Confidence intervals and coverage probabilities

A **confidence interval (CI)** is an interval of the form (a, b), that is constructed from the data. The purpose of a CI is to cover an unknown population parameter with “high probability” (we use probability here since it is not possible to construct an interval that is guaranteed to always cover the population parameter of interest). For example, if we want to estimate the mean body mass index (BMI) in a population of people, where the true mean BMI is 25.5, then the interval (24, 26.2) would cover the target, while the interval (26.1, 28) would not.

The confidence interval is defined through its **lower confidence bound (LCB)** and its **upper confidence bound (UCB)**, which are both functions of the data. The population parameter of interest, denoted here by θ (*theta*), is an unknown constant. In the example above, θ = 25.5. The “coverage probability” of the confidence interval is the probability that LCB <= θ <= UCB, written P(LCB <= θ <= UCB). The coverage probability is set by the researcher, and in most cases will be set to 95%.

The coverage probability is defined in terms of (hypothetical) repeated sampling of multiple data sets from the population of interest. Over many such repeated samples, constructing one CI from each sample, there will be a fraction of the confidence intervals that cover the target. **This fraction is the coverage probability.**

A wider confidence interval will have an easier time covering the target than a narrower one. On the other hand, a very wide interval is not very informative (imagine if we reported the fraction of voters supporting a particular candidate in an election as 55%, with a 95% CI spanning from 2% to 98%). Thus, the primary goal when constructing an interval is to “adapt to the data”, yielding a wider interval when the power is low and uncertainty is high, and a narrower interval when the power is high and uncertainty is low.

Ideally the “actual” coverage probability of a confidence interval obtained in practice will match the intended or “nominal” coverage probability. But a CI may fail to perform as desired. This is because a CI may be used in a setting where the conditions under which it was derived are violated. Here we will explore some common reasons why this may occur.

The actual coverage probability of a confidence interval may be either less than the nominal coverage level (yielding an “anti-conservative” interval), or greater than the nominal coverage level (yielding a “conservative” interval). Although a conservative interval is often viewed slightly more favorably than an anti-conservative interval, both of these outcomes are undesirable -- we wish to obtain an interval whose actual coverage is as close as possible to the nominal coverage probability.

It is important to reiterate that in practice, we obtain one confidence interval from one sample. This CI either covers or fails to cover the target value. For a specific data set, we do not know whether the CI derived from it actually covers the target value, but this is something that is either true or false - there is no probability involved when discussing whether one specific CI covers the target value.

We rarely have multiple independent samples from the same population, so we cannot usually verify that a confidence interval attains its intended coverage probability. To reassure ourselves that the desired coverage is attained, we can study the theoretical properties that would be guaranteed to result in the intended coverage rate being achieved. We can also use computer simulations to assess how a given method for constructing CIs performs in various hypothetical settings. Statisticians make use of both of these approaches when assessing the performance of confidence intervals in particular settings.

The confidence intervals we have seen so far are all constructed using two key quantities:

1. an unbiased estimate of a population parameter, and
2. the standard error of this estimate.

For example, if we are interested in estimating the population mean based on an independent and identically distributed (iid) sample of data, the unbiased estimate is the sample mean ($\bar{x}$ or x_bar), and the standard error of this estimate is s/sqrt(n) (or $\sigma/\sqrt{n}$), where *s* is the standard deviation of the data, and *n* is the sample size.

Many confidence intervals are constructed using the form “point estimate +/- K standard errors." For example, when working with the sample mean $\bar{x}$ (x_bar), the interval is $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$. The constant K is chosen to give the desired level of coverage. Specifically, we need the “Z-score” $\sigma/\sqrt{n}$ * ($\bar{x}$ - mu) / s to fall between -K and K with probability alpha. As long as this holds, then the interval $\bar{x}$ +/- K *$\sigma/\sqrt{n}$ will have the intended coverage probability. The constant K plays a very important role in determining the properties of a CI, and will be discussed in more detail below.

