Sampling Distributions

Grinnell College

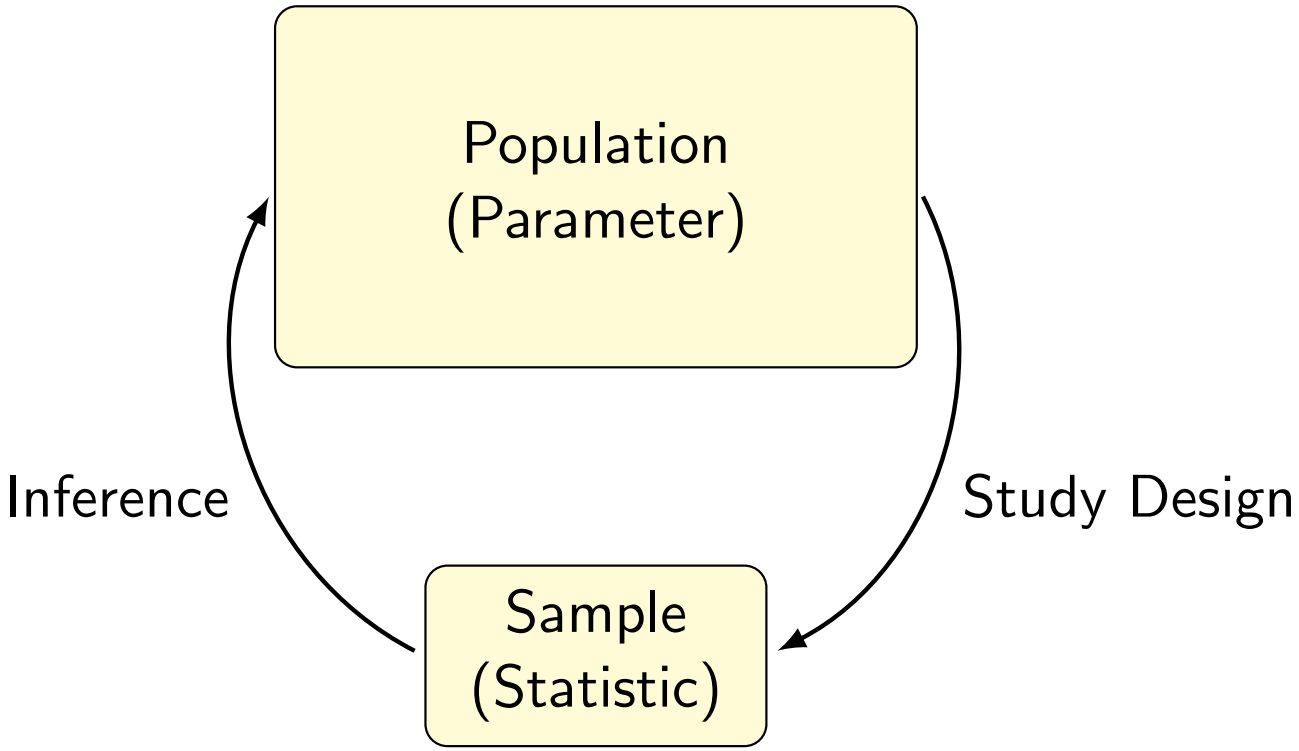
March 3, 2025

We have already spent a bit of time discussing the relationship between populations and samples, and, in particular, the importance of a sample being *representative*

While much of that discussion then oriented around *bias*, today we will be concerning ourselves with **variability**

More specifically, we will be investigating today the relationship between **parameters** and **statistics**

The Statistical Framework



Recall that a *distribution* tells us

- What values
- How frequently

A **parameter**, most generally, is a type of numerical summary of a distribution (i.e., the maximum, the median, the standard deviation)

A subset of these, known as **distributional parameters**, tend to be of special interest. If we know the distributional parameters, we know everything we possibly can about a random process

