FIT2086 Lecture 4 Notes, Part II
Confidence Intervals

Dr. Daniel F. Schmidt*

September 7, 2020

1 Introduction

Imagine that we wish to fit a parametric distribution $p(\mathbf{y}|\theta)$ to some data. In Lecture 3 we learned about the method of maximum likelihood which gave us a procedure for finding a single good guess $\hat{\theta}$ of the model parameters given the data. This process of guessing a single value for the parameters is called point estimation, as we estimate a single "point" in the parameter space. However, we know that our best guess will never be exactly equal to the population parameter – even for very large sample sizes, just due to randomness in the sampling procedure, and in our data. Therefore, in almost all settings it is of crucial importance to be able to quantify how certain/uncertain we are about our single best guess. We can do this by specifying a range of *plausible values* for our population parameters. This process is called interval estimation.

A point estimator takes some sample of data $\mathbf{y}$ and returns a single best guess of the parameter(s); i.e., $\hat{\theta}(\mathbf{y}) \equiv \hat{\theta}$. An interval estimator takes a sample of data and returns an *interval*, say T, of values in the parameter space, for example:

$$T(\mathbf{y}) = \left(\hat{\theta}^{-}(\mathbf{y}),\ \hat{\theta}^{+}(\mathbf{y})\right) \subset \mathbb{R}$$

which says that given our sample $\mathbf{y}$, a plausible range of values for the population parameter θ we are trying to estimate is anywhere between $\hat{\theta}^{-}(\mathbf{y})$ and $\hat{\theta}^{+}(\mathbf{y})$. The size of this interval can be thought of as implicitly quantifying how *uncertain* we are about the single best guess that we have made using the point estimation procedure. The narrower the interval, the smaller the range of plausible values and the more certain we are about our best guess; conversely, if the interval is very wide, the range of plausible values is much greater and we are less certain about our single best guess.

The obvious question is: how should we go about choosing such an interval so that it captures this uncertainty in an objective fashion using just the data? Unsurprisingly, it turns out that there are a number of methods in the statistical literature for doing this. The one we will examine is perhaps the most commonly used technique in industry/research, and is called the method of confidence intervals.

2 Confidence intervals

As a concept, the idea of generating an interval to capture the uncertainty or measure the accuracy of an estimate is highly intuitive. Despite this, confidence intervals do have a reputation for being difficult

to understand at a technical level. The concepts are closely tied to the idea of repeated sampling that we examined in Lecture 3, and build directly on the idea of the *sampling distribution*, $p(\hat{\theta})$, of a point estimate $\hat{\theta}$.

Let us begin by "ripping the bandaid" off, so to speak, and diving straight into the technical definition of a confidence interval. Imagine that we have been given some procedure/algorithm, $T(\mathbf{y})$, that takes a sample of data $\mathbf{y}$ and returns an interval of the parameter space Θ.

Definition 1. We say that the interval $T(\mathbf{y})$ is an $100(1-\alpha)\%$ confidence interval for population parameter θ if and only if

$$\mathbb{P}(\theta \in T(\mathbf{y})) = 1 - \alpha,$$

where $\alpha \in (0, 1)$ and the probability is with respect to the population distribution $p(\mathbf{y} \mid \theta)$ over all possible data samples of size n.

In practice it is very common to consider $\alpha = 0.05$, i.e., 95% confidence intervals. Let us now consider this definition in words. Imagine that we drew a very large (essentially infinite) number of samples of size n from our population, and for each of these samples we asked our procedure $T(\mathbf{y})$ to generate a confidence interval for the unknown population parameter θ. If for 95% of the samples the associated interval generated by our procedure $T(\mathbf{y})$ contained, or "covered", the unknown population parameter, then the procedure $T(\mathbf{y})$ is said to generate a 95% confidence interval. Figure 1 illustrates this idea. For each of the possible samples we could draw from the population, we can calculate an interval using our procedure; if, for 95% of these samples, the interval calculated by our procedure includes the true population parameter θ, we can say that this procedure generates a 95% confidence interval. More generally, as above, we can talk about $100(1 - \alpha)\%$ confidence intervals. For $\alpha = 0.05$ we have a 95% confidence interval, which is the most common interval used in statistical analysis. However, one should note that this choice is primarily driven by convention rather than any distinguishing property of the choice $\alpha = 0.05$.

The idea of "confidence" can initially be confusing. The important thing to recognise is what a confidence interval procedure does is give you a special guarantee under *repeated sampling* from the population. For example, for $\alpha = 0.05$, we can say that *before* seeing a sample $\mathbf{y}$ drawn from our population, we know that there is a 95% chance the random sample we draw will lead to a 95% confidence interval that contains the true value of the population parameter. We accept that 5% of the samples we might draw will result in a confidence interval that does not contain the population parameter.

