Confidence Intervals

In the last two lectures, one of the items we discussed was how to estimate different parameters, *e.g.*, the sample mean $\bar{X}$ for the population mean μ, and the sample proportion $\hat{p}$ for the success probability p. These are called **point estimates**, because they are single numbers. Usually, in order for a point estimate to be useful, it is important to describe a likely range of values around that estimate that the true value could take.

In this lecture, we will look at constructing **confidence intervals** for parameters, which is an interval of values that will contain the true parameter with some level of certainty. These are statements like "we are 95% confident that the true diameter of the piston is in the interval $(13.7, 14.3)$." Intuitively, for a higher confidence level, our interval will be wider, and for a smaller confidence interval, it becomes narrower.

1. Large-Sample Intervals for a Population Mean

To construct a confidence interval around a population mean, the first step is to grab a random sample $X_1, ..., X_n$ from the population. From that, we calculate $\bar{X}$ and s, *i.e.*, the sample mean and standard deviation.

Now, remember, the Central Limit Theorem tells us that $\bar{X}$ is approximately normal for sufficiently large n; in particular, $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$, where μ and σ are the parameters of the population. Knowing this, computing an α-level confidence interval around μ boils down to asking: For what interval $(\bar{X} - z\sigma, \bar{X} + z\sigma)$ is there less than $1 - \alpha$ of a chance that μ is *outside* of this interval?

Figures 1 and 2 show the situation for a 95% confidence interval. 95% of the population density is covered within $\mu \pm 1.96\sigma_{\bar{X}}$, because 1.96 is the z-score corresponding to $P(-1.96 < Z < 1.96) = 0.95$.[1] If $\bar{X}$ happens to be

[1] For a z-table, see Table A.2 in the book or `http://www.z-table.com`.

in the middle 95%, then the interval $(\bar{X} - 1.96\sigma, \bar{X} + 1.96\sigma)$ will succeed in covering μ (Figure 1). If it is not, then it won't (Figure 2). How confident are we that it will? 95%, which is why this is a **95% confidence interval** for the population mean μ.

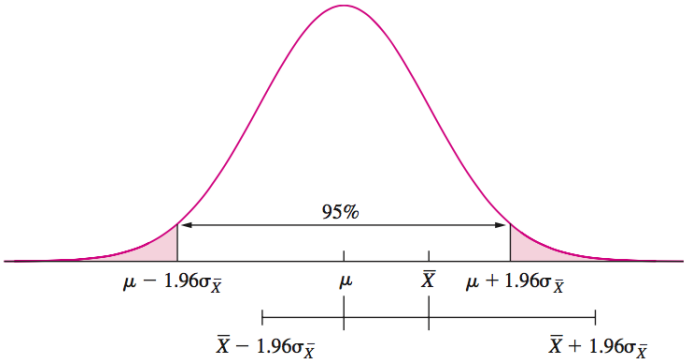

Figure 1: $\bar{X}$ is within the middle 95% of the distribution, so the 95% confidence interval $\bar{X} \pm 1.96\sigma_{\bar{X}}$ succeeds in covering μ.

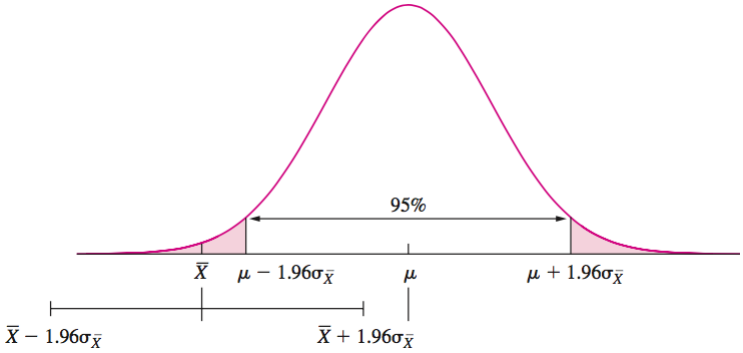

Figure 2: $\bar{X}$ is not within the middle 95% of the distribution, so the 95% confidence interval $\bar{X} \pm 1.96\sigma_{\bar{X}}$ fails in covering μ.

Determining a Confidence Interval

95% is probably the most common confidence level, but it's important to generalize this procedure to a **level $100(1-\alpha)\%$** confidence interval. Let $X_1, ..., X_n$ be a large ($n > 30$) random sample from a population, so that $\bar{X}$ is approximately normal. Then a level $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{X} \pm z_{\alpha/2}\sigma_{\bar{X}}$$

where $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. When the value of σ is unknown, it can be replaced with the sample standard deviation s.