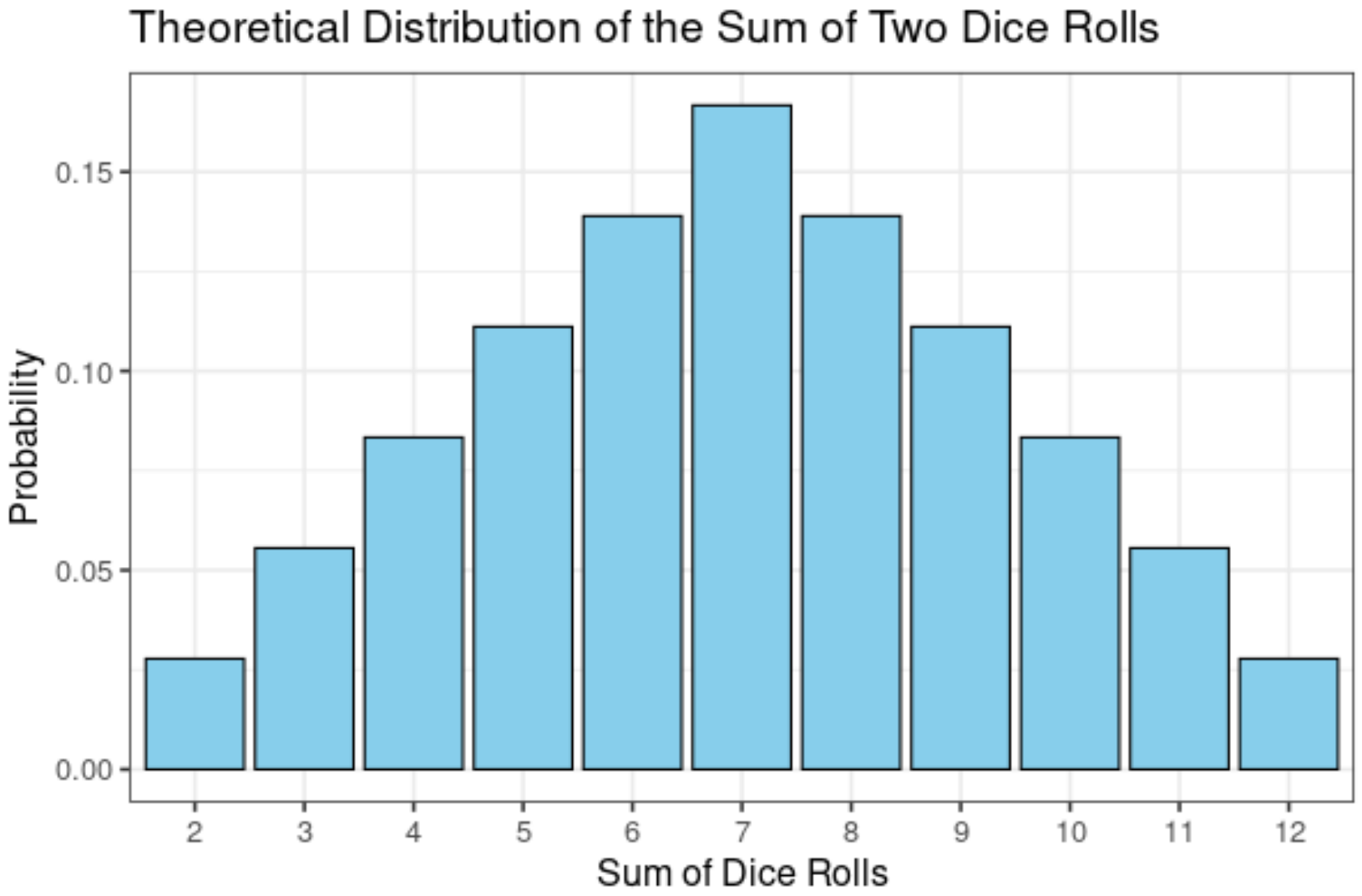
Examples of distributional parameters include:

- Probability of flipping heads in a single coin toss
- Knowing the true number of red and blue marbles in an urn
- Knowing the proportion of times a fair dice will roll a 1

For example, knowing the distribution of dice rolls allowed us to find the distribution of sums of two rolled dice added together

In this sense, we can say that the distribution describes the *random process* that generates the data

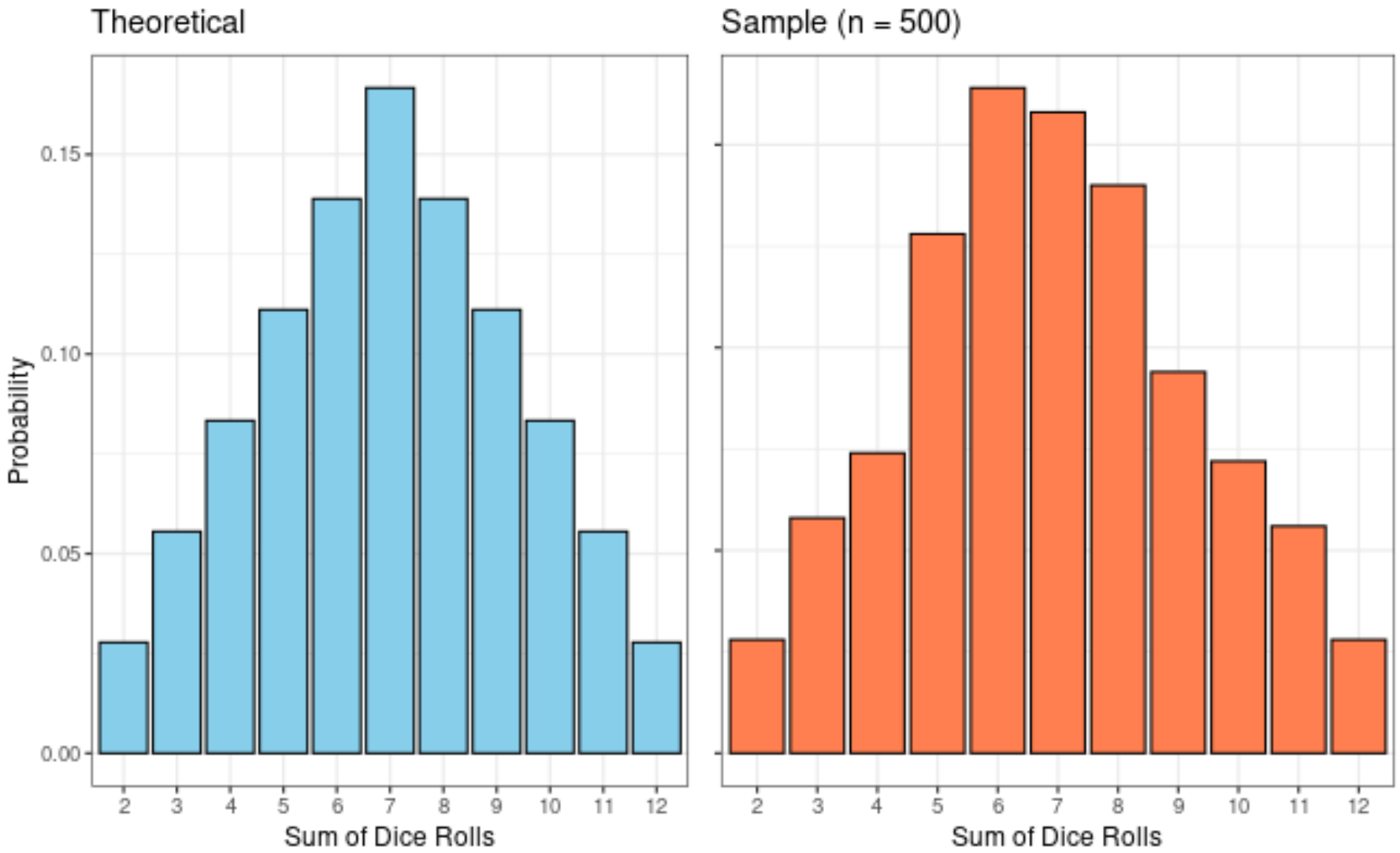
Dice Sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



