Confidence intervals

Contents

Introduction

As a general rule, a random sample will provide sample statistics that are close to the equivalent population parameters. But we have to recognise that there is certainly the chance that a random sample will generate sample statistics that are distant from the equivalent population parameters. Thus, the sampling error could be low or high. (For example, say the true proportion of voters who will vote for the Australian Labor Party in the next election is 45%. From a random sample we could get exactly 450 people, or 45%, saying 'ALP', but we could also get 443 (that is, 44.3%) or 456 (or 45.6%) or 501 (or 50.1%).

Since samples are likely to be wrong and we do not know if this error will be small or large, a correction factor should be employed in the form of a *margin of error*.

We frequently use margins of error ourselves. For example, we all make statements like:

> 'I will be there in half an hour, give or take a few minutes.'
>
> 'At that restaurant, expect to pay $85 for two, give or take a few dollars.'
>
> 'We expect the company's share price to end the year around the $4.50 mark'.

Confidence intervals are not a dissimilar concept. If a public opinion poll showed 620 out of 1000 were in favour of a proposed new government initiative, then it means the true proportion of all voters in favour is 'round about the 62% mark'. But we need to be more precise than that: could it be more than 62%, could it be *less* than 62%. How much more (or less)?

Confidence intervals are a formal procedure for correcting random sampling error. The sample estimator—such as the sample mean, which is a point estimator of the population mean—is converted to an interval estimator by adding and subtracting a correction factor known as the margin of error. Further, we can provide a degree of confidence for the interval constructed. The degree of confidence chosen is typically 90%, 95%, 98% or 99%.

An interval is not as precise as we would like: a single figure is preferable to a range. But in many cases the information needed by a business about the population parameter does not need to be extremely precise. In a marketing proposal, for example, we might be quite happy to assume that the average income of our target segment is between $63 500 and $65 500, especially if we are 95% or 99% confident that the true figure is in that range.

In summary, we tend to use confidence intervals if we have no idea about the population parameters (such as μ or p) which we are trying to estimate. A random sample is taken, the sample statistic ($\bar{X}$ or p) determined, then a margin of error placed either side of this value. The margin of error is calculated in conjunction with the normal distribution (and later with the t distribution). The end result is that the margin of error and the interval estimate have a degree of confidence (say 95%) attached.

Confidence intervals

A key point you *must* realise is the temptation and danger of using a point estimate of a sample statistic as a definitive measure of the equivalent population parameter. Thus, population parameters could be estimated by point estimators which give us a single estimate. That is, $\overline{X}$ = 56 is a point estimate of some unknown population *mean* while *p* = 0.08 is a point estimate of some unknown population *proportion*.

However, the point estimates typically come from one sample taken from the population. Because we cannot guarantee that the sample chosen will be totally representative of the population, there is a danger that the point estimate may not be accurate as a substitute for the population parameter. The sample statistic may be close to the population parameter, but it might also be distant; we must allow for this possibility in the inferential process. Sampling theory suggests that a statistic could underestimate or overestimate the population parameter, in particular. Our knowledge of the normal distribution is that the sample mean or proportion could be up to three standard errors above or below the true figure (based on three standard deviations and 99.73% probability).

We should incorporate this information, allow for a margin of error and construct an *interval* estimate for the population parameter. We use sampling theory to provide the basis for constructing the interval; thus, we can associate probabilities with the intervals constructed.

For many decision-making processes, an interval estimate is more than adequate. For example, it is generally not necessary to estimate the average income (μ) of our target market exactly, but it may suffice if we estimate that the target market income is between $55,000 and $60,000—especially if we could attach a high degree of *confidence* to the interval estimate. We cannot attach a degree of confidence to a point estimate. In the same way it may suffice to estimate that the proportion of customers who will pay by cash in our store is between 15–20% with a confidence 95%.

We now proceed to construct confidence intervals under different conditions. All confidence intervals have the same basic features:

- a margin of error either side of the sample statistics, and
- the margin of error depends on the degree of confidence desired (90%, 95%, etc.).

