AN INTRODUCTION TO STATISTICS

Article 4

An introduction to estimation—1. Starting from Z

P Driscoll, F Lecky, M Crosby

Objectives

- Estimating an element from a population with known standard deviation
- Estimating a statistic from a population with known standard deviation

In covering these objectives we will introduce the following terms:

- Standard normal distribution
- z statistic
- Confidence intervals

In the previous article the term inferential statistic was introduced.¹ This form of numeric manipulation is often used to estimate a population's parameter from a sample's statistic. For example, inferential statistics would be used to estimate a population's mean from a sample's mean. It can also be used to do the opposite—that is, estimate a sample statistic from a population's parameter. This is not commonly done because it requires the population's mean and standard deviation to be known and this is rarely the case.

In both calculations the values obtained are only estimations because of the normal variation that occurs. We can however work out the probability of a particular value based upon information from either the sample or population. Central to this is converting the original data to a standard normal distribution so that these estimations can be made.

Standard normal distribution

A standard normal distribution is a particular type of normal distribution that has the following properties:

- Symmetrical bell shaped curve
- Mean equal to zero
- Standard deviation equal to 1
- Total area under the curve equal to 1

It is possible to convert any normal distribution to a standard normal distribution by adjusting it such that the population mean becomes zero and the standard deviation is equal to 1. To do this, each of the data points (elements) is modified by:

- Subtracting the population mean
- Dividing the result by the standard deviation of the population

The final value is known as the z statistic. The z statistic is therefore describing the size of the difference between the element (X) and population's mean (μ) value in multiples of the population's standard deviation (σ). For exam-

ple, a z statistic of +1.5 means the difference between the element and mean value is +1.5 times the size of the standard deviation of the population.

Key points

- It is possible to convert any normal distribution to a standard normal distribution
- The Greek letters μ and σ are often used to respectively represent the population's mean and standard deviation
- Therefore for any element the z statistic is equal to $[X-\mu]/\sigma$

When a normal distribution is plotted as a standard normal distribution, the original values on the horizontal axis are converted to their equivalent z statistic. To demonstrate this consider the continuing work of Egbert Everard in the Emergency Department of Deathstar General. Following his previous work, he suspects that the men in the department could be unfit. He therefore decides to explore this further by looking at the resting heart rate. He finds in a physiology textbook that resting heart rate for fit men (μ) is 60/min with a standard deviation (σ) of 10 (fig 1). Two of the staff members he finds are Charge Nurse "Pot belly" Pete the hospital's darts and wine tasting expert and "Skier" Sphen the locum SHO from Sweden. They have resting heart rates of 90 and 40/min respectively.

Consequently the z statistic for "Pot belly" Pete's resting heart rate is:

$$90-60/10 = 3$$

and the z statistic for "Skier" Sphen's resting heart rate is:

$$40-60/10 = -2$$

Plotting these on a standard normal distribution curve, "Pot belly"'s result lies three standard deviations above the mean of fit men whereas Sphen lies two standard deviations below (fig 2).

Key points

- The z statistic is the number of standard deviations an element is from the mean
- A positive z statistic indicates an element value greater than the mean
- A negative z statistic indicates the value of the element is less than the mean

Accident and
Emergency
Department, Hope
Hospital, Salford
M6 8HD, UK

Correspondence to:
Mr Driscoll, Consultant in
A&E Medicine (pdriscoll@
hope.srht.nwest.nhs.uk)

Figure 1 Normal distribution of resting heart rate in fit men and standard normal distribution using the same data.

Figure 2 Standard normal distribution of resting heart rate along with "Pot belly" Pete and "Skier" Sphen's results.

The above examples show that it is possible to convert a normal distribution to a standard distribution by altering the value of the element to a z statistic. This is important because z statistic tables can then be used to estimate the probability of the particular value occurring. These tables show a column for the z value adjacent to

the corresponding area under the distribution curve between the mean and the z point (fig 3).

For example a z value of +1.96 demarcates 47.5% of the area under the curve. This extends from the midline to a vertical line drawn through +1.96 on the horizontal axis (fig 4 (A)). Therefore the total proportion to the left of this z value equals $0.5 + 0.475 = 0.975$ (97.5%). This is because each half of the curve represents a proportion of 0.5 (50% of the area under the curve). Consequently the chances of getting a z value of +1.96 (or less) is 97.5%.

Figure 3 lists proportions for the right half of the curve. However, as the standard normal distribution is symmetrical, the right and left halves are equal. Therefore the area under the curve between the midline and -1.96 is also 47.5%. Be aware also that some z tables give the area from the particular z value to the tip of the tail (fig 4(B)). Tables therefore either have a description or a diagram on the top to ensure you are reading them correctly (fig 3).

Key points

- A standard normal distribution is important because the proportion under the curve demarcated by any value of z has already been calculated and is available in z statistic tables.
- If the numerical value of z and -z are equal, the proportions (percentage of the area under the curve) are the same but in opposite directions.
- Z statistic tables can differ in which the area under the curve is described

It follows that by changing to a standard normal distribution it is possible to determine:

- The probability of getting an element greater than or equal to a particular value
- The range of element values that have a particular probability of occurring
- The probability of getting an element between two particular values.

We now need to consider how these probabilities can be calculated when we are dealing with three different situations (box 1).

Box 1 Three common estimations

- A single value from a population with known μ and σ
- A sample mean from a population with known μ and σ
- A sample mean from a population with unknown μ and σ

This article will deal with the first two situations. The third scenario will be dealt with in article 5 of this series.

Estimating a single value from a population with known μ and σ

THE CHANCE OF GETTING AN ELEMENT GREATER THAN OR EQUAL TO A PARTICULAR VALUE
Consider "Pot belly" Pete, what would be the chance of getting a resting heart rate equal or

greater than 90 and still be part of the fit male population?

Using the z statistic tables (fig 3), the area demarcated by $z = 0$ (that is, the mean) and $z = 3$ is 0.4987.

