MAT220 Lecture 12: An Introduction to Estimation

Contents

In this lecture, we will use our theory of the normal distributions for *estimation.* The basic question is simple: if we cannot measure an entire population, can we still make a reasonable guess about an unknown population parameter, such as μ? As we will see, the answer is yes. We will assume, for now, that the population standard deviation σ is already know. That way, we can use the Central Limit Theorem to connect observed sample data to the true population mean. Using the machinery of normal distributions, we will then create *estimates* for the value of μ by constructing tolerable error bounds around sample means $\bar{x}$ – a construction known as a *confidence interval.*

1 A Review of the Central Limit Theorem

1.1 The sampling distribution

Suppose that we have some population distribution with mean μ and standard deviation σ. If we take a simple random sample of size n, then we can compute its sample mean:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

If we repeat this process again and again, we get many possible sample means. The collection of all of these sample means forms a new distribution, called a *sampling distribution*.

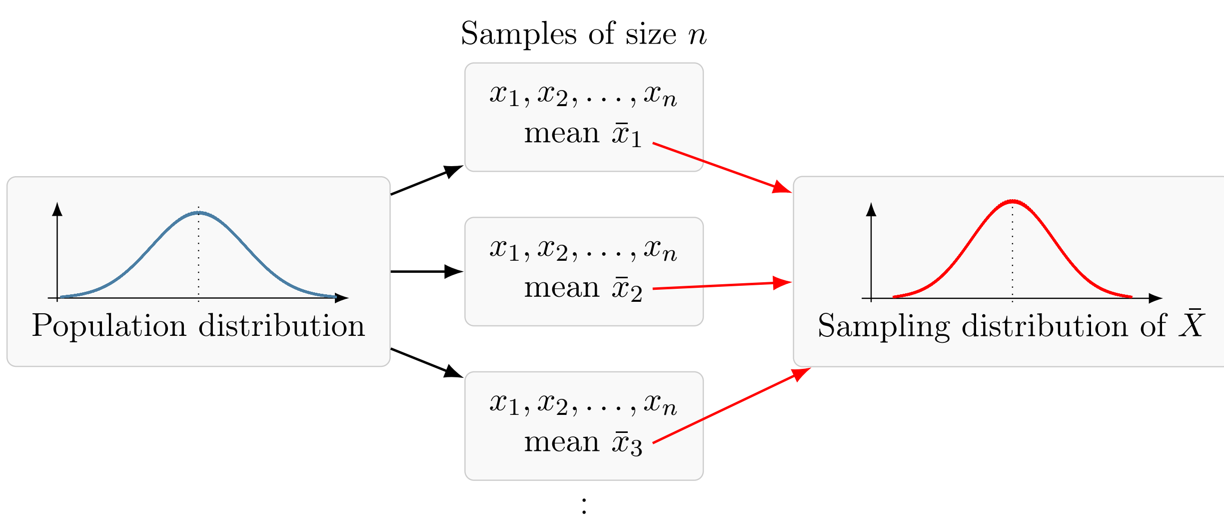
Sampling distribution of the sample mean

Given a probability distribution, the **sampling distribution of $\bar{X}$ for samples of size n** is the probability distribution formed from all sample means taken from all simple random samples of size n.

The important point here is that the samples are taken randomly. This means that the exact value of $\bar{x}$ is also going to be random. Before we take a particular sample we don't know which actual values we will get, and therefore we also don't know what the exact value of $\bar{x}$ will be. So, we can upgrade the sample mean into a random variable, written $\bar{X}$. This is used as a placeholder for all the possible values that the sample mean can take, depending on all of the possible random samples we could take from the population.