In particular:

$$\begin{aligned}
\bar{X} &\pm \frac{s}{\sqrt{n}} \text{ is a 68\% confidence interval for } \mu \\
\bar{X} &\pm 1.645\frac{s}{\sqrt{n}} \text{ is a 90\% confidence interval for } \mu \\
\bar{X} &\pm 1.96\frac{s}{\sqrt{n}} \text{ is a 95\% confidence interval for } \mu \\
\bar{X} &\pm 2.58\frac{s}{\sqrt{n}} \text{ is a 99\% confidence interval for } \mu \\
\bar{X} &\pm 3\frac{s}{\sqrt{n}} \text{ is a 99.7\% confidence interval for } \mu
\end{aligned}$$

Suppose we are loading boxes of a particular size from a factory onto several trucks. Each truck has space for 45 boxes. We want to get a confidence interval around the mean box weight, so we sample all 45 on one truck, finding $\bar{X} = 12$ lbs and $s = 1$ lb.

Since $45 > 30$, we can construct *e.g.*, a 90% confidence interval as $12 \pm 1.645 \cdot 1/\sqrt{45} = 12 \pm 0.2452$, or $(11.755, 12.245)$.

A 99% interval would be $12 \pm 2.58 \cdot 1/\sqrt{45}$, or $(11.615, 12.385)$.

Notice that the higher confidence level has a broader range. There is always a tradeoff between confidence and precision.

Determining a Confidence Level

We can also work in reverse, starting off with an interval and figuring out what confidence level it would correspond to. From the prior example, maybe we want to know how confident we can be that $\mu \in (11.9, 12.1)$. To figure this out, we need to solve for the z-score corresponding to these boundaries. We can use either side:

$$\begin{aligned}
12.1 &= \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}} \\
12.1 &= 12 + z_{\alpha/2} \cdot 1/\sqrt{45}
\end{aligned}$$

So $z_{\alpha/2} = 0.67$. From the z-table, we find that $\alpha/2$, the area to the right of $z = 0.67$, is $1 - 0.7486 = 0.2514$, so $\alpha \approx 0.50$. That means we can be roughly 50% confident of this.

Determining a Necessary Sample Size

Sometimes, we may want to make a confidence interval narrower for the same confidence level. For example, we may want to have the precise $(11.9, 12.1)$ interval above with a higher level of confidence. The way to do this is to increase the number of samples taken.

For example, suppose we want an 80% confidence interval of $(11.9, 12.1)$, *i.e.*, of width ± 0.1. $\alpha = 0.2$ corresponds to $z_{\alpha/2} = 1.28$, because $P(Z > 1.28) \approx 0.1$. We solve for n: $0.1 = 1.28 \cdot 1/\sqrt{n}$, so $n = 163.8$ which rounds up to $n = 164$.

Note that there's no guarantee that $\bar{X}$ and s will stay the same as we draw more samples; technically, we would need to re-compute the confidence interval with the new values. Nonetheless, this gives us a chance at obtaining what we are looking for.

Probability vs. Confidence

If we have a 95% confidence interval of $(12, 13)$, is it correct to say that the *probability* μ is between 12 and 13 is 95%? Actually, no. The statement "μ is in $(12, 13)$" is binary, either right or wrong. This is why we say that we have 95% *confidence*, rather than probability.

On the other hand, if we are discussing the *method* that we used to come up with the confidence interval, then speaking of probabilities is valid. Each time we draw a random sample and apply the method, we have a 95% chance of covering the mean, and 5% chance of not.

One-Sided Confidence Intervals

We've looked at **two-sided** confidence intervals so far, specifying a lower and upper bound for μ. Sometimes, we are only interested in one side or the other, making **one-sided** intervals appropriate.

The $100(1-\alpha)\%$ lower confidence interval for μ is given by

$$(\bar{X} - z_\alpha \sigma_{\bar{X}}, +\infty)$$

while the $100(1-\alpha)\%$ upper confidence interval is

$$(-\infty, \bar{X} + z_\alpha \sigma_{\bar{X}})$$

The key difference in computing a one-sided interval is that we search for the z-score corresponding to the *full* area $1-\alpha$, rather than splitting it on both sides.