The area under the curve represents the total probability and is equal to 1.0. As the curve is symmetrical, the area under each half is 0.5. Consequently the probability of getting a value greater than or equal to 3.0 is represented by area under the curve to the right of the shaded area. This is:

$$0.5 - 0.4987 = 0.0013.$$

Therefore the chances of being a fit man and having a resting heart rate greater than or equal to 90/min is 0.0013 or 0.13%.

THE RANGE OF ELEMENT VALUES WITH A PARTICULAR PROBABILITY OF OCCURRING

Following on from this result Egbert considered the right side of the distribution curve showing the resting heart rates of fit men. He wanted to know the resting heart rate that separated the upper 2.5% of the population from the remaining 97.5%. This is determined by carrying out the reverse of the method described above.

1 Convert the percentage to a proportion by dividing by 100.

2.5% is 0.025 when expressed as a proportion.

2 Determine the proportion of a standard normal distribution curve from the midline to 0.025.

Using figure 4(B) as a reference, it can be seen that the answer is equal to:

$$0.5 - 0.025 = 0.475$$

3 Convert the proportion 0.475 to a z statistic.

Using the z statistic tables, 0.475 gives a z statistic of 1.96 (fig 3).

4 Using this value for z , determine the element value.

Remembering that:

$$z = \frac{\text{Element value} - \text{Population mean}}{\text{Population standard deviation}}$$

$$1.96 = \frac{\text{Element value} - 60}{10}$$

Therefore the element value is $19.6 + 60 = 80/\text{min}$ (rounded up)

Consequently the top 2.5% of the population of fit men have a resting heart rate of approximately 80/min or greater.

THE CHANCE OF GETTING AN ELEMENT BETWEEN TWO PARTICULAR VALUES

Egbert has often seen in published work, ranges of values for the middle 95% of the population. He therefore would like to know what is the range of resting heart rates of the middle 95% of the population of fit men. Looking at figure 5, you can see the middle 95% leaves two tails each representing 2.5% of the area under the curve (fig 5).

The value demarcating the upper 2.5% has already been calculated (that is, 80/min). Therefore Egbert only has to repeat the procedure for the lower 2.5%.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.000	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0359	0.0359
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4773	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4812
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Figure 3 Extract of a table of z scores. An entry in the table is the proportion under a standard normal distribution curve from $z = 0$ to a positive value of z (for example, $z = 3$). Areas for the negative z values are obtained by symmetry. The curve accompanies the table to show which areas were used to derive the z values.

Figure 4 Standard normal distribution showing two ways of calculating areas for the same value of z .

Figure 5 Standard normal distribution showing two 2.5% tails.

Table 1 Practising with z statistics

Element value	Population mean (μ)	Population standard deviation (σ)	z statistic	Area under the curve to the right of the z statistic
8.0	12.0	2.0	?	?
?	12.0	2.0	-3.0	?
16.0	14.0	4.0	?	?
?	14.0	4.0	1.5	?

1 Convert the percentage to a proportion by dividing by 100.

2.5% is 0.025 when expressed as a proportion

2 Determine the proportion of a standard normal distribution curve from the midline to 0.025.

This is equal to:

$0.025 - 0.5 = -0.475$. The negative value indicates we are dealing with the left side of the curve.

3 Convert the proportion -0.475 to a z statistic.

Using the z statistic tables, -0.475 gives a z statistic of -1.96 (fig 3).

4 Using this value for z , determine the element value.

$-1.96 = \text{Element value} - 60/10$

Consequently the element value is $-19.6 + 60 = 40/\text{min}$ (rounded down). In other words the lower 2.5% of the population of fit men have a resting heart rate of approximately 40/min or less. Egbert therefore concludes that the middle 95% range is from approximately 40 to 80/min.

To check if you are confident with these calculations, take a moment to fill in the gaps in table 1 (answers on page 415).

Key points

- It is possible to convert elements in a normal distribution to a z statistic so that they can be plotted on a standard normal distribution
- This conversion allows the probability of the element occurring to be determined

It is pertinent to notice at this time that the middle 95% of the population is represented by the mean ± 1.96 standard deviations. This is usually rounded up to the mean ± 2 standard deviations. Using this same system it is possible to calculate the percentage of the area of the curve covered by various standard deviations on either side of the mean (table 2). You can now see how, in article 2, the percentages of the area under the curve were derived.²

Table 2 Percentage of the normal distribution

Range of z statistic	Approximate area under the curve (%)
Mean ± 1 standard deviation	68
Mean ± 2 standard deviations	95
Mean ± 3 standard deviations	99

Another way of calculating the percentage (or proportion) of the area between two z statistics is to subtract the area under the curve to the left of the smaller z value from the larger one (fig 5). For example, the area to the left of a z statistic equal to -1.96 is 0.025 (2.5%). Conversely the area to the left of z of 1.96 is $0.5 + 0.475$ (97.5%). Consequently, the probability of a value lying between these two z statistics is:

$$0.975 - 0.025 = 0.95 \text{ (95\%)}$$

Figure 6 Skewed original population with a normal distribution of sample means around the population mean.

Estimating a sample mean from a population with known μ and σ

To tackle this problem we need to remind ourselves about the properties of the standard error of the mean (SEM). In article 2 we described how normally distributed data from a sample could be summarised by recording the mean and standard deviation.² We also went on to show how it is possible to use the mean calculated from a particular study (called the sample mean) to estimate the overall mean (called the population mean). Samples selected randomly from the same population are unlikely to have exactly the same means. This is attributable to chance variation and is known as random error. The SEM quantifies this by indicating the range, above and below the sample mean, in which the population mean lies. Another factor affecting the SEM is the size of the sample from which it is estimated. As sample size increases, the SEM variation caused by random error decreases.

Key point

SEM = Population standard deviation (σ) / number in the sample (n)

It can be shown that if random samples are taken, which are sufficiently large, their mean will form a normal distribution irrespective of the original population's distribution (fig 6). Furthermore, if an infinite number of samples were taken, the mean of the distribution and population would be the same. The mathematical proof for these statements comes from the central limit theorem.