Confidence intervals for the population mean μ (σ known)

In general the formula for the confidence interval for μ where σ is known is given by:

$$\overline{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

Z corresponds to the level of confidence. For example, for 95% confidence the corresponding value of Z is 1.96 (this can be obtained from either statistical software or the standardised normal table). The value of Z increases/decreases

depending on whether a higher/lower level of confidence is required. As the confidence coefficient increases (or decreases) then the width of the confidence interval also increases (or decreases).

In the tutorial, you will practice calculating confidence intervals using software. You also need to know the basics of calculating confidence intervals by hand as illustrated in the following application.

APPLICATION: LIGHT BULBS

The quality control manager at a light-bulb factory needs to estimate the average life of a large shipment. The process standard deviation (σ) is known from past experience to be 100 hours. A random sample of 50 light bulbs indicated a sample average life of 350 hours.

(a) Construct a 95% confidence interval for the population average lifetime (μ).

For 95% confidence, z = 1.96 (see Normal distribution table)

Applying the formula

95% confidence interval for μ (σ known).

$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

$$350 \pm 1.96 \frac{100}{\sqrt{50}}$$

or 322.28 to 377.72 (Interval estimate)

We estimate, with 95% confidence, that the mean lifetime of the population of light bulbs is somewhere between 322.28 and 377.72 hours. In particular, we would say '**we are 95% confident the mean lifetime of all light bulbs is somewhere in the range 322.28 hours and 377.72 hours**'. This information gives management a range of values used to estimate the population average lifetime of the bulbs. Management could use this information to decide if the quality of the bulbs was meeting standards.

(b) Suppose we wish for the same circumstance to construct a 98% confidence interval.

We determine for 98% confidence, Z = 2.33.

Substitution of the relevant values into the equation yields:

$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

We obtain:

$$350 \pm 2.33 \frac{100}{\sqrt{50}}$$

= 350 ± 32.9511

or 317.05 to 382.95

Note that the 98% interval is wider than the 95% interval constructed earlier. We have more confidence in the interval estimate but the interval is less precise.

Finally, the confidence interval procedure outlined above implicitly assumes that the sampling distribution has a normal distribution (or approximate). The Central Limit Theorem tells us that the sampling distribution is normal provided that the sample size is large enough (sometimes put at $n = 30$ or more).

Confidence interval for the population mean μ (σ unknown)

If the population standard deviation σ is not known, we have to modify our approach to constructing confidence intervals. The two key modifications are using the sample standard deviation, *S*, instead of σ, and using the *t* distribution instead of the normal distribution.

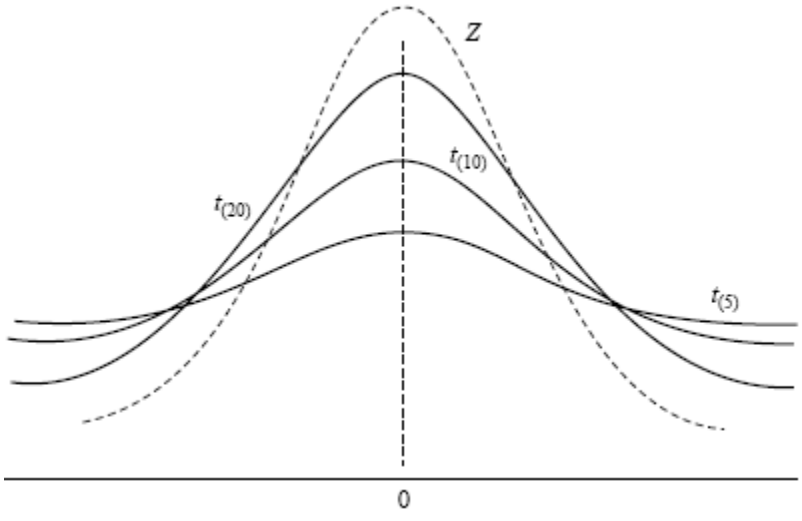
In practice, we need to use *S* as σ is generally not known, and *S* will generally provide us with a very good approximation to σ. And we use the *t* distribution, not the *Z* distribution. If we use the sample standard deviation we are introducing greater sampling variation for the standardised statistic. The *t* variable allows for this whereas the *Z* variable doesn't.