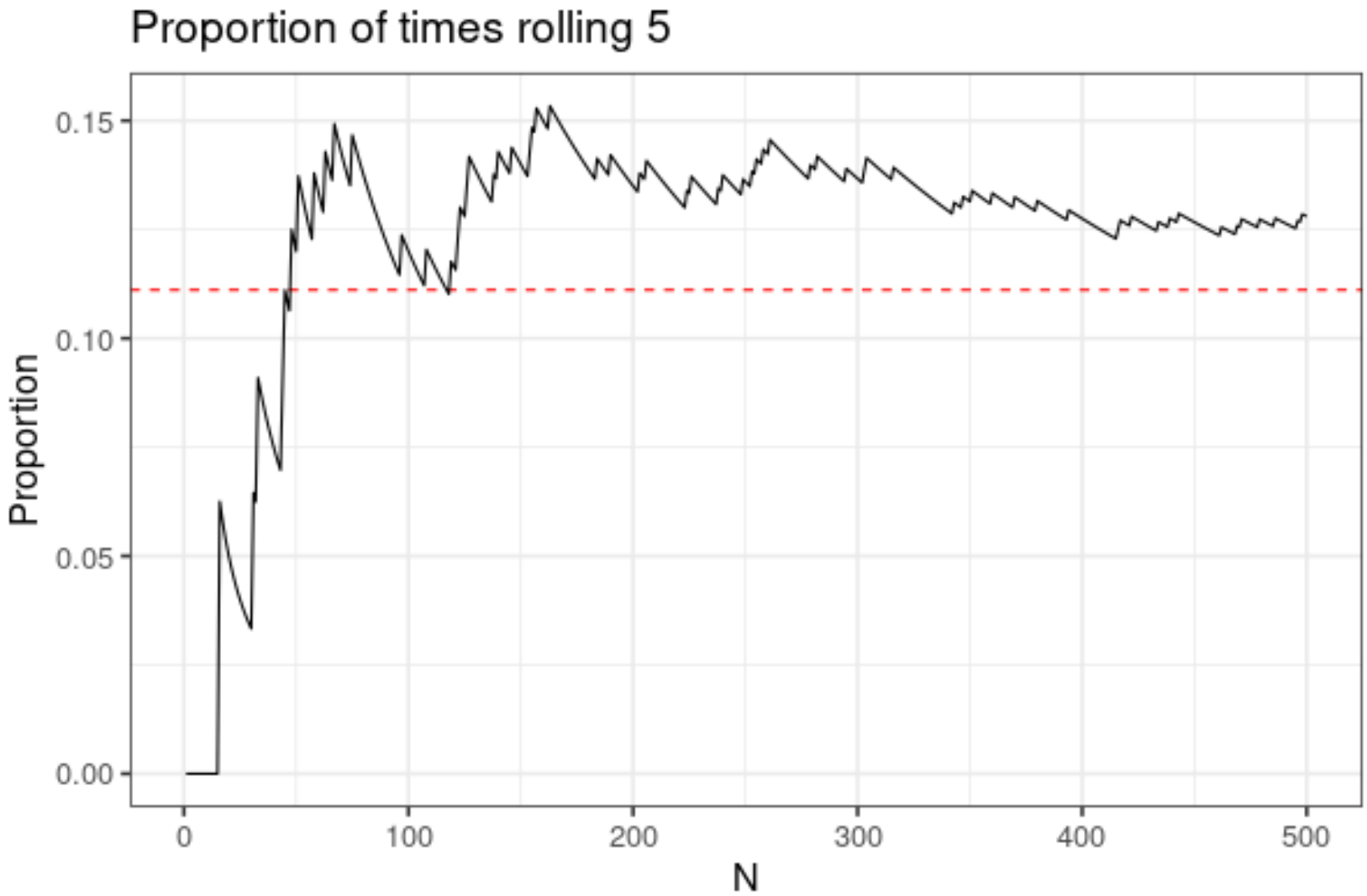
Often, we don't have information on the full population (thus, on parameters), and we are required to collect a *sample* from the population instead.