1.2 The Central Limit Theorem

Generally speaking, the Central Limit Theorem is a mathematical result that tells us what the sampling distribution of $\bar{X}$ looks like. If the original population is normal, then the sampling distribution of $\bar{X}$ is exactly normal:



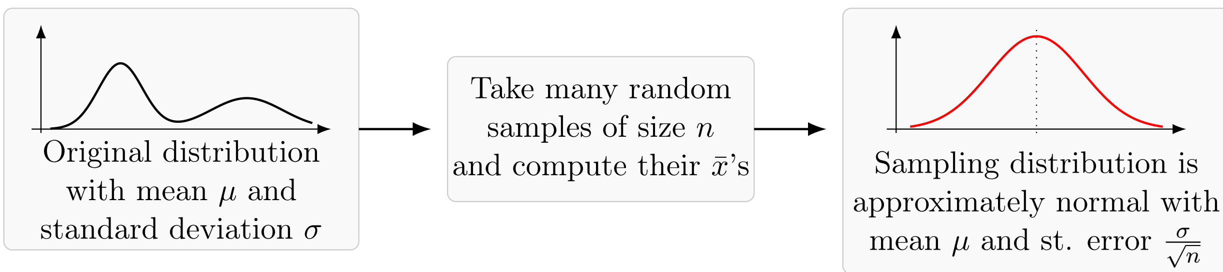
Moreover, if the original population is *not* normal, then the Central Limit Theorem says that the resulting sampling distribution of $\bar{X}$ is approximately normal, provided that n is large enough. Formally, one statement of the Central Limit Theorem is as follows.

Central Limit Theorem

If a population has mean μ and standard deviation σ, then for large enough n, the sampling distribution of $\bar{X}$ satisfies:

$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

This means that the sampling distribution is centred at the true population mean μ, but its spread is smaller than the spread of the original population. The larger the sample size n, the smaller the spread of the sample means.



In the statement of the Central Limit Theorem, the variance of the sampling distribution is equal to the particular quantity $\frac{\sigma^2}{n}$. The square root of this, $\frac{\sigma}{\sqrt{n}}$, is the standard deviation of the sampling distribution. To avoid confusing this with the standard deviation σ of the original population, we use different terminology.

Standard error

The standard deviation of the sampling distribution of $\bar{X}$ is

$$\frac{\sigma}{\sqrt{n}}.$$

This is called the **standard error**.

2 The Idea of Estimation

2.1 Estimation with Known σ

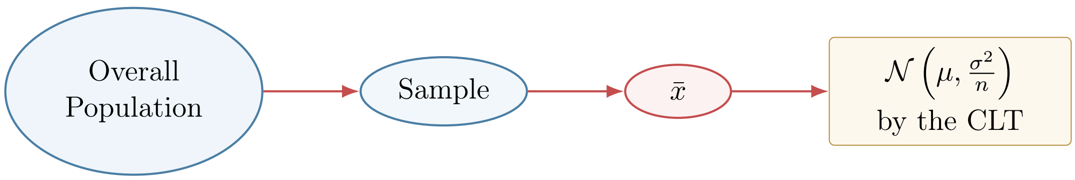
In real statistical problems, we usually do not know the true population mean μ. Knowing something like the true average of a population is extremely useful, and it is certainly a common parameter that we would like to know. However, in practice, we cannot go out into the world and measure every single object in a population in order to calculate the true mean.

In this lecture, we are going to assume that:

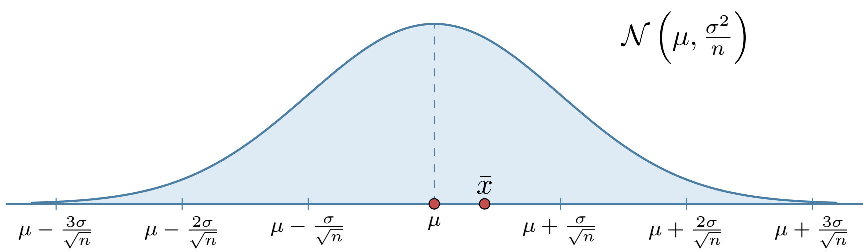
- the population mean μ is unknown, but
- the population standard deviation σ happens to be *known*.

This is obviously an unrealistic scenario. However, theoretically it is the simplest setup to estimate μ. Next lecture we will relax this assumption and estimate μ in the realistic scenario in which σ is also unknown.