Note that there are many similarities of the *t* with the standard normal distribution (Z distribution). It is generally flatter than the *Z* curve and longer in the tails, that is, more area or probability is contained in the tail area.

In determining a confidence interval for the population mean (μ) (σ unknown), the degrees of freedom for the *t* distribution are equal to ($n - 1$) where n = sample size.

Varying the degrees of freedom alters the shape of the curve. A comparison of some *t* curves is shown below. Also shown for comparison is the standard normal (*Z*).

Exhibit: ***t* distribution**



As the degrees of freedom increase, the *t* curve becomes more closely dispersed around 0. As *n* increases, the *t* curve approaches *Z (standard normal distribution)*.

It should be stressed that in using the *t* distribution for small samples (when $n < 30$) in the context of estimating the population mean we need to assume that the underlying population distribution is normal. Since *n* is small we cannot rely on the central limit theorem and must assure ourselves that the population is normal.

In situations where σ is unknown and is estimated by *S* there are two basic situations:

1 If $n < 30$ we use *S* as a substitute for σ and use the *t* distribution, not the *Z* distribution. However, we must be confident that we are able to assume the population is approximately normal.

2 If $n \geq 30$ then we use *S* as a substitute for σ and use the *t* distribution, not the *Z* distribution. However, we do not have to be concerned about whether the population is normal.

In either situation, the general formula for calculating a confidence interval is:

$$\bar{X} \pm t\frac{S}{\sqrt{n}}$$

The following application shows you how to apply the formula.

APPLICATION: LENGTH OF CALLS

(a) Suppose Telephone Company A wishes to estimate the mean length of telephone calls for private residences between Sydney and Melbourne from 6 pm to 9 pm every day. A random sample of 16 calls reveals a mean length of 350 seconds with a standard deviation of 100 seconds. Based on the data given, estimate a 95% confidence interval for the mean duration of a call between the times indicated.

Since $n < 30$, we need to check our sample to see if we can assume the population is approximately normal. If this assumption is not valid, then the confidence interval procedure below is invalid.

If the assumption is valid, the confidence level is 95%, the degrees of freedom are $n - 1 = 15$, and therefore $t = 2.1315$.

The relevant formula is then:

$$\bar{X} \pm t\frac{S}{\sqrt{n}}$$

Substituting the relevant values from the sample and from the *t* tables, we calculate:

$350 \pm 2.1315 \left(\frac{100}{4}\right)$

$= 350 \pm 2.1315\ (25)$

$= 350 \pm 53.29$

296.71 to 403.29

297 to 404 seconds.

We interpret this interval as follows:

We are 95% confident that the population average length of calls is somewhere in the interval stated, that is, between 296 and 404 seconds. Management could use this information to decide on charges for certain durations.

Note how the small sample size has made the interval imprecise. The reason is that there is a deal of variability in the sample (population) itself. This should indicate to you that confidence intervals from very small samples will be generally of little use unless the standard deviation of the sample (population) is small.

(b) Suppose Telephone Company B took a survey of 50 calls using its network from private residences from Sydney to Melbourne between the hours of 6 pm to 9 pm. They found a sample mean of 350 seconds with a sample standard deviation S = 40 seconds. Determine a 95% confidence interval for the average length of call in the circumstance stated for the Company B network.

$$\bar{X} \pm t\frac{S}{\sqrt{n}}$$

$$= 350 \pm 2.0096\ \frac{40}{\sqrt{50}}$$

$$= 350 \pm 11.368$$

338.632 to 361.368

Thus, we are 95% confident that the true population average length of call is somewhere between 339 seconds and 361 seconds. This information would be of some use as the confidence level is 95% and the interval is reasonably precise.

