is *interval estimate* (of μ). This is in distinction to the single-number estimate *m*, which would be referred to as a *point estimate*. This terminology reflects the fact that while a single value might have advantages as an estimate of a parameter, a single value cannot adequately acknowledge the uncertainty in the estimate. For example, in a clinical trial of two agents intended to reduce blood pressure, it would be difficult to interpret the outcome that mean blood pressure on treatment A is 1 mmHg lower than on treatment B. If a confidence interval on this difference was (-3, 5) mmHg then it could be reasonably concluded that there was no material difference between the effect on blood pressure of A and B. On the other hand if the interval was from (-30, 32) mmHg then there could well be an important difference between the treatments, even though its precise nature has not been adequately elucidated by this trial. The difference between these alternative outcomes has only become apparent through the use of confidence intervals. It is for this reason that many medical journals now insist on the use of confidence intervals in the presentation of results.

Degree of Precision

It should be noted that the SE decreases as the sample size increases, because the denominator in the ratio $\frac{\sigma}{\sqrt{n}}$ gets larger. This is in distinction to the standard deviation, which does not have a tendency to get larger or smaller as *n* increases – the sample standard deviation, *s*, simply becomes a better estimate of σ as *n* increases. However, the square root in the formula means that the SE does not decrease with sample size as quickly as might be hoped: in order to halve the SE the sample size must quadruple.

It follows that an experimenter prepared to collect a sufficiently large sample could have a SE that was as small as they liked. However it is important to ensure that an

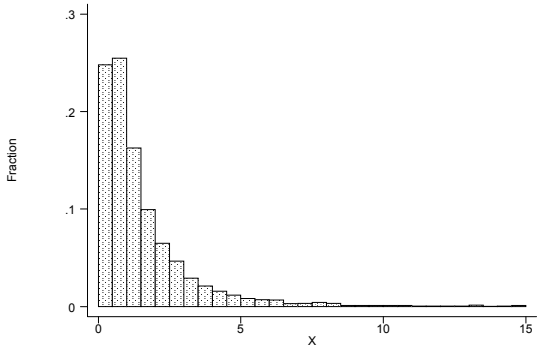
[§] This is more or less true, but a slight technicality has been overlooked here

estimate has a precision that is appropriate to the purpose of the investigation, rather than one that is arbitrarily high. Collecting data on sufficient patients to determine the mean blood pressure of a group to within 1 mmHg is likely to be a waste of time and money.

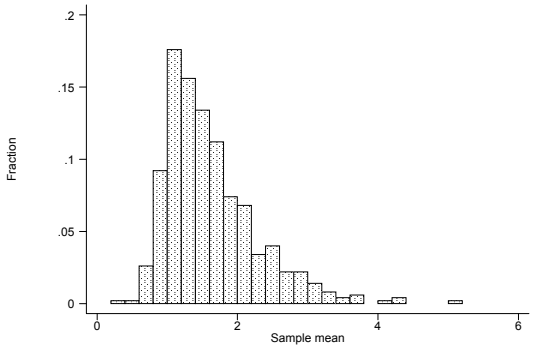
Distribution of sample mean

This is a useful place to explain an important feature of the sample mean. Figures 1 and 2 show that the means of samples of Normally distributed variables do themselves have a Normal distribution. It is this fact that makes our definition of a confidence interval work: only because the sample mean is Normally distributed can we assert that 95% of sample means are with 2 SEs of the population mean.

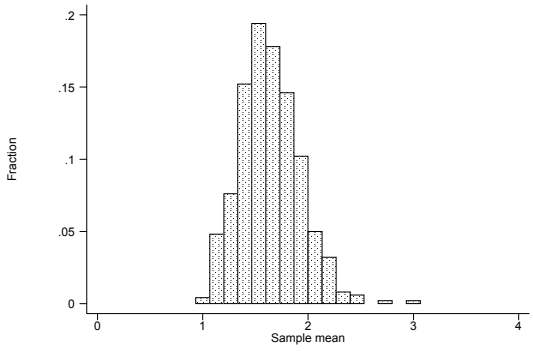
However, even when the samples are of variables that are not Normally distributed, the sample means have a distribution that is often very close to a Normal distribution. This is a phenomenon explained by the *Central Limit Theorem* (CLT) and is illustrated in figure 3.