In particular:

$$\begin{aligned}
\bar{X} &+ 1.28\frac{s}{\sqrt{n}} \text{ is a 90\% upper confidence bound for } \mu \\
\bar{X} &+ 1.645\frac{s}{\sqrt{n}} \text{ is a 95\% upper confidence bound for } \mu \\
\bar{X} &+ 2.33\frac{s}{\sqrt{n}} \text{ is a 99\% upper confidence bound for } \mu
\end{aligned}$$

And we replace + by − to get the lower confidence bounds.

2. Intervals for Proportions

We can use methods similar to those before to find confidence intervals around a population proportion (rather than a population mean). In this case, our samples $X_1, ..., X_n$ are Bernoulli trials, either 0 or 1, marking failure or success. The total number of successes $X = \sum_i X_i$ is then a Binomial random variable, *i.e.*, $X \sim \text{Bin}(n, p)$. The estimate for p and its uncertainty from the sample are:

$$\hat{p} = \frac{X}{n} \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

When the sample size n is large, then, the Central Limit Theorem tells us that

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

So, computing a confidence interval for p is is very similar to computing one for μ. In particular, a $100(1-\alpha)\%$ confidence interval for p is

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$$

If the lower limit is less than 0, we replace it with 0, and if the upper limit is greater than 1, we replace it with 1.

What are $\tilde{p}$ and $\tilde{n}$? The convention for a long time was to use $\hat{p}$ and n. Interestingly, it turns out that this tends to make the confidence interval

too short in some cases, especially for small n. Recent research has shown that the particular choices

$$\tilde{n} = n + 4$$
$$\tilde{p} = \frac{X + 2}{\tilde{n}}$$

are much more accurate in practice. So, we should add 4 to the sample size, and add 2 to the number of successes, to get a better result.

As an example, suppose we want to estimate the success probability on a quiz question. For each student i, we let $X_i = 0$ if the student answers incorrectly and $X_i = 1$ if the student answers correctly. In a sample of 50 students, we find 30 of them answer successfully. Then $\tilde{n} = 50 + 4 = 54$ and $\tilde{p} = (30 + 2)/54 = 0.5926$. So, a 95% confidence interval for p is

$$0.5926 \pm 1.96\sqrt{\frac{0.5926(1 - 0.5926)}{54}} = 0.5926 \pm 0.1311 = (0.4615, 0.7237)$$

One Sided Confidence Intervals

For a one-sided interval, a level $100(1-\alpha)\%$ lower confidence bound for p is

$$\left(\tilde{p} - z_\alpha\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}},\ 1\right)$$

An upper confidence bound for p is

$$\left(0,\ \tilde{p} + z_\alpha\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}\right)$$

3. Small-Sample Intervals for a Population Mean

In studying confidence intervals for population means so far, we have required that the sample size be large ($n > 30$), for the central limit theorem to hold. When n is small, those same methods won't apply.

What can we do in these situations? We need some more information about the population. In particular, if we know the population is *approximately normal*, then it turns out that the sample mean $\bar{X}$ will follow a distribution called the **Student's *t* distribution**. In these cases, we can use this to get a confidence interval around the population mean instead.

The Student's t Distribution

Let $X_1, ..., X_n$ be a small ($n < 30$) sample from a normal population with mean μ. Then the quantity

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has a Student's t distribution with $n - 1$ degrees of freedom, denoted t_{n-1}. When n is large, the distribution of this quantity is very close to normal, so the normal curve can be used instead of the Student's t.

What does the t distribution look like? In Figure 3, the probability density function for three different degrees of freedom – t_2, t_5, t_{10} – are shown, as is the z-distribution $N(0, 1)$. The curves all have a similar shape to z, and as n increases, t_{n-1} becomes closer and closer to z. For smaller values of n, the t distribution has more area in its tails (*i.e.*, is more spread out), which should make sense because it is based on a smaller sample size. Mathematically, having $s/\sqrt{n}$ in the denominator accounts for this, because we expect s to be much smaller than σ for a small sample size.