What a confidence interval procedure does not do is give you is a guarantee for the *particular sample* you have observed. To understand this, remember that the population parameter θ, while unknown, is *not* a random variable. It is fixed at some particular value for our particular population. Therefore, *after* observing a sample $\mathbf{y}$ from our population, the 95% confidence interval we generate will either contain the true value, or it won't. There is nothing random about this after we have observed our sample. All we know is that for 95% of possible samples that we could observe, our procedure will give us an interval that covers the true parameter value – but we have no idea whether the sample we *have observed* is one of those samples that does produce a CI that covers θ.

2.1 Confidence Intervals for the Mean of a Normal Distribution

Let us begin our examination of confidence intervals with a simple and important problem: deriving a confidence interval for the maximum likelihood estimate of the mean of a normal distribution. We start with this problem for two reasons. The first is that it is simple enough to allow us to cleanly observe the technical steps required to derive a confidence interval. The second reason is that this is a problem of great importance; estimation of means (averages) form the basis of an incredible number

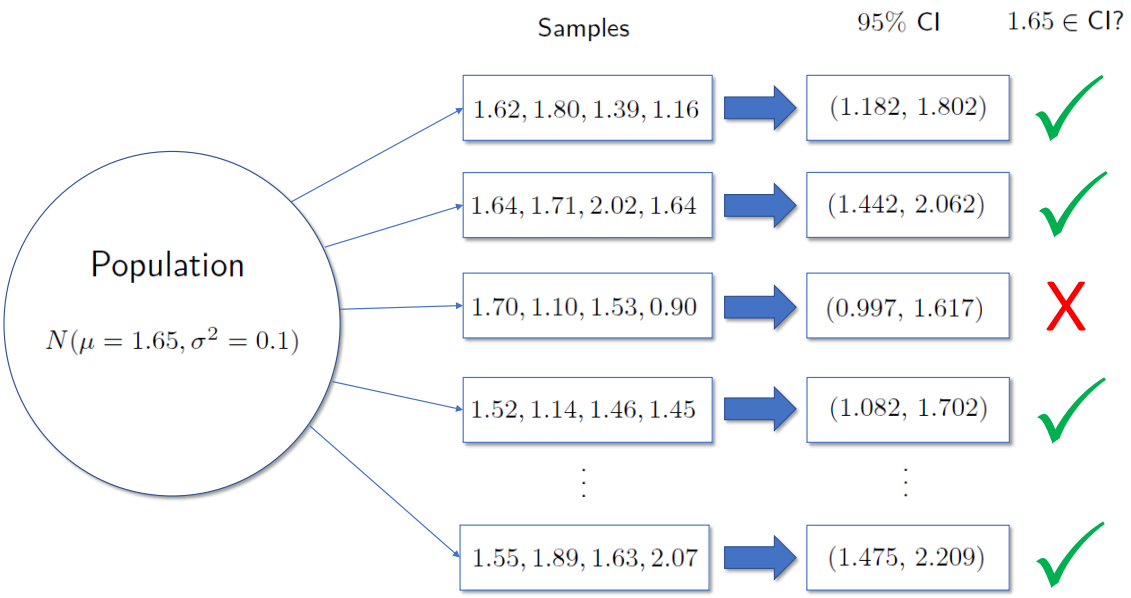

Figure 1: Cartoon showing multiple samples drawn from a $N(\mu = 1.65, \sigma^2 = 0.1)$ population, along with the 95% confidence intervals for each sample. 5% of possible samples will result in CIs that do not include $\mu = 1.65$.

of analyses performed around the world each day. Think about the number of times you have seen newspapers, media or websites quoting "the average rainfall/speed/amount of ..."; and now think how many times these numbers are accepted at face value despite the fact that there are no measures of accuracy attached to them. Your readiness to accept a particular statistic may drastically shift if you learn that the uncertainty is greater than the number quoted, and this is why confidence intervals are such a crucial element of good data science practice. We will now learn how to use the information contained in the sampling distribution of the sample mean to derive a confidence interval with which we can formally quantify the uncertainty in our estimate.

2.2 CI for Normal Mean with Known Variance

To begin with we assume that the population is normally distributed with unknown mean μ and known variance σ^2. As per Lecture 3, we know that the maximum likelihood estimator for the mean is

$$\hat{\mu} \equiv \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

which is equivalent to the sample mean. Further, under the assumptions about the population that we have made, we know from Lecture 3 that the sampling distribution for our estimate $\hat{\mu}$ is

