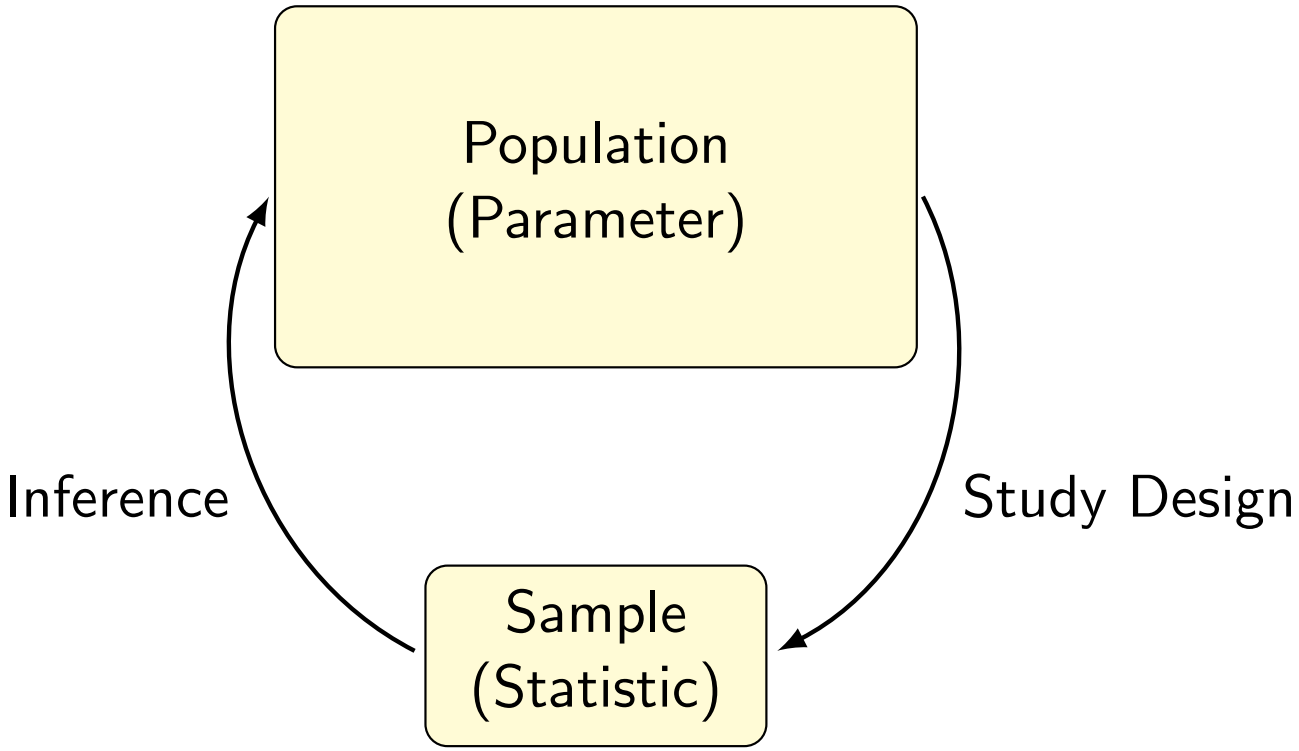
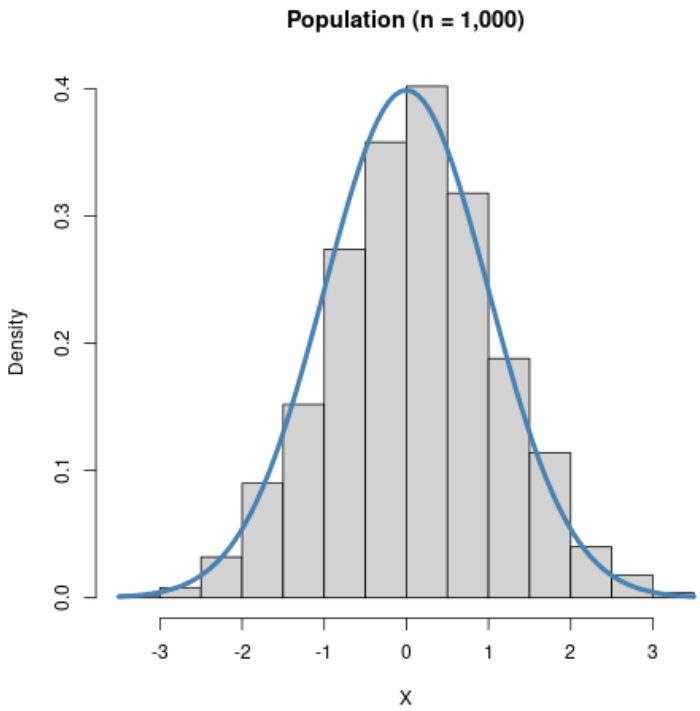
Sampling Distribution

