Randomness and Distributions

Collin Nolte

January 25, 2022

Review

In the previous class, we considered topics in the domain of descriptive statistics:

- Measures of centrality
- Measures of dispersion
- Measures of association
- Histograms, box plots, and scatter plots

Random Processes

To say that something is "random" is to say that its outcome is not predictable

While this is true in itself, it obscures a key pillar of statistics, namely that randomness itself is not necessary random

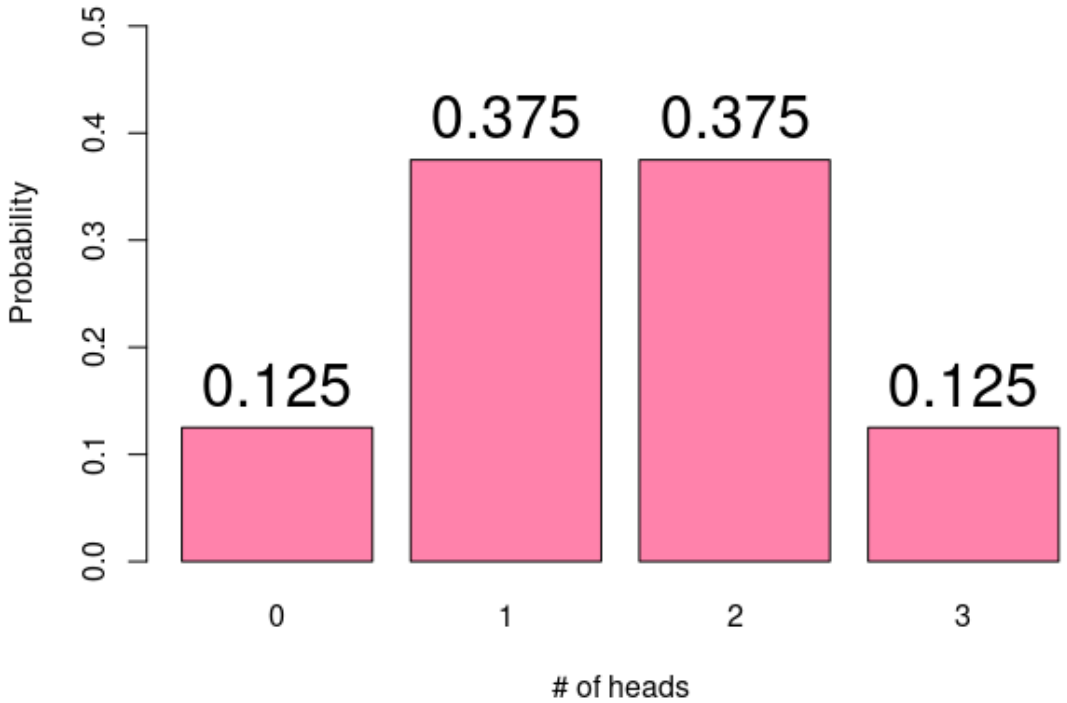
When flipping a coin, for example, while we may not be able to predict the outcome of a *particular* flip, we do generally feel safe in asserting that the coin can only take the values of heads or tail, each with equal probability

What we have here is an example of a *random process*, a situation in which individual events themselves are unpredictable, yet the distribution of outcomes over repeated events follows a predictable pattern

Flipping Coins

For example, suppose we consider flipping an individual coin three times. Here, there are a finite number of possible outcomes:

$$\mathcal{S} = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$$



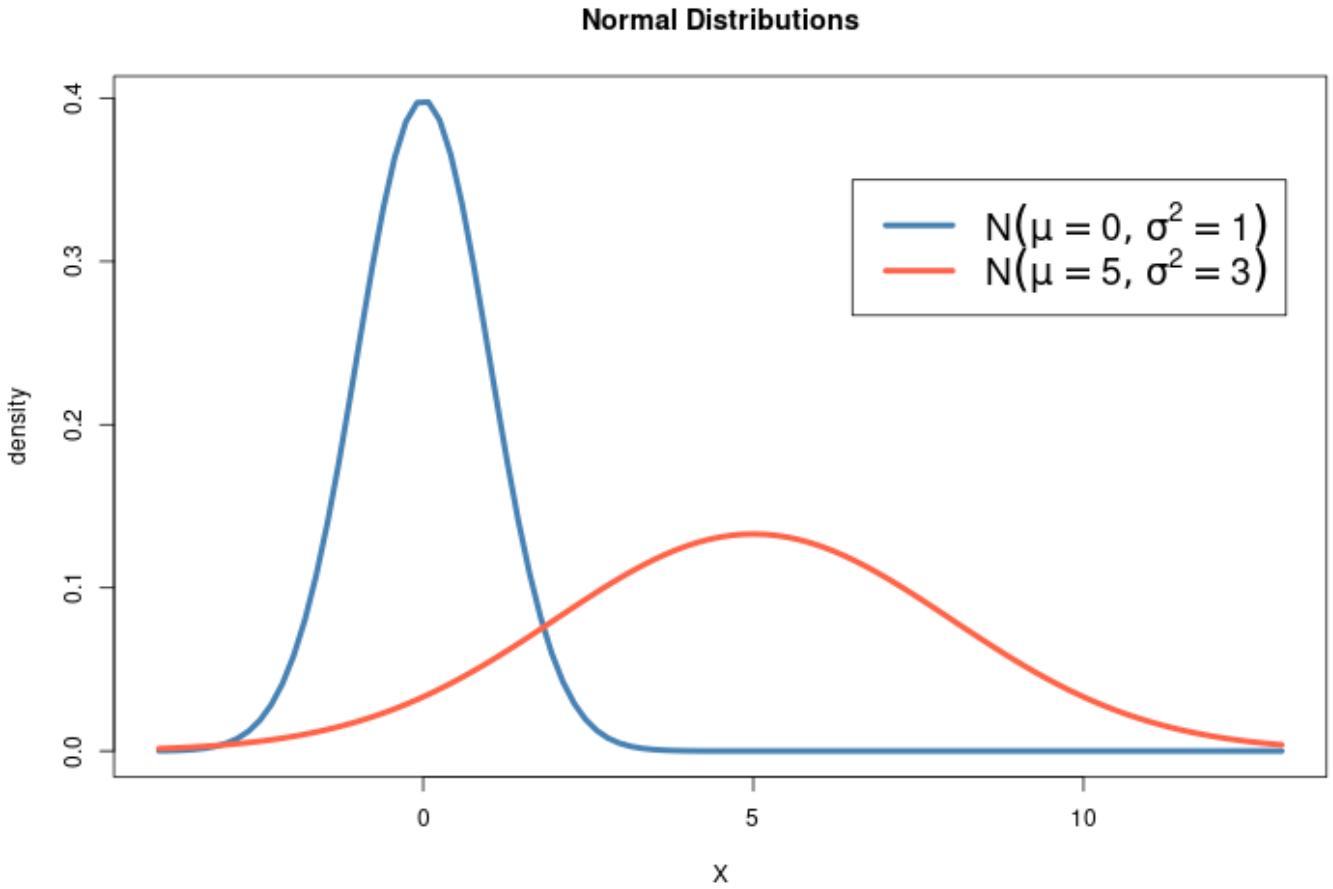
Distributions

At its core, a *distribution* is a functional representation of a random process, taking "events" as input and returning an associated probability

- Convenient to think of a physical process, i.e., "data generating mechanism"
- Governed by a set of parameters
- Continuous or discrete
- When we say X follows a particular distribution, we mean that our observations were generated by a mechanism following a particular form