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Remember, the above statement means that if we repeatedly drew samples of size n from our population and calculated the estimate $\hat{\mu}$ for each of these samples, the different values of the estimates we obtained

would be normally distributed with a mean of μ and a variance of σ^2/n. Using the sampling distribution we can construct a 95% confidence interval. The key step is to note that the z-score (see Lecture 2) for $\hat{\mu}$

$$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}$$

follows an $N(0,1)$ distribution, by definition. The quantity $(\sigma/\sqrt{n})$ in the denominator is often called the standard error, and it measures the variability of the estimator under repeated sampling. Using this information we can write the following probability statement

$$\mathbb{P}\left(-1.96 < \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

by exploiting the properties of normal distributions (i.e., that 95% of samples from an $N(0,1)$ distribution fall within the interval -1.96 to 1.96) (see Lecture 2). By recalling that the normal distribution is symmetric around zero, we can multiply all terms of the equation inside the $\mathbb{P}(\cdot)$ by $-\sigma/\sqrt{n}$ and obtain

$$\mathbb{P}\left(-1.96\frac{\sigma}{\sqrt{n}} < \mu - \hat{\mu} < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

The final step is to add $\hat{\mu}$ to all sides of the equation inside the $\mathbb{P}(\cdot)$, which yields

$$\mathbb{P}\left(\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

We can interpret the above equation as saying that the probability that the estimate $\hat{\mu}$ from a random sample drawn from the population will be no further away from the true, population mean μ than 1.96 standard deviations is 0.95; or, conversely, we can think of it as saying that for 95% of all possible samples we could see from our population, the true unknown population mean μ will lie within $1.96\,\sigma/\sqrt{n}$ of the sample mean. This information then lets us build our 95% confidence interval as

$$\text{CI}_{95}(\hat{\mu}) = \left(\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}},\ \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right) \tag{1}$$

or more generally, a $100(1-\alpha)\%$ confidence interval is given by

$$\left(\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{2}$$

where $z_{\alpha/2}$ is the value that satisfies

$$\mathbb{P}\left(-z_{\alpha/2} < \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = (1-\alpha),$$

and is given by the $100(1-\alpha/2)$-th percentile of the standard normal distribution, i.e., the value $z_{\alpha/2}$ that solves

$$\mathbb{P}(Z < z_{\alpha/2}) = 1 - \alpha/2.$$

where $Z \sim N(0,1)$. Looking at the confidence interval for the normal mean with known variance (2) we can observe that the width of the interval depends crucially on the population variance σ^2, the sample size n and the desired level of coverage α. In particular, the width of the interval:

- increases with increasing population standard deviation σ. This is because the greater the population variance, the more variability there is within the population, the harder it is to nail down the exact value of the population mean and the less accurate an estimate will be for a given sample size.

- decreases proportionally to the square-root of the sample size. That is, the more data we collect the more accurate our estimates becomes and the shorter the confidence interval will be. Note that the inverse-square-root relation implies that to reduce the width of the confidence interval by a factor of c we need to collect c^2 times as many samples.
- increases with increasing confidence level $(1 - \alpha)$. That is, the greater the desired level of confidence, the wider the interval must be to provide the appropriate guarantee. In the limit, a 100% confidence interval will cover the entire real line, and in contrast, a 0% confidence interval will cover the single point $\hat{\mu}$ (i.e., it will reduce to the point estimate).

Again, it is important to stress that a $100(1 - \alpha)\%$ confidence interval only gives you a probability guarantee, regarding the likelihood of the resulting interval containing the true unknown population parameter (in this case, μ) *before you see the data*. Once you have observed a sample, the confidence interval gives you no further guarantee, in the sense that the resulting interval either contains μ, or it does not, and we of course have no way of verifying whether it does or does not contain the true value.

Example 1: CI for Normal Mean with Known Variance

To demonstrate the above procedure, consider the follow sample of body mass indices (BMI) measured on people with diabetes drawn from a study of the Pima ethnic group in the United States:

$$\mathbf{y} = (53.2, 33.6, 36.6, 42.0, 33.3, 37.8, 31.2, 43.4)\, kg/m^2.$$

Imagine that we are told that the population variance is 43.75, a figure that has been estimated from another very large study of Pima people. Let us construct a 95% confidence interval for the population mean based on the above sample. First, we calculate the sample mean, which is $\hat{\mu} = 38.88 kg/m^2$. Using (1) (i.e., equation (2) with $\alpha = 0.05$) yields the interval

$$\left(38.88 - 1.96\sqrt{43.75/8},\ 38.88 + 1.96\sqrt{43.75/8}\right)$$

which is equal to

$$(34.3,\ 43.47)\, kg/m^2.$$

In words, we might summarise the results of our analysis by a paragraph such as:

> *"The estimated mean BMI of people from the Pima ethnic group with diabetes (sample size $n = 8$) is $38.88\, kg/m^2$. We are 95% confident the population mean BMI for this group is between $34.3\, kg/m^2$ and $43.75\, kg/m^2$."*