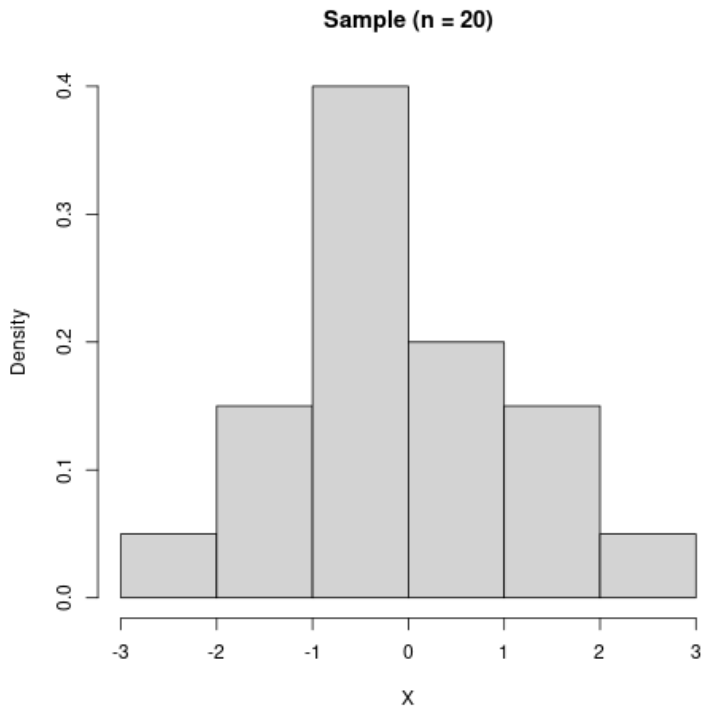
February 16, 2021

Sampling Process



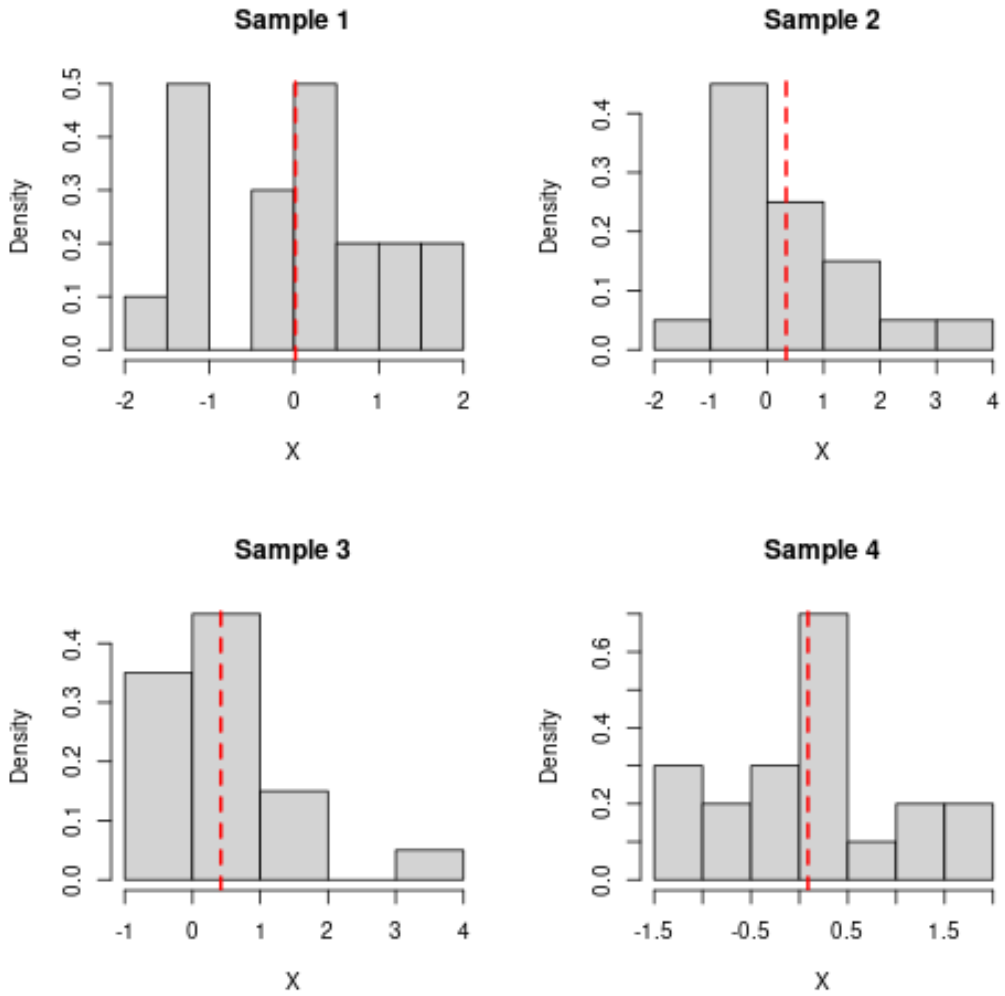
Distribution → Population → Sample





More Samples ($n = 20$)