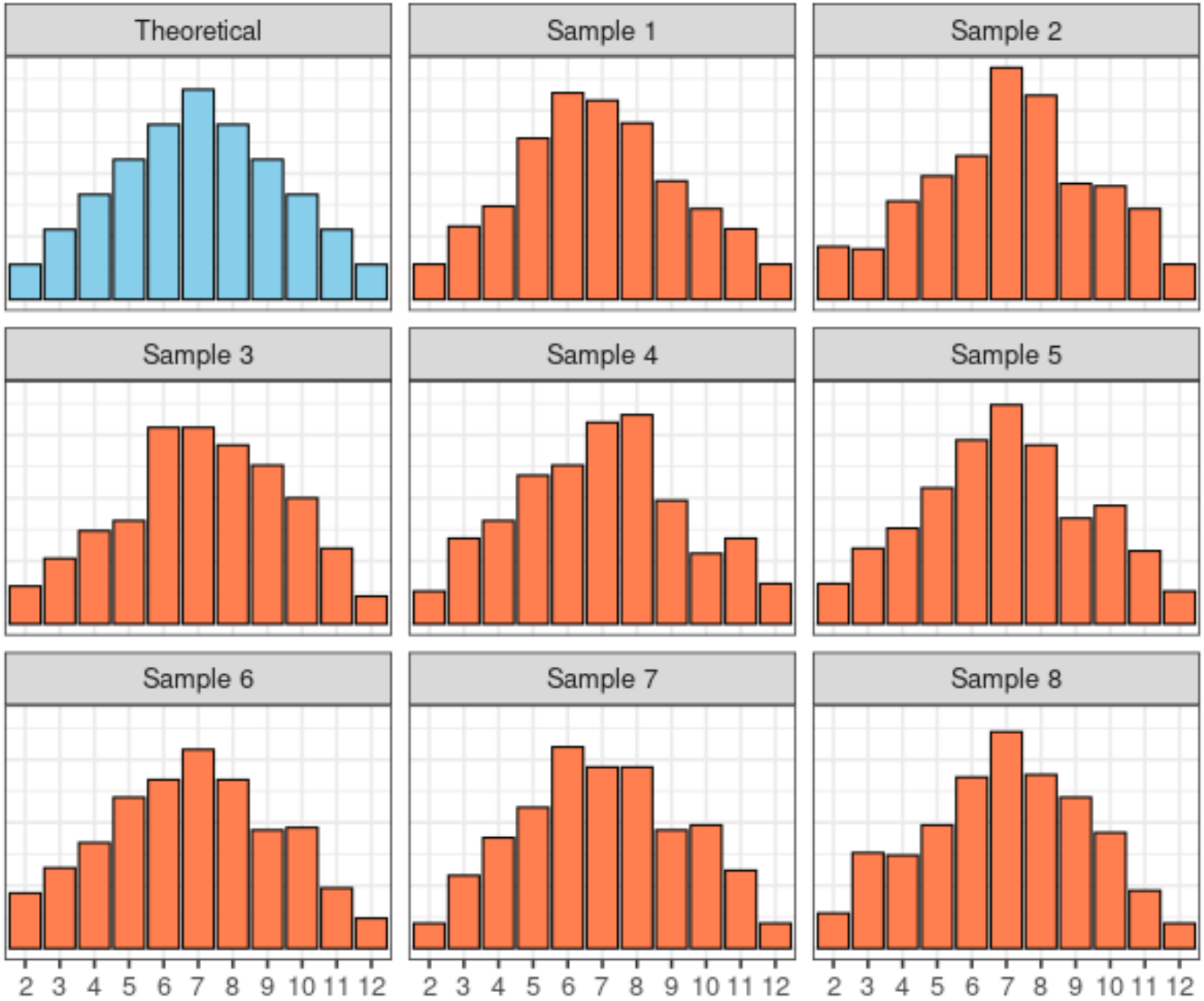
From our sample, we can then compute a **statistic** as an estimate of our population parameter. However, collecting a sample is itself a random process. The randomness present in the sampling process will also be present in our statistic.