Note the structure of the above summary: (i) first, we state the observed quantity (in this case, the estimated mean) and clarify explicitly the details of our population (BMI, ethnic Pima people with diabetes, sample size). Note that we always include units of measurement; (ii) second, we state the confidence interval with the statement "we are 95% confident that ...". Again, note the use of units. Summarising the results of statistical analyses using these types of statement makes it very clear exactly what our analysis is showing, and is extremely important in ensuring that the findings are communicated clearly and in context. □

2.3 CI for Normal Mean with Unknown Variance

The assumption that the population variance σ^2 is known is in general unrealistic. However, it turns out that even if we do not assume that the variance is known it is still possible to construct a 95%

confidence interval in a similar manner to the case when σ^2 is known. The key difference, of course, is that we now need to estimate σ^2 from our data sample. An obvious approach to this problem would be to estimate σ^2 using the unbiased estimate of variance

$$\hat{\sigma}^2 = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(y_i - \bar{y})^2$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ is the sample mean. Once we have this estimate it is tempting to simply plug $\hat{\sigma}$ into (2) in place of the unknown population standard deviation σ and compute our confidence interval as before. Unfortunately, while tempting, this simple approach does not quite work: plugging our estimate $\hat{\sigma}$ into σ will *not* lead to an exact 95% confidence interval. For smaller sample sizes the resulting confidence interval *will not* cover the true parameter value for 95% of possible samples (i.e., it will not give 95% coverage). The reason that this approach does not work is that the statistic

$$\frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}} \tag{3}$$

is no longer normally distributed if we use the estimate $\hat{\sigma}$ in place of the (unknown) population σ. This is because the variance is being estimated from the data, and we need to take into account the fact that this quantity is an estimate, with its own variability, and not an exact value. Instead, it was shown in the early 1900s that the quantity (3) follows what is known as a Student-t distribution with $n-1$ degrees-of-freedom. The Student-t distribution on the surface looks similar to the norml distribution, but has an important difference. Figure 2 shows two different t distributions along with the standard normal distribution. It is clear that both the t-distributions and the normal distribution share a number of similar features: they are symmetric and unimodal (one peak), and both tail off to zero in a monotonic (strictly decreasing) fashion either side of the peak. However, they differ in the way in which they *spread* their probability over the number line. The Student-t distribution spreads more of its probability over larger values than the normal distribution does. The Student-t distribution is an example of a "heavy-tailed" distribution, as its "tails" go towards zero at a slower rate than the normal. The rate at which the tails decrease to zero is determined by the degrees-of-freedom parameter. The smaller the degrees-of-freedom, the heavier the tails are. For very large degrees-of-freedom, the standard normal distribution and t distribution are virtually the same.

An important property of the Student-t distribution is that it is symmetric and self-similar in the same fashion as the normal distribution. This is important because it lets us utilise exactly the same steps as we used in deriving our confidence interval for the mean with known variance to arrive at the $100(1-\alpha)\%$ confidence interval for μ in the case that σ^2 is unknown:

$$\left(\hat{\mu} - t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}},\ \hat{\mu} + t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}}\right), \tag{4}$$

The difference is that $t_{\alpha/2,n-1}$ is now the $100(1-\alpha/2)$-th percentile of the standard Student-t distribution with $n-1$ degrees-of-freedom. Most statistical packages will have a command or function to compute this quantity; in R we can find it using the command $\texttt{qt}(\texttt{p} = 1 - \texttt{alpha}/2, \texttt{n} - 1)$. The interval (4) will achieve exactly $100(1-\alpha)\%$ coverage if the population from which our sample is drawn follows a normal distribution. Examination of the intervals (4) and (2) reveal close similarities: both construct the interval as a multiple of the standard error and both are centred on the sample mean $\bar{y}$. They differ in two aspects: (i) the fact that (4) uses the estimate $\hat{\sigma}$ in place of the population parameter σ to determine the standard error, and (ii) in the different way in which the multipliers are calculated. To compare with the known variance case, we can compute the values of $t_{\alpha/2,n-1}$ when $\alpha = 0.05$ (i.e., 95% CI) for several different sample sizes:

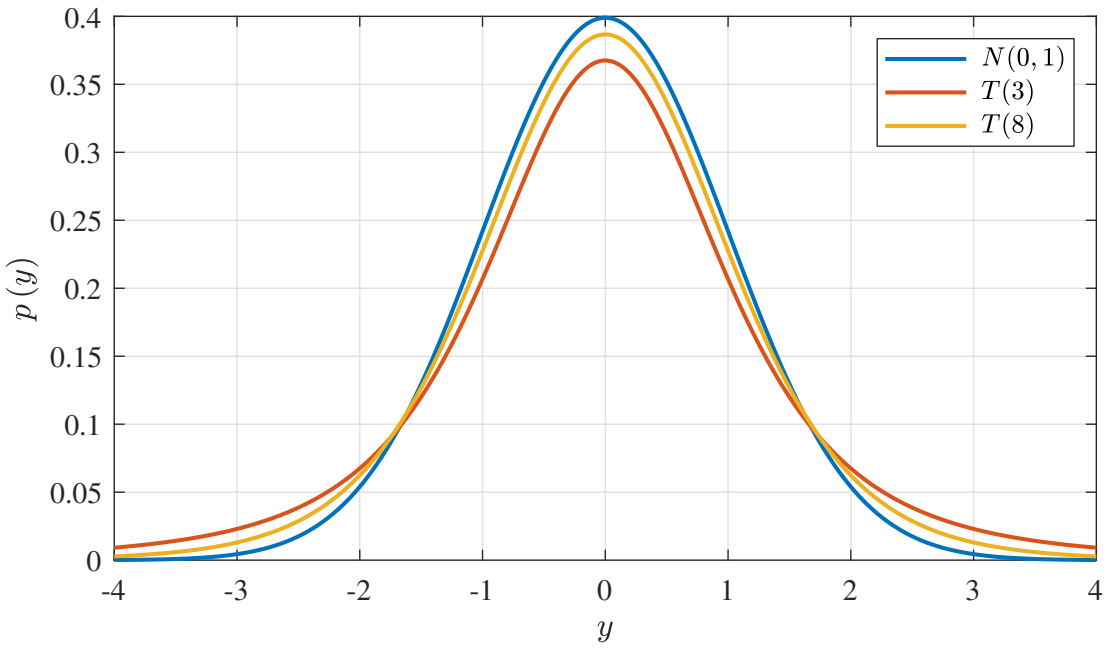

Figure 2: Plot of a standard normal $N(0,1)$ distribution and two Student-t distributions, one with degrees-of-freedom (DOF) of 3, and one with DOF of 8. Note how the t-distributions spread the probability out more and tail off to zero slower than the normal distribution.

- for $n = 3$, $t_{0.025,2} \approx 4.3$;
- for $n = 6$, $t_{0.025,2} \approx 2.57$;
- for $n = 11$, $t_{0.025,10} \approx 2.22$.

In comparison to $z_{0.025} = 1.96$ we see that the multipliers in the case that the variance is unknown are always larger than in the case that the variance is known. This is because $\hat{\sigma}$ is an estimate of σ and carries with it its own inherent variability. Taking this variability into account leads us to a larger multiplier than if the quantity σ was exactly known. However, if we take the sample size to be very large we will find that $t_{\alpha/2,n-1} \to z_{\alpha/2}$ as $n \to \infty$. Therefore, for very large sample sizes using the estimated variance makes very little difference to our confidence interval, but can make a large different for smaller sample sizes. This is because for large sample sizes we expect our estimate of variance, $\hat{\sigma}^2$, to be quite close to the true, unknown population variance σ^2.

Example 2: CI for Normal Mean with Unknown Variance

Let us revisit our Pima BMI data

$$\mathbf{y} = (53.2, 33.6, 36.6, 42.0, 33.3, 37.8, 31.2, 43.4),$$

and this time assume that we do not have access to a good value for the population variance as Example 1. As above, our estimate for the mean is $\hat{\mu} = 38.88$; this time, however, we need to estimate the variance from the data. Using the unbiased estimate of variance yields

$$\hat{\sigma}^2 = \frac{1}{7}\sum_{i=1}^{n}(y_i - 38.88)^2 \approx 51.37.$$

To use the interval (4) we also need to determine $t_{\alpha/2,n-1}$. Let us construct a 95% confidence interval by taking $\alpha = 0.05$. Our sample size is $n = 8$, so using R we can find our multiplier as $t_{0.025,7} = \texttt{qt}(\texttt{p} = 1 - 0.05/2, \texttt{n} = 7) \approx 2.36$. Using this in (4) we can find our interval to be

$$\left(38.88 - 2.36\sqrt{51.37/8},\ 38.88 + 2.36\sqrt{51.37/8}\right)$$

which is equal to $(32.9, 44.86)\, kg/m^2$. Comparing this to the "known variance" CI we calculated previously, $(34.4, 43.47)$ we see the CI assuming the variance is unknown is wider. It is natural to ask, given the fact that $t_{\alpha/2,n-1} > z_{\alpha/2}$ for any finite sample size n, will our interval using (4) *always* be wider? The answer is: not always. While the multipliers may be larger there will be some proportion of samples of data from our $N(\mu, \sigma^2)$ population which lead to an estimate $\hat{\sigma}^2$ that is sufficiently smaller than the true population σ^2 to result in a shorter confidence interval. However, what is guaranteed by using (4) is that 95% of samples will result in CIs that cover the true population μ, irrespective of the value of the true (unknown) population variance. □