Each random sample will have a different sample mean, $\overline{x}$



Sample Size and Sampling Distribution

Suppose we have a population of size $N = 1,000$ with $X \sim N(\mu, \sigma^2)$, and we have a sample of size n.

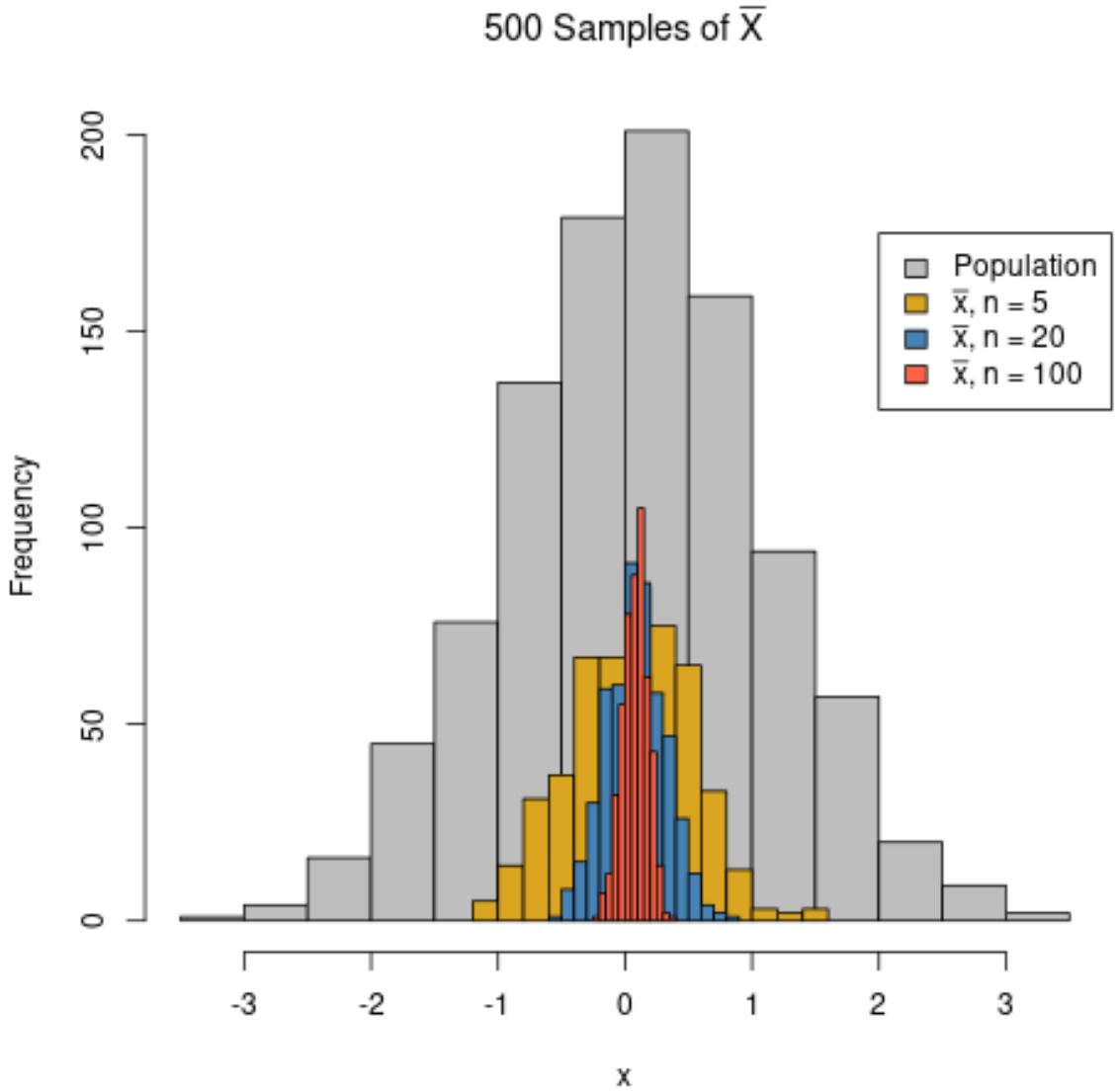
- If $n = 1,000$, i.e., I sample every person in my population, what would we expect the value of $\overline{X}$ to be? How much variability will there be between samples?
- What if $n = 1$, i.e., if $\overline{X} = X_i$?
- $n = 20$?