The benefit of assuming that σ is known is that we may use the Central Limit Theorem to connect sample means to the true population mean:



Graphically, if we take a simple random sample, then the Central Limit Theorem tells us that the observed value of $\bar{x}$ will sit somewhere in this normal distribution:

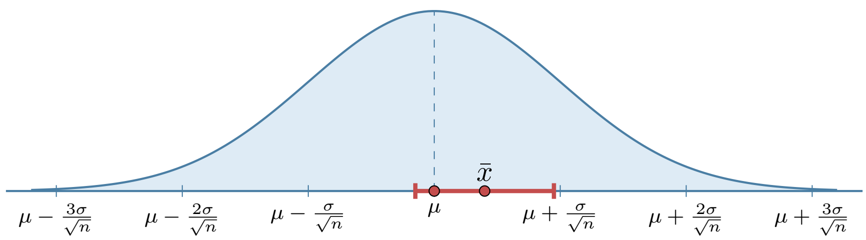


Assuming that σ is known allows us to exploit the structure of the sampling distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. For instance, we don't know where μ is, but we know that $\bar{X}$ has about a 68% chance of landing within one standard error of μ. Why? Because the randomness of $\bar{X}$ is encoded in the shape of the distribution – z-score calculations from Lecture 10 tell us that 68% of the area of the distribution is within one standard error of the center.

Importantly, these sorts of ideas are still true *even if we don't know where μ actually is.* According to the Central Limit Theorem, the randomness of $\bar{X}$ follows a normal distribution with the true population mean at the center. As we will see in a moment, even if we don't know what μ is, we can at least *relate* it to an observed value $\bar{x}$.

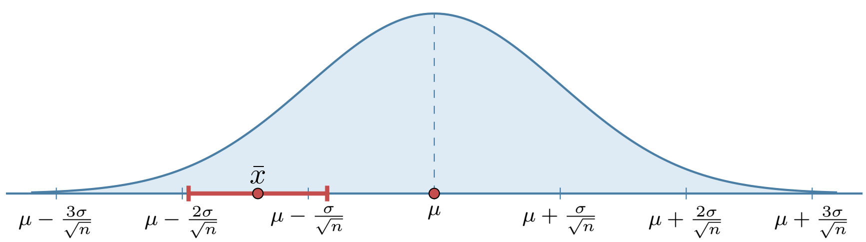
2.2 Intervals around sample means

If we pick a sample at random, then there is no guarantee that the resulting mean $\bar{x}$ will be exactly the same as μ. As a matter of fact, we would generally expect that every sample mean is slightly different from the true mean, ultimately due to random noise. However, at the very least, we can always define a sense in which $\bar{x}$ is *close enough* to μ. Given that we don't know what μ is, instead of saying "this is where μ is exactly", it seems the best we can say is "μ is probably here, within some acceptable margin of error". Mathematically, this "acceptable margin of error" will look like an interval:



In the above picture, our randomly chosen value of $\bar{x}$ has an interval around it of a fixed length. Although μ is different from $\bar{x}$, it still lies in this neighbourhood of $\bar{x}$.

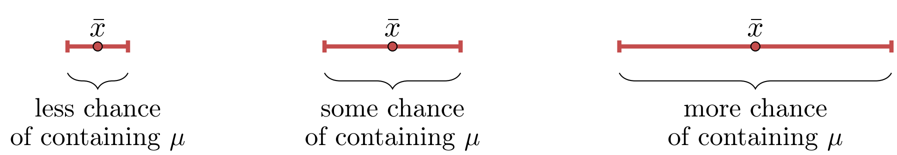
Of course, sometimes the sample will be statistically unlikely, which will cause the resulting value of $\bar{x}$ to be far away from the center:



This is bound to happen due to the nature of randomness. In fact, given some random value of $\bar{X}$, the best we can say is something like:

> We think with $c\%$ confidence that this interval contains μ.