3 Confidence Intervals for Difference of Normal Means

One of the most important fundamental problems in data science is the estimation of the difference in population means between two distinct populations. This problem appears repeatedly across scientific disciplines. As an example, imagine we have a cohort of people in a medical trial for a weight-loss drug. At the start of the trial, all the weights of all participants are measured and recorded. Call this sample $\mathbf{y}_A$, and assume it has an unknown population mean of μ_A. The participants are then administered a weight-loss drug for 6 months, and at the end of the trial period, we re-measure the participant's weights; call this sample $\mathbf{y}_B$, with population mean μ_B. To see if the drug had any real effect on weight-loss we can try to estimate the population mean difference in the weights pre- and post-trial, i.e., $\mu_A - \mu_B$. If there is no difference at a *population* level, $\mu_A = \mu_B \Rightarrow \mu_A - \mu_B = 0$.

To estimate this difference, we first need to estimate the mean for both samples, say $\hat{\mu}_A = \bar{y}_A$ and $\hat{\mu}_B = \bar{y}_B$. The estimated difference is then $\hat{\mu}_A - \hat{\mu}_B$. Of course, even if there is no difference at the population level (i.e., the drug had no effect), the observed difference will never be exactly zero, due to random chance and variability in our sampling. Therefore, it is of crucial importance to quantify how accurate our estimate of the difference is by calculating an appropriate confidence interval. One can no doubt quite easily imagine a huge number of variations of problem settings in which the estimation of differences in means is relevant (e.g., performance of different fuel types, differences in the quality of goods under different manufacturing procedures, and so on).

3.1 CI for Difference of Normal Means with Known Variances

To construct a confidence interval for the difference in means between two samples we will first assume that both samples come from normal populations with *unknown* means μ_A and μ_B, and *known* variances σ_A^2 and σ_B^2, respectively. We do not require that $\sigma_A^2 = \sigma_B^2$; rather, we simply require that they both be known. Then, if we estimate both population means by their respective sample means, we know that

$$\hat{\mu}_A \sim N\left(\mu_A, \frac{\sigma_A^2}{n_A}\right) \quad \text{and} \quad \hat{\mu}_B \sim N\left(\mu_B, \frac{\sigma_B^2}{n_B}\right),$$

where n_A and n_B are the sizes of the two samples, respectively. To put this in words, we know that our sample means both follow a normal distribution, in the sense that if we repeatedly draw new samples from the two populations and computed the corresponding sample means, they would follow the normal distributions shown above. Using the property of additivity of normally distributed random

variables (see Lecture 2) it is therefore easy to show that that the estimated difference $\hat{\mu}_A - \hat{\mu}_B$ also follows a normal distribution; in particular

$$\hat{\mu}_A - \hat{\mu}_B \sim N\left(\mu_A - \mu_B,\ \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right).$$

We can then use properties of the normal distribution to find the statistic

$$\frac{(\hat{\mu}_A - \hat{\mu}_B) - (\mu_A - \mu_B)}{\sqrt{\dfrac{\sigma_A^2}{n_A} + \dfrac{\sigma_B^2}{n_B}}} \sim N(0,1).$$

This quantity is a "z-score" for the difference, i.e., it is a standardised measure of difference between the sample means of the two samples, with the standardisation determined by the size of the samples and the variability of the two populations (characterised by σ_A and σ_B, respectively) from which they were drawn. Now that we have a statistic that we know follows an $N(0,1)$ distribution, we can use exactly the same procedure as in Section 2.2 (for deriving the CI for the mean with known variance) to arrive at the interval

$$\left(\hat{\mu}_A - \hat{\mu}_B - z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}},\ \ \hat{\mu}_A - \hat{\mu}_B + z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}\right),$$

which is a $100(1-\alpha)\%$ confidence interval for $\hat{\mu}_A - \hat{\mu}_B$ when the two samples we are comparing follow normal distributions with known variances. In general, when summarising confidence intervals of differences, it is important to consider the follow three scenarios:

1. Is the interval entirely negative? If so, it is suggestive of a negative difference at population level, with the suggestion being stronger the higher the confidence (i.e., the closer α is to one).
2. Is the interval entirely positive? If so, it is suggestive of a positive difference at population level, with the suggestion being stronger the higher the confidence (i.e., the closer α is to one).
3. Does the interval contain zero (i.e., is the lower end of CI negative and the upper end of CI positive)? If this is the case, it means that we cannot rule out the possibility that there is actually no difference at the population level.

The difference in two means is one of the basic tools in data analysis, providing a way to answer a very wide range of scientific questions, and its applications extend well beyond the relatively straightforward setting we have examined here.