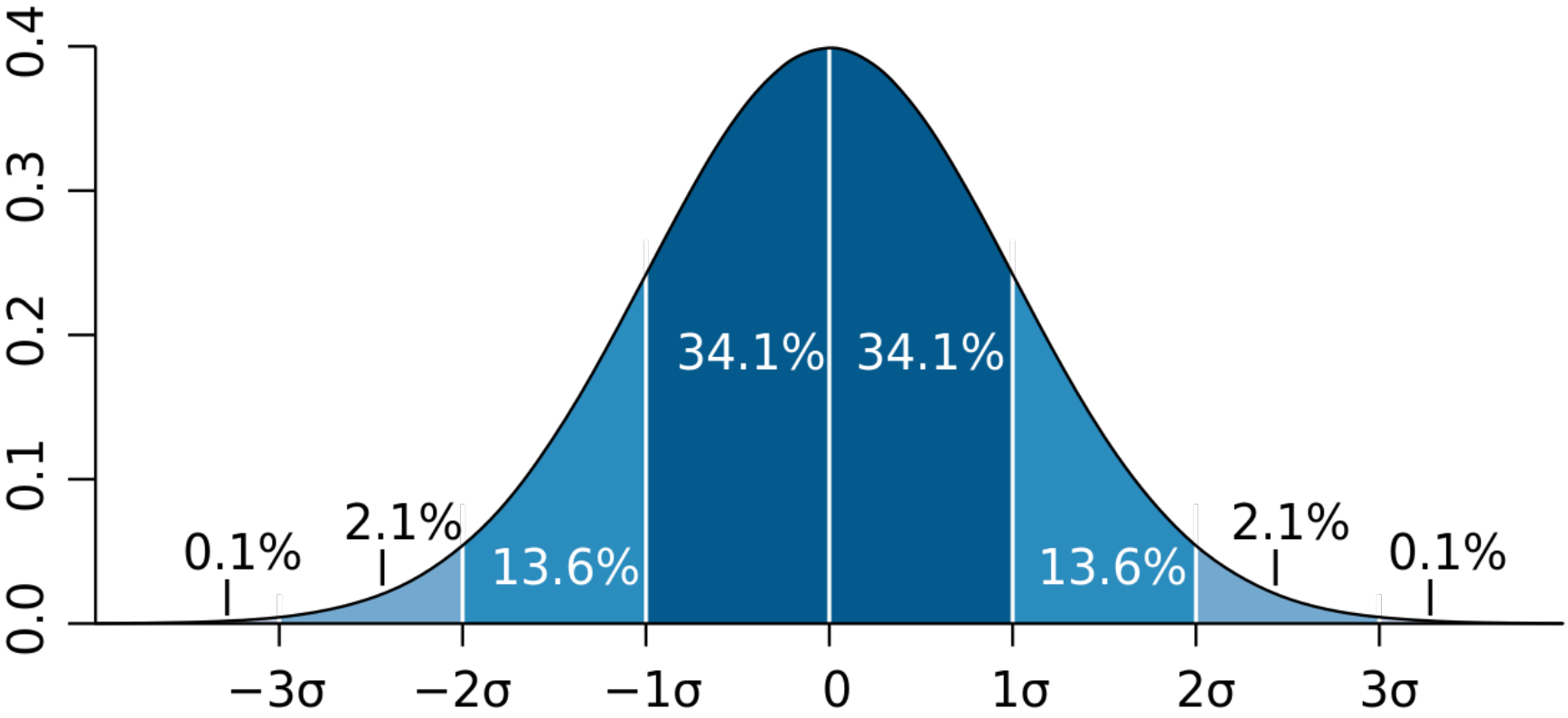
Normal Distribution

$$X \sim N(\mu, \sigma^2) \quad \Longleftrightarrow \quad f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Normal Distribution, cont.

Parameters of a distribution tell us a great deal about centrality and dispersion