This is ultimately a statement about uncertainty: we don't *know* where μ is, but instead can say that intervals of this particular length would capture it a certain proportion of the time. Of course, if we made the interval longer, then it would be more likely to hit μ, and if we made the interval shorter then it would be less likely to contain μ:



So, it seems that our degree of confidence is directly related to the length of the interval: if we wanted to be very certain that the interval around our randomly-obtained $\bar{x}$ contained μ, then we should make the interval very large. Conversely, if we make the interval very small, then we are less certain that it will contain μ. This gives us the idea of a *confidence interval*.

Confidence interval

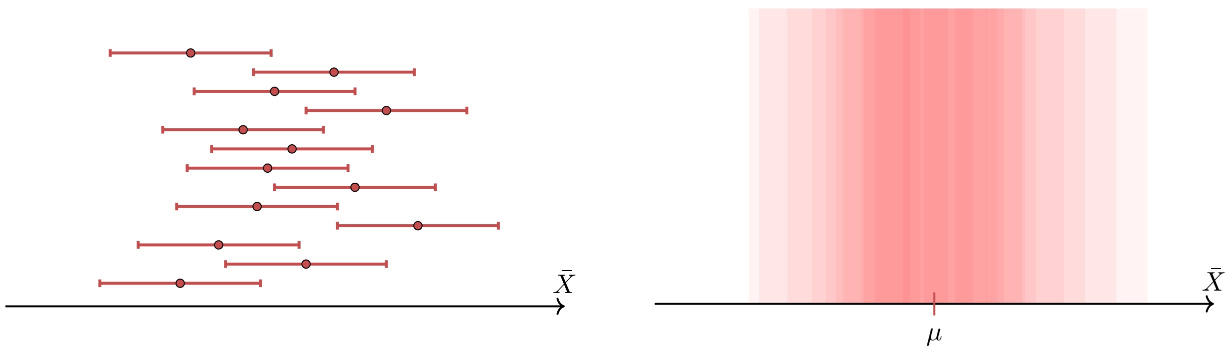
A **confidence interval** is an interval estimate for an unknown population parameter. The confidence level describes how often the method captures the true parameter in repeated sampling.

We often write the confidence level with the letter c. The idea here is that c can be written as a percentage, and there are different confidence intervals corresponding to different levels of confidence. If we want to be more confident, then we increase c and this will increase the length of the resulting interval.

2.3 Interpretation of the Confidence Level

For the sake of illustration, let's assume that we have constructed a confidence interval with confidence level $c = 95\%$. By construction, this means that the random interval produced by the method will contain the true population mean μ 95% of the time. In other words, if we were to take 100 simple random samples and compute 95% confidence intervals around their sample means, then on average we would expect that 95 of these intervals would contain μ and 5 would not.

The picture below helps illustrate the long-run meaning of the confidence level: if we kept taking samples over and over again, then most of the resulting confidence intervals would contain μ:



In the image on the right, we have taken the various confidence intervals from the left-hand side and overlaid them. As you can see, many of the intervals overlap near the true population mean μ.

Before discussing the construction of confidence intervals, we must first dispel a common misconception.

Common misconception

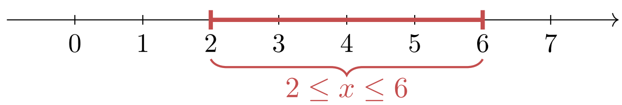
Suppose that we construct a 95% confidence interval around some observed $\bar{x}$. Here, the "95%" does **not** mean that there is a 95% chance that the interval contains μ. Instead, it refers to the long-run success rate of the overall method.

Saying something like "there is a 95% chance that this interval contains μ" implies that μ is somehow changing its value randomly. This is not true: μ is unknown, but its value is absolutely fixed. For any interval around a chosen $\bar{x}$, the mean μ is either in the interval or it is not. Instead, it is the value of $\bar{x}$ that is randomly moving around in the distribution (as represented with the notation $\bar{X}$). The confidence interval moves around with the value of $\bar{x}$, and the confidence level refers to *this* movement.

