MAKING SENSE OF DATA

Intermediate series

CONFIDENCE INTERVALS

In this topic you find out how to build, calculate, and interpret confidence intervals, and to work through the formulas involving one or two population means or proportions. Other points of confidence intervals are covered: what makes them narrow or wide, what makes you more or less confident in their results, and what they do and don't measure.

Estimation

A *confidence interval* (abbreviated CI) is used for the purpose of estimating a population parameter (a single number that describes a population) by using statistics (numbers that describe a sample of data). For example, you might estimate the average household income (parameter) based on the average household income from a random sample of 1,000 homes (statistic). However, because sample results will vary you need to add a measure of that variability to your estimate. This measure of variability is called the margin of error, the heart of a confidence interval. Your sample statistic, plus or minus your margin of error, gives you a range of likely values for the parameter — in other words, a confidence interval.

For example, say the percentage of children who like football is 40 percent, $\pm$ 3.5%. That means the percentage of children who like football is somewhere between 36.5% (40 - 3.5%) and 43.5% (40 + 3.5%). The lower end of the interval is your statistic minus the margin of error, and the upper end is your statistic plus the margin of error.

To estimate a parameter with a confidence interval:

1. Choose your confidence level and your sample size (see details later in this topic).
2. Select a random sample of individuals from the population.
3. Source or collect reliable and relevant data from the individuals in the sample.
4. Summarize the data into a statistic (e.g. a sample mean or proportion).
5. Calculate the margin of error (details following).
6. Take the statistic plus or minus the margin of error to get the final estimate of the parameter.

   This is called a *confidence interval* for that parameter.

Margin of error

The ultimate goal when making an estimate using a confidence interval is to have a small margin of error. The narrower the interval, the more precise the results are.

How do you go about ensuring that your confidence interval will be narrow enough? You want to think about this issue before collecting any data; after the data are collected, the width of the confidence interval is set.

Three factors affect the size of the margin of error:

- ✓ The confidence level
- ✓ The sample size
- ✓ The amount of variability in the population

These three factors all play important roles in influencing the width of a confidence interval.

Confidence level

Variability in sample statistics is measured in standard errors. A *standard error* is very similar to the standard deviation of a data set or a population. The difference is that a standard error measures the variation among all the possible values of the statistic (for example all the possible sample means) while a standard deviation of a population measures the variation among all possible values within the population itself.

The *confidence level* of a confidence interval corresponds to the percentage of the time your result would be correct if you took numerous random samples. Typical confidence levels are 95% or 99%. The confidence level determines the number of standard errors you add and subtract to get the percentage confidence you want.

When working with means and proportions, if the proper conditions are met, the number of standard errors to be added and subtracted for a given confidence level is based on the standard normal (Z-) distribution, and is labelled z*. The higher the confidence level, the more standard errors need to be added and subtracted, hence a higher z*-value. For 95% confidence, the z*-value is 1.96, and for 99% confidence, z*-value is 2.58.

Percentage Confidence	z*-value
80	1.28
90	1.64
95	1.96
98	2.33
99	2.58

Table 1: z*-values for selected (percentage) confidence levels

Using statistical notation, you can write a confidence level as $(1 - \alpha)$, where α represents the percentage of confidence intervals that are incorrect (don't contain the population parameter by random chance). So if you want a 95 percent confidence interval, $\alpha = 0.05$. This number is also related to the chance of making a Type I error in a *hypothesis test*.

Sample size

The relationship between margin of error and sample size is simple: as the sample size increases, the margin of error decreases. Or put another way, the more information you have, the more accurate your results are going to be.

Looking at the formula for standard error for the sample mean, $\frac{\sigma}{\sqrt{n}}$ (from "Sampling distributions" topic) notice that as n increases, the denominator of this fraction increases, which makes the overall fraction get smaller. That makes the margin of error, $z^* \frac{\sigma}{\sqrt{n}}$, smaller and results in a narrower confidence interval.