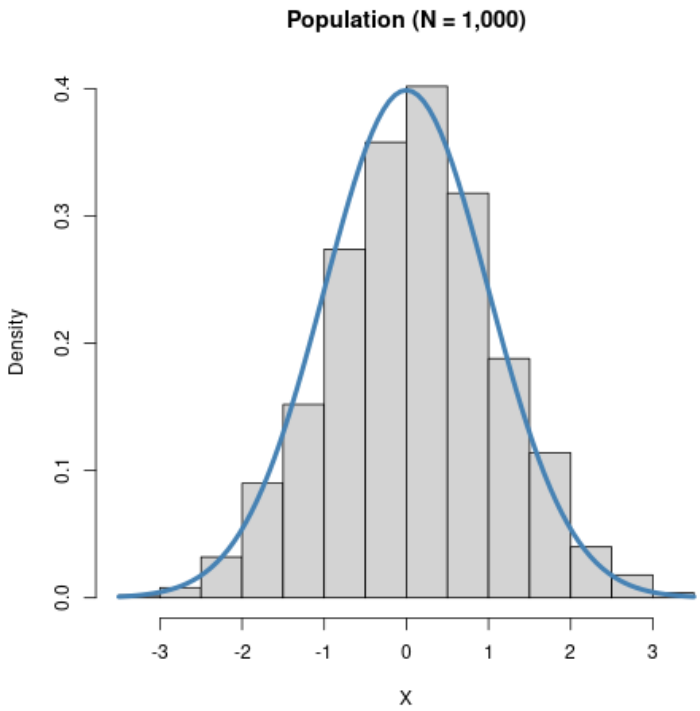
Populations and Samples

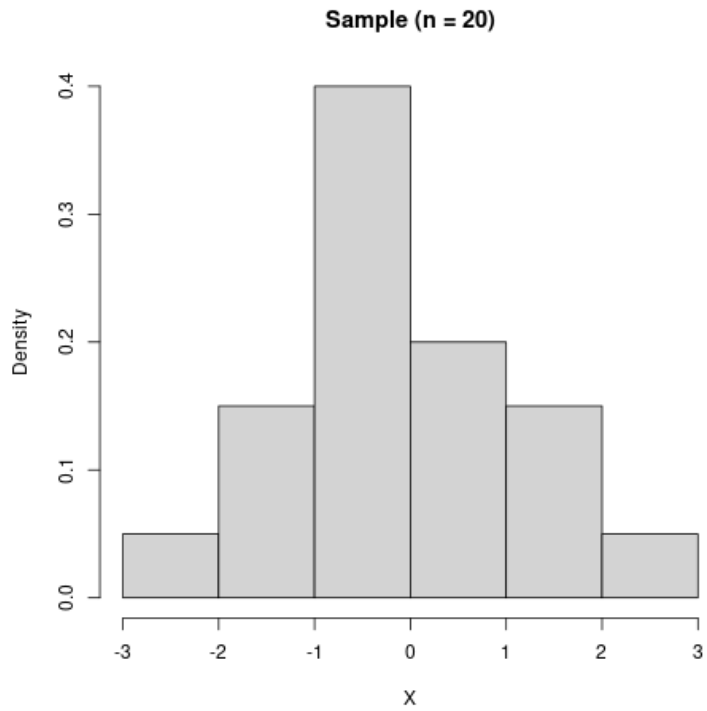
Typically in our research, we are interested in studying and answering questions about a *population*, which consists of all individuals with a particular set of characterisics (age, sex, disease status)

As studying all members of a particular population is usually prohibitive, we limit study to a *random sample* of the population

Done correctly, the makeup of the sample should match that of the population. This allows statements about a sample to be generalized to the population as a whole

Distribution → Population → Sample





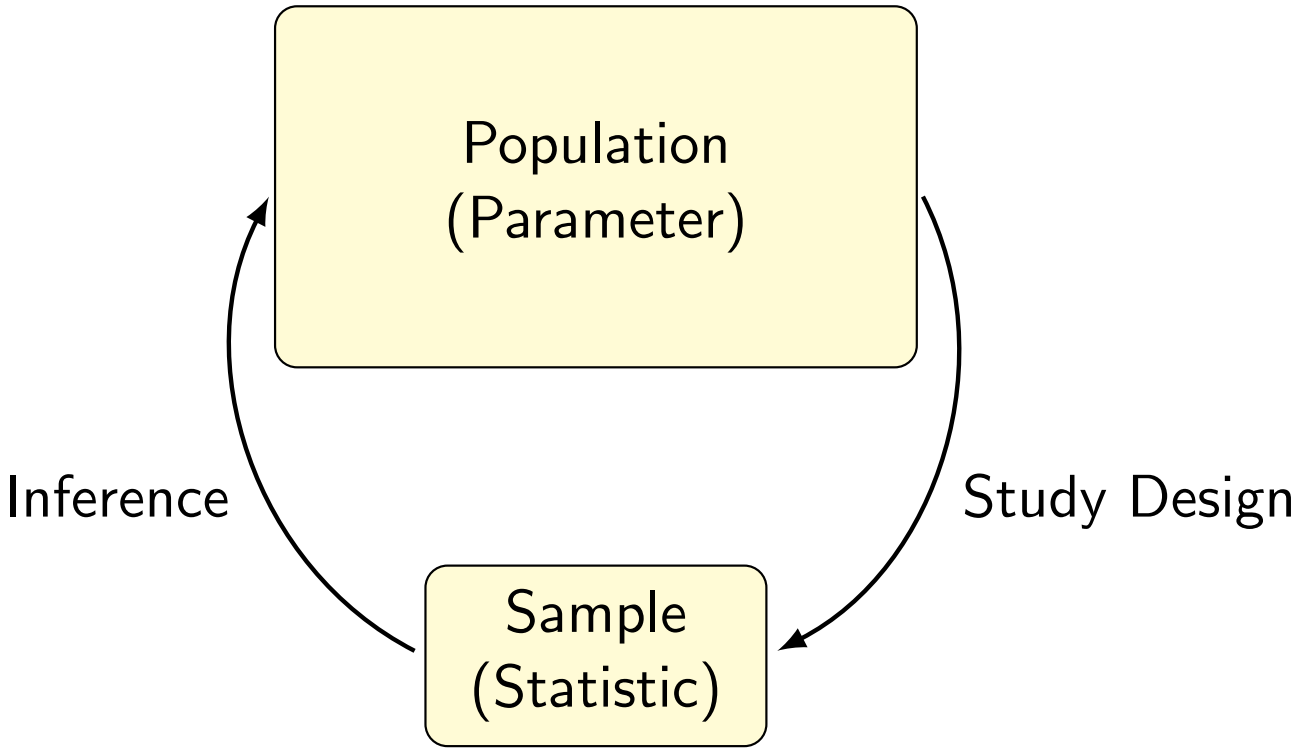
Populations and Samples, cont.