3 Constructing Confidence Intervals

3.1 A reminder about intervals

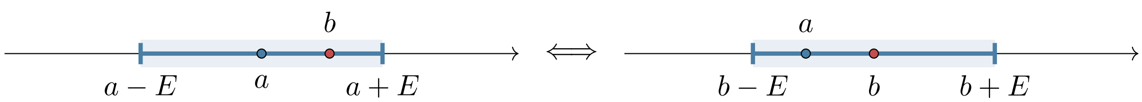
Generally speaking, a connected interval is fully described by its endpoints: we simply write where the interval starts and finishes, and from there it is understood that the interval itself is everything in between. For example, the interval $[2, 6]$ means all numbers x satisfying $2 \leq x \leq 6$:



Suppose now that a and b are two points on the real line. Saying that "b is within distance E of a" means the same thing as saying that "a is within distance E of b". In symbols:

$$a - E \leq b \leq a + E \qquad \Longleftrightarrow \qquad b - E \leq a \leq b + E.$$

Geometrically, we see that if b is in an interval of length $2E$ centered at a, then a will also be in the interval of length $2E$ centered at b:



When constructing our confidence intervals, our goal will be to specify the two endpoints of each of the intervals. We are going to be looking for two endpoints $\bar{x}_1$ and $\bar{x}_2$ on either side of an observed $\bar{x}$, such that $\bar{x}$ is exactly in the center. That way, in symbols we have:

$$\mu \text{ is in the interval centered at } \bar{x} \qquad \Longleftrightarrow \qquad \bar{x} - E \leq \mu \leq \bar{x} + E.$$

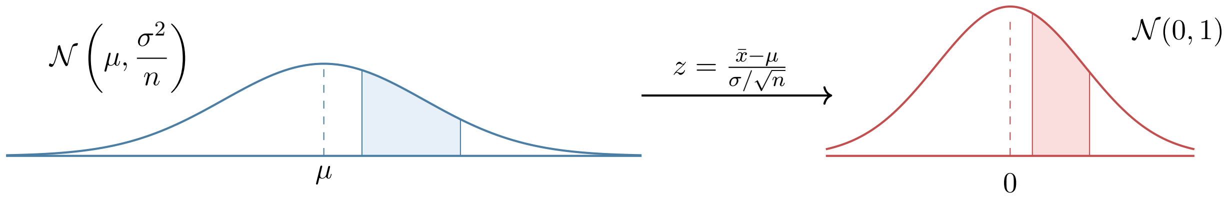
This observation is the key idea behind confidence intervals: if $\bar{X}$ is close to μ, then μ is also close to $\bar{X}$.

3.2 The z-score for the sampling distribution

According to the Central Limit Theorem, we know that the sampling distribution of $\bar{X}$ can for large enough n be approximated by: $\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Since this is a normal distribution, we can relate it to the Standard Normal Distribution $\mathcal{N}(0, 1)$ by creating a z-score. The formula for the z-score measures the distance between a value and the center, and then divides by the standard deviation of the distribution. In the case of the sampling distribution, the center is μ and the distribution is measured in terms of the standard error $\frac{\sigma}{\sqrt{n}}$. Therefore, the z-score formula that we have to use is:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}.$$

This converts the sampling distribution of $\bar{X}$ into the standard normal distribution, and allows us to compute areas in the sampling distribution using a z-table:

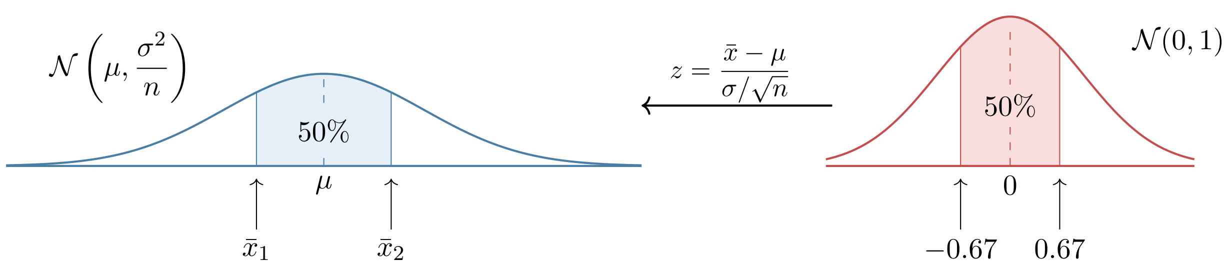


3.3 Finding a 50% confidence interval

Suppose we want $\bar{X}$ to be close to μ with probability 50%. We can describe confidence intervals around the values of $\bar{X}$ by first looking for 50% of the area centered around 0 in the standard normal distribution. From there, we can use the z-score formula and *work backwards* to figure out the bounds of the same area, this time in the sampling distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. According to the z-table, we can see that the values of z_1 and z_2 that give equal tail areas and satisfy

$$P(z_1 < Z < z_2) = 0.5$$

are precisely $z_1 = -0.67$ and $z_2 = 0.67$. We can now use these two values to work backwards through the z-score formula and find the corresponding values in the sampling distribution:



These values will bound 50% of the area around μ, meaning that $\bar{X}$ will have a 50% chance of landing in this region.

We can find the values of $\bar{x}_1$ and $\bar{x}_2$ by substituting in the values of ± 0.67 for z and solving algebraically. Using $z = 0.67$, we have:

$$0.67 = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \quad \Longrightarrow \quad \bar{x} = \mu + \frac{0.67\sigma}{\sqrt{n}}.$$

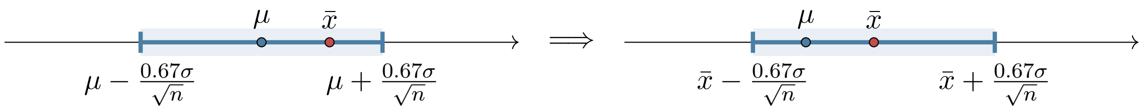
Similarly, using $z = -0.67$ gives:

$$-0.67 = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \quad \Longrightarrow \quad \bar{x} = \mu - \frac{0.67\sigma}{\sqrt{n}}.$$

These two endpoints can now be used to describe an interval around μ for which $\bar{X}$ has a 50% chance of being inside. In symbols, by construction our interval satisfies:

$$P\left(\mu - \frac{0.67\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{0.67\sigma}{\sqrt{n}}\right) = 0.5.$$

Here is where we use our observation from Section 4.1: if $\bar{X}$ is nearby μ 50% of the time, then that means that μ is also nearby $\bar{X}$ with probability 0.5:



Therefore, if we keep the same length of the interval, we can center it around $\bar{X}$ and by construction this interval will contain μ 50% of the time. In other words, we have just created a 50% confidence interval around $\bar{X}$. We write this as:

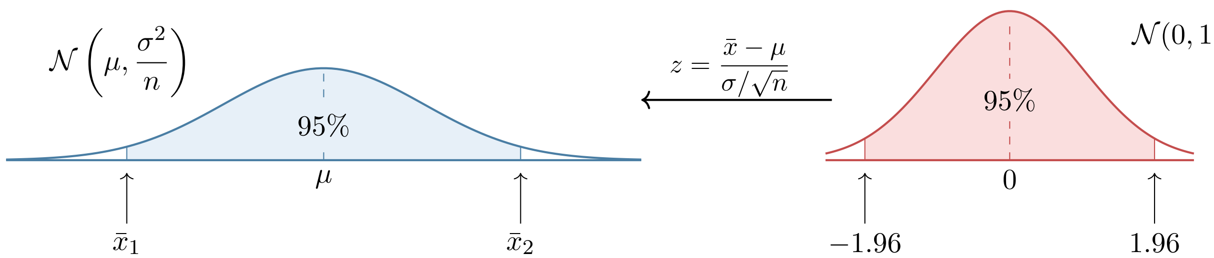
$$\left[\bar{X} - \frac{0.67\sigma}{\sqrt{n}},\ \bar{X} + \frac{0.67\sigma}{\sqrt{n}}\right].$$

