# FIT2086 Lecture 4 Summary
# Central Limit Theorem & Confidence Intervals

Dr. Daniel F. Schmidt

August 23, 2019

## 1 Part I: The Central Limit Theorem

**Central Limit Theorem**. The Central Limit Theorem (CLT) is often called the most important theorem, or result, in statistics. Why? It is because the CLT tells us that many of the quantities we study as part of statistics and data science are asymptotically distributed, and importantly, tells us that they are approximately distributed as per a normal (Gaussian) distribution (recall that asymptotically means "for large values"). This means that we can often approximate the exact distribution of many quantities we study by a normal distribution, which is much easier to work with.

Let's start with a simple statement of the central limit theorem: let $Y_1, \ldots, Y_n$ be random variables (RVs) with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$. Then, for large $n$, the distribution of the sum $\sum_{i=1}^{n} Y_i$ is approximately normally distributed, with a mean of $n\mu$ and a variance of $n\sigma^2$. Note the weakness of our assumptions: we assumed only that the random variables had a mean of $\mu$ and a variance of $\sigma^2$ – as long as they satisfy those properties, then regardless of their distribution, their sum for large $n$ will be approximately normally distributed. This is obviously a powerful result. More formally, we can write that

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\mu, n\sigma^2) \tag{1}$$

as $n \to \infty$, where "a $\xrightarrow{d}$ b" means that the quantity $a$ converges in distribution to the quantity $b$ in the limit.

In words, the CLT says that sums of many RVs, each with a finite mean and variance, are approximately normally distributed, with the approximation getting better and better the greater the number of RVs being added together.

**Implication 1: natural phenomena are often normally distributed**. One interesting implication of the central limit theorem is that it helps explain why so many natural phenomena appear to be normally distributed. As an example, consider the heights of adults in a homogenous population; they have been shown to be well approximated by a normal distribution in many studies. The central limit theorem helps answer why this may be the case. As an example, a person's height is known to be determined by a sum of many different factors: there are genetic determinants, in which millions of genetic markers across the genome have been shown to be associated with height, and there are environmental factors, such as dietary choices and behavioural patterns. If we treat these factors for an individual as the realisations of many RVs, we see that a persons height is determined by a sum

of the effects of many RVs, and appealing to the CLT we see that this sum will be (approximately) normally distributed.

**Implication 2: normality of some distributions for large parameter values**. Another interesting implication of the CLT is that many common distributions become essentially equivalent to the normal distribution for large values of one (or more) of their parameters. This means that for these cases, the distributions can be approximated by the normal distribution. For example, recall the binomial distribution:

$$p(M = m \,|\, \theta) = \binom{n}{m} \theta^m (1-\theta)^{(n-m)}.$$

This models the number of successes, $M$, that occur in $n$ Bernoulli trials, with each trial having a probability of success of $\theta$. If $Y_1, \ldots, Y_n$ are the Bernoulli trials (binary RVs), then the number of successes can be written as the sum

$$M = \sum_{i=1}^{n} Y_i.$$

As $M$ is the sum of $n$ RVs, and as each $Y_i$ has mean $\mathbb{E}[Y_i] = \theta$ and $\mathbb{V}[Y_i] = \theta(1-\theta)$ (by properties of Bernoulli RVs, see Lecture 2), the CLT (1) tells us that for large values of $n$ the number of successes is approximately

$$M \sim N(n\theta, n\theta(1-\theta)).$$

This result is actually quite astonishing: it says that if you tossed a coin $n$ times and recorded the number of successes, then repeated this many times and produced a histogram of the number of successes, you would find that for large values of $n$ that this distribution is approximately normal. That is, adding together 0s and 1s in sufficient number produces a normally distributed random variable!