For obvious reasons we cannot take such a large number of samples. Nevertheless we can use the SEM to estimate the limits within which the population mean will lie. As the distribution is normal, there is a 95% chance that the population's mean will lie somewhere between $1.96 \times \text{SEM}$ below the sample mean to $1.96 \times \text{SEM}$ above the sample mean.

Key points

- The SEM is a measure of the deviation of sample means from the population mean
- The accuracy of the SEM is dependent upon the samples being representative of the population and being randomly selected
- The size of the SEM is dependent upon the size and variability of the sample
- As the sample size increases, the SEM will get smaller and there will be closer agreement between the sample and population means

To help demonstrate these points consider the following hypothetical example. Suppose you wanted to find the mean presenting peak expiratory flow rate (PEFR) for adult asthmatic patients attending your emergency department. If you took the measurements in a sample of 20 patients you could work out the mean for that sample. However, if you then repeated this process in another 20 patients you would not expect exactly the same value. This is because of the chance variation between patients. Nevertheless if you continued to carry on making these recordings in more samples of similar size you would find eventually that there would be a normal distribution of the sample means. The central limit theorem indicates that the overall mean of these samples would be the same as the population mean. The standard deviation of this normal distribution is called the SEM. Consequently 95% of samples of adult asthmatic patients attending your emergency department will have a mean presenting PEFR lying between the population's mean value ± 2 SEM.

In summary therefore when a large number of randomly selected and representative sample means are plotted they take on a normal distribution with the mean being the population mean and the standard deviation being the SEM. It is therefore possible to convert this distribution to a standard normal type so that you can determine:

- The chance of getting a sample mean greater than or equal to a particular value
- The range of a sample means with a particular chance of occurring
- The chance of getting a sample mean between two particular values

THE CHANCE OF GETTING A SAMPLE MEAN GREATER THAN OR EQUAL TO A PARTICULAR VALUE

To demonstrate this, consider Egbert's result when he has measured the resting heart rate in all of the 25 male members of the department. He finds that the sample mean is 67/min and wants to know what would be the chance of getting a resting heart rate equal or greater to this and still be part of a normal fit male population?

The z statistic for this resting heart rate is:

$$[\text{Sample mean} - \text{Population mean}] / \text{SEM}$$
 Where:

SEM = Population standard deviation (σ) / size of the sample (n)

Therefore in this case:

$$\text{SEM} = 10/\sqrt{25} = 2.0$$

Therefore the z statistic is:

$$[67-60]/2.0 = +3.5$$

Using the z statistic table, the area between $z = 0$ and $z = +3.5$ is 0.4989. Therefore the probability of getting a z value greater than or equal to 3.5 is:

$$0.5-0.4989 = 0.0011$$

Consequently the chances of a similar sample of fit men having a resting heart rate greater than or equal to 67/min is 0.0011 or 0.11%.

THE VALUE OF SAMPLE MEAN WITH A PARTICULAR CHANCE OF OCCURRING

As before Egbert then turns his attention to the right tail of the distribution of mean values. Assuming a sample consists of 25 fit men who have been randomly selected, Egbert wants to know what sample mean demarcates the upper 2.5% of the population. This is carried out in a similar manner to that used when an element was considered.

1 Convert the 2.5% to the proportion 0.025

2 Determine the proportion of a standard normal distribution curve from the midline to 0.025. This is equal to $0.5-0.025 = 0.475$

3 Convert the proportion 0.475 to a z statistic.

Using the z statistic tables, 0.475 gives a z statistic of 1.96 (fig 3).

4 Using this value for z, determine the sample mean.

Remembering that:

$$z = [\text{Sample mean} - \text{Population mean}] / \text{SEM}$$

$$1.96 = [\text{Sample mean} - 60] / 2$$

Therefore the element value is $3.92 + 60 = 64/\text{min}$ (rounded up). Consequently 2.5% of these samples that have been randomly selected would have a mean resting heart rate of 64/min or greater.

THE CHANCE OF GETTING A SAMPLE MEAN BETWEEN TWO PARTICULAR VALUES

Looking at the middle of the distribution, Egbert then wants to know what range of means from similar samples demarcate the middle 95% of the population.

As the upper 2.5% has already been calculated (64/min), Egbert calculates the value for the lower 2.5%. Using the same system as before he works out that the proportion of a standard normal distribution curve from the midline to 0.025 is:

$$0.025-0.5 = -0.475$$

He then uses the z statistic tables to convert the proportion -0.475 to a z statistic and finds this to be -1.96 (fig 3).

Using this value for z, the sample mean can be determined by:

$$-1.96 = [\text{Sample mean} - 60] / 2$$

Therefore the element value is $-3.92 + 60 = 56/\text{min}$ (rounded up).

Consequently the middle 95% of random samples of 25 fit men would have a mean resting heart rate between 56 and 64/min.

As with the situation with elements and standard deviations, we can see that the middle 95% of all the sample means is represented by the population mean ± 1.96 standard errors of the mean. This is usually rounded up to the mean ± 2 SEM. Using this same system it is possible to determine the percentage of the area of the curve covered by various multiples of the SEM on either side of the mean (table 3).

Table 3 Percentage of the normal distribution

Range of z statistic	Area under the curve (%)
Mean ± 1 SEM	68
Mean ± 2 SEM	95
Mean ± 3 SEM	99

From the calculations carried out above, we know that 95% of all sample means will lie within approximately 2 SEM of the population mean. In other words knowing the population mean, we can calculate the range in which the sample means will lie 95% of the time. For example, if the population mean is 50 cm with a SEM of 5 cm, 95% of all sample means will lie between:

$$[\text{Mean} - 2 \times \text{SEM}] \text{ to } [\text{Mean} + 2 \times \text{SEM}]$$

This is equal to:

$$[50-10] \text{ to } [50+10] = 40 \text{ cm} - 60 \text{ cm}$$

At first glance this may not seem very useful because we usually do not know the population mean. However if the 95% of the sample means are within 2 SEM of the population mean, it follows that the population mean lies within 2 SEM of a sample mean 95% of the time. For example, if the mean of a randomly selected sample is found to be 120 mm Hg and the SEM known to be 6, there is a 95% chance that the true population mean lies between:

$$\text{Mean } [\pm 2 \text{ SEM}] = 120 [\pm 12] = 108 \text{ to } 132 \text{ mm Hg}$$

This range of values is known as the 95% confidence interval. In other words we are 95% confident that population's mean lies between 108 and 132 mm Hg.