3.2 CI for Difference of Normal Means with Unknown Variances

Of course, assuming σ_A^2 and σ_B^2 are known is not realistic; if we make the assumption that they are unknown but both the same, then we can derive an exact confidence interval for the difference in means using the t-distribution. However, even this assumption is not particularly realistic. In general, we cannot assume either that σ_A^2 or σ_B^2 are known, or that both are the same. Deriving an exact confidence interval if $\sigma_A^2 \neq \sigma_B^2$, and both are unknown, is actually quite difficult. We now briefly discuss an *approximate* procedure. Essentially, we use the unbiased estimates $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$ of the population variances in place of the known population variances to derive the approximate $100(1-\alpha)\%$ confidence interval

$$\left(\hat{\mu}_A - \hat{\mu}_B - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}},\ \ \hat{\mu}_A - \hat{\mu}_B + z_{\alpha}\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}\right). \tag{5}$$

We acknowledge that as in the case of the CI for mean with unknown variance, simply replacing the variance by its estimate and treat the resulting statistic as a z-score will not lead to exact $100(1-\alpha)\%$ coverage. However, we do note that the approximation gets better in the sense that the coverage achieved by (5) gets closer to $100(1-\alpha)\%$ as the sample sizes n_A and n_B get larger. For moderate sample sizes (roughly $n_A, n_B > 50$) this interval is quite accurate, though it is important to acknowledge the fact that it is only approximate and does not give exact coverage for any finite sample size.

Example 3: CI for Difference of Normal Means

As an example, let us return our example involving diabetic Pima people. Imagine now that we have also obtained a sample of non-diabetic Pima people; the two samples of body-mass index (BMI) are:

$$\begin{aligned} \mathbf{y}_N &= (34.0, 28.9, 29.0, 45.4, 53.2, 29.0, 36.5, 32.9) \\ \mathbf{y}_D &= (53.2, 33.6, 36.6, 42.0, 33.3, 37.8, 31.2, 43.4) \end{aligned}$$

where $\mathbf{y}_N$ denotes non-diabetics and $\mathbf{y}_D$ denotes diabetics. The estimates of the population means, as well as the unbiased estimates of population variances, for these two groups are:

$$\begin{aligned} \hat{\mu}_N &= 36.11, \qquad & \hat{\sigma}^2_N &= 78.05 \\ \hat{\mu}_D &= 38.88, \qquad & \hat{\sigma}^2_D &= 51.37, \end{aligned}$$

and the observed difference is in BMI between the two groups is

$$\hat{\mu}_N - \hat{\mu}_D = 36.1 - 38.8 = -2.77\, kg/m^2.$$

Using this estimates in (5) yields an approximate 95% confidence interval

$$\left(-2.77 - 1.96\sqrt{\frac{78.05}{8} + \frac{51.37}{8}},\ -2.77 + 1.96\sqrt{\frac{78.05}{8} + \frac{51.37}{8}}, \right)$$

which is $(-10.65,\ 5.11)$. We could summarise these results using a statement such as:

> *"The estimated difference in mean BMI between people from the Pima ethnic group without (samples size $n = 8$) and with diabetes (sample size $n = 8$) is $-2.77\, kg/m^2$. We are 95% confident the population mean difference in BMI is between $-10.65\, kg/m^2$ (BMI is lower in people without diabetes) up to $5.11\, kg/m^2$ (BMI is greater in people without diabetes). As the interval includes zero, we cannot rule out the possibility of there being no difference at a population level between Pima people with and without diabetes."*

Again, note the structure of the above summary. The first part states exactly what was observed, and for what variable (in this case BMI), in what units (in this case kg/m^2), from what population (Pima ethnic people with and without diabetes) and what sample sizes ($n = 8$ for both samples). The second part summarises the confidence interval; it states what the lower and upper ends of the interval are, and how they could be interpreted (BMI lower/higher in people without diabetes, as appropriate). Finally, we note that the interval for the difference includes the number 0 (no difference at population level), and state that this suggests we cannot rule out the possibility there is no difference at the population level. □

4 Approximate CIs for sample means

This section gives a small insight into the power of the central limit theorem. We have seen that quite a few estimators for different model parameters (Poisson rate parameter, Bernoulli success probability parameter) are equivalent to the sample mean. As the population is not normally distributed for these models, the sampling distribution of the estimators is therefore not exactly normal distributed. However, from the central limit theorem we know that for large sample sizes n all sample means from populations with finite means and variances are *approximately* normally distributed. We can use this fact to obtain approximate CIs in this case.