**The CLT and distribution of sample means**. In Lecture 3 we examined the sample mean, and derived a general result regarding the mean and variance of the sample mean under quite weak assumptions. Let $Y_1, \ldots, Y_n$ be RVs with mean $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$; then, the sample mean

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{2}$$

was shown to have

$$\mathbb{E}\left[\bar{Y}\right] = \mu, \;\; \mathbb{V}\left[\bar{Y}\right] = \frac{\sigma^2}{n}.$$

That is, the mean of the sample mean is equal to the mean of any one observation from our population, and the variance of our sample mean is equal to the variance of any one sample from our population, divided by the sample size $n$. These results are exact under our assumptions. These results were useful for two reasons: (i) many estimators are equivalent to the sample mean (for example, maximum likelihood estimate of normal $\mu$ parameter, or ML estimation of Poisson rate parameter $\lambda$), and (ii) they give us some idea of the behaviour and variability of the sample mean when data is assumed to come from different populations. But what about the distribution of $\bar{Y}$? This is much more difficult to derive and depends on the exact distribution that we assume for the population (remember, the population is the infinite large source of data from which we are draw our sample). However, as the sample mean is the sum of the RVs $Y_1, \ldots, Y_n$, we see that we can use the CLT to get an asymptotic (in the sample size $n$) distribution for our sample mean, once again under very weak assumptions.

Let us assume our RVs are independent and satisfy $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$; then from the CLT we know that

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\mu, n\sigma^2).$$

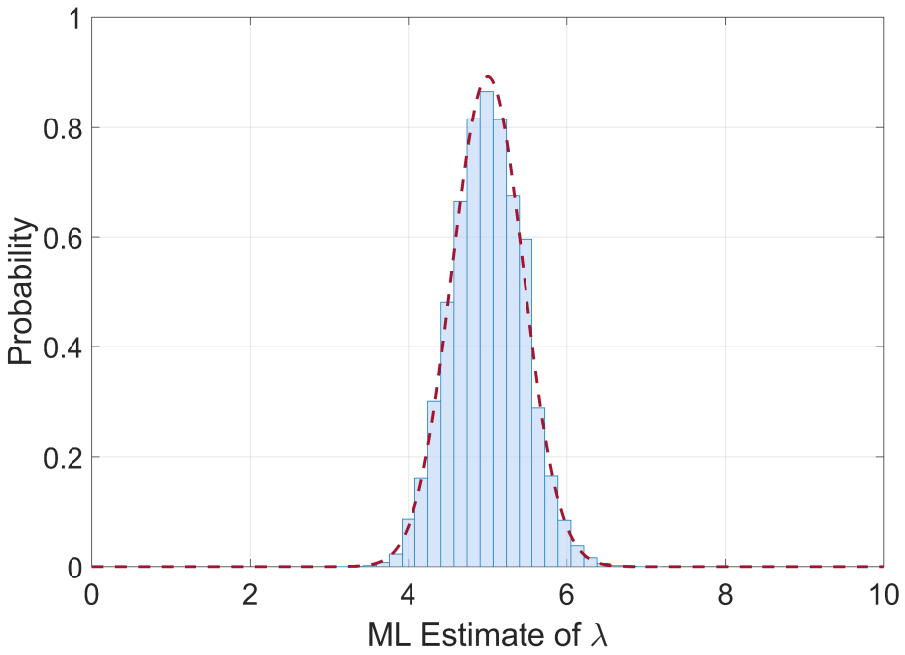


Figure 1: Histogram of $\hat{\lambda}_{\text{ML}}$ from $1,000,000$ data samples, each of size $n = 25$ and generated from a Poi(5) distribution. Also plotted is the normal $N(5, 0.2)$ approximation to the sampling distribution.

We can then use the fact that $\mathbb{V}[Y_i/n] = \mathbb{V}[Y_i]/n^2$ to conclude that

$$\bar{Y} \xrightarrow{d} N(\mu, \sigma^2/n)$$

as $n \to \infty$. So this says that under quite weak assumptions (finite mean and variance, independently distributed RVs) then for large sample sizes $n$ we can get an approximate sampling distribution of the sample mean $\bar{Y}$, and that this approximate sampling distribution is normal with mean equal to the mean of a single observation from the population (i.e., $\mathbb{E}[Y_i]$) and variance equal to the variance of a single observation from the population (i.e., $\mathbb{V}[Y_i]$) divided by the sample size $n$. This approximation improves in accuracy as $n$ gets larger and larger.