Generally the confidence interval range is equal to the sample mean plus/minus [z statistic obtained from the z table for the required percentage level of confidence (z_c) multiplied by the SEM]. As an equation:

$$\text{Confidence interval} = \text{Sample mean } \pm [z_c \times \text{SEM}]$$

In this way we do not need to measure every single blood pressure in a particular population. Instead the sample is used to make an estimation of the range of possible values. In considering this range it is important to bear in mind that the best single estimation of the population's mean is the sample mean.

Key point

In general the CI is equal to sample mean $\pm [z \times \text{SEM}]$. Where the z statistic is appropriate to the level of confidence you require. As 95% CI are often used this is approximately equal to the sample mean $\pm [2 \text{ SEM}]$.

Now that you have seen how CI are derived you will appreciate the range is dependent upon the SEM. As the latter is equal to the σ/n , the range can only be reduced by increasing the sample size. However, the range is proportional to the square root of the sample size. Consequently increasing the sample size by fourfold will only halve the confidence interval.

Summary

It is possible to convert a normal distribution to a standard type. This enables you to use the z statistic table to calculate the proportion of particular areas under the distribution curve. If the population's mean and standard deviation are known, it is possible to estimate:

- The probability of getting an element or statistic greater than or equal to a particular value
- The value of an element or statistic with a particular probability of occurring
- The probability of getting an element or statistic between two particular values.

In addition, the latter range can be used to determine the confidence intervals for the estimated values.

Table 4 Answer to table 1: practising with z statistics

Element value	Population mean	Population standard deviation	z statistic	Area under the curve to the right of the z statistic
8.0	12.0	2.0	-2.0	$0.4772 + 0.5 = 0.9772$
6.0	12.0	2.0	-3.0	$0.4987 + 0.5 = 0.9987$
16.0	14.0	4.0	0.5	$0.5 - 0.1915 = 0.3085$
20.0	14.0	4.0	1.5	$0.5 - 0.4332 = 0.0668$

Quiz

(1) Assuming a standard normal distribution, what is the probability of a value being:

- Greater than, or equal to, 2 standard deviations above the mean?
- Less than, or equal to, 1.5 standard deviations below the mean?
- Between 3 standard deviations below the mean to 1 standard deviation above it?

(2) Assuming a standard normal distribution, what is the value of X if you are told that the probability of having this figure, or greater, is 0.15 (that is, 15%)?

(3) What is the probability of getting a Hb concentration of 16.0 g, or higher, if the Hb distribution is normal and has a mean of 15.0 g with a standard deviation of 0.3?

(4) Assume that a previous study has shown that the mean presenting PEFr for asthmatic men attending your emergency department is 685 l/min with a standard deviation of 30 l/min. What are the chances of a sample of 36 asthmatic men having a mean PEFr of 700 l/min?

(5) In another department the mean presenting PEFr in a sample of women is found to be 400 l/min with a SEM of 20 l/min. What is the best single estimation of the population mean and what is its 95% confidence intervals?

Answers

(1) Assuming a standard normal distribution, the probability of a value being:

- Greater than, or equal to, 2 standard deviations above the mean is $0.5 - 0.4772 = 0.0228$

- Less than, or equal to, 1.5 standard deviations below the mean is $0.4332 + 0.5 = 0.9332$

- Between -3 standard deviations and +1 standard deviation from the mean is $0.4987 + 0.3413 = 0.84$

(2) If the probability of getting a value an element value (X), or greater, is 0.15, the probability of a value less than this is 0.85. As this standard normal distribution the area under the curve from this value to the mean is $0.85 - 0.5 = 0.35$. The z statistic for 0.35 is 1.04.

We know that:

$$z \text{ statistic} = [X - \mu]/\sigma$$

Therefore:

$$X = 1.04 [\sigma] + \mu$$

Assuming a standard normal distribution, σ equals 1 and μ zero, X equals 1.04.

(3) We know that:

$$z \text{ statistic} = [X - \mu]/\sigma$$

Therefore:

$$z \text{ statistic} = [16 - 15]/0.3$$

Therefore a Hb of 16 g is 3.3 standard deviations from the mean. The normal distribution tables show that the area to the left of this point is 0.999. Therefore the probability of getting a value of 16 g or higher is $1 - 0.999 = 0.001$ (that is, one in a thousand).

(4) $z = \text{Sample mean} - \text{Population mean} (\mu) / \text{SEM}$

$$\text{SEM} = \text{Population standard deviation} (\sigma) / \sqrt{n} = 30/6 = 5$$

Therefore:

$$z = 700 - 685/5 = 3.$$

Using the z table, a z score of 3 corresponds to a probability of 0.4987. Therefore the chances of a sample having a mean PEFr of 700 l/min, or greater is $0.5 - 0.4987 = 0.0013$ (that is, just over one in a thousand).

(5) The best single estimation comes from the sample mean—that is, 400 l/min.

The 95% confidence intervals = Sample mean $\pm 1.96 \times \text{SEM}$

$$\text{Therefore the 95\% CI} = 400 \pm 1.96 \times 20 = 361 - 439 \text{ l/min.}$$

The authors would like to thank Sally Hollis, Jim Wardrope and Iram Butt for their invaluable suggestions.

- 1 Driscoll P, Lecky F, Crosby M. An introduction to inferential statistics. *J Accid Emerg Med* 2000;17:357-62.
- 2 Driscoll P, Lecky F, Crosby M. An introduction to everyday statistics—2. *J Accid Emerg Med* 2000;17:274-81.

