Confidence Intervals

We now know how to find outcomes from a sampling distribution that are the most likely, given a particular set of assumptions about the world. The goal now is to use that approach to formalize our uncertainty about estimates of the world.

To make this concrete, let's return to the idea of a categorical character with two values. Let *Xi* be the value of this categorical character. Since there are only two outcomes, *Xi* can be modeled as a Bernoulli random variable with parameter *p*. The parameter *p* represents the fraction of the population with one of the two possible outcomes for this character. For example, if we are interested in how much of a population has received some vaccine, *p* is the fraction of the population that is vaccinated. We don't know what it is; it is our goal to estimate *p*.

Now that we have an individual probability distribution, we can generate a set of expectations based on an assumption of random sampling. If we have a random sample of *n* individuals from the population, then the number of successes in the sample (call it *S*), follows a binomial distribution with parameters *n* and *p*.

$X_i \sim \text{Bernoulli}(p)$ ← Individual Probability

(success or failure)

$S \sim \text{Binomial}(n, p)$ ← Population Probability.

(# of successes)

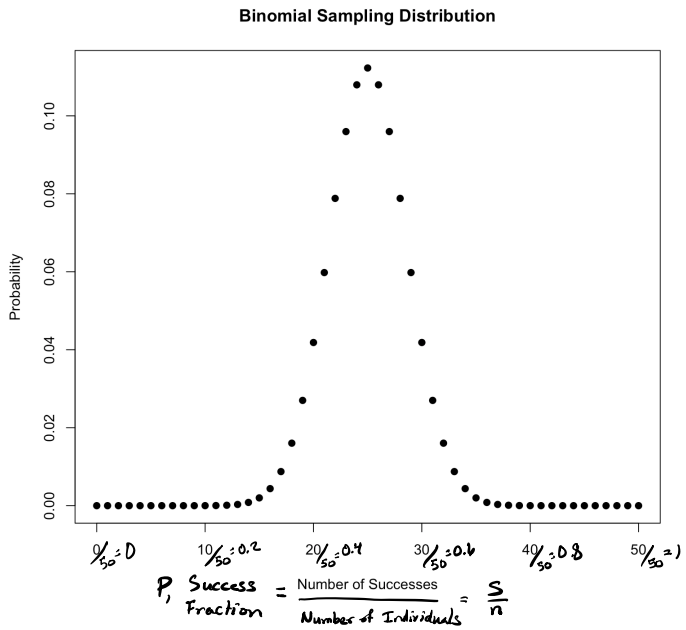
Our goal is to estimate the value of *p* **from a particular outcome of S.** *S* is the number of successes, not a fraction. So something that we would be interested in is *S*/*n*, which we could call *P*.

This means that we can rephrase the probability distribution of *S* as an estimate of the **fraction of successes,** *P*. It looks like this in math

$S \sim \text{Binomial}(n, p) \rightarrow \underbrace{n \cdot P}_{=S} \sim \text{Binomial}(n, p)$

because $\frac{S}{n} = P \rightarrow S = nP$

It looks like this in graphical form.



This is an example with $n = 50$ individuals in the sample and where the probability of success is $p = 0.5$.