In particular, we want to ask ourselves two questions:

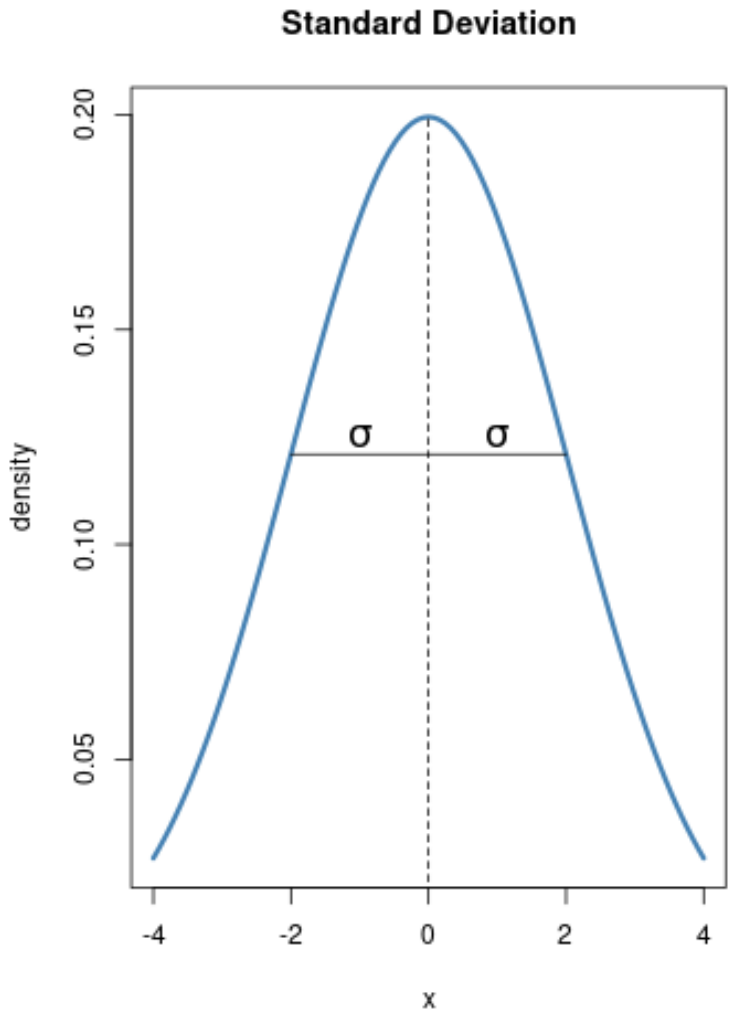
1. What is the expected value of $\overline{X}$? In other words, if I randomly sample $\overline{X}$ a large number of times, what would I expect the average to be?
2. How much variability will we see if we randomly sample $\overline{X}$? How does this relate to sample size?