Further reading

- Bowyers D. Measuring spread. In: *Statistics from scratch*. Chichester: Wiley, 1996:114-49.
- Glaser A. Descriptive statistics. In: *High yield statistics*. Baltimore: William and Wilkins, 1995:1-18.
- Glaser A. Inferential statistics. In: *High yield statistics*. Baltimore: William and Wilkins, 1995:9-30.
- Koosis D. Populations and samples. In: *Statistics—a self teaching guide*. 4th ed. New York: Wiley, 1997:40-76.
- Norman G, Streiner D. Statistical inference. In: *PDQ statistics*. 2nd ed. St Louis: Mosby, 1997:17-36.

AN INTRODUCTION TO STATISTICS

Article 5. An introduction to estimation—2: from z to t

P Driscoll, F Lecky

Objectives

- Comparing a large sample with a population with unknown standard deviation
- Using a large sample to estimate a population's probability value
- Comparing a small sample with a population with unknown standard deviation

In covering these objectives we will introduce the following terms:

- Estimated standard error of the mean
- Degrees of freedom
- t statistic

Introduction

In the previous article we found that it was possible to estimate the probability of getting an element greater than or equal to a particular value (X) in a population with the known parameters, mean (μ) and standard deviation (σ).¹ In these cases the z statistic is calculated to locate the position of X in a standard normal distribution where:

$$z = (X - \mu) / \sigma \quad (1)$$

A similar process can be used when dealing with sample means. If a sufficient number of samples have been taken, and their means plotted, then they begin to take up a normal distribution. It can be shown mathematically that the mean of this distribution (μ_x) is the same as the population mean (μ). Furthermore, the standard deviation of the distribution is equal to $\sigma/\sqrt{n}$, where n is the number of cases in the sample. This is known as the standard error of the mean (SEM). To estimate the probability of getting a value greater than or equal to a particular sample mean (x), in a population with a known mean (μ) and standard deviation (σ), we again calculate the z statistic. However, as we are dealing with the distribution of the means, we use the SEM rather than the population's standard deviation:

$$z = [(x - \mu) / (\sigma / \sqrt{n})] \quad (2)$$

You will have noticed that both of these calculations are dependent upon knowing the population's mean and standard deviation. In clinical and experimental practice this is rarely the case. However, we know that the best single estimate we have for the parameter μ is our sample mean.¹ Unfortunately the same does

not apply to the sample's standard deviation. We get around this problem by using the estimated standard error of the mean.

Estimated standard error of the mean

If we simply replaced the sample's standard deviation for σ to determine the SEM, we would end up with an underestimation of its true value. To overcome this we use an estimation of the population's standard deviation (s) using the following formula:

$$s = \sqrt{[\Sigma(x - \text{sample mean})^2 / (n - 1)]} \quad (3)$$

where:

s is the estimate of the population standard deviation based upon the sample data

n is the number in the sample

$n-1$ is called the degrees of freedom

Key point

When a formula is dealing with descriptive statistics the degrees of freedom are equal to n . In contrast, when they are dealing with inferential statistics the degrees of freedom are smaller (for example, $n-1$). This is to compensate for the formula's tendency to under estimate the parameter being derived from the statistic.

To test our understanding so far, consider Egbert Everard's continuing assessment of staff in the Emergency Department of Deathstar General. Egbert selects five female night nurses at random and weighs them (50 kg, 60 kg, 60 kg, 60 kg, 70 kg). What is Egbert's best estimate of the population mean and standard deviation based upon this sample?

The best estimate of the population mean is the sample mean:

Estimated population mean = (sum of all measures/ n) = $300/5 = 60.0$ kg

The best estimation of the population's standard deviation is s where:

$$\begin{aligned} s &= \sqrt{[\Sigma(x - 60)^2 / 4]} \\ &= \sqrt{[200 / 4]} \\ &= 7.07 \text{ kg} \end{aligned} \quad (4)$$

Accident and
Emergency
Department, Hope
Hospital, Salford
M6 8HD, UK

Correspondence to:
Mr Driscoll, Consultant in
Accident and Emergency
Medicine (pdriscoll@
hope.srht.nwest.nhs.uk)

Key point

We use s as a substitute for σ when trying to estimate the chances of getting a particular sample mean in a population with an unknown standard deviation. In these cases, rather than using the SEM:

$$\text{SEM} = \sigma/\sqrt{n}$$

we use the estimated SEM:

$$\text{ESEM} = s/\sqrt{n}$$

Comparing a large sample mean with a population with unknown SD

The ESEM will provide a close approximation of the SEM if the sample size is 100 or greater. Consequently, using the method described in the previous article,¹ it is possible to determine:

- The chance of getting a sample mean greater than or equal to a particular value
- The value of a sample mean with a particular chance of occurring
- The chance of getting a sample mean between two particular values

THE CHANCE OF GETTING A SAMPLE MEAN GREATER THAN OR EQUAL TO A PARTICULAR VALUE

To demonstrate this consider the following example. Egbert wonders whether the female staff in the Emergency Directorate are as unfit as their male counterparts. To test this he measures the weight in 100 randomly selected female medical and nursing staff. The sample mean is 60 kg and s is 20 kg. From actuarial tables he finds that the mean weight for fit females is 55 kg, but the standard deviation is unknown. Egbert therefore wants to know what is the chance of getting a mean weight equal or greater to 60 kg from a sample that is still part of a normal fit female population?

The sample size is large enough to allow the normal probability distribution to be used even though the standard deviation of the population is not known. The z statistic for this weight is therefore:

$$z = (\text{sample mean} - \text{population mean})/\text{ESEM}$$

Where the ESEM is $s/\sqrt{n} = 20/\sqrt{100} = 2$

Therefore the z statistic is:

$$(60 - 55)/2 = 2.5$$

Using the z statistic table, the area between $z = 0$ and $z = 2.5$ is 0.4938.