When you need a high level of confidence, you have to increase the z*-value and, hence, the margin of error. This makes your confidence interval wider (not good). But we can offset this wider confidence interval by increasing the sample size (n) and bring the margin of error down, thus narrowing the confidence interval.

When your statistic is a sample proportion or percentage (such as the proportion of females) a "quick-and-dirty" way to figure margin of error is to take 1 divided by the square root of n (the sample size).

Approximately what sample size is needed to have a narrow confidence interval with respect to opinion polls? Using this formula, you can make some quick comparisons. A survey of 100 people will have a margin of error of about $\frac{1}{\sqrt{100}} = 0.10$ or $\pm 10\%$ (which is fairly large). However, if you survey 1,000 people, your margin of error decreases dramatically, to $\pm \frac{1}{\sqrt{1000}}$ or about 3%.

Population variability

Another factor influencing variability in sample results is the variability (standard deviation) within the population itself. For example, in a population of houses in a large city like London you see a large amount of variability in price. This variability in house price over the whole city will be higher than the variability in house price if your population was limited to a certain Borough in London (where the houses are likely to be similar to each other).

As a result, if you take a sample of houses from the entire city of London and find the average price, the margin of error will be larger than if you take a sample from a single Borough in London. So you'll need to sample more houses from the entire city of London in order to have the same amount of accuracy that you would get from a single Borough.

Variability is measured in terms of standard errors/deviations. Notice that the population standard deviation, σ appears in the numerator of the standard error of the sample mean, $\frac{\sigma}{\sqrt{n}}$. As σ (numerator) increases, the standard error (entire fraction) increases. A larger standard error means a larger margin of error and a wider confidence interval.

More variability in the original population increases the margin of error, making the confidence interval wider. However, this increase can be offset by increasing the sample size. (Remember the sample size, n, appears in the denominator of the standard error formula, $\frac{\sigma}{\sqrt{n}}$, so an increase in n results in a decrease in the margin of error.)

Confidence Intervals for …

A Population Mean

When the characteristic that's being measured (such as income, IQ, price, height, quantity, or weight) is *numerical*, people often want to estimate the mean (average) value for the population. You estimate the population mean by using a sample mean plus or minus a margin of error. The result is a *confidence interval for a population mean*, μ.

The formula for a CI for a population mean is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

where $\bar{x}$ is the sample mean; σ is the population standard deviation; n is the sample size; n and z* is the appropriate value from the Z-distribution for your desired confidence level.

When your sample size is small (under 30), you use the appropriate value on the *t–distribution* with n–1 degrees of freedom instead of z*.

A Population Proportion

When a characteristic being measured is categorical — for example, opinion on an issue (support, oppose, or are neutral), or type of behaviour (do/don't shop on-line), people often want to estimate the proportion (or percentage) of people in the population that fall into a certain category of interest. Examples include the percentage of staff in favour of flexible working, or the proportion of voters in favour of the death penalty. In each of these cases, the object is to estimate a population proportion using a sample proportion plus or minus a margin of error. The result is called a *confidence interval for a population proportion*, p.

The formula for a CI for a population proportion, p, is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ is the sample proportion; n is the sample size; and z* is the appropriate value from the standard normal (Z-) distribution for your desired confidence level. (Note that a sample proportion is the proportion of individuals in the sample that had the characteristic of interest.)

For example, suppose you want to estimate the percentage of the time you get a red light at a certain junction. If you want a 95% confidence interval, your z*-value is 1.96. Let's say you take a random sample of 100 different trips through this junction, and you find that you hit a red light 53 times, so $\hat{p} = 53/100 = 0.53$. Now calculate 0.53 * (1 - 0.53) and divide by 100 to get $0.249/100 = 0.00249$. Take the square root of this result to get 0.0499 (or 0.05). The margin of error is given as $\pm\ 1.96 * 0.05 = 0.098$.

Therefore you can conclude the overall percentage of the times you expect to hit a red light at this junction is somewhere between 43.2% and 62.8%, based on this sample, with a confidence level of 95%.