EXERCISE 1

1 A market researcher states that the average weekly consumption of beer per household is between 2.7 and 3.1 litres with 95% confidence. Explain the meaning of this statement.

2 Students should attempt the following exercise by hand and verify the results using a computer where appropriate.

At a soft-drink filling factory the filling machine is set to an average of 1250 ml with a standard deviation of 50 ml. Regularly, 16 bottles from the production line are examined and the sample average noted. For one such sample, the sample average was 1223 ml with a sample standard deviation of 42 ml.

(a) Calculate 90% and 99% confidence intervals for the average fill of the machine. Compare your answers.

(b) What assumption was necessary for your calculations in (a)?

(c) Based on your results in (a) would you say that the machine is producing to the set standard, that is, average = 1250 ml? Explain.

Confidence intervals for the population proportion, π

In many circumstances the variable of interest is not the mean of a population but the proportion possessing a particular attribute. Again, there are dangers in using the sample proportion as a point estimator of the population proportion. We need to allow for sampling error, and thus allow for a margin of error. To do so, we need the equivalent to the central limit theorem to understand the sampling distribution of sample proportions.

Sampling distribution for the sample proportion, *p*

The sampling distribution for *p* is basically an extension of the binomial distribution when large samples are being taken. If we were taking a random sample of 10 items to determine the number of defectives, we would use the binomial distribution. But if we can take a random sample of 100, or even 500, we can concentrate on the proportion (rather than the number) of defectives in each sample and use the normal distribution rather than the binomial. The reason is that, provided the sample size is large enough (see below), the binomial distribution can be approximated by the normal distribution, which is a far easier distribution to work with.

The sample size is considered to be large enough if:

$n\pi \geq 5$, and $n(1-\pi) \geq 5$

Then the distribution of *p* will be approximately normally distributed about a mean of π and with a standard deviation (or standard error) of:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

We then proceed to calculate probabilities associated with given ranges by using the *Z* transformation because the variable is normal.

It also allows us a framework on which we can build a formal procedure for inference (confidence intervals) about π with associated probability statements and error statements.

Calculating confidence intervals for a population proportion, π

The confidence interval procedure illustrated earlier for population means is applicable in the situation where the population proportion is the parameter of interest. For example, we may wish to use sample data to estimate the proportion of all staff who are satisfied or the proportion of all customers who feel service levels can be improved.

The confidence intervals for proportions are constructed and interpreted using the same basic procedure as for the quantitative variables, except that proportions are qualitative (categorical) variables and the Z distribution is always used.

The general formula for calculating a confidence interval for a proportion is:

$$p \pm Z\sqrt{\frac{p(1\pm p)}{n}}$$

The following application shows you how to apply the formula.

APPLICATION: AUDIT

The credit manager of a large department store is concerned at the number of customers that have been complaining about errors in their credit accounts. To estimate the proportion of erroneous accounts, a sample of 400 accounts are audited. In the sample, 80 accounts are found to be in error. Determine a 95% confidence interval for the true proportion of accounts that are erroneous.

Firstly, estimate the sample proportion *p*.

$$p = \frac{X}{n} = \frac{\text{Number of accounts in error}}{\text{Sample size}}$$

$$= \frac{80}{400} = 0.2 \text{ or } 20\%$$

$$p \pm Z\sqrt{\frac{p\ (1 \pm p\ )}{n}}$$

$$0.2 \pm 1.96\ \sqrt{\frac{(0.2)(0.8)}{400}}$$

$$0.2 \pm 1.96\ (0.02)$$

$$0.2 \pm 0.0392$$

$$0.1608 \text{ to } 0.2392$$

We estimate with 95% confidence that the true proportion of accounts in error is somewhere between 16.08% and 23.92%. Management can use this information to consider if action is necessary, or if more stringent accounting procedures should be implemented.