Therefore the probability of getting a z value greater than or equal to 2.5 is:

$$0.5 - 0.4938 = 0.0062.$$

Consequently the chances of a similar sample of fit women having a mean weight greater than or equal to 60 kg is 0.0062 or 0.62%.

THE VALUE OF A SAMPLE MEAN WITH A PARTICULAR CHANCE OF OCCURRING

Using the same process described in article 4, Egbert can also determine what sample mean demarcates the top 2.5% of the population.¹

- 1 Convert the 2.5% to the proportion 0.025
- 2 Determine the proportion of a standard normal distribution curve from the midline to 0.025. This is equal to $0.5 - 0.025 = 0.475$
- 3 Convert the proportion 0.475 to a z statistic.

Using the z statistic tables, 0.475 gives a z statistic of 1.96.

4 Using this value for z , determine the sample mean.

Remembering that:

$$z = (\text{sample mean} - \text{population mean})/\text{ESEM}$$

$$1.96 = (\text{sample mean} - 55)/2$$

Therefore the element value is $3.92 + 55 = 59$ kg (rounded up). Consequently there is a 2.5% chance that a randomly selected sample of 100 fit women would have a mean weight of 59 kg or greater.

THE CHANCE OF GETTING A SAMPLE MEAN BETWEEN TWO PARTICULAR VALUES

Looking at the middle of the population, Egbert then wants to know what range of means from similar samples would demarcate the middle 95% of the population of fit women.

As the upper 2.5% has already been calculated, Egbert calculates the value for the lower 2.5%. Using the same system as shown above he finds the lower 2.5% is:

$$-3.92 + 55 = 51 \text{ kg (rounded down).}$$

Consequently the middle 95% of random samples of 100 fit women would have a mean weight between 51 to 59 kg. This range of values is known as the 95% confidence interval.¹ In other words we are 95% confident that a random sample of 100 fit women from this population would have a mean weight between 51 and 59 kg.

Therefore, provided the sample is large enough, the z statistic can be used to calculate confidence intervals when the population's standard deviation is not known. In these cases the confidence interval is equal to the sample mean plus/minus the z statistic appropriate for the level of confidence (z_c) multiplied by the ESEM:

$$\text{Confidence interval} = \text{sample mean} \pm (z_c \times \text{ESEM})$$

Key points

- Provided the sample is big enough the ESEM can be used as a close approximation of the SEM.
- It is therefore possible in these circumstances to determine the CI of the estimation of the population's mean (μ) when the population standard deviation (σ) is not known
- Confidence interval = sample mean $\pm (z_c \times \text{ESEM})$ where z_c is the z statistic for the appropriate level of confidence

Estimating a population's probability values from a large sample

We have seen in the above example that it is possible to determine the confidence interval of the estimation of the population's mean when σ is not known. It is also possible to do the same thing with respect to determining the confidence interval for the population's probability (P) using the sample's probability value (p). This is because the binomial probability

distribution becomes approximately normal in shape when the sample is large. Consequently the z statistic can again be used to determine confidence intervals.

To demonstrate this consider the following example. In the midst of his study on emergency staff, Egbert has been asked by his consultant to determine the proportion of patients who are covered for tetanus. Ever keen to help, he teams up with Dr Endora Lonely, an SpR in the Emergency Department at the neighbouring hospital St Heartsinc. Together they survey 700 patients at random and find 550 have adequate immunisation against tetanus infection.

The probability (p) of adequate tetanus immunisation is therefore:

$$p = 550/700 = 0.786$$

This represents the proportion of adequate tetanus immunisation in the sample. Egbert therefore now needs to estimate what the proportion would be in a population of similar patients (P). As with the situation described previously, the best estimate for the population's probability is p —that is, 0.786. What is now needed is the confidence interval of this estimation.

The formula for calculating the confidence interval for P from a large sample is:

$$p \pm [z_o \times \sqrt{(pq/n)}] \quad (5)$$

where:

- z_o is the z statistic appropriate for the confidence interval
- p is the probability we are concerned with (that is, tetanus covered)
- q is the probability we are not concerned with (that is, not tetanus covered)

This formula assumes that the sample is large and that the smaller of the two groups must have at least 10 cases. As these both apply in this example, Egbert calculates the 95% confidence interval to be:

$$\begin{aligned} &= 0.786 \pm [1.96 \times \sqrt{(0.786 \times 0.214/700)}] \\ &= 0.757 \text{ to } 0.815 \end{aligned} \quad (6)$$

He therefore reports to his consultant that the proportion of patients adequately immunised against tetanus is 0.79 with a 95% confidence interval of 0.76 to 0.82.

Comparing a small sample mean with a population with unknown SD

In clinical practice we commonly deal with sample sizes smaller than 100 from populations with unknown standard deviations. When dealing with such samples to make inferences about the population it is no longer valid to use the z statistic. To overcome these difficulties, W S Gossett derived a replacement known as the t statistic.

Statistics trivia (2)

Gossett carried out his work while working in the Guinness Brewery in Dublin. It was based upon samples taken from a population made up of the heights of 3000 criminals. At the time the company would not allow employees to publish their own work. He therefore had to have his findings printed under the pseudonym “Student” in 1908. Hence the name “Student’s t distribution” and “Student’s t test”.

THE t STATISTIC

The t statistic is derived in a similar fashion to the z statistic:

$t = (\text{sample mean} - \text{population mean})/\text{ESEM}$
Consequently the t statistic is the number of estimated SEM a particular sample mean lies above or below the population mean.

t DISTRIBUTION CURVES

The t statistic tables show the area under the curve between a particular t value and the tip of the tail (fig 1). Along the horizontal axis is the t value. These are equivalent to the SD seen with the normal distribution plots. Therefore the same principle applies regarding set areas under the curve representing particular probabilities.

The curve from which the z statistics were derived remains constant irrespective of the number in the sample. Consequently z values of mean ± 1.96 will always mark out the middle 95% of the population. In contrast, the t statistics vary with sample size because the shape of the distribution changes. It is always symmetrical but with small sample sizes the curve is flatter and has longer “tails”. This is a result of the variation in ESEM as the sample size changes.