After observing a particular sample mean $\bar{x}$, we write the confidence interval around it as:

$$\left[\bar{x} - \frac{0.67\sigma}{\sqrt{n}},\ \bar{x} + \frac{0.67\sigma}{\sqrt{n}}\right].$$

3.4 Finding a 95% confidence interval

Suppose now that we would like to construct a 95% confidence interval instead. The same logic still holds, except that this time, we simply seek out the values of z that will give a 95% area centered around 0 in $\mathcal{N}(0,1)$. Again, by using a z-table, these happen to be ± 1.96. We now use the z-score formula to find the corresponding values of the endpoints in the sampling distribution:



Now using $z = \pm 1.96$ instead of ± 0.67, we find:

$$1.96 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \Longrightarrow \quad \bar{x} = \mu + \frac{1.96\sigma}{\sqrt{n}}$$

$$-1.96 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \Longrightarrow \quad \bar{x} = \mu - \frac{1.96\sigma}{\sqrt{n}}$$

By the same logic as before, the region between $\bar{x}_1$ and $\bar{x}_2$ will contain $\bar{X}$ 95% of the time:

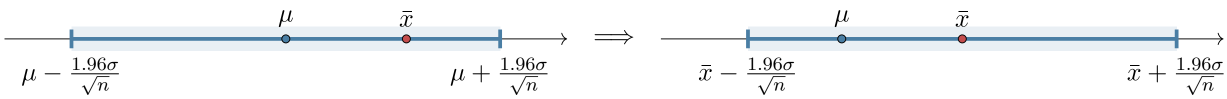
$$P\left(\mu - \frac{1.96\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95.$$

Therefore, if we take the same interval and instead center it at $\bar{X}$, then μ will be contained in this new interval 95% of the time as well:

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

This time, geometrically we have created a wider interval around $\bar{x}$:[1]

[1] This image is not to scale. In fact, the real interval should be about 3 times larger than the interval drawn in the previous section.



In symbols, a 95% confidence interval around $\bar{X}$ can be written:

$$\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}},\ \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right],$$

or alternatively:

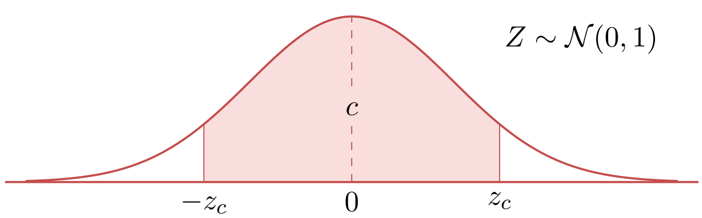
$$\left[\bar{x} - \frac{1.96\sigma}{\sqrt{n}},\ \bar{x} + \frac{1.96\sigma}{\sqrt{n}}\right],$$

if a value of $\bar{x}$ has already been observed.

4 The General Formula for Confidence Intervals

4.1 Critical values

Suppose now that we want to construct a confidence interval with some confidence level c. Keeping with the same ideas as the previous section, we need to first look for the area of the standard normal distribution that corresponds to the confidence level c:



In general, c can take different values. Therefore, the exact values of $\pm z_c$ will change depending on how large or small you would like to make the area. This number z_c has a special name: it is called the *critical value* for confidence level c.

Critical value

For a confidence level c, the **critical value** z_c is the positive number satisfying