Dice Sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

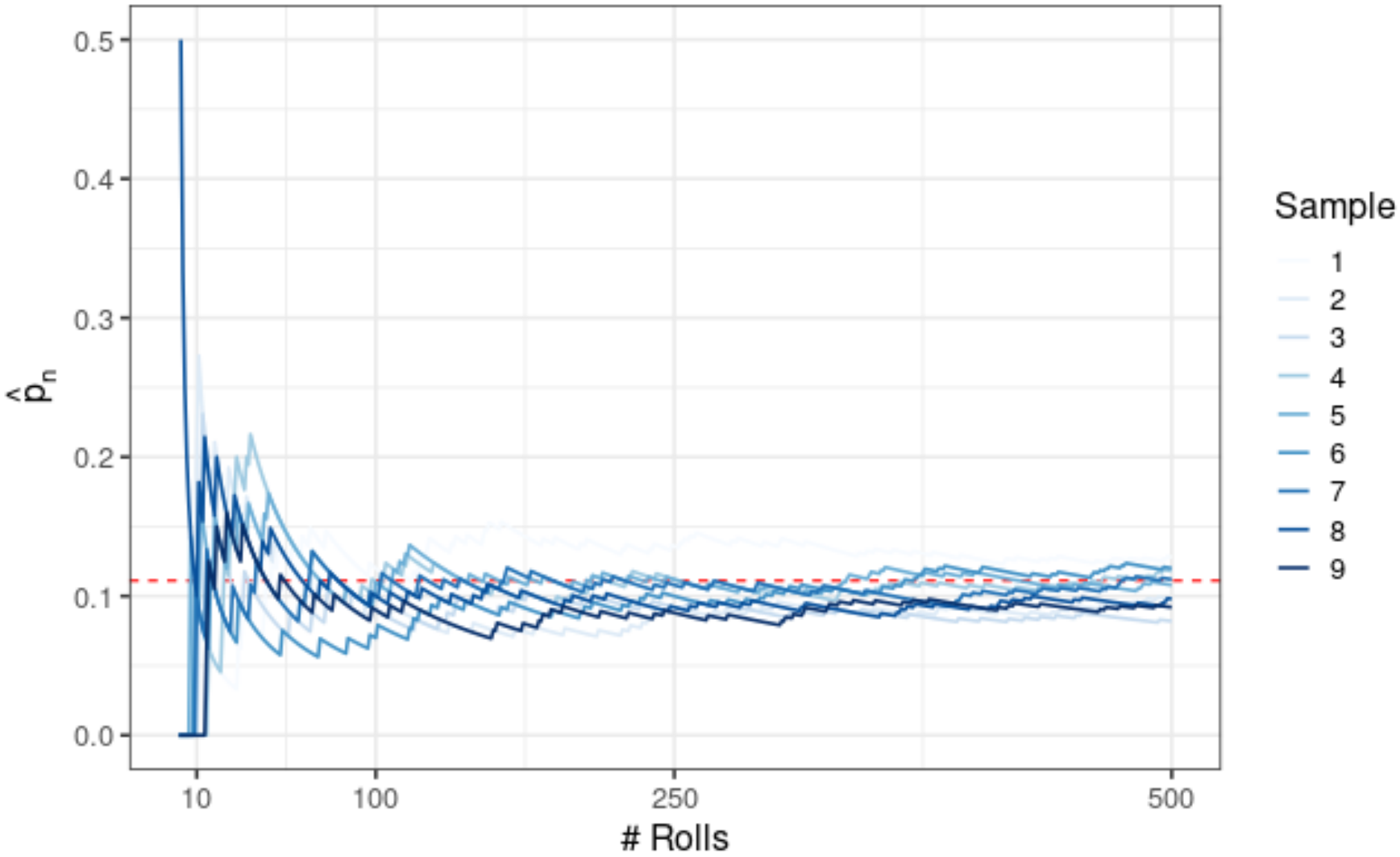


Recall the **Law of Large Numbers**: the number of observations in our sample will play a key role in how closely our statistic approximates our parameter




Theoretical
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
2 3 4 5 6 7 8 9 10 11 12

The true probability that the sum of two dice will equal 5 is $\frac{4}{36}$

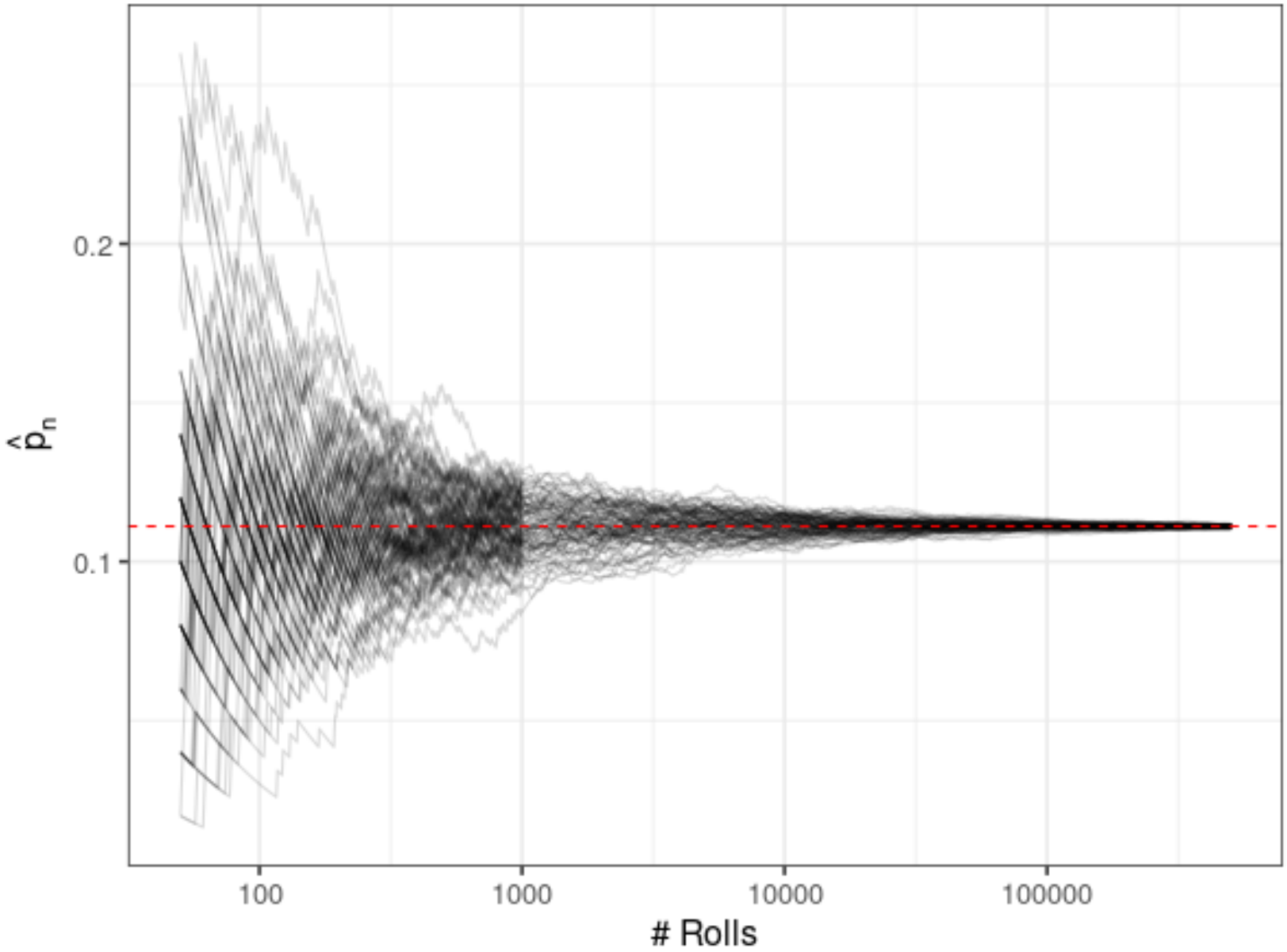


As we can see, not only is there variation associated within the sample itself (as a function of the sample size), there is variation *between* samples as well, directly impacting our statistics

From this, a natural next step is to ask ourselves: what if I were to collect samples an infinite number of times?

Does it make sense to ask about a *distribution* of statistics?

I collected samples of 500,000 observations 100 times. Are there any attributes of this distribution that we notice?



The **sampling distribution** refers to the the distribution of a statistic.

This gets at the idea – if I were to sample an infinite number of times:

- What values would my statistics take?
- How frequently do they appear?