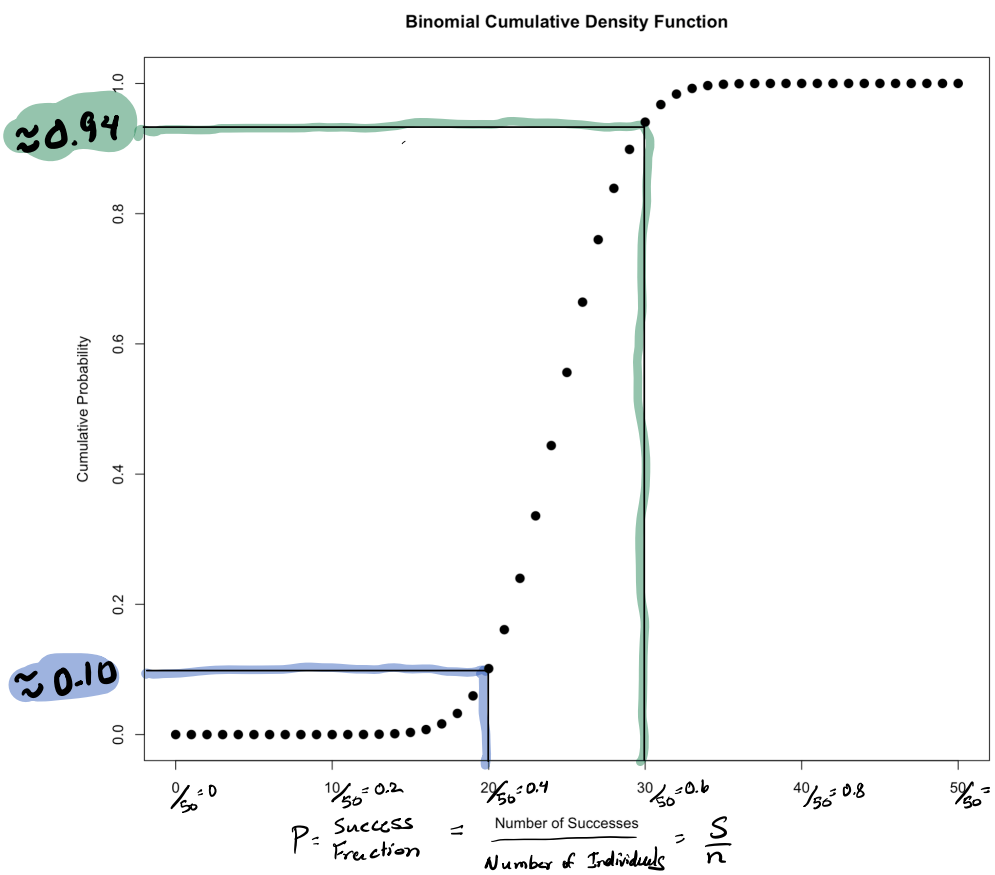
Confidence intervals give a measure of how much our estimate of p might change under repeated sampling. Here, we know the value of $p = 0.5$.

Let's create a "rule" that tells us how far away from p we are for a particular outcome of P. This "rule" corrects for how wrong a particular sample estimate of P is. For example, if we sampled 30 successes in 50, we would be wrong by $P - p = 30/50 - 0.5 = 0.6 - 0.5 = 0.1$. This says we <u>overestimated</u> the probability of success by 0.1 units.

We would also be wrong by 0.1 in the other direction if we sampled 20 successes in our 50 individuals. In that case we would <u>underestimate</u> p by $P - p = 20/50 - 0.5 = 0.4 - 0.5 = -0.1$.

<u>A "rule" that makes up for these errors is to add and subtract 0.1 to our estimate of P.</u>

That is, whatever our estimate of P, an interval that makes up for errors of this sort is $[P - 0.1, P + 0.1]$. Clearly this interval encapsulates the true values of p some of the time, but not other times. We can figure exactly how frequently by asking what the probability is that P is between the values of $p - 0.1 = 0.4$ and $p + 0.1 = 0.6$. We get this probability from cumulative density functions, as it can tell us the probability of being in this range. Here is the cumulative density function for this sampling distribution and the associated cumulative probabilities.



≈0.94

≈0.10

$Pr(P \le 0.6)$

$= Pr\left(\frac{S}{n} \le 0.6\right)$

$= Pr(S \le 0.6 \cdot n)$

$= Pr(S \le 30)$

≈ 0.94

$Pr(P \le 0.4)$

$= Pr\left(\frac{S}{n} \le 0.4\right)$

$= Pr(S \le 0.4 \cdot n)$

$= Pr(S \le 20)$

≈ 0.10

$Pr(p - 0.1 \le P \le p + 0.1)$

$= Pr(0.4 \le P \le 0.6) = Pr(P \le 0.6) - Pr(P \le 0.4)$

$= Pr(S \le 30) - Pr(S \le 20)$

$= pbinom(30, 50, 0.5) - pbinom(20, 50, 0.5)$

$\approx 0.94 - 0.10$

$= 0.84$

Thus, it is the case that this rule of adding and subtracting 0.1 from the fraction of successes in a sample will make up for any errors about 84% of the time. **What we have done is create an 84% confidence interval.** Any *P* we calculate from our sample, we add and subtract 0.1 from it and we end up with an interval that includes the population value of *p* 84% of the time.

Confidence intervals always work this way. **Confidence intervals give a range of values where we can make a probabilistic statement about the chance that the population parameter is in that range.** In this case, we have picked a "rule" and figured out how widely it applies. In other cases, we want to find a role that applies to a specific fraction of all possible samples.

For example, what should the rule be that applies in 95% of samples? To do this, we just do the entire process backwards. We first pick the probability that we care about, say 95%. Then we find the range of most likely outcomes this applies to by finding quantiles corresponding to 97.5% and 2.5% cumulative densities. And the difference of these boundaries from the population parameter *p* gives the "rule" to be applied to a specific estimate of *p*.