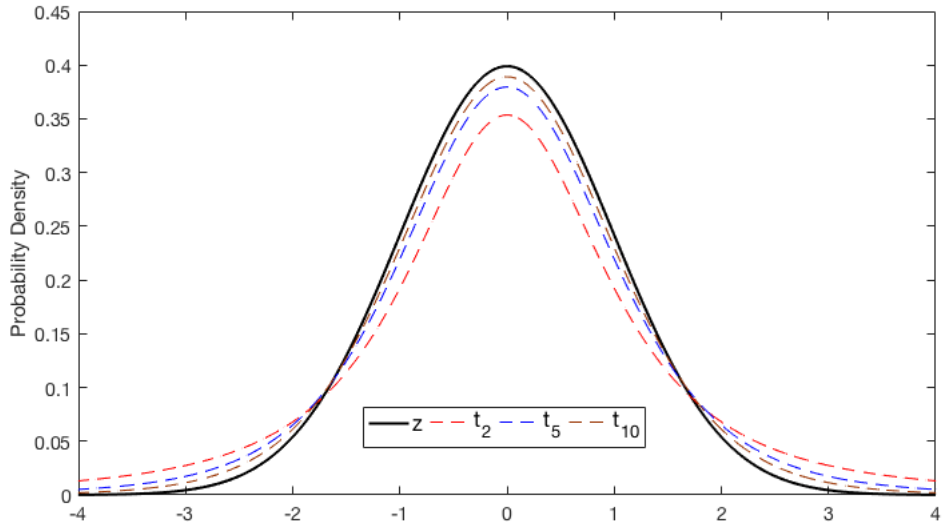

Figure 3: Plot of the Student's t_{n-1} curve for different degrees of freedom. The z curve is plotted here for comparison.

Table A.3 in the book, called a ***t* table**, provides probabilities associated with the t distribution, similar to how Table A.2 gives those for the z distri-

bution.[2] The difference here is that the degrees of freedom is also a factor in the table. Naturally, then, the table will be less precise in reporting probabilities. What do we do if the value we need does not appear in the table exactly? We can either report a range of possible values, or we can use linear interpolation.[3]

Confidence Intervals with *t*

Let $X_1, ..., X_n$ be a small random sample from a normal population with mean μ. Then a level $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{X} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

where $t_{n-1,\alpha/2}$ is the value of t_{n-1} which cuts off an area of $\alpha/2$ in the right-hand tail.

Let's take the example of five data points $56.3, 65.4, 58.7, 70.1, 63.9$ that come from an approximately normal population. How can we determine a 95% confidence interval for the population mean μ? With $\alpha = 0.05$, the first step is to find the value of $t_{5-1,0.025}$, which is 2.776. So, our confidence interval will be

$$\bar{X} \pm 2.776 \cdot \frac{s}{\sqrt{n}}$$

Then, we can find $\bar{X}$ and s from the sample:

$$\bar{X} = 62.88 \qquad s = 5.484$$

So, the 95% confidence interval is

$$62.88 \pm 2.776 \cdot 5.484/\sqrt{5} = (56.072, 69.6882)$$

A key point here is that before we apply the t distribution to a population mean, we need to be sure that the population is approximately normal. Unfortunately, with small samples, departures from normality can be hard to detect. A reasonable way to proceed is to see whether the sample drawn

[2] For a t-table, see Table A.2 in the book or `http://albertskblog.blogspot.com/2010/08/student-t-distribution-table.html`.

[3] For a quick tutorial on linear interpolation, see `http://www.eng.fsu.edu/~dommelen/courses/eml3100/aids/intpol/`.

contains any statistical outliers: if it does, then this procedure is probably not valid, because it is likely that it comes from a non-normal population.

We can also come up with one-sided confidence intervals based on the t distribution, analogous to the case of large samples. The $100(1-\alpha)\%$ upper confidence bound for μ is

$$\left(-\infty,\ \bar{X} + t_{n-1,\alpha}\frac{s}{\sqrt{n}}\right)$$

And the lower confidence bound is

$$\left(\bar{X} - t_{n-1,\alpha}\frac{s}{\sqrt{n}},\ +\infty\right)$$

Use z or t?

To summarize, when the sample size is large enough ($n \geq 30$), we use the z-distribution to come up with a confidence interval around μ. The Central Limit Theorem tells us that the sample mean is approximately normally distributed around μ, and that's all we need to know.

When the sample is not large enough ($n < 30$), we can only construct a confidence interval if it is reasonable to assume the population is normally distributed. We can tell right away that this is *not* valid if the sample contains any outliers. When it is valid, we can compute $\bar{X}$ and s, and then use the t-distribution to get a confidence interval around μ.

What if we have a small sample but we know what the population standard deviation σ is? In these cases, we would use the z-distribution: the purpose of t is to adjust for the fact that we're estimating σ from s, so if we know σ, it's better to use z. Note that either way, with a small sample, we still need to know that the population is normal; that is, regardless of whether we know σ or not.