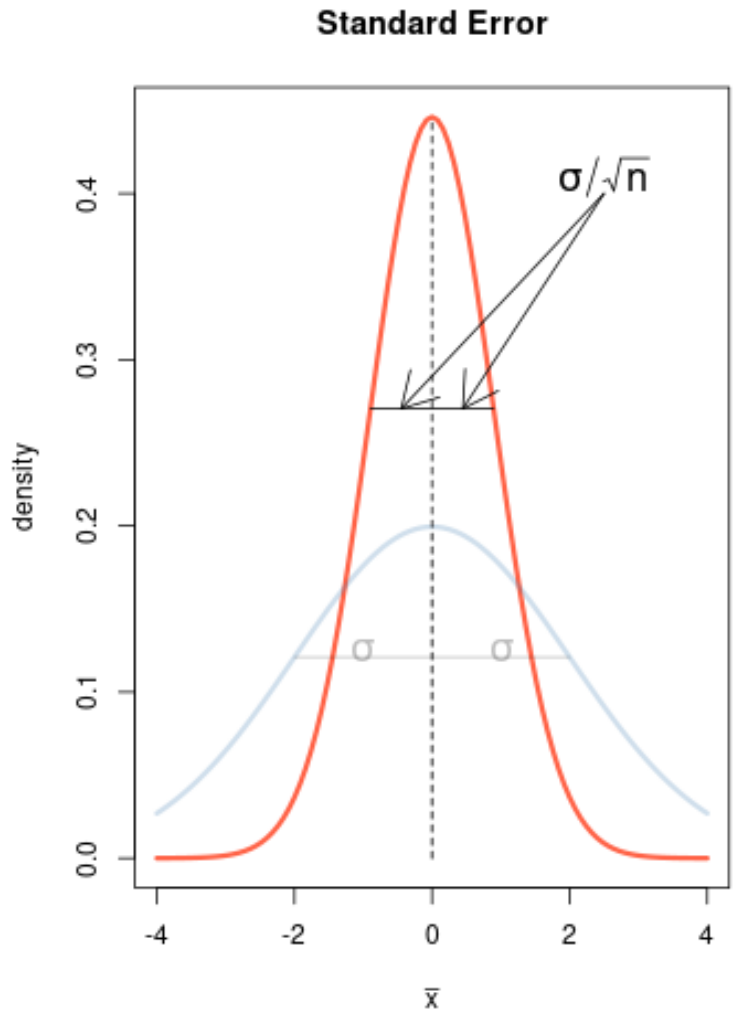
Sampling Distribution

Just as our population follows a distribution, so does our test statistic



Standard Deviation vs. Standard Error





Central Limit Theorem

This is perhaps the most single most amazing thing in the field of statistics.

If a collection of random samples $X_1, X_2, \ldots, X_n$ are drawn from a population that has mean μ and finite variance σ^2, then

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right) \sim N(0,1)$$

In other words, for large enough n, we have a close approximation $\overline{X} \sim N(\mu, \sigma^2/n)$.

N.B. the distribution $N(0,1)$ is important enough that it has its own name: we call this a *standard normal distribution*, written $Z \sim N(0,1)$.

Frequentist Paradigm

Frequentist statistics assumes that our population parameters are fixed but unknown quantities, i.e., they are not random. It does not make sense, for example, to say: "There is a 80% probability that μ is equal to 0". It either is or it isn't.

Probabilistic statements are made with the assumption that an experiment can be repeated an infinite number of times.

Questions we can ask are more akin to: "If we do this experiment an infinite number of times, with what frequency will our statements about it be correct?"

It is often in this sense that we talk about the *expectation* of a random variable. When we say $E(\overline{X}) = \mu$, we mean that the more frequently we sample $\overline{X}$, the closer it's long run average should be to μ.

Quick Recap

$X \sim N(\mu, \sigma^2)$, where μ and σ^2 are fixed but unknown quantities.

We hypothesize about the value of μ, $H_0 : \mu = \mu_0$. Collecting a sample x of n independent observations, we compute the statistic $\overline{x}$ and use it as evidence in our consideration of H_0

There are two mistakes we can make:

- Type I error (α): We could incorrectly *reject* H_0 when $\mu = \mu_0$
- Type II error (β): We could *fail to reject* H_0 when $\mu \neq \mu_0$

Our primary concern right now is going to be to controlling Type I error. (how are we going to do that?) That is, if $\mu = \mu_0$, we want to be sure we do not incorrectly reject. (why type I)

Controlling Error with size of (L, U)

As $\overline{X}$ is random, we should rarely suspect that it will be exactly equal to μ. What perhaps makes more sense is to offer an interval of plausible values centered around $\overline{X}$ (why centered?), say, (L, U)

At one extreme, our interval is just a point, the shortcomings of which were noted above. However, at the other extreme, if our intervals are excessively large, are we really saying much about the value of μ?

Here, again, is the tension between Type I and Type II error. Whether or not μ_0 falls in our interval (L, U) will determine if we either reject or fail to reject H_0.

As our focus for now is in controlling the Type I error at level α, we will posit that however we construct our interval, it should contain the true value of μ (1-α)% of the time.