Let's find a 95% confidence interval. We know that particular sample outcomes *P* should cover 95% of outcomes of *P*.

$$Pr(q_{0.025} \leq S \leq q_{0.975}) = 0.975 - 0.025 = 0.95$$

These *q* values are quantiles at specific cumulative probabilities. Let's find them using the quantile function in R (which goes from y-axis to the x-axis) on a cumulative probability plot.

$q_{0.025} = qbinom(0.025, size = 50, prob = 0.5) = 18$ (S) $\quad 18/50 = 0.36$ (F)

$q_{0.975} = qbinom(0.975, size = 50, prob = 0.5) = 32$ $\quad 32/50 = 0.64$

How far away are these $\hat{P}$ values from the population value?

$\frac{18}{50} - p = 0.36 - 0.5 = -0.14$ ↖

$\frac{32}{50} - p = 0.64 - 0.5 = 0.14$ ↙

They are 0.14 units away from P.

Now we can create a "rule."

Calculate $\hat{P}$ from a sample.

Add and subtract 0.14 to that value.

95% of samples with this interval will cover the population value p. The probability it doesn't is 5%.

Example

$n = 25$

successes = 10

$\hat{p} = 10/25 = 0.4$

↑ Fraction of successes in sample.

95% confidence interval.

	$\hat{p}$ ↓	# successes	fraction successes
qbinom(0.025, size=25, prob=0.4) =		5	$5/25 = 0.2$
qbinom(0.975, size=25, prob=0.4) =		15	$15/25 = 0.6$

This says that our 95% confidence interval for the true population fraction of successes is [0.2, 0.6]. Only 5% of the random samples we get from this population when we apply this rule will be wrong.

The “rule” here is adding and subtracting 0.2 in this case because 0.2 is 0.2 units away from our sample estimate of 10/25 = 0.4 and 0.6 is 0.2 away from our sample estimate of 10/25 = 0.4.

We have just estimated it directly. If instead we wanted a 99% percent confidence interval, we do the same thing but finding different quantiles.

$q_{0.005} = qbinom(0.005, size=25, prob=0.4) = 4$

$q_{0.995} = qbinom(0.995, size=25, prob=0.4) = 16$

In fraction of successes, this is

$$\left[\frac{4}{25}, \frac{16}{25}\right] = [0.16, 0.64]$$

↖ 99% CI for p.

The rule here is to add and subtract 0.24. The interval is larger because it includes a larger set of possible samples we could get.

Estimation of the Mean

A common thing that biologists are interested in is the mean of a population. The mean represents the central tendency of a distribution and so is often the first thing under study.

How do we estimate the mean? Well, we can estimate the mean of a sample, which, we call “x bar” and write as

$$\bar{X}$$

The question is, what is the sampling distribution of the sample mean? Once we have a sampling distribution, we can do what we did above with the binomial distribution to find the quantiles that correspond to a confidence interval with a particular alpha level.

For the mean, we can rely on a fundamental result in probability theory called **the central limit theorem (CLT)**. The central limit theorem is the justification for what seems like an unwarranted focus in statistics on the normal distribution. You might ask “where does the normal distribution come from? Who decided we should have a distribution with a mean and a variance? Why?” The central limit theorem gives the answer.

Here it is, in its most simple form.

<u>The Central Limit Theorem</u>
For very many independent and identical random variables, the sum of these variables is normally distributed.

This result is so fundamental in probability theory that it is how normal distributions were derived.

Why is this valuable? It’s valuable because the sample mean is essentially a sum of many random variables. And if you have sampled randomly, then they are by definition independent and identically distributed random variables. It doesn’t matter what the population distribution looks like. It could be continuous or discrete. It could be skewed. It could be multimodal. It could be odd shaped. None of that matters. If there are enough random variables, their sum is distributed like a normal distribution.

Why does this matter for statistics and estimation of the mean? The Central Limit Theorem gives us means that we can always use the normal distribution as a sampling distribution for the sample mean. To make use of this, it's worth spending some time understanding how the normal distribution works.

The Normal Distribution

Normal distributions have outcomes that are continuous. This is unlike the binomial distribution that has discrete numerical outcomes (# of successes) and unlike the Bernoulli distribution that has only categorical outcomes. Like Binomial and Bernoulli distributions, it has some parameters. The parameters that are needed to describe a normal distribution are the mean and variance.