(a) original variable



(b) means of samples of size 10



(c) means of samples of size 50

Figure 3

Figure 3 (a) is a histogram of 5000 observation from a population that has a skewed distribution. Figure 3 (b) is a histogram of the means of 500 samples, each of size 10 and figure 3 (c) is the corresponding histogram for samples of size 50. It is clear that the act of taking a mean of just ten observations has produced a quantity that has a much less skewed distribution and taking the mean of 50 observations gives a distribution that appears very close to Normal.
A technical description of why this is requires the reader to assimilate the details of the CLT. However, something of a heuristic explanation is as follows. The mean of

the population shown in figure 3 (a) is 1.649 and the skewness can be thought of as arising from the fact that a necessarily positive the variable has substantial variation. Consequently, large departures from mean with this value can only occur through values larger than the mean, so an asymmetric distribution is inevitable. On the other hand, sample means will tend to be distributed close to the population mean (at least for sufficiently large samples), so they will deviate from the population mean by amounts that are small enough that the shape of the distribution is not rendered asymmetric by the fact that the sample means must be positive.

Appendix: exact form of the confidence interval (Not Assessed)

The 95% confidence interval was introduced as $m \pm 2 \times s/\sqrt{n}$. While this version of the confidence interval is adequate for many practical purposes it is, in fact, only an approximation to a more exact version. An explanation of why this is so is given below, so that the reader will not be confused by differences between the above definition and definitions found in textbooks. In addition, most statistics programs will use the exact method, so this will explain discrepancies between intervals calculated by widely used programs and using $m \pm 2 \times s/\sqrt{n}$.

There are two reasons why $m \pm 2 \times s/\sqrt{n}$ is only approximate. The first is a minor point and stems from the fact that 95.45% of a Normal population lies between two standard deviations below and two above the mean. Changing the multiplier '2' to the somewhat less memorable '1.96' gives an interval that encloses 95% of the population.

The second point is more subtle and involves a point which has been glossed over in the above discussion. The SE is actually $\sigma/\sqrt{n}$ which, because it depends on a *parameter*, σ, is unknown. The calculated confidence interval $m \pm 2 \times s/\sqrt{n}$ has substituted the sample standard deviation, s, for σ in the formula for the standard error. However, just as m is not an exact estimate of μ, neither does s give the exact value for σ. Moreover, s tends vary about σ more in smaller than in larger samples. A consequence is that the interval $m \pm 2 \times s/\sqrt{n}$ tends to be too narrow and gives less than 95% confidence that it encloses μ. The solution is to replace the multiplier '2' by a larger value[¶]. However, the value chosen will depend on the sample size, with larger values being needed for smaller samples. With large samples, typically in excess of 100, s will be a good estimate of σ and the exact 95% confidence interval will be very close to $m \pm 1.96 \times s/\sqrt{n}$. With smaller samples the discrepancy is greater. This is illustrated in the following table, in which the approximate 95% confidence interval, $m \pm 2 \times s/\sqrt{n}$, is compared with the exact version, for both the full sample of 99 heights from 'Data Description, Populations and the Normal Distribution' and for the sample of just the first ten heights.

	Approximate	Exact
Sample of 10 heights	(104.21, 111.33)	(103.74, 111.80)
Sample of 99 heights	(107.29, 109.39)	(107.30,109.38)

It can be seen that the exact interval is noticeably wider for the smaller sample. For samples of size 99 the difference is unimportant (and, to two decimal places, not apparent).

[¶] It is, in fact, the 97.5% point of a Student's *t* distribution with *n*-1 degrees of freedom