With larger samples the t distribution becomes indistinguishable from a normal distribution. Consequently in these cases the z and t statistic values are the same. Therefore, a relevant question at this stage is how small does the sample need to be before the use of the t statistic is necessary. There is no definite answer because it depends upon several factors, including the distribution of the data. For example, when the data are normally distributed the z statistic can be used when there is as little as 30 subjects in the sample. In general however, it is recommended that the t test should be used when dealing with ESEM derived from samples sizes that are less than 100.²

Key points

- As the ESEM varies with sample size, the t statistic value will also vary with sample size
- Smaller samples have the biggest differences between the z and t statistics
- As the sample size increases the t distribution takes on a normal distribution

df \ p	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.683	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.160	2.650	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
24	1.318	1.711	2.064	2.492	2.797

Figure 1 Extract of the *t* table. The first column lists the degrees of freedom ($n - 1$). The remaining columns give the probabilities (P) for t to exceed the values listed. Symmetry is used for negative t values.

USING THE *t* TABLE

As there is a family of *t* distribution curves, depending upon the sample size, the *t* table does not look initially like the *z* statistic table (fig 1). However, each line of the table represents the equivalent of a whole *z* table for a particular sample size. The left column deals with the size of the sample. It is labelled the “degrees of freedom” rather than sample number because, for mathematical reasons, we need to use a value one less than the number in the sample. For example, the *t* statistics for a sample of 15 would be found along the line whose degree of freedom was 14. Therefore, for this sample size, 2.5% of the total area under the curve lies between a *t* value of +2.145 to the right tail tip.

As described above, the *t* statistic allows estimations of the population’s standard error of the mean to be made from the sample data. This enables you to determine, in a population with an unknown standard deviation:

- The chance of getting a sample mean greater than or equal to a particular value
- The value of a sample mean with a particular chance of occurring
- The chance of getting a sample mean between two particular values

THE CHANCE OF GETTING A SAMPLE MEAN GREATER THAN OR EQUAL TO A PARTICULAR VALUE

To demonstrate this, consider Egbert’s result when he measured the resting heart rate in all the 25 male members of the department. He found the sample mean to be 70/minute with an *s* of 16. The population mean for fit men was found to be 60/minute.

Therefore:

$$\text{ESEM} = s/\sqrt{n} = 16/\sqrt{25}$$

Without knowing the population’s SEM, Egbert must use the ESEM to determine the chance of getting a resting heart rate equal or greater to 70/minute if his department’s men were part of a fit male population.

The *t* statistic for this resting heart rate is: (sample mean – population mean)/ESEM
Therefore the *t* statistic is:
 $(70 - 60)/3.2 = 3.13$

Using the *t* statistic table, for a sample size of 25, the area between $t = 3.13$ and the tip of the tail is less than 0.005. Consequently the chances of a sample having a resting heart rate greater than or equal to 70/minute in a fit male population is less than 0.5%.

Key points

- When a sample is less than 100, the *t* statistic should be used (rather than *z*) when making inferences about populations that are based upon the ESEM
- You must use ESEM in these cases even if the population’s standard deviation is known

THE VALUE OF SAMPLE MEAN WITH A PARTICULAR CHANCE OF OCCURRING

Using the ESEM, Egbert then determines what random, 25 male person sample mean demarcates the top 2.5% of the population of fit men. This is carried out in a similar manner to before, but this time using the *t* statistic.

Using the *t* statistic table, the proportion 0.025 is equal to 2.064.

This value for *t* can then be used to determine the sample mean by remembering that:

$$t = \text{sample mean} - \text{population mean}/\text{ESEM}$$

Therefore:

$$2.064 = (\text{sample mean} - 60)/3.2$$

Consequently the sample mean is 6.6 + 60 = 67/minute (rounded up)

Therefore 2.5% of random samples of 25 fit men would have a mean resting heart rate of 67/minute or greater.

THE CHANCE OF GETTING A SAMPLE MEAN BETWEEN TWO PARTICULAR VALUES

Again with the ESEM, Egbert can determine the value of the sample means demarcating the middle 95% of the population of fit men. Using the same system, the lower 2.5% of the curve is demarcated by the *t* statistic –2.064 for a sample size of 25.

With this value for *t*, the sample mean can be determined by:

$$-2.064 = (\text{sample mean} - 60)/3.2$$

Therefore the element value is –6.6 + 60 = 53/minute (rounded down).

It follows that the middle 95% of random samples of 25 fit men would have a mean resting heart rate between 53 to 67/minute. This also represents the 95% confidence interval—that is, we are 95% confident that a random sample of 25 men from this population would have a mean resting heart rate between 53 and 67 beats/minute.

The t statistic can therefore be used to calculate confidence intervals. When using data from a sample, the confidence intervals are equal to the sample mean plus/minus the t statistic appropriate for the level of confidence (t_α) multiplied by the ESEM:
 Confidence interval = sample mean $\pm$ ($t_\alpha \times$ ESEM)

Key points

- The t statistic enables the CI of the estimation of the population's mean (μ) to be determined when σ is not known
- When using t to establish a confidence interval the population is assumed to be normally distributed
- Confidence interval = sample mean $\pm$ ($t_\alpha \times$ ESEM) where t_α is the t statistic for the appropriate level of confidence
- As a rough guide, the t statistic for the 95% confidence interval is usually around 2. Therefore, as an approximation, the true mean will lie within a range 2 ESEM above and below the sample mean.

Summary

Provided the sample size is large enough (that is, n greater than 100), the z statistic can be used to determine the confidence interval estimation of the population mean even when the σ is not known. In these cases the estimation of the standard error of the mean is used. The z statistic is also valid when determining the population's proportion based upon a large sample.

However, when dealing with smaller samples, the z statistic is replaced by the t statistic. This makes it possible to estimate, in a population with an unknown standard deviation:

- The probability of getting a sample mean greater than or equal to a particular value
- The value of a sample mean with a particular probability of occurring
- The probability of getting a sample mean between two particular values

The confidence interval for the estimation of the population mean can also be determined using the t statistic.