While there may be a number of things that we wish to know about the population, it usually involves estimating a *population parameter*, a numeric quantity associated with a population

For example, suppose we are interested in knowing the average height of the student population at the University of Iowa

Rather than measuring each individual, we collect a random sample from the population and use the average height of the sample as our estimate of the population parameter. The estimate derived from the sample is called a *statistic*

The Statistical Framework



Sampling as a Random Process

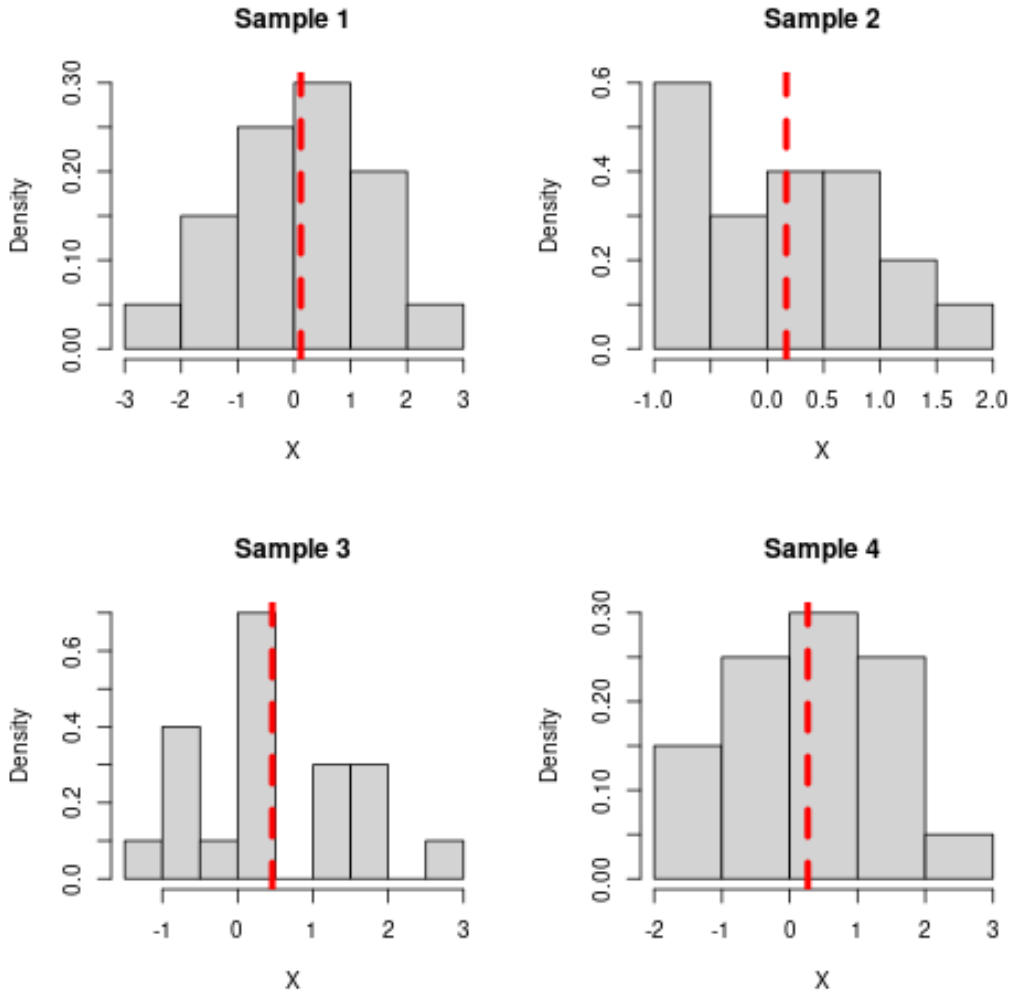
Usually, we are limited to collecting a single random sample from a population. As a consequence, we are left with a single statistic to estimate our parameter