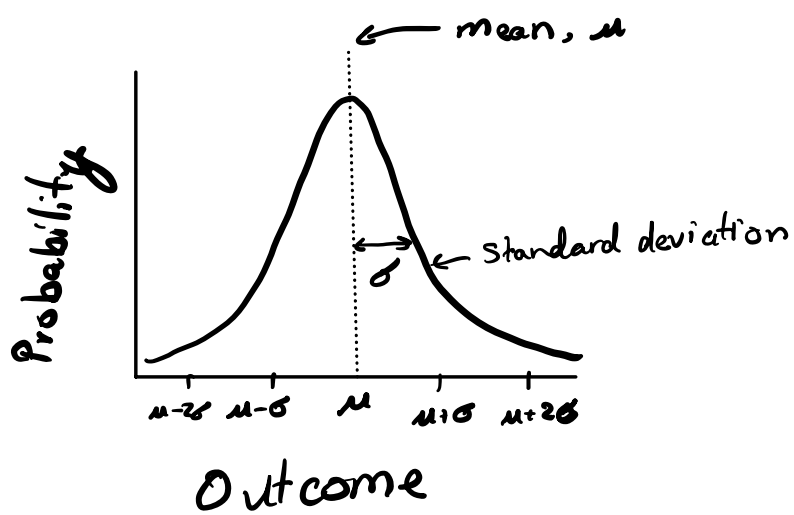
We say a distribution is normally distributed like this.

$$X \sim \text{Normal}(\mu, \sigma^2)$$

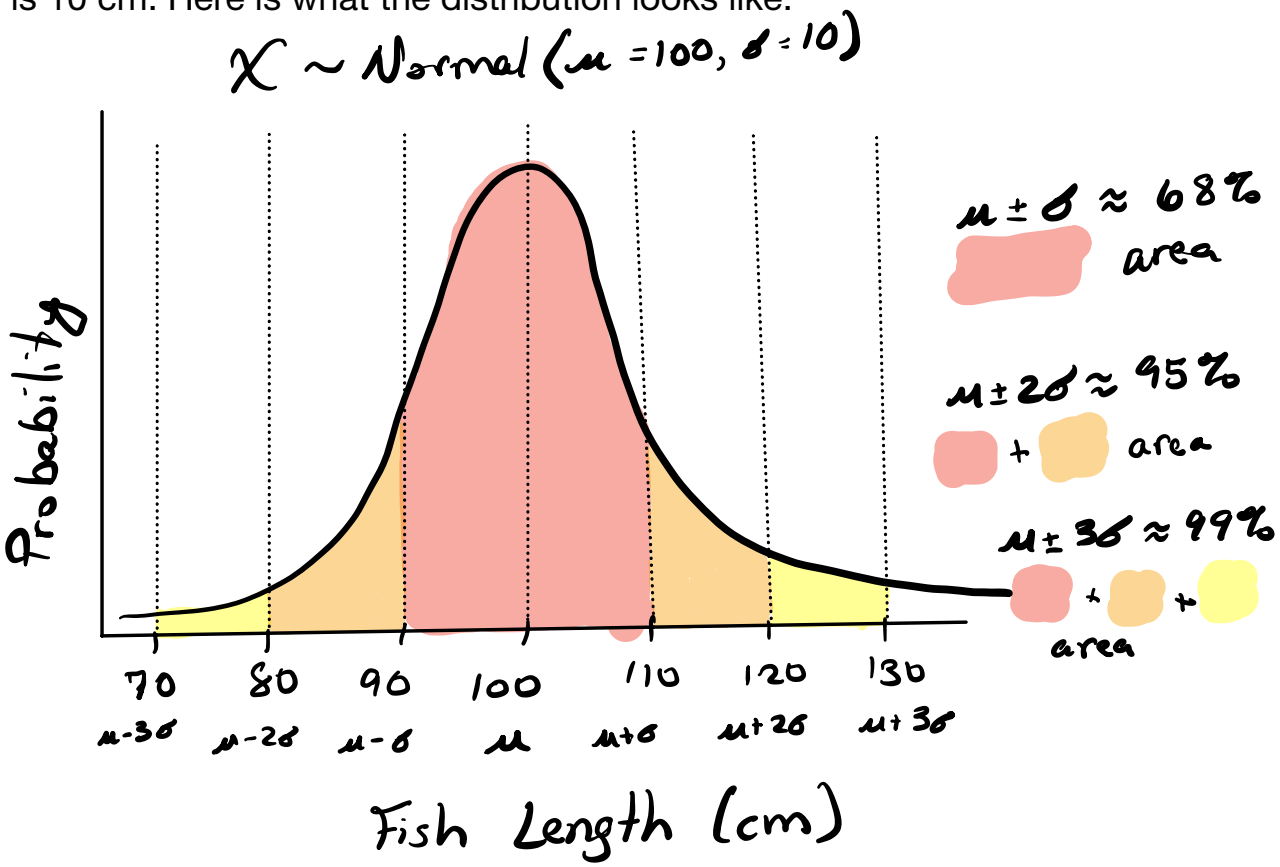
μ: mean; σ^2: variance

$\sigma = \sqrt{\sigma^2}$ is the standard deviation.