The Difference of Two Means

The goal of many surveys is to compare two populations, such as Conservative versus Labour. When the characteristic being compared is numerical (for example, height, weight, or income) the object of interest is the amount of difference in the means (averages) for the two populations. For example, you may want to compare the difference in average incomes of Conservative versus Labour voters. You estimate the difference between two population means by taking a sample from each population and using the difference of the two sample means, plus or minus a margin of error. The result is a *confidence interval for the difference of two population means*, $\mu_1 = \mu_2$.

The formula for a CI for the difference between two population means is

$$\bar{x} - \bar{y} \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $\bar{x}$ and $\bar{y}$ are the sample means, respectively; n_1 and n_2 are the sample sizes; σ_1 and σ_2 are the population standard deviations; and z* is the appropriate value from the standard normal (Z-) distribution for the desired confidence level.

If one or both of the sample sizes are small (less than 30) you use the appropriate value on the *t-distribution* with $n_1 + n_2 - 2$ degrees of freedom instead of z*.

The Difference of Two Proportions

When two populations are compared regarding some categorical variable (such as comparing male to female staff regarding their opinion on flexible working) you estimate the difference between the two population proportions. We do this by taking the difference in their corresponding sample proportions (one from each population) plus or minus a margin of error. The result is called a *confidence interval for the difference of two population proportions*, $p_1 - p_2$.

The formula for a confidence interval for the difference between two population proportions is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $\hat{p}_1$ and n_1 are the sample proportion and sample size of the first sample; $\hat{p}_2$ and n_2 are the sample proportion and sample size of the second sample; and z* is the appropriate value from the standard normal (Z-) distribution for your desired confidence level.

Interpreting confidence intervals

The big idea of a confidence interval is that it presents a range of likely values for the population parameter, based on one random sample, with a certain confidence level (such as 95%). However, there are some intricacies that can lead to incorrect interpretation of the results.

Consider a survey conducted by Ipsos MORI (a national leader in the survey research). Suppose they sample 1,000 people at random from the UK, and the results show that 520 people (52%) think the Prime Minister (PM) is doing a good job. Ipsos MORI reports this survey has a margin of error of plus or minus 3%. So far, you know that a majority of the 1,000 people in this sample approve of the PM, but can you say this opinion carries over to a majority of *all* residents in the UK?

If 52% of those sampled approve of the PM, you can expect the percentage of all UK residents who approve of the president to be 52%, plus or minus 3.0%. That is, a range of likely values is between $52\% - 3\% = 49\%$ and $52\% + 3\% = 55\%$. To report the results from this poll, you would say, "Based on this sample, 52% of all UK residents approve of the Prime Minister, plus or minus a margin of error of 3.0 percent, with a confidence level of 95%."

The subtle but very important point with how to interpret a confidence interval is when one particular confidence interval is calculated; do not include a probability statement about your particular result when you draw your conclusions. That is, it's *wrong* to say "I am 95% confident that the population mean is between XXX and XXX." The confidence level (in this case 95%) does not apply to a single confidence interval.

So how do you interpret the 95%? A *confidence level* is the percentage of all possible samples of size n whose confidence intervals contain the population parameter. When taking many random samples from a population, you know that some samples (in this case, 95% of them) will represent the population, and some won't (5% of them) just by random chance. Random samples that represent the population will result in confidence intervals that contain the population parameter (they are correct); and those that do not represent the population will result in confidence intervals that are not correct.

Confidence level (such as 95%) represents the percentage of all possible random samples of size n that typify the population and hence result in correct confidence intervals. It isn't the probability of a single confidence interval being correct.

Misleading confidence intervals

There are two possible reasons that a confidence interval is incorrect (does not contain the population parameter). First, it can be incorrect by random chance because the random sample it came from didn't represent the population; or second, it can be incorrect because the data that went into it weren't any good.

No matter how well presented and scientific someone's confidence interval may look, the formula that was used to calculate it doesn't have any idea of the quality of the data that went into it. For example, if the data for the confidence interval was based on a biased sample (one that favoured certain people over others); a bad design; bad data-collection procedures; or misleading questions, the margin of error is suspect.