Determining sample size

Determining the sample size required for a particular sampling situation is one of the most essential steps in statistics. This section illustrates how the sample size can be determined objectively based on predefined confidence levels and levels of required precision. Sample size determination can be done for both means and proportions.

Sample size determination for the mean (μ)

The relevant formula for this situation is:

$$n = \frac{z^2\sigma^2}{ME^2}$$

The following application shows you how to apply the formula.

APPLICATION: AMERICAN TOURISTS

In order to determine the average amount of money American tourists spend in Australia per week, the Australian Tourist Board wishes to commission a survey. They wish to sample enough tourists so that the level of confidence is 98% and the precision level is within $200 from the population average. (In other words, we wish to be 98% confident that the value of $\bar{X}$ in our sample will be no more than $200 from the true mean μ.) Based on a pilot study, and survey results from earlier years, we can assume that the population standard deviation is $800. What sample size is needed for the survey?

The confidence level is 98%. Thus:

$Z = 2.33$ (From the Z tables)

The given population standard deviation is $800.

The level of precision required is ME = $200

$$n = \frac{(2.33)^2(800)^2}{(200)^2}$$

= 86.86, rounded up to 87

Therefore, a sample of at least 87 is needed to achieve the precision objectives.

(Note that we round up to the next integer **in all cases** to ensure we take a sample that is at least large enough to achieve the survey objectives. From a practical point of view, perhaps 100 would be surveyed for this problem.)

Sample size determination for a proportion

This type of analysis is very common and undertaken by the familiar polling firms (Morgan Gallup, AGB McNair etc.) to help them determine the required sample size.

Clearly, whether for estimating μ and π, the specified confidence and margin of error must be balanced against the cost of collecting the sample data.

As a matter of interest, when you next read the results of a poll in a magazine or newspaper, read the fine print which typically outlines the sample size and the precision level. In many cases, the fine print will indicate the 'margin for error' (for example, 2–3%). This is an alternative term for the precision level. Also, typically the level of confidence is not stated, but it is usually 90% or 95% or 99%.

The relevant formula for this situation is:

$$n = \frac{z^2\pi(1-\pi)}{ME^2}$$

The following application shows you how to apply the formula.

APPLICATION: POLLING

In many instances, market research firms undertake polls to estimate, among other things, the proportion of the population that favours some proposition (such as gun control). Suppose we wish to commission a survey to estimate the proportion of the population in favour of restoring tariffs at a level of 20% across the board. The confidence level required is 90% and the precision level of ± 0.025 is needed. This means that we want to be 90% confident the sample proportion is within 0.025 of the true proportion. What sample size is needed?

The confidence level is 90%. Thus $Z = 1.645$ (From the Z table)

The level of precision required is $ME = 0.025$.

$$n = \frac{(1.645)^2 \times \pi(1-\pi)}{(0.025)^2}$$

In this case we don't know π. As a conservative rule use $\pi = 0.5$.

$$n = \frac{(1.645)^2 (0.5)(0.5)}{(0.025)^2}$$

$= 1082.41$

We would need to poll at least 1083 people to achieve the nominated precision and confidence levels. From a practical point of view we might take 1100.

Note you also could use the sample proportion p, as an estimate of π, if you had no prior knowledge of π.

Summary

Confidence interval estimates are procedures for estimating population parameters by making an allowance for sampling error through the calculation of the margin of error. The key advantage of confidence interval procedures is that they allow probability statements to be associated with the procedure, whereas point estimation does not permit this.

Further resources

Black, K 2008, *Business statistics for contemporary decision making*, 5th edn, Wiley, NJ.

Anderson, DR, Sweeney, DJ & Williams, TA 2008, *Statistics for business and economics*, 10th edn, South-Western Thomson Learning, Cincinnati, Mass.

Selvanathan, A, Selvanathan, S, Keller, G & Warrack, B 2006, *Australian business statistics*, 4th edn, Nelson Thomson Learning, Melbourne.