Normal Distributions look like this.



One example of a feature of an individual that is often modeled using a normal distribution is size. Consider the case of fish length. Assume the mean length of fish in some population is 100 cm and the standard deviation is 10 cm. Here is what the distribution looks like.



To understand how to interpret the normal distribution, it's helpful to think about how much probability is in different ranges. The normal distribution is so common in science, it's worth remembering these ranges.

About 68% of the outcomes fall within 1 standard deviation of the mean.
About 95% of the outcomes fall within 2 standard deviations of the mean.
About 99% of the outcomes fall within 3 standard deviations of the mean.

All that really matters for the normal distribution is where an outcome fall relative to the mean. For example, I might say a fish is 2 cm away from the mean. That's pretty close to the population mean for this population. But if another population had a standard deviation of 0.5 cm, then being 2 cm away from the mean is 4 standard deviations away, suggesting a gargantuan fish for that population.

To make these kinds of comparison clear, we use something called a **z-score**. Z-scores measure how different an outcome is from the mean in units of standard deviations. The equation for a z-score is

$$Z = \frac{X - \mu}{\sigma}$$

A z-score of 1 means that the outcome is 1 standard deviation above the mean. A z-score of -1 means that the outcomes is 1 standard deviation below the mean.

A z-score of 2.5 means the outcomes is 2.5 standard deviations above the mean.

This is helpful for in understanding how exceptional a particular outcome is. From the probability density figure above, we can make some qualitative statements about how likely any outcomes are. Outcomes more than 3 standard deviations from the mean are rare and should only occur ~1% of the time. Hence, fish with lengths above 130cm or below 70cm only comprise ~1% of the population. Fish larger than 2 standard deviations above the mean (120 cm) and below 2 standard deviations below the mean (80cm) comprise only about 5% of the population. They are also rare, but slightly less so.

Z-scores are helpful for making comparisons across different systems. Exceptionally big fish (z-scores > 3) and exceptionally small fish (z-scores < -3) have the same z-score regardless of the population you look at (so long as the populations can reasonably be described by a normal distribution).

You can also see why this might be valuable for making confidence intervals. If the sampling distribution is normally distributed, the 95% confidence interval is about 2 standard deviations away from the mean!

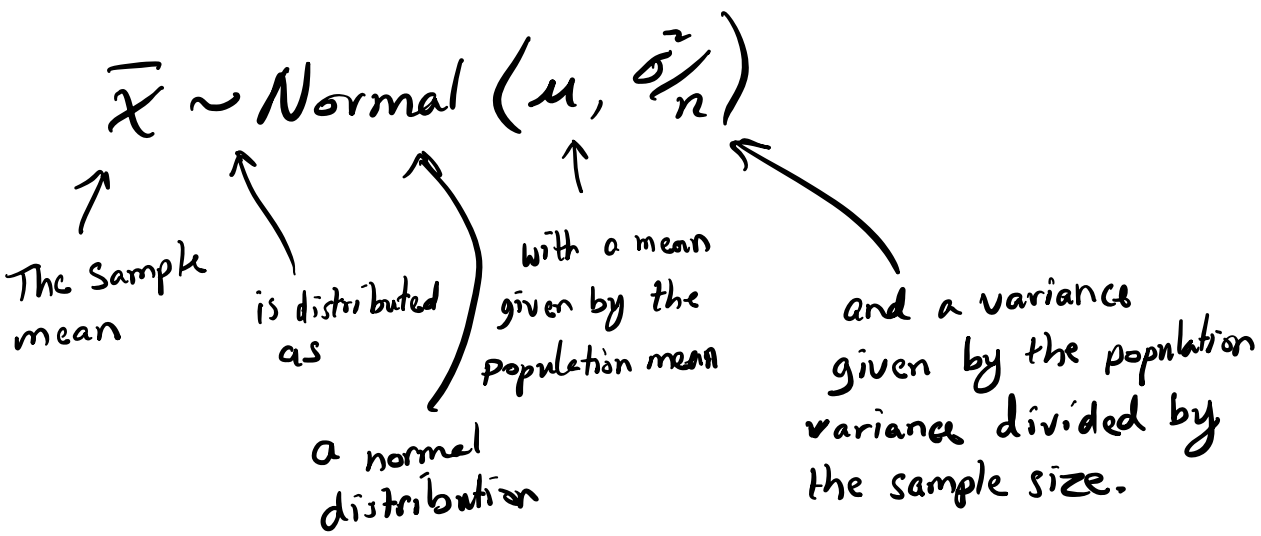
But to make use of this, we need to know the parameters of the sampling distribution of the mean. At this point, we have only stated that is normally distributed. But what are the parameters?