Typically, we will find that, on average, the sampling distribution will tend to be centered around the true value of our parameter

What we would like to know is: *on average, how far can we expect our given statistic to be from the center?*

The **Central Limit Theorem** states:

1. For a *population* with mean value μ and standard deviation σ
2. And a sample with n observations
3. The sample mean, $\overline{x}$, has a *sampling distribution* with mean μ and standard deviation $\sigma/\sqrt{n}$ that is approximately normal, with

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

The **normal distribution**, also called the bell-curve, is a distribution that has as its *distributional parameters* the mean and the standard deviation; if we know these, we know everything about the distribution.

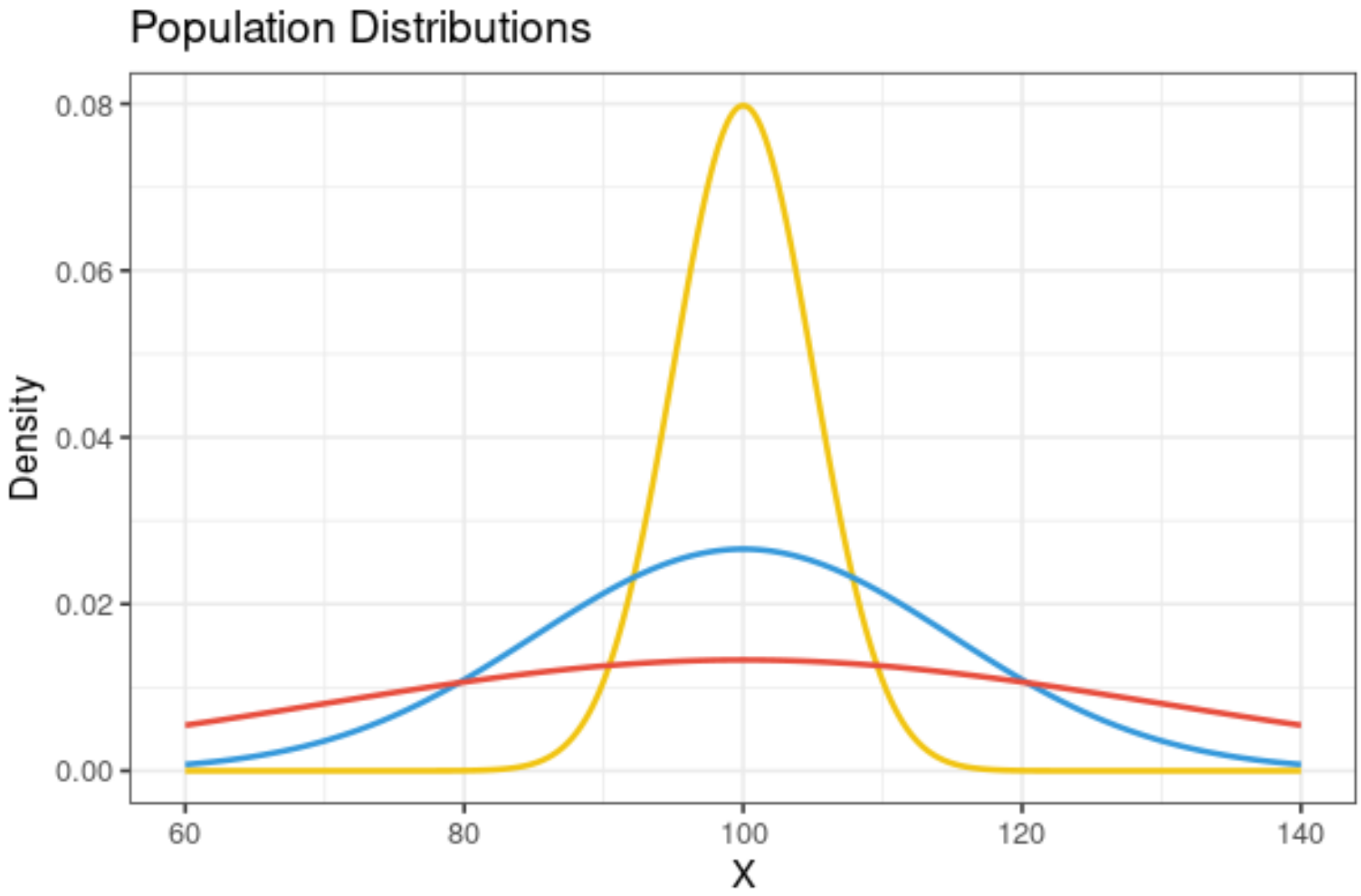
We indicate that a random variable X is normally distributed with the notation:

$$X \sim N(\mu,\ \sigma)$$

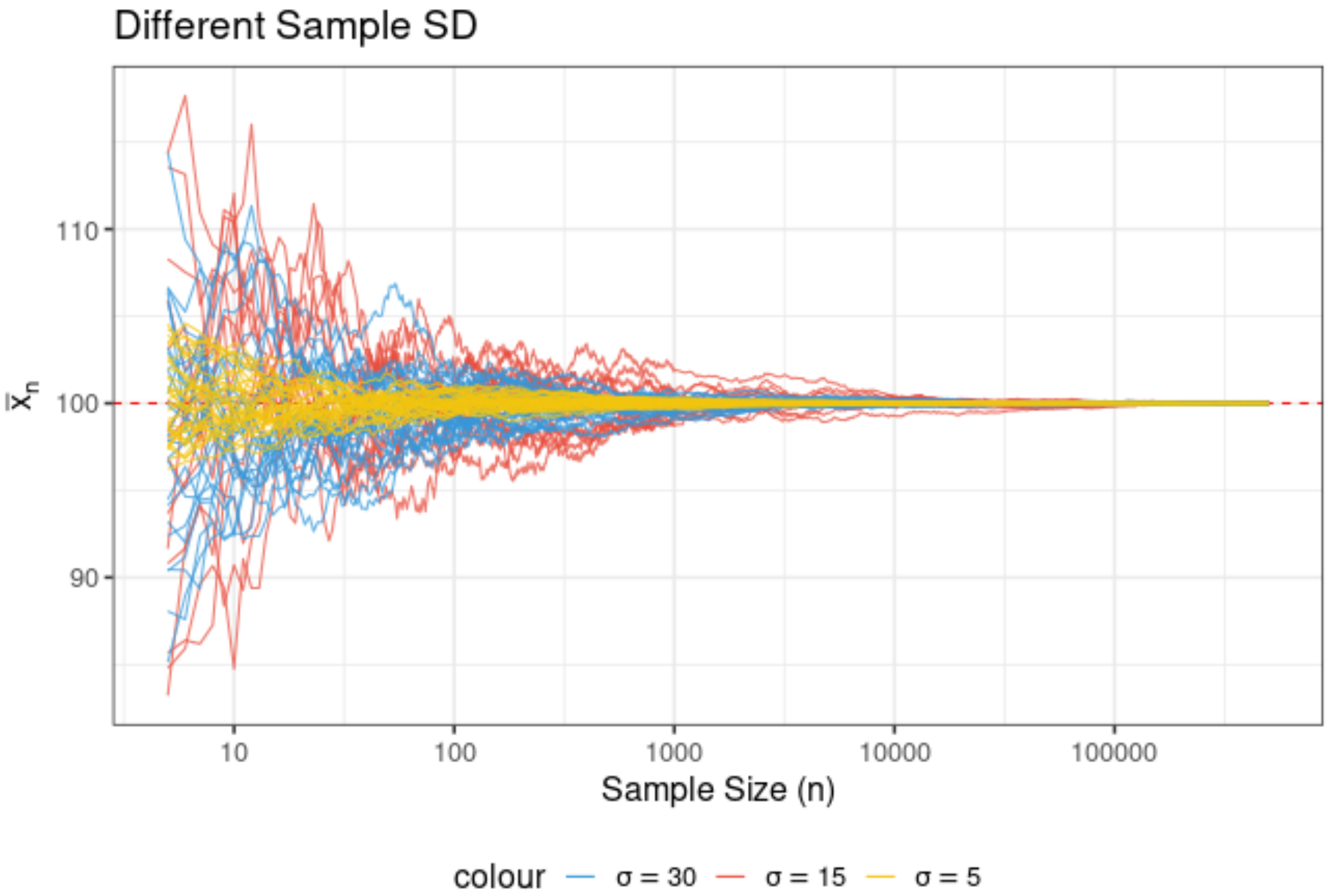
where μ represents the mean value and σ represents the standard deviation.

We read this as: The *random variable* X is *normally distributed* ($\sim$) with mean value μ (mew) and standard deviation σ (sigma)

To illustrate, say that we start with some population that we know has the following distribution $N(100, \sigma)$ with $\sigma = \{5, 15, 30\}$



Observe that the sampling distribution for $\overline{X}$ has standard error $\sigma/\sqrt{n}$



Sampling Distribution and the CLT

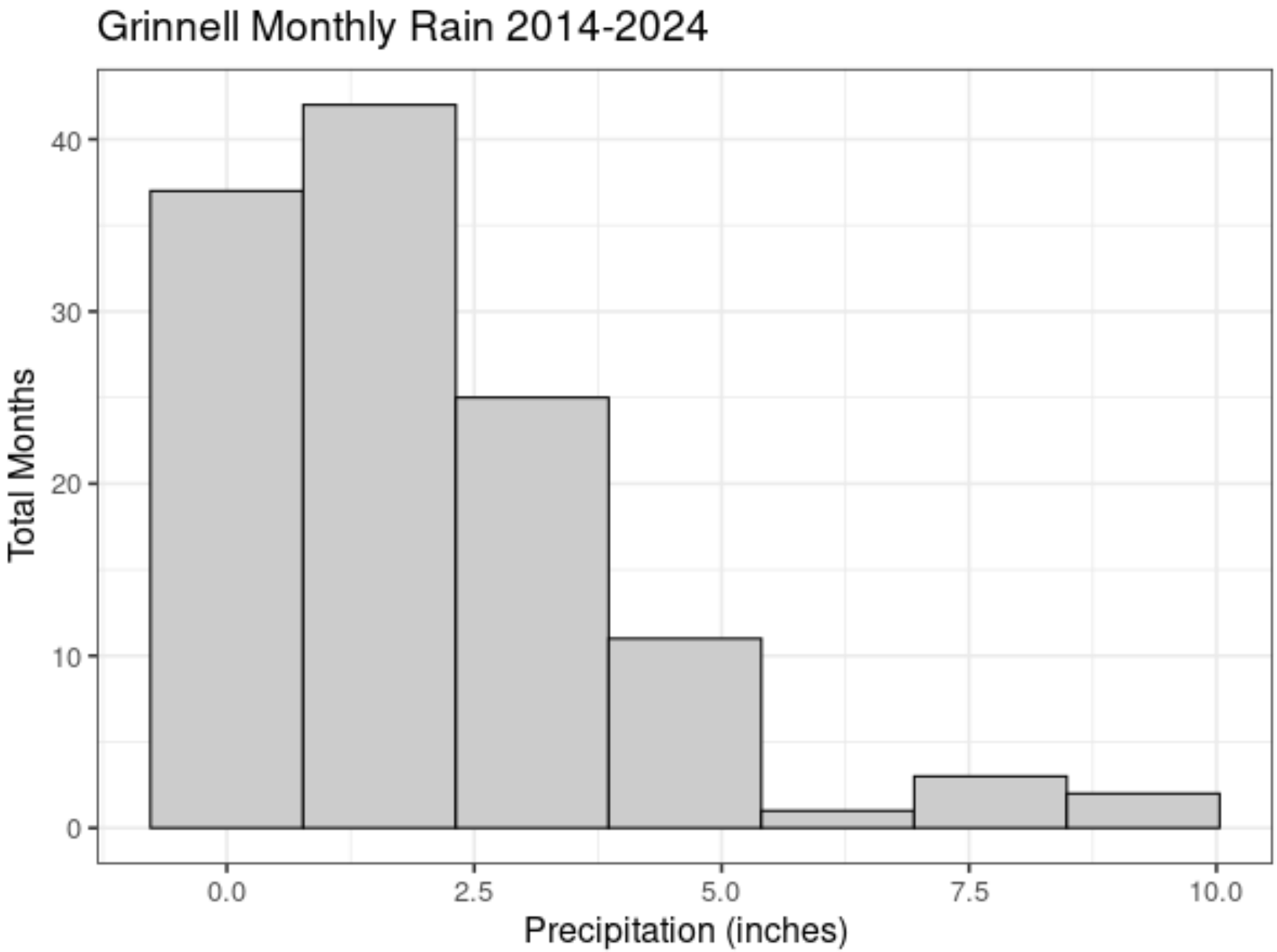
There are a few important things to note:

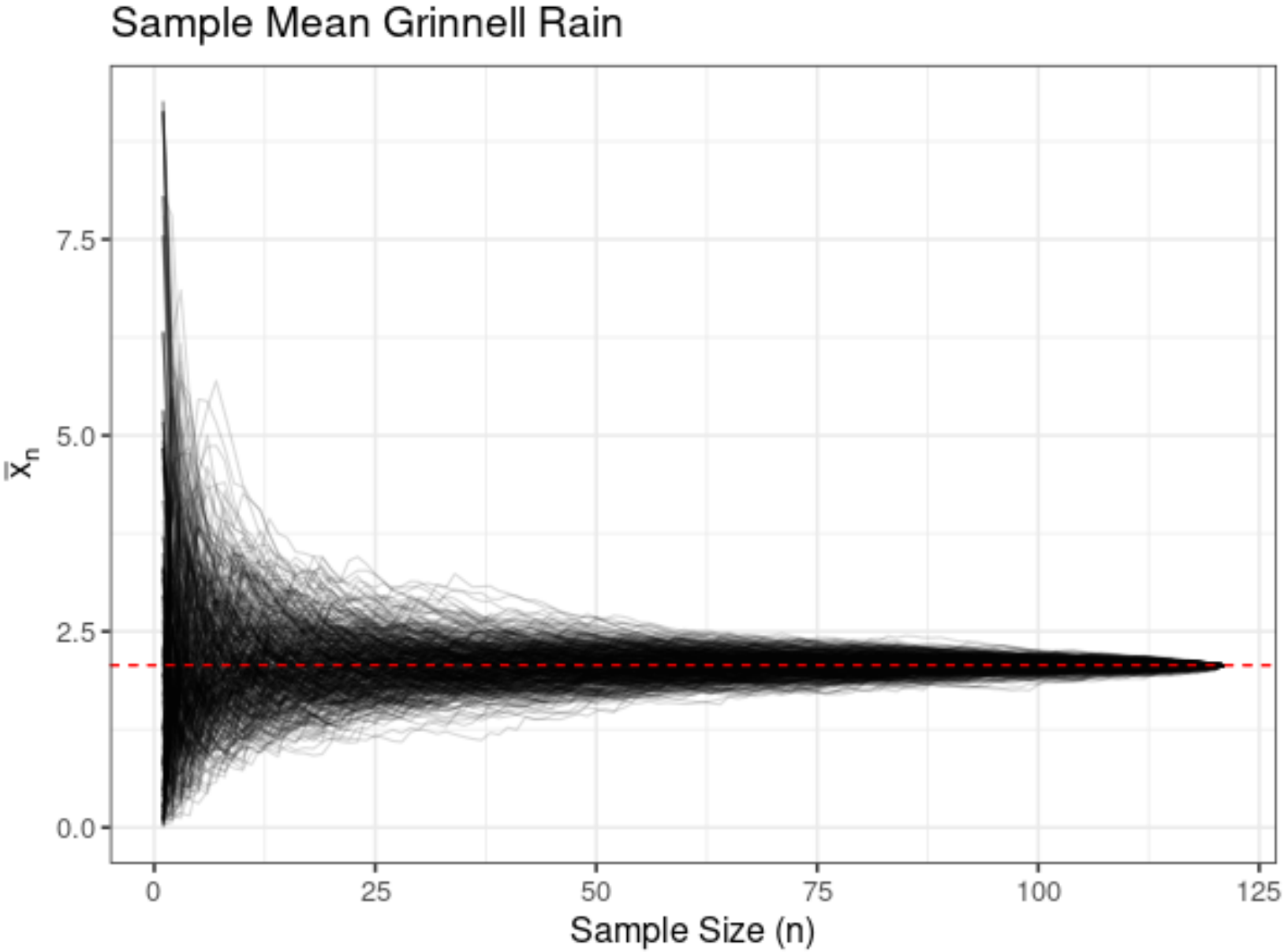
- The CLT encapsulates the idea of a statistic being a *random variable* as it is the result of a random process. As such, it makes sense to have a distribution
- In real life, we will usually *not* have a multitude of samples. We will have on sample and, as a result, only a single sample statistic
- However, even with only a single statistic, the CLT allows us to identify the distribution of values from which this statistic was drawn
- Specifically, this tells me:
  - Where is the distribution of my sample centered?
  - How much variability should I expect in my statistic, i.e., how much does it reliably tell me about the population parameter?

Note that the Central Limit Theorem *doesn't* require that either our population or our sample be normally distributed, though the more skewed our population is, the larger the number of samples we will need for our approximation to be useful

To demonstrate this, consider a dataset containing the average monthly rainful in Grinnell, IA from 2014-2024