It's important to understand *the sampling process itself* to be a random process. That is, while the makeup of a particular sample is itself random and unpredictable, the process of sampling results in a pattern that is predictable

Our primary goal in this class is to identify this pattern and leverage its properties to make inference

Samples are Random ($n = 20$)

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



Sample Statistics

As we saw in the previous slide, the values of $\overline{x}$ computed for each sample are random, so their values are slightly different

In light of this, it would be helpful to know about the *distribution* of our sample statistic. Knowing the distribution of a sample statistic will tell us two critical pieces of information:

1. Where is the sampling distribution centered? (mean)
2. How dispersed is the sampling distribution about its center? (variance)

Central Limit Theorem

This is perhaps the most profound result in the field of statistics

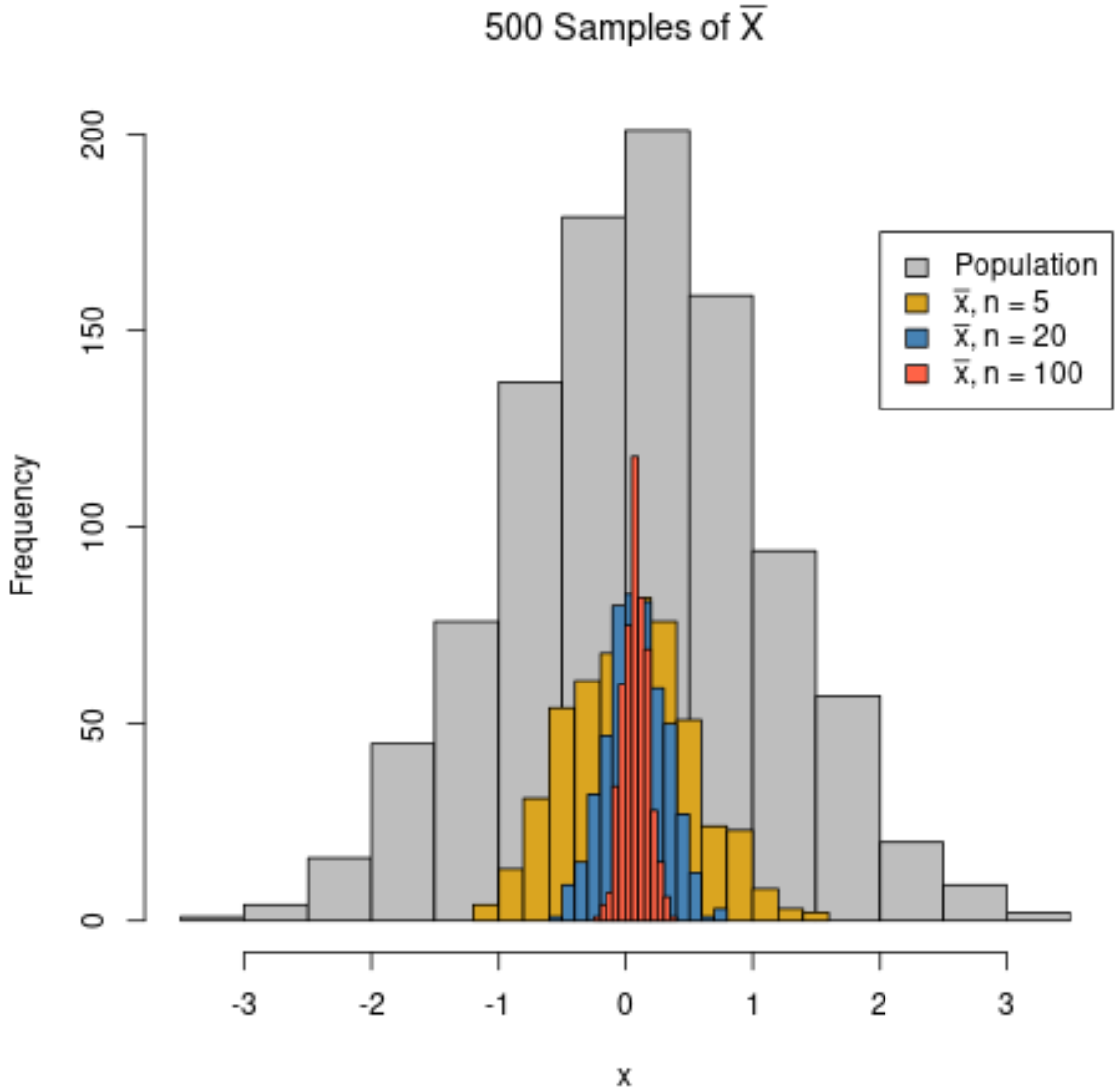
The theorem states that if a population X has a mean value of μ and variance of σ^2, with a sample $X_1, X_2, \ldots$, then the sample mean approximately follows a normal distribution,