(L, U) in Frequentist Paradigm

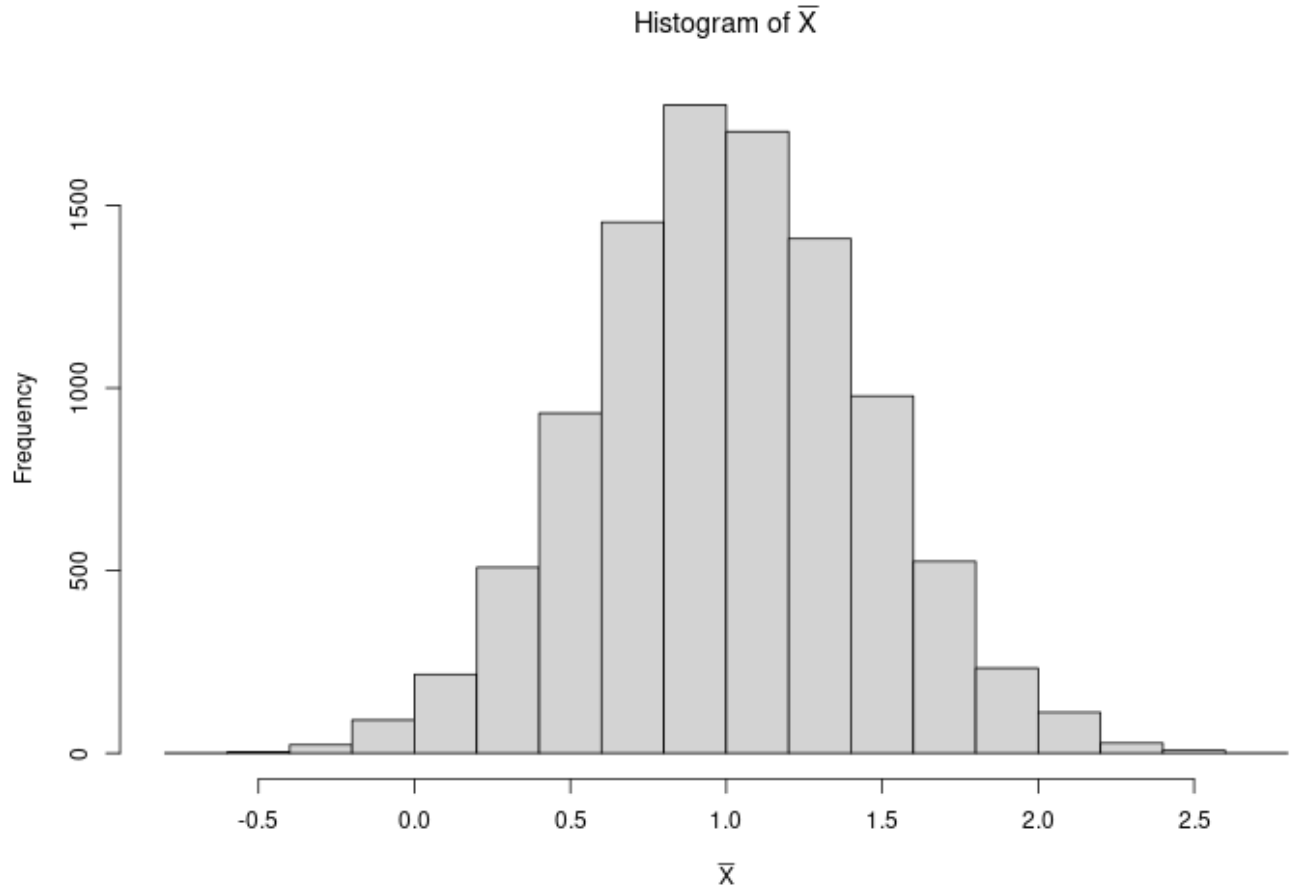
Under the frequentist paradigm, we can restate our objective as such:

Let's assume that we can sample a value $\overline{X}_i$ an infinite number of times, and for each sample, we will construct an interval (L_i, U_i). To control the Type I error rate at level α, we will construct this interval in such a way that $(1 - \alpha)\%$ of our constructed intervals will contain the true value of μ.

The question then becomes how to best determine the values for (L, U).

