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 ($\mu_{\bar{x}}$) is the same as the population mean (μ). Furthermore, the standard deviation of the distribution is equal to σ/√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:

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

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:
SEM = σ/√n
we use the estimated SEM:
ESEM = s/√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 = (sample mean − population mean)/ESEM
Where the ESEM is = s/√n = 20/√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 = (sample mean − population mean)/ESEM
1.96 = (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 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_o) multiplied by the ESEM:

Confidence interval = sample mean +/- ($z_o \times$ 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 +/- ($z_o \times$ ESEM) where z_o 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:

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

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 over come 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 = (sample mean − population mean)/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

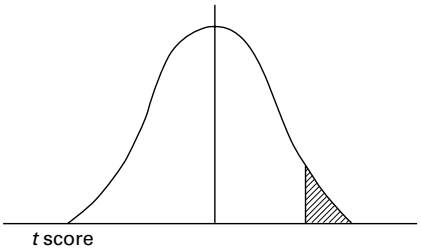
The *t* statistic tables show the area under the curve between a particular t value and the tip of the tail (fig 1). Along the horizontal axis is the *t* value. These are equivalent to the SD seen with the normal distribution plots. Therefore the same principle applies regarding set areas under the curve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

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 = sample mean − population mean/ESEM

Therefore:

2.064 = (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 = (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_o) multiplied by the ESEM:
Confidence interval = sample mean +/− (t_o x 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 +/− ($t_o \times$ ESEM) where t_o 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/ n :
where:

$$s = \sqrt{[\Sigma(x-2)^2/4]} = \sqrt{[2/4]} = 0.707 \quad (7)$$

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:
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 +/− ($z_o \times$ ESEM)
z_o for a 95% confidence interval is 1.96. Therefore:
95% CI = 140 +/− (1.96 × 2.5)
= 140 +/− 4.9
= 135 to 145 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:

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

For a 99% confidence interval z_o is 2.58. Therefore:
99% confidence intervals = 0.8 +/− [2.58 × √ (0.75 × 0.25/200)]
= 0.8 +/− 0.07
= 0.73 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:

ESEM = $s/\sqrt{n}$
= 2.5/5 = 0.5

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 +/− ($t_o \times$ 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:

95% confidence interval = 8.1 +/− (2.064 × 0.5)
= 8.1 +/− 1.03
= 7.1 to 9.1 mmol/l

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)/√number in the sample (n)

On p 414, 1st column, the 1st and 2nd lines should read:
SEM = Population standard deviation (σ)/√size of the sample (n)

On p 415, 1st column, the 2nd sentence should read:
As the latter is equal to the √(σ)/n, the range can only be reduced by increasing the sample size.