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

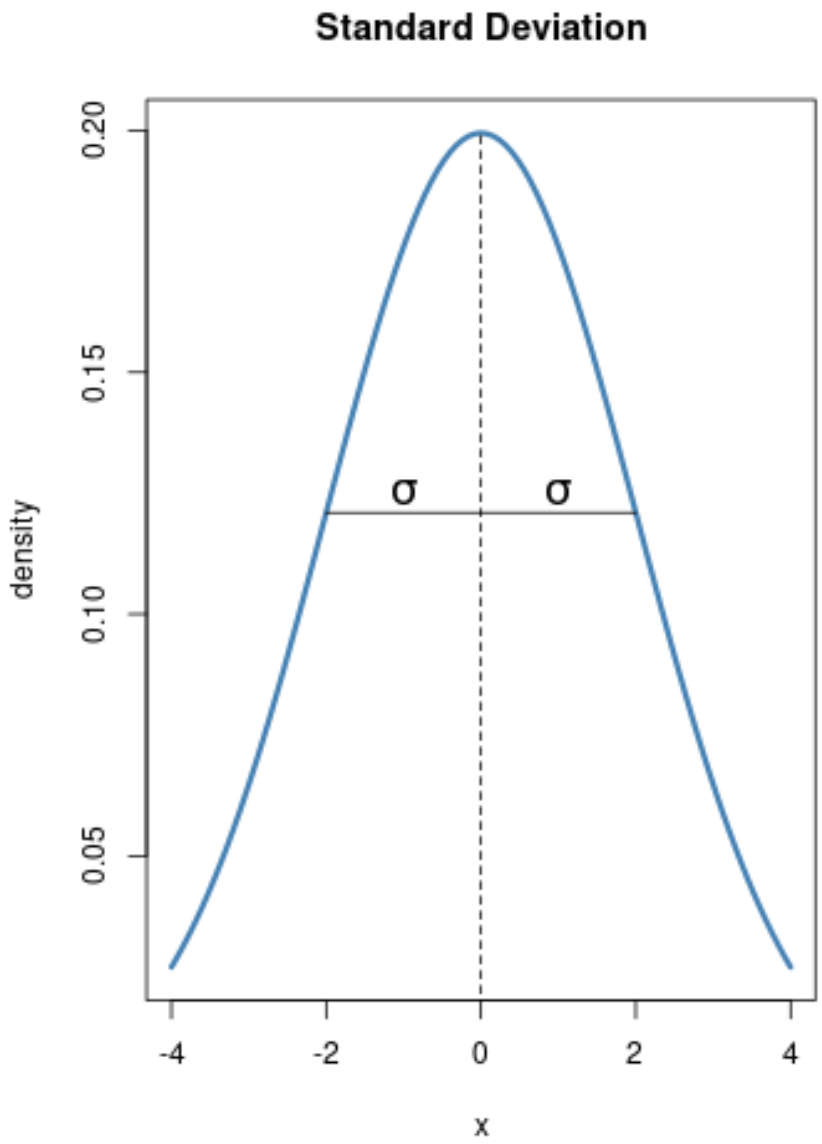
where n is the size of the collected sample