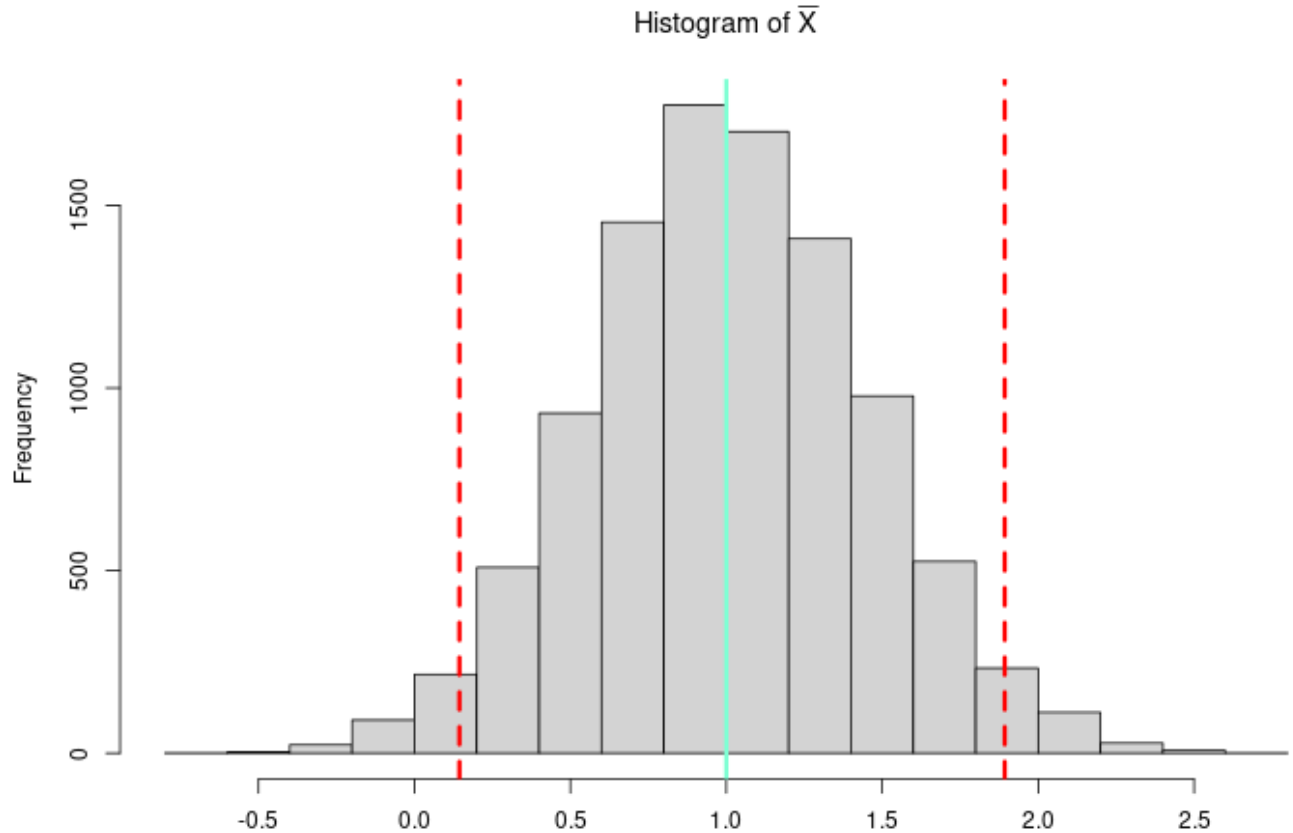
Simulation of $\overline{X}$

While we are limited in real life, we can often investigate these types of questions through simulation. Here, let's assume $X \sim N(1, 4)$ (with σ^2 known), and we will compute the mean from a sample of size $n = 20$. We will repeat this experiment $N = 10,000$ times.