Sampling distribution of the sample mean by the CLT

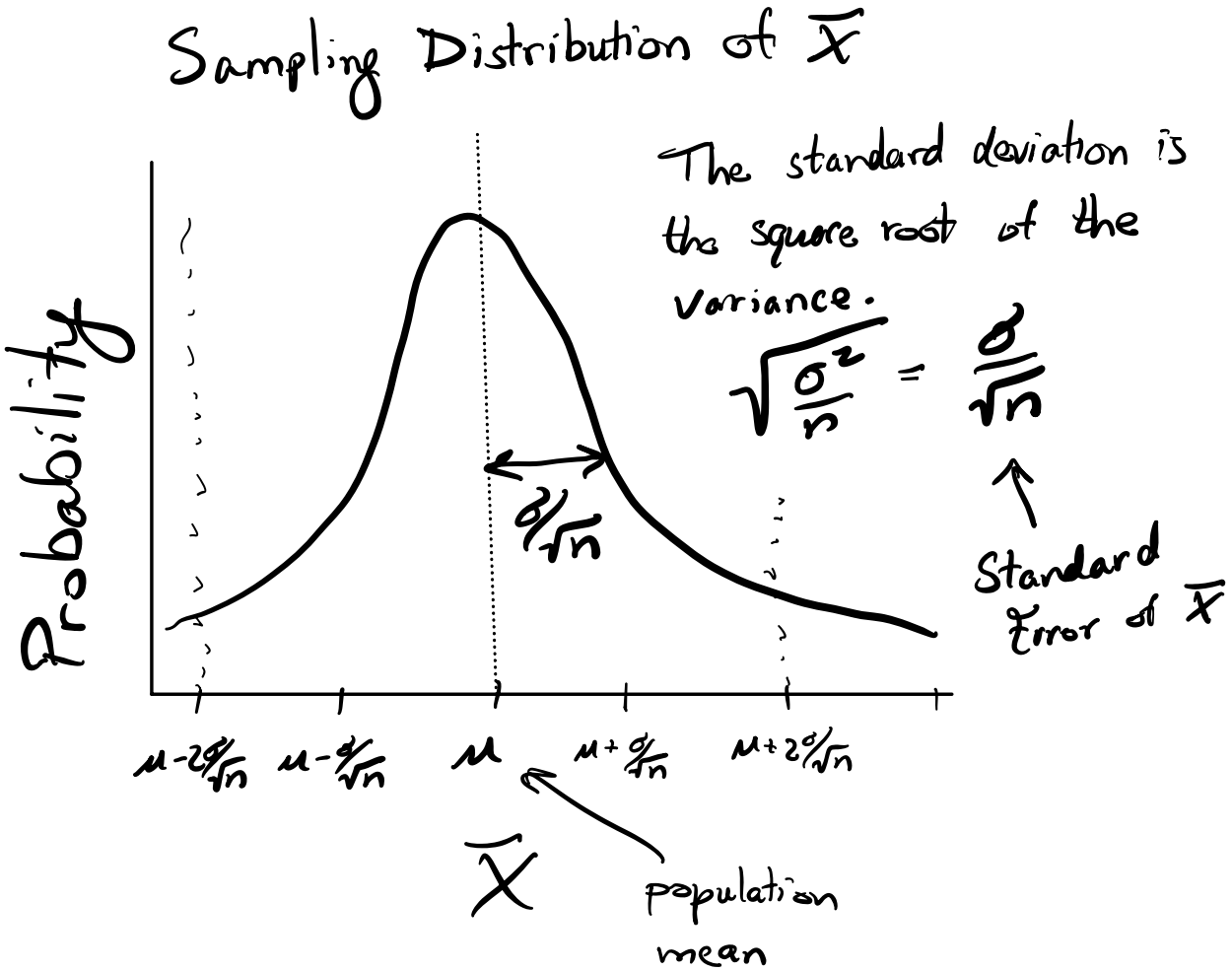
To understand the sampling distribution of the sample mean, let's review the objective.