$$P(-z_c < Z < z_c) = c,$$

where

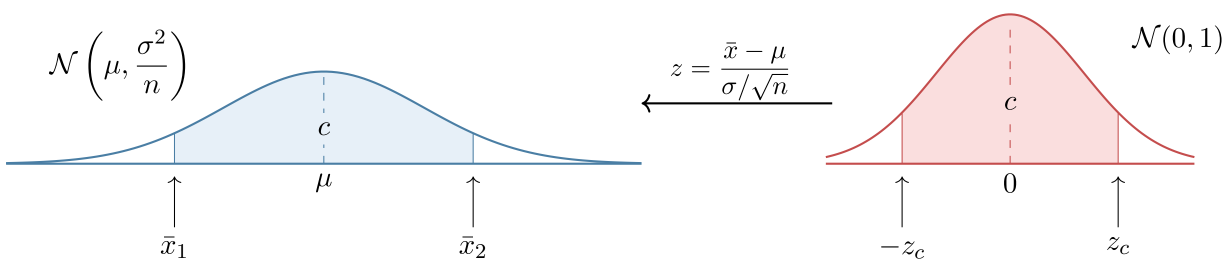
$$Z \sim \mathcal{N}(0,1).$$

The precise value of z_c can presumably be found from a z-table or a calculator. However, in practice, we often compute confidence intervals for common levels of confidence such as 90% or 99%. For convenience, some common critical values are listed in the table below.

Level of confidence c	**Critical value** z_c
0.70, or 70%	1.04
0.75, or 75%	1.15
0.80, or 80%	1.28
0.85, or 85%	1.44
0.90, or 90%	1.645
0.95, or 95%	1.96
0.98, or 98%	2.33
0.99, or 99%	2.58

4.2 The general confidence interval

Suppose now that we want to construct a confidence interval with some confidence level c. Keeping with the same ideas as the previous section, we need to first look for the central area c of the standard normal distribution. Once we have found the critical value that corresponds to c, we can solve for the corresponding endpoints of $\bar{X}$:



Solving

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

for $\bar{x}_1$ and $\bar{x}_2$, we find that the endpoints in the sampling distribution are:

$$\mu - \frac{z_c \sigma}{\sqrt{n}} \qquad \text{and} \qquad \mu + \frac{z_c \sigma}{\sqrt{n}}.$$

Therefore, by construction we have found an interval centered at μ that contains $\bar{X}$ with probability c. In symbols:

$$P\left(\mu - \frac{z_c \sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{z_c \sigma}{\sqrt{n}}\right) = c.$$

By the symmetry of being "nearby", we can take this interval and center a random interval around the random variable $\bar{X}$, and this interval will therefore contain μ with the same probability:

$$P\left(\bar{X} - \frac{z_c \sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{z_c \sigma}{\sqrt{n}}\right) = c.$$

This gives the general confidence interval formula.

Confidence interval for μ with known σ

Given a sample mean $\bar{x}$, known population standard deviation σ, sample size n, and confidence level c, the confidence interval for μ is:

$$\left[\bar{x} - z_c \frac{\sigma}{\sqrt{n}},\ \bar{x} + z_c \frac{\sigma}{\sqrt{n}}\right].$$

4.3 Margin of error

Notice that the quantity

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

appears in the general formula for confidence intervals. This distance E tells us the acceptable error bound of our estimate. For that reason, E is called the **margin of error**. Using this terminology, a confidence interval can be written simply as follows.

Margin of error

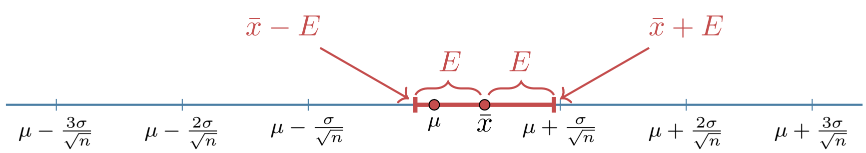
For a confidence interval for μ with known σ, the margin of error is

$$E = z_c \frac{\sigma}{\sqrt{n}}.$$

The confidence interval is

$$[\bar{x} - E, \bar{x} + E].$$

Geometrically, the length E is one-half of the total length of the confidence interval:



Therefore, to find a confidence interval for μ with known σ, we follow these steps:

Step 1: Use the confidence level c to find the critical value z_c.
Step 2: Calculate the margin of error:

$$E = z_c \frac{\sigma}{\sqrt{n}}.$$

Step 3: Write the interval:

$$[\bar{x} - E, \bar{x} + E].$$