Constructing confidence intervals

There are two ways we can obtain values of K to use in constructing the CI. One approach is based on making the very strong assumption that the data are independent and identically distributed, and follow a normal (Gaussian) distribution. If this is the case, then the Z-score follows a Student-t distribution with n-1 degrees of freedom. If we set K equal to the 1 - (1 - ɑ)/2 quantile of the Student-t distribution with n-1 degrees of freedom, then the resulting interval will have the intended coverage rate. Values of K constructed from the Student t-distribution will range from 2 to 2.5 for 95% coverage intervals if the sample size is greater than 5 (samples smaller than 5 observations are rare in practice). Thus, most CIs will be constructed by taking a “margin of error” around the point estimate that is between 2 and 2.5 times the standard error.

An alternative and much more broadly applicable basis for obtaining a value for K is to use the “central limit theorem” (CLT). The CLT states that the sample mean of independent and identically distributed values will be approximately normally distributed. The CLT also implies that the Z-score will be approximately normally distributed. Importantly, the CLT provides these guarantees even when the individual data values have distributions that are not normal, as long as the sample size is “sufficiently large.” There are some additional technical conditions needed for the CLT to be applicable, but we will not discuss them here.

Unfortunately, there is no universal rule that defines how large the sample size should be to invoke the central limit theorem. In general, if the data distribution is close to being normal, then the Z-scores will be close to normally-distributed even when the sample size is quite small (e.g. around 10). If the individual data values are far from being normally distributed (e.g. they are strongly skewed or have heavy tails), then the CLT may not be relevant until the sample size is larger, say around 50.

As long as we can justify invoking the CLT, it is appropriate to use the 1 - (1 - ɑ)/2 quantile of the normal distribution to define K, which leads to setting K=1.96 in order to achieve an (approximate) 95% coverage probability. Thus, normality of the individual data values is not needed for a CI to have good coverage properties. It is good practice to inspect the distribution of a sample before proceeding to construct a confidence interval for its mean, for example, by looking at a histogram or quantile plot of the data. But it is not necessary that this show a nearly-normal distribution in order for the confidence interval to be meaningful, unless the sample size is very small and the data are strongly non-normal.

Another common practice is to use K as calculated from the Student-t distribution, even when the data are not taken to be normal. The rationale for doing this is that even though the Z-scores do not follow a Student-t distribution in this setting, the values of K obtained using the t-distribution will always be slightly larger than 1.96. Thus, the coverage will be slightly higher when using the t-distribution to calculate K compared to when using the normal distribution. Using a slightly larger value of K helps compensate for several possible factors that could lead to the Z-scores being slightly heavier-tailed than predicted by a normal distribution. As the sample size grows, the values of K obtained from the normal and t-distributions will become very similar. The distinction between using these two approaches is therefore mainly relevant when the sample size is smaller than around 50.

Alternative procedures for challenging situations

There are a few ways to reduce the risk that strong non-normality will lead to confidence intervals with poor performance. In order to provide some exposure to the types of procedures that statisticians use to conduct inference in challenging situations, we discuss two of these approaches next.

When working with the sample proportions, it is common to add two extra “successes” and two extra “failures” to the data before calculating the proportion. Thus, if we observe 5 successes and 7 failures, instead of estimating the success rate as 5 / (5 + 7), we estimate it as 7 / (7 + 9). The standard error is also estimated using this adjustment. The resulting confidence interval generally has better coverage properties than the usual CI when the sample size is small. This interval is often called the “Agresti-Coull” interval, after its inventors.

When working with strongly skewed data, another practical technique for improving the coverage properties of intervals is to transform the data with a skew-reducing transformation, e.g. a log transformation, then calculate the interval in the usual way (as described above) using the transformed data. The resulting interval can be transformed back to the original scale by applying the inverse transformation to the LCB and UCB. For example, if the transformation is the natural logarithm, the inverse transformation would be to exponentiate (anti-log) the LCB and UCB.

Conclusion

In summary, although normality of the data can play a role in determining the coverage properties of a confidence interval, it is generally not a major factor unless the sample size is quite small (much smaller than 50), or if the data are strongly non-normal. In most cases, other factors besides Gaussianity of the individual data values are more likely to give rise to sub-optimal coverage. Two such factors that can cause major problems with CI coverage probabilities are clustering or other forms of dependence in the data, and overt or hidden pre-testing or multiplicity in the analysis. Clustering will be discussed extensively in Course 3. We will discuss multiplicity in Week 3 of this course.