1. We have a population. The population has a mean, which we call μ.
2. We don't know the population mean, so we sample it. We typically call this sample our "data".
3. That data has a mean. We call this mean the "sample mean" and write it as $\bar{X}$.
4. The sample mean may or may not be close to the population mean. To characterize how certain we are about the relationship between the sample and population mean, we need to know the probability of getting any particular sample mean. We do this with the sampling distribution.

For a given population mean and population variance, the central limit theorem tells us the following:



Here's a visual.



Now that we have a sampling distribution, we can use the properties of the normal distribution to make statements about how likely we are to sample a particular sample mean.

For example, we know that the probability that the sample mean (Xbar) is within 2 standard deviations of the population mean is 95%. That is

$$Pr\left(\mu - 2\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

$\mu - 2\frac{\sigma}{\sqrt{n}}$: z-score

$\mu + 2\frac{\sigma}{\sqrt{n}}$: z-score

We can rewrite this in terms of z-scores as

$$Pr\left(-2 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 2\right) \approx 0.95$$

↑ Definition of z-score

We can turn this into a statement about the distance of the sample mean from the population mean.

$$Pr\left(-2\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

Probability that the sample mean is within 2 standard errors of the population mean is about 95%.

Reminder: The Standard error is simply the standard deviation of a sampling distribution. Since $\bar{X} \sim Normal(\mu, \sigma^2/n)$, $SE_{\bar{x}} = \sigma/\sqrt{n}$.

The only thing that keeps us from getting to an actual confidence interval is the fact that we don't know the population variance (sigma^2). We don't know the population variance because we don't know the properties of the population. Again, this is the whole point of estimation. We want to know the properties of the population.

What we do in this case is just estimate the population variance with the sample variance, which we write as *s*^2. The sample standard deviation is *s*.

This gives us an estimate of the standard error of the sampling distribution of the mean

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

s ← Sample standard deviation

n ← sample size

And now we know that $\bar{x}$ has probability of $\approx 95\%$ of being within $2SE_{\bar{x}}$ of μ.

In math

$$Pr(-2SE_{\bar{x}} < \bar{x} - \mu < 2SE_{\bar{x}}) \approx 0.95$$

which rearranges to

$$Pr(\bar{x} - 2SE_{\bar{x}} < \mu < \bar{x} + 2SE_{\bar{x}}) \approx .95$$

95% Confidence Interval.

⇒ The population mean μ has probability 0,95 of being in the range

$$[\bar{x} - 2SE_{\bar{x}},\ \bar{x} + 2SE_{\bar{x}}]$$

Example

We collect $n=25$ fish and measure their lengths.

We calculate a mean of $\bar{x} = 108\text{cm}$

and a sample variance of $s^2 = 144\text{cm}^2$.

What is the 95% confidence interval?

$$s = \sqrt{s^2} = \sqrt{144\text{cm}^2} = 12\text{cm}.$$

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{12\text{cm}}{\sqrt{25}} = \frac{12\text{cm}}{5} \approx 2.4\text{cm}$$

The 95% confidence interval is approximately

$$\left[\bar{x} - 2 \cdot SE_{\bar{x}},\ \bar{x} + 2SE_{\bar{x}}\right]$$

$$\left[108\text{cm} - 2 \cdot 2.4\text{cm},\ 108 + 2 \cdot 2.4\text{cm}\right]$$

$$\left[108\text{cm} - 4.8\text{cm},\ 108 + 4.8\text{cm}\right]$$

$$\left[103.2\text{cm},\ 112.8\text{cm}\right]$$

What this means is that the probability that the population mean (the thing we are trying estimate) is within the bounds of 103.2 cm to 112.8 cm is approximately 95%.

Taking a step back

Our approach here was to recognize that the sampling distribution of the sample mean (Xbar) follows a normal distribution and that this normal distribution has a standard deviation given by the equation

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

We then used an approximate property of the normal distribution to find 95% of the most likely outcomes. This is that **approximately** 95% of the probability is within 2 standard deviations of the mean. This is typically called the **2 standard error rule-of-thumb**, because it gets you pretty close to the 95% confidence interval.

To find the **exact** 95% percent of most likely outcomes (or any other outcomes), we would need to use the cumulative distribution function to find the appropriate quantiles.