Let $Y_1, \ldots, Y_n$ be RVs from our population, and let us assume our parameter θ of interest can be estimated using the sample mean $\bar{Y}$. Further, assume that $\mathbb{E}[Y_i] = \theta$, and that $\mathbb{V}[Y_i] = v(\theta)$; that is, we assume that the mean of any observation from our population is θ and the variance of any observation from our population is some known function of θ. If our estimate $\hat{\theta}$ is equivalent to the sample mean, then from the CLT we know that

$$\hat{\theta} \xrightarrow{d} N\left(\theta, \frac{v(\theta)}{n}\right)$$

as the sample size $n \to \infty$. This implies that the statistic

$$\frac{\hat{\theta} - \theta}{\sqrt{v(\theta)/n}} \xrightarrow{d} N(0, 1),$$

which can be used to derive an approximate confidence interval using the basic procedure outlined previously (see Section 2.2). The problem with this approach is that we do not know the population value of θ (otherwise we would not be estimating it!), and therefore we do not know the value of the population variance $v(\theta)$. To circumvent this problem we can use our estimate $\hat{\theta}$ in $v(\cdot)$, i.e., $v(\hat{\theta})$, to estimate the population variance, and therefore obtain an *approximate* $100(1-\alpha)\%$ confidence interval

$$\left(\hat{\theta} - z_{\alpha/2}\sqrt{v(\hat{\theta})/n},\ \hat{\theta} + z_{\alpha/2}\sqrt{v(\hat{\theta})/n}\right). \tag{6}$$

The quantity $\sqrt{v(\hat{\theta})/n}$ can be viewed as an approximate standard error of the estimate (i.e., a measure of how much we might expect it to change if we drew a new sample from the same population and re-estimated $\hat{\theta}$ based on this new sample). The accuracy and coverage achieved by the above approximation improves as the sample size n increases, though for a number of problems quite good results can be obtained even for small sample sizes. We now examine the utility of this approximation through several example applications.

4.1 Approximate CI for the Poisson Distribution

To demonstrate how useful this result is, let us first consider the problem of constructing an approximate confidence interval for the Poisson rate parameter λ. In this case,

$$Y_1, \ldots, Y_n \sim \text{Poi}(\lambda),$$

and therefore $\mathbb{E}[Y_i] = \lambda$ and $\mathbb{V}[Y_i] = \lambda$, so that $v(\lambda) = \lambda$ (see Lecture 2). The maximum likelihood estimate of λ is (see Lecture 3)

$$\hat{\lambda}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

which is equivalent to the sample mean. Therefore, we can apply formula (6) to construct an approximate 95% CI for λ as

$$\left(\hat{\lambda}_{\text{ML}} - 1.96\sqrt{\hat{\lambda}_{\text{ML}}/n},\ \hat{\lambda}_{\text{ML}} + 1.96\sqrt{\hat{\lambda}_{\text{ML}}/n}\right). \tag{7}$$

This confidence interval shows that as the estimate of the rate parameter $\hat{\lambda}_{\text{ML}}$ increases, the confidence interval grows in width; this is because $v(\lambda) = \lambda$, so that the variability in the population is larger for larger values of the rate parameter λ. A natural question to ask is: how accurate is the interval produced by the approximate formula (7)? How close to 95% coverage does it get? The last question of Studio 4 examines this, and the solutions show that for $\lambda \geq 5$ and n as small as $n = 10$ the approximation is basically exact.

4.2 Approximate CI For the Bernoulli Distribution

As a further example of the usefulness of the procedure detailed in Section 4 we consider the problem of constructing an approximate CI for the probability of the success parameter θ for a Bernoulli distribution. In this case

$$Y_1, \ldots, Y_n \sim \text{Be}(\theta)$$

and therefore $\mathbb{E}[Y_i] = \theta$ and $\mathbb{V}[Y_i] = \theta(1-\theta)$, so that $v(\theta) = \theta(1-\theta)$ (see Lecture 2). The maximum likelihood estimate of θ can be shown to be $\hat{\theta}_{\text{ML}} = \bar{Y}$, i.e., the sample mean. Therefore, we can apply formula (6) to construct the following approximate 95% CI for θ:

$$\left(\hat{\theta}_{\text{ML}} - 1.96\sqrt{\frac{\hat{\theta}_{\text{ML}}(1-\hat{\theta}_{\text{ML}})}{n}},\ \hat{\theta}_{\text{ML}} + 1.96\sqrt{\frac{\hat{\theta}_{\text{ML}}(1-\hat{\theta}_{\text{ML}})}{n}}\right).$$

The above formula shows that as θ gets close to zero, or one, the interval becomes smaller (as the variance in the observations becomes smaller). The interval is largest when $\theta = 1/2$ as the variability in binary data is greatest in this case. The above interval works reasonably well for moderate n and θ not too close to zero or one. If n is small and $\hat{\theta}$ is close to zero (one), then it can occur that the lower (upper) end of the interval is less than zero (greater than one), and therefore that the interval will cover values of θ that are outside the legal parameter space. This demonstrates both the usefulness of the formula (6), but also serves to demonstrate potential weaknesses of the approximation when the parameter space is constrained. There are ways to improve the above approximation by appropriate transformation of the parameter space, but these are beyond the scope of this unit.