**Example 1: CLT and sample mean for normal populations**. If we assume our population is distributed as per a $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$, i.e., it is normally distributed, then using the CLT we find

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\mu, n\sigma^2)$$

so that the sample mean satisfies

$$\bar{Y} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$$

as $n \to \infty$. In fact, from Lecture 3, we know that if our population is normally distributed then the sample mean is *exactly* normally distributed for all sample sizes $n$, so in this case the CLT is not necessary.

**Example 2: CLT and sample mean for Poisson populations**. A second example in which the CLT is much more useful is when the population is distributed as a per a Poisson distribution with rate $\lambda$, i.e., $Y_1, \ldots, Y_n \sim \text{Poi}(\lambda)$. Recall that the maximum likelihood estimator for $\lambda$ is:

$$\hat{\lambda}(Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

which is equivalent to the sample mean (2). Under our assumed population we know that $\mathbb{E}[Y_i] = \lambda$ and $\mathbb{V}[Y_i] = \lambda$ (see Lecture 2), and from the CLT (1) we know that

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\lambda, n\lambda)$$

as $n \to \infty$; therefore,

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

as $n \to \infty$. This gives us an approximate distribution of the estimate of the Poisson rate parameter which gets better and better as $n$ gets larger. In fact, for $\lambda > 2$ the approximation is very good even for sample sizes as small as $n = 10$. Figure 1 shows the distribution of $\hat{\lambda}$ as calculated by simulation and plotted against the normal approximation. As an be seen for $n = 25$ it is virtually indistinguishable from the normal approximation.

**Example 3: CLT and other estimators**. The CLT can be applied to many other estimators to derive approximate sampling distributions. For example, recall the ML estimator of the variance parameter $\sigma^2$ of a normal distribution:

$$\hat{\sigma}^2(Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

This estimator is not exactly equivalent to the sample mean (2) so how do we apply the CLT? If we define the RVs

$$E_i = (Y_i - \bar{Y})^2$$

we can write the above estimator as

$$\hat{\sigma}^2(Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{i=1}^{n} E_i,$$

which is clearly a sample mean of these new RVs $E_1, \ldots, E_n$, and if they have a finite mean and variance (which they do, if we assume the population RVs $Y_i$ have finite mean and variance) we can appy the CLT, and $\hat{\sigma}^2$ will be approximately normally distributed for large $n$. In fact, this result holds for many estimators that on the surface don't appear to be equivalent to the sample mean, although detailed study of these estimators is beyond the scope of the subject.

## 2 Part II: Confidence Intervals

**Interval estimation**. Imagine we want 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 best guess $\hat{\theta}$ of the model parameters given the data. This process of finding a single, best value of 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. It is important 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 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 of values in the parameter space, for example:

$$\Big(\hat{\theta}^{-}(\mathbf{y}),\ \hat{\theta}^{+}(\mathbf{y})\Big)$$

which says that given our sample $\mathbf{y}$, a plausible range of values for the population parameter $\theta$ 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 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 an interval that captures this uncertainty in an objective fashion using just the data? The are a number of methods in the statistical literature for doing this: the one we will examine is one of the most commonly used techniques in industry/research, and is called the method of confidence intervals.

**Confidence intervals**. Imagine we have a procedure/algorithm that takes a sample $\mathbf{y}$ and spits out an interval $(\hat{\theta}^{-}_{0.05}(\mathbf{y}),\ \hat{\theta}^{+}_{0.05}(\mathbf{y}))$. If for 95% of all possible samples we could draw from our population the interval that is generated by this procedure contains ("covers") the population value of the parameter $\theta$, then we say that the procedure generates a 95% confidence interval (CI). We can then say that "we are 95% confident that the true population parameter $\theta$ lies between $\hat{\theta}^{-}_{0.05}(\mathbf{y})$ and $\hat{\theta}^{+}_{0.05}(\mathbf{y})$". Figure 2 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 $\theta$, we can say that this procedure generates a 95% confidence interval. More generally we can talk about a $100(1-\alpha)\%$ confidence interval; for $\alpha = 0.05$ we have a 95% confidence interval, which is generally the most common interval used in statistical analysis.