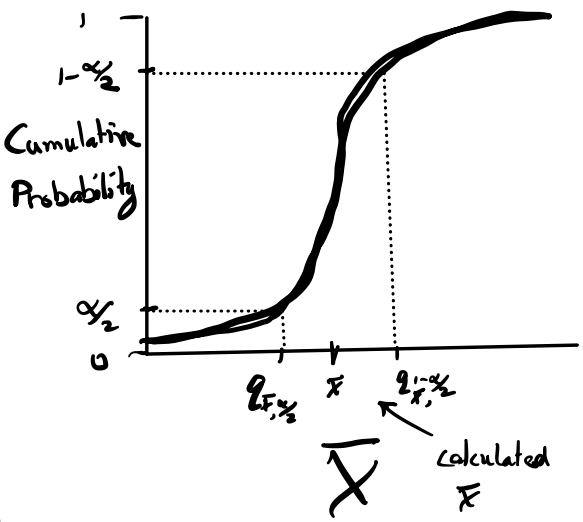
Let's first start by making a statement about how comfortable we are being wrong. When we make a 95% confidence interval, we are saying there is a 5% chance the population parameter is *not contained in the interval*. That is, there is a chance we are wrong, and that chance is 5%. That is our level of tolerance for being wrong. We call this probability of being wrong an ***alpha-level*.**

Let's be general and specify that we are willing to be wrong with probability alpha. To find the 1-alpha most likely outcomes from a normal distribution, we have two approaches. The first is to specify the distribution in R and just ask for the appropriate corresponding quantiles. The other is to look for quantiles of a standardized normal distribution, i.e., the z-scores that correspond to 1-alpha probability of most likely outcomes. Here are both.

Finding $1-\alpha$ confidence intervals

Direct Approach

1. State α
2. Calculate $\bar{X}$ an $SE_{\bar{X}} = \frac{s}{\sqrt{n}}$
3. Find quantiles of interest.



$q_{\bar{X},\alpha/2}$ ← $\alpha/2$ quantile of $\bar{X}$

$q_{\bar{X},1-\alpha/2}$ ← $1-\alpha/2$ quantile of $\bar{X}$

$$qnorm\left(\alpha/2, mean=\bar{X}, sd=SE_{\bar{X}}\right) = q_{\bar{X},\alpha/2}$$

$$qnorm\left(1-\alpha/2, mean=\bar{X}, sd=SE_{\bar{X}}\right) = q_{\bar{X},1-\alpha/2}$$

Standardized Approach

1. State α level.
2. Calculate z-score quantiles corresponding to $1-\alpha$ of most likely outcomes.

$$Z = \frac{\bar{X}-\mu}{SE_{\bar{X}}}$$

$$qnorm(\alpha/2) = z_{\alpha/2}$$

$$qnorm(1-\alpha/2) = z_{1-\alpha/2}$$

$$Pr\left(z_{\alpha/2} < \frac{\bar{X}-\mu}{SE_{\bar{X}}} < z_{1-\alpha/2}\right) = 1-\alpha$$

$z_{\alpha/2}$ ← $\alpha/2$ quantile of standardized normal

$z_{1-\alpha/2}$ ← $1-\alpha/2$ quantile of standardize normal.

3. Use quantiles to go from z-scores to confidence interval

The z-score quantiles have this property.

$$Pr\left(z_{\alpha/2} < \frac{\bar{X}-\mu}{SE_{\bar{X}}}\right) = \alpha/2 \longrightarrow Pr\left(\mu < \bar{X} - SE_{\bar{X}} z_{\alpha/2}\right) = \alpha/2$$

Rewrite this part as

$$SE_{\bar{X}} z_{\alpha/2} < \bar{X} - \mu \rightarrow \mu < \bar{X} - SE_{\bar{X}} z_{\alpha/2}$$

$$Pr\left(z_{1-\alpha/2} < \frac{\bar{X}-\mu}{SE_{\bar{X}}}\right) = 1-\alpha/2 \longrightarrow Pr\left(\mu > \bar{X} - z_{1-\alpha/2} \cdot SE_{\bar{X}}\right) = 1-\alpha/2$$

Together, this says that

$$Pr\left(\bar{X} - z_{1-\alpha/2} \cdot SE_{\bar{X}} < \mu < \bar{X} - z_{\alpha/2} \cdot SE_{\bar{X}}\right) = 1-\alpha.$$

$1-\alpha$ confidence interval.

Assumptions involved in using the Central Limit Theorem

Using this approach to finding confidence intervals involves some assumptions. The main one is that you have enough samples so that the sum used to create a mean is “very many” in the central limit theorem.

How many is “very many”? That depends on how close the population distribution is to a normal distribution. If it’s pretty close, very many can actually be quite small, such as 30 individuals. If it’s very different from a normal distribution, then very many may be hundreds.

The main take away here is that applying the Central Limit Theorem is really an approximation. It’s an approximation that works very well with large samples and may not work so well for small samples. To deal with small samples, we will use a slightly different approach.