Quiz

- 1 A sample of five patients with fractured necks of femurs was studied. The trolley waiting times were: 1, 2, 2, 2, 3 hours respectively. What is the best estimate for the population's mean and estimated standard deviation of the mean?
- 2 The systolic blood pressure (SBP) is measured in 144 randomly selected, elderly (over 70 years) male patients, presenting to Deathstar's Emergency Department. The mean SBP is 140 mm Hg and $s = 30$ mm Hg. What is the 95% confidence interval for the mean SBP for this population of patients?
- 3 Egbert and Endora are asked to determine the proportion of asthmatics who had their inhaler technique assessed before discharge

from their emergency departments. After a year's study 160 were appropriately assessed out of a random sample of 200 asthmatics. What is the 99% confidence interval for the proportion assessed in the population?

- 4 Egbert is interested in the total cholesterol concentrations of patients presenting with chest pain. He finds the mean concentration is 8.1 mmol/l in a sample of 25 randomly selected patients. s is calculated to be 2.5 mmol/l. Assuming that the population is normally distributed, what is the 95% confidence interval for the population's mean cholesterol level?
- 5 One for you to try on your own. Endora repeats Egbert's resting heart rate study with 16 female nurses in the Emergency Department of St Heartsinc.
 60, 66, 66, 62, 68, 70, 70, 70, 72, 72, 76, 76, 78, 78, 80, 80 beats/minute
 What is the 95% confidence interval for the population's mean resting heart rate?

Answers

- 1 The best estimate of the population mean is the sample mean:
 Estimated population mean = (sum of all measures/ n) = $10/5 = 2$ hours
 The ESEM is $s/\sqrt{n}$:
 where:

$$s = \sqrt{[\Sigma(x - 2)^2/4]} \quad (7)$$

$$= \sqrt{[2/4]} = 0.707$$

therefore;

ESEM = $0.707/\sqrt{5} = 0.32$ hours (approximately)

- 2 The best estimate for the population's mean SBP is the sample mean (that is, 140 mm Hg). The SEM of the population is not known but the estimated standard error of the mean can be calculated:

$$\text{ESEM} = s/\sqrt{n}$$

$$= 30/12 = 2.5$$

As the sample is over 100 it is reasonable to assume the z statistic will be valid. Therefore the confidence interval for the estimated mean is:

Sample mean $\pm$ ($z_\alpha \times$ ESEM)

z_α for a 95% confidence interval is 1.96. Therefore:

$$95\% \text{ CI} = 140 \pm (1.96 \times 2.5)$$

$$= 140 \pm 4.9$$

$$= 135 \text{ to } 145 \text{ mm Hg}$$

- 3 Again the best estimation of the population's proportion is the samples proportion. This is: $160/200 = 0.8$

As the smallest group is greater than 10, and the sample is large, it is valid to determine the confidence interval for this estimated proportion by the following formula:

$$p \pm [z_\alpha \sqrt{(pq/n)}] \quad (8)$$

For a 99% confidence interval z_α is 2.58. Therefore:

$$99\% \text{ confidence intervals} = 0.8 \pm [2.58 \times \sqrt{(0.75 \times 0.25/200)}]$$

$$= 0.8 \pm 0.07$$

$$= 0.73 \text{ to } 0.87$$

4 The best estimation for the population mean is the sample mean—that is, 8.1 mmol/l.

We do not know the standard deviation of the population but it is possible to calculate an estimation of the standard error of the mean:

$$\begin{aligned} \text{ESEM} &= s/\sqrt{n} \\ &= 2.5/5 = 0.5 \end{aligned}$$

The sample size is less than a 100 but we know the population is normally distributed. Consequently the confidence intervals should be determined using the *t* statistic:

Confidence interval = sample mean $\pm$ ($t_o \times \text{ESEM}$)

In this case we are interested in the 95% confidence intervals. The sample size is 25, which means the degrees of freedom = 24. Using the *t* table this gives a t_o of 2.064. Therefore:

$$\begin{aligned} 95\% \text{ confidence interval} &= 8.1 \pm (2.064 \times 0.5) \\ &= 8.1 \pm 1.03 \\ &= 7.1 \text{ to } 9.1 \text{ mmol/l} \end{aligned}$$

The authors would like to thank Sally Hollis, Jim Wardrope and Iram Butt for their invaluable suggestions.

- 1 Driscoll P, Lecky F, Crosby M. An introduction to estimation—1. *J Accid Emerg Med* 2000;17:409–15.
- 2 Bland M. Analysis of the means of small samples using the *t* distribution. In: *An introduction to medical statistics*. Oxford: Oxford University Press, 1989:165–87.

Further reading

- Glaser A. Inferential statistics. In: *High yield statistics*. Baltimore: William and Wilkins, 1995:9–30.
- Norman G, Streiner D. Statistical inference. In: *PDQ statistics*. 2nd ed. St Louis: Mosby, 1997:17–36.
- Koosis D. Estimating. In: *Statistics*. 4th ed. New York: John Wiley, 1997:77–103.
- Philips J. Description to inference: a transition. In: *How to think about statistics*. 6th ed. New York: WH Freeman, 2000:108–24.

Correction

We regret that three errors occurred in the previous paper in the series “An introduction to statistics” (Driscoll P, Lecky F, Crosby M. Article 4. An introduction to estimation—1. Starting from Z. *J Accid Emerg Med* 2000;17:409–15).

On p 413 the key point, 1st column, should read:

SEM = Population standard deviation (σ)/ $\sqrt{\text{number in the sample (n)}}$

On p 414, 1st column, the 1st and 2nd lines should read:

SEM = Population standard deviation (σ)/ $\sqrt{\text{size of the sample (n)}}$

On p 415, 1st column, the 2nd sentence should read:

As the latter is equal to the $\sqrt{(\sigma/n)}$, the range can only be reduced by increasing the sample size.