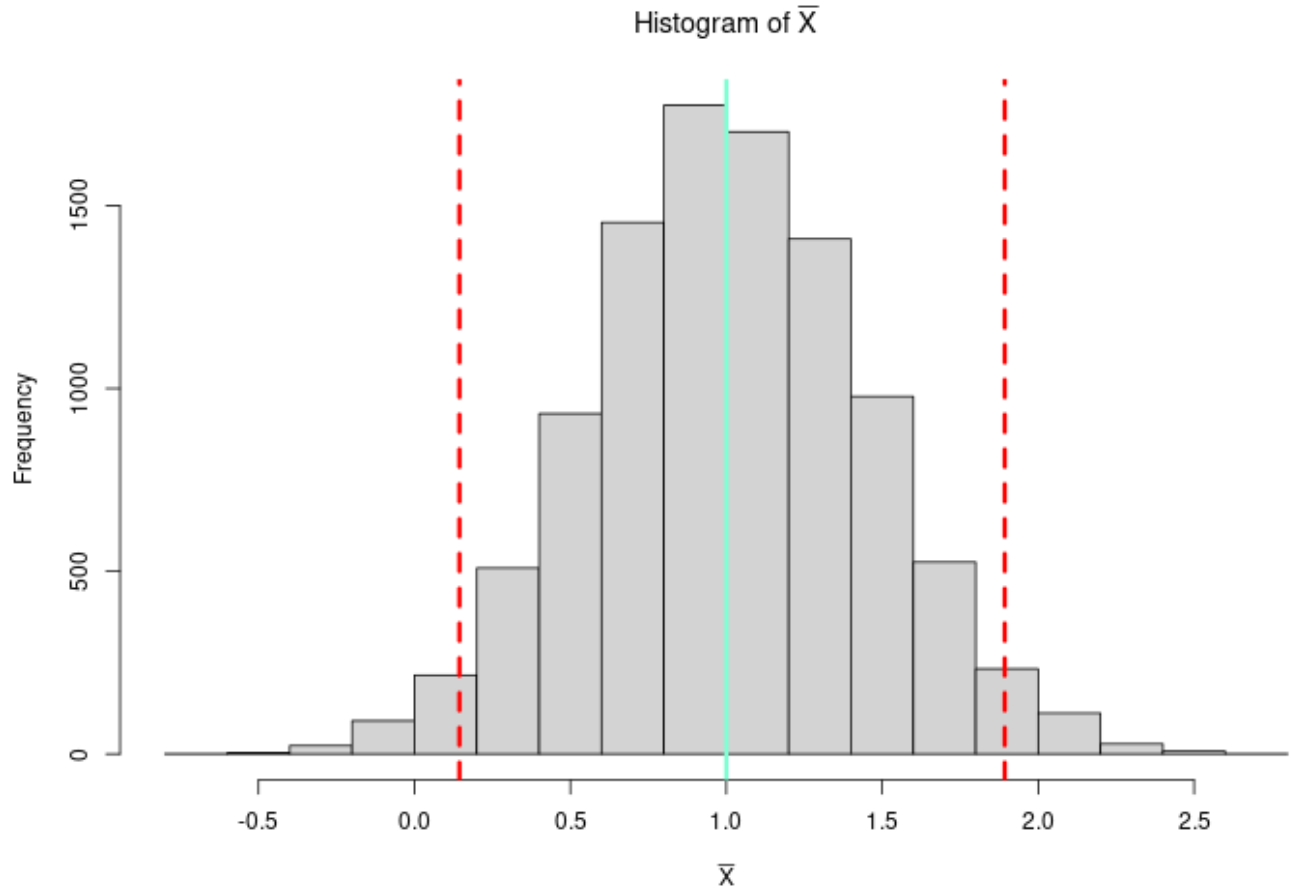
Quantiles of $\overline{X}$

Since we expect that $E(\overline{X}) = \mu$, we know that most of our samples should be pretty close. If we allow a Type I error rate of $\alpha = 0.05$, it is not unreasonable to then consider the interval into which 95% of our samples fall, bounded by the 2.5% and 97.5% quantiles. For our example here, this creates the interval $(0.1231, 1.8725)$.



Quantiles of $\overline{X}$

Since we expect that $E(\overline{X}) = \mu$, we know that most of our samples should be pretty close. Using the 2.5% and 97.5% quantiles of our sample, we can quickly find the interval where 95% of our samples for $\overline{X}$ fall. For our example, this creates the interval $(0.14425, 1.8927)$.



An analytic approach

Unfortunately, we are rarely in a situation where we can collect thousands of independent samples and are instead left with a single value of $\overline{X}$.

Recall, however, from the CLT that

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right) \sim N(0,1)$$

Let's substitute our expression above so that $Z = \lim_{n\to\infty} \sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right)$. As Z is a standard normal, it will prove very easy for us to work with.

An analytic approach cont.

Let's say we want to find an interval for Z such that, when randomly sampled, $(1-\alpha)\%$ of our observations lie in this interval. As this is symmetric about zero, we might say we are looking for the value z_α such that

$$P(-z_\alpha \leq Z \leq z_\alpha) = 1 - \alpha$$

Being that $Z \sim N(0,1)$, we know everything there is to know about Z; computing these values is relatively simple to do on a computer for any given value of α.

Intervals from CLT

With this in place, let's make another substitution

$$
\begin{aligned}
1-\alpha &= P(-z_\alpha \leq Z \leq z_\alpha) \\
&= P\left(-z_\alpha \leq \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right) \leq z_\alpha\right) \\
&= P\left(-z_\alpha \frac{\sigma}{\sqrt{n}} \leq \overline{X}-\mu \leq z_\alpha \frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\overline{X}-z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}+z_\alpha \frac{\sigma}{\sqrt{n}}\right)
\end{aligned}
$$

And violá, we have an interval $\overline{X} \pm z_\alpha \frac{\sigma}{\sqrt{n}}$ that will contain μ $(1-\alpha)\%$ of the time.

CLT in Practice

It's important to note what exactly is being stated regarding the intervals on the previous slide. Specifically, what is presented is a method for computing the interval (L, U) such that, if done a large number of times, $(1 - \alpha)\%$ of those will contain the true interval

If we were to substitute $\overline{X} = 1$ and $\sigma = 2$, we would get the interval $(0.1235, 1.8765)$. As the number of simulations increases, our empirical interval should approach this theoretical one

In actuality, we will often only have one value of $\overline{x}$ from which to compute this interval. Doing so with the first observation from the simulation gives the interval $(-0.2204, 1.5326)$

How do we explain what is going on here?