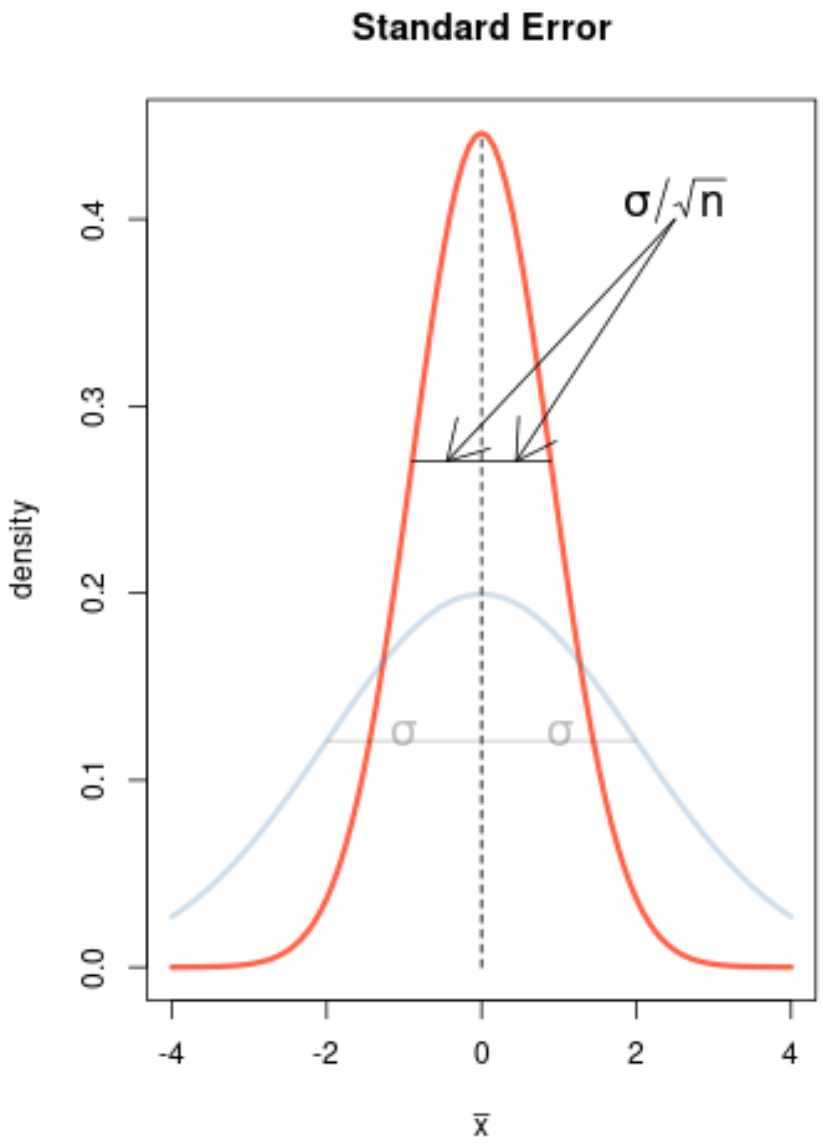
In other words, the center of the sampling distribution of $\overline{X}$ will be equal to the population mean, μ, and the amount of variability in the sampling process will be equal to the population variance, divided by the number of samples

Sampling Distribution



Standard Deviation vs. Standard Error





A few notes

First, we see that the sampling distribution centers around the population mean, telling us that the expected value of the sample mean is $E(\overline{X}) = \mu$

Next, the variability in our sample statistic is related to two properties:

1. The variance in the population
2. The size of the sample

And while it wasn't mentioned above, the results of the CLT hold in all cases – even if the distribution of the population is not normal, we can still expect normality in our sample statistic

Going Forward

The value of a sample statistic, along with its variance, will give us an ability to identify a range of plausible values for our population parameter

The range of these values also serves as a quantification of certainty, with larger ranges being associated with less certainty and smaller ranges, more

We will consider this in the context of hypothesis testing, with our sample either providing evidence to reject our null hypothesis or failing to do so