The idea of "confidence" can initially be confusing. The important thing to recognise is that a confidence interval *procedure* gives you guarantees 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 we have a 95% chance of drawing a sample that when fed into our procedure will lead to a 95% confidence interval that contains the true value of the population parameter.

What a confidence interval does not do is give you a guarantee for the particular sample we have observed. To understand this, remember that the population parameter $\theta$, while unknown, is *not* a random variable. It is fixed to 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 we could see 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

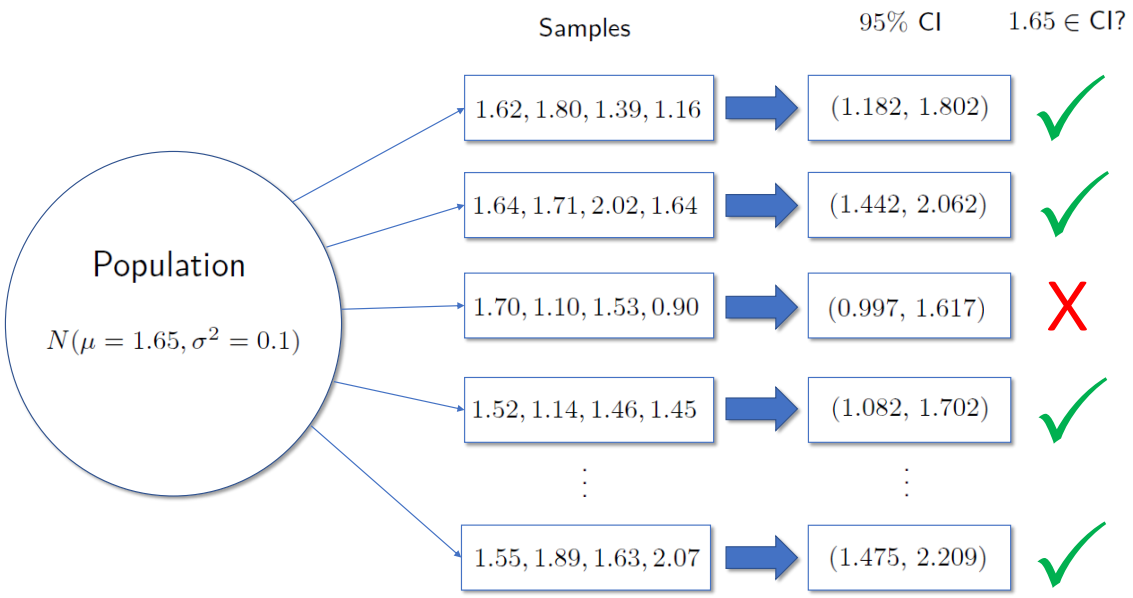


Figure 2: Cartoon showing multiple samples drawn from a $N(\mu = 1.65, \sigma^2 = 0.1)$ population, along with the 95% confidence intervals for each sample. 5% of possible samples will result in CIs that do not include $\mu = 1.65$.

one of those samples that does produce a CI that covers $\theta$.

**CI for normal mean, known variance**. Let us examine confidence intervals for the maximum likelihood estimate of the mean of a normal distribution. We assume that the population is normally distributed with unknown mean $\mu$ and known variance $\sigma^2$. 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. Under the assumptions of our population, we know (see Lecture 3) that our estimate $\hat{\mu}$ is distributed as per

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

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 $\mu$ and a variance of $\sigma^2/n$ (see Lecture 4 and Lecture 4 summary for refresher). Knowing this sampling distribution we can construct a 95% confidence interval. The important step is to note that the $z$-score for $\hat{\mu}$

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

follows an $N(0, 1)$ distribution; 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

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

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

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

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

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

We can interpret the above equation as saying that for 95% of possible samples we could see from our population, the true unknown population mean $\mu$ will be within $1.96\sigma/\sqrt{n}$ of the sample mean. This information then lets us build our 95% confidence interval as

$$\left(\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}},\ \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right) \tag{3}$$

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

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

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

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

and is given by the $100(1-\alpha/2)$-th percentile of the standard normal distribution (see Studio 4 for more details). Looking at the CI for the normal mean with known variance (4) we can observe that the width of the interval:

- increases with increasing population standard deviation $\sigma$; this is because the more variability in the population, the harder it is to nail down the exact value of the population mean;
- decreases proportionally to the square-root of the sample size; that is, the more data we collect the more accurate our estimates become;
- increases with increasing confidence level $(1-\alpha)$; that is, the more confident we require our interval to be, the wider the interval must be to guarantee this level of confidence.

**Example: CI for normal mean, 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:

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

We are told that the population variance is 43.75, which has been estimated from another very large study of Pima people. Let us construct a 95% confidence interval for the population mean. First, we

calculate the sample mean, which is $\hat{\mu} = 38.88 kg/m^2$. Using (3) (i.e., equation (4) 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

$$(34.3,\ 43.47)$$

In words, we can summarise our analysis by:

> *"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. This type of statement makes it very clear exactly what our analysis is showing.

**CI for normal mean, unknown variance**. The assumption that the population variance $\sigma^2$ is known is generally unrealistic. Even if we assume that the variance is unknown we can still construct a 95% confidence interval in a similar manner to the case when $\sigma^2$ is known. The difference is that we now need to estimate $\sigma^2$ from our data sample. An obvious approach would be to estimate $\sigma^2$ using the unbiased estimate of variance

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

where $\bar{y}$ is the sample mean. We could then plug $\hat{\sigma}$ into (4) in place of the unknown population standard deviation $\sigma$. Unfortunately, this approach does not lead to an exact 95% confidence interval – for smaller sample sizes it will not cover the true parameter value for 95% of possible samples (i.e., does not give 95% coverage). The reason is that the statistic

$$\frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}} \tag{5}$$

is no longer normally distributed when we use the estimate $\hat{\sigma}$ in place of the unknown population $\sigma$. This is because the variance is being estimated from the data, and we need to take into account the fact that this is an estimate and not an exact value. Instead, (5) follows a something called a Student-$t$ distribution with $n - 1$ degrees-of-freedom. Figure 3 shows two different $t$-distributions along with the standard normal distribution. We see that both the $t$-distributions and the normal distribution have similarities: they are symmetric and unimodal (one peak), and tail off to zero either side of the peak. The difference is that the $t$-distribution spreads more probability over larger values than the normal distribution does – it is called a "heavy-tailed" distribution as its "tails" go towards zero at a slower rate than the normal. The smaller the degrees-of-freedom, the heavier these tails. For very large degrees-of-freedom, the standard normal distribution and $t$-distribution are virtually the same.

The Student-$t$ distribution is symmetric and self-similar in the same fashion as the normal distribution. This lets us use exactly the same steps as we used in deriving our CI for the mean with known variance to arrive at the $100(1 - \alpha)$ confidence interval for $\mu$ when $\sigma^2$ is unknown (see Ross, Chapter 7 for details):

$$\left(\hat{\mu} - t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}},\ \hat{\mu} + t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}}\right) \tag{6}$$

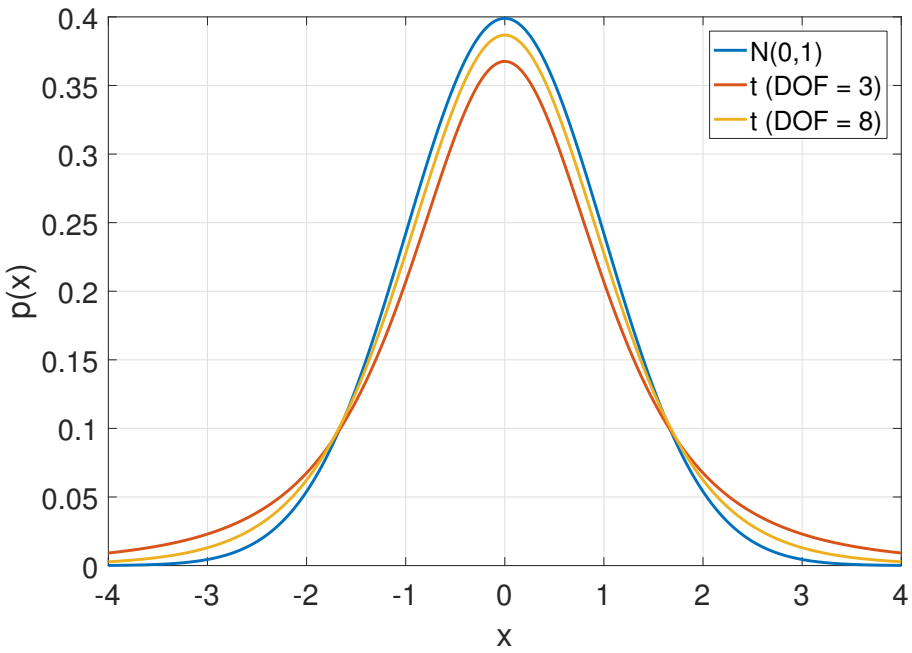


Figure 3: Plot of a standard normal $N(0,1)$ distribution and two Student-$t$ distributions, one with degrees-of-freedom (DOF) of 3, and one with DOF of 8. Note how the $t$-distributions spread the probability out more and tail off to zero slower than the normal distribution.

where $t_{\alpha/2,n-1}$ is now the $100(1-\alpha/2)$-th percentile of the standard $t$-distribution with $n-1$ degrees-of-freedom, which we can calculate in R using the command $\texttt{qt}(\texttt{p} = \texttt{1} - \texttt{alpha/2}, \texttt{n} - \texttt{1})$. This interval will achieve exactly $100(1-\alpha)\%$ coverage if the population from which our sample is drawn is normally distributed. The intervals (6) and (4) are very similar; both build the interval as a multiple of the standard error. They differ in the fact that one uses the estimate $\hat{\sigma}$ in place of the population parameter $\sigma$ to determine the standard error, and 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:

- 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. If we take the sample size to be very large we will see that $t_{\alpha/2,n-1} \to z_{\alpha/2}$ as $n \to \infty$; so 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 accounting for the fact that we need to estimate the variance from the data, which brings with it a degree of uncertainty.

**Example: CI for normal mean, 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 assume we do not have access to a good value of the population variance. Our estimate of the mean was $\hat{\mu} = 38.88$; we now need to estimate the variance from the data:

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

To use the interval (6) 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 (6) 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 $(32.9, 44.86)$. 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, that given the fact that $t_{\alpha/2,n-1} > z_{\alpha/2}$ for any finite sample size $n$, will our interval using (6) *always* be wider? The answer, surprisingly, is not always – while in general it may be wider, there will be at least some samples of data from our $N(\mu, \sigma^2)$ population which lead to an estimate $\hat{\sigma}^2$ sufficiently smaller than $\sigma^2$ to result in a shorter confidence interval. However, what is guaranteed by using (6) is that 95% of samples will result in CIs that cover the true population $\mu$.

**CI for difference of normal means**. One of the most important estimates we are often interested in is the difference in population means between two populations. 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 participant's are measured and recorded. Call this sample $\mathbf{y}_A$, and assume it has an unknown population mean of $\mu_A$. 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 $\mu_B$. 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 estimate the mean from both samples, say $\hat{\mu}_A = \bar{y}_A$ and $\hat{\mu}_B = \bar{y}_B$. The estimated difference is then $\hat{\mu}_A - \hat{\mu}_B$; 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. So it is of value to quantify how accurate our estimate of the difference is using a confidence interval. To do this, first assume both samples come from normal populations with unknown means $\mu_A$ and $\mu_B$, and known variances $\sigma_A^2$ and $\sigma_B^2$, respectively. Then, if we estimate both means by their respective sample means, we know that

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

where $n_A$, $n_B$ are the sizes of the two samples. Then, we know that

$$\mathbb{E}\left[\hat{\mu}_A - \hat{\mu}_B\right] = \mu_A - \mu_B$$

and if we assume the samples are independent, we have from independence of RVs (see Lecture 2)

$$\mathbb{V}\left[\hat{\mu}_A - \hat{\mu}_B\right] = \mathbb{V}\left[\hat{\mu}_A\right] + \mathbb{V}\left[\hat{\mu}_B\right],$$

so that the estimated difference satisfies

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

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

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

Now that we have an $N(0,1)$ distributed statistic, we can use exactly the same procedure as when deriving the CI for the mean with known variance to arrive at the interval

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

is a $100(1-\alpha)\%$ confidence interval for $\hat{\mu}_A - \hat{\mu}_B$.

Of course, assuming $\sigma_A^2$ and $\sigma_B^2$ 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 (see Ross, Chapter 7.4, pp. 257–260). However, even this assumption is not particularly realistic. Deriving an exact confidence interval if $\sigma_A^2 \neq \sigma_B^2$, and both are unknown is quite difficult; we briefly discuss an *approximate* procedure that we can use, though R does implement some more exact procedures. 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

$$\left( \hat{\mu}_A - \hat{\mu}_B - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}},\ \ \hat{\mu}_A - \hat{\mu}_B + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}} \right). \tag{7}$$

This approximation gets better in the sense that the coverage is 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 often decent enough, though we acknowledge the fact that it is only approximate and does not give exact 95% coverage for finite sample sizes.

**Example: 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:

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

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 (7) 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. In general, when summarising confidence intervals of differences, consider the follow three scenarios:

- 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 (closer $\alpha$ is to one).
- 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 (closer $\alpha$ is to one).
- Does the interval contain zero (i.e., lower end of CI is negative, upper end of CI is positive)? This means we cannot rule out the possibility that there is actually no difference at the population level.

**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 also 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 to get approximate CIs in this case.

Let $Y_1, \ldots, Y_n$ be RVs from our population, and let us assume our parameter $\theta$ of interest can be estimated using the sample mean $\bar{Y}$. Assume only that $\mathbb{E}[Y_i] = \theta$, and that $\mathbb{V}[Y_i] = v(\theta)$; that is, we assume that the mean of any datapoint from our population is $\theta$ and the variance of any datapoint from our population is some function of $\theta$. 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 could be used to derive an approximate CI using the basic procedure outlined previously (CI for mean of normal population with known variance). The problem is that we don't know the population value of $\theta$ (otherwise we would not be estimating it), and therefore we don't know the value of the population variance $v(\theta)$. To get around 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

$$\left(\hat{\theta} - z_{\alpha/2}\sqrt{v(\hat{\theta})/n},\ \hat{\theta} + z_{\alpha/2}\sqrt{v(\hat{\theta})/n}\right). \tag{8}$$

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).

To demonstrate how useful this result is, let us consider the problem of constructing an approximate CI for the Poisson rate parameter $\lambda$. 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 ML estimate of $\lambda$ is

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

so we can use (8) to construct an approximate 95% CI for $\lambda$ as

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

This CI shows that as the estimate of the rate parameter $\hat{\lambda}$ 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 $\lambda$. The last question of Studio 4 examined how good the coverage obtained by this approximation is; the solutions show that for $\lambda \geq 5$ and $n > 10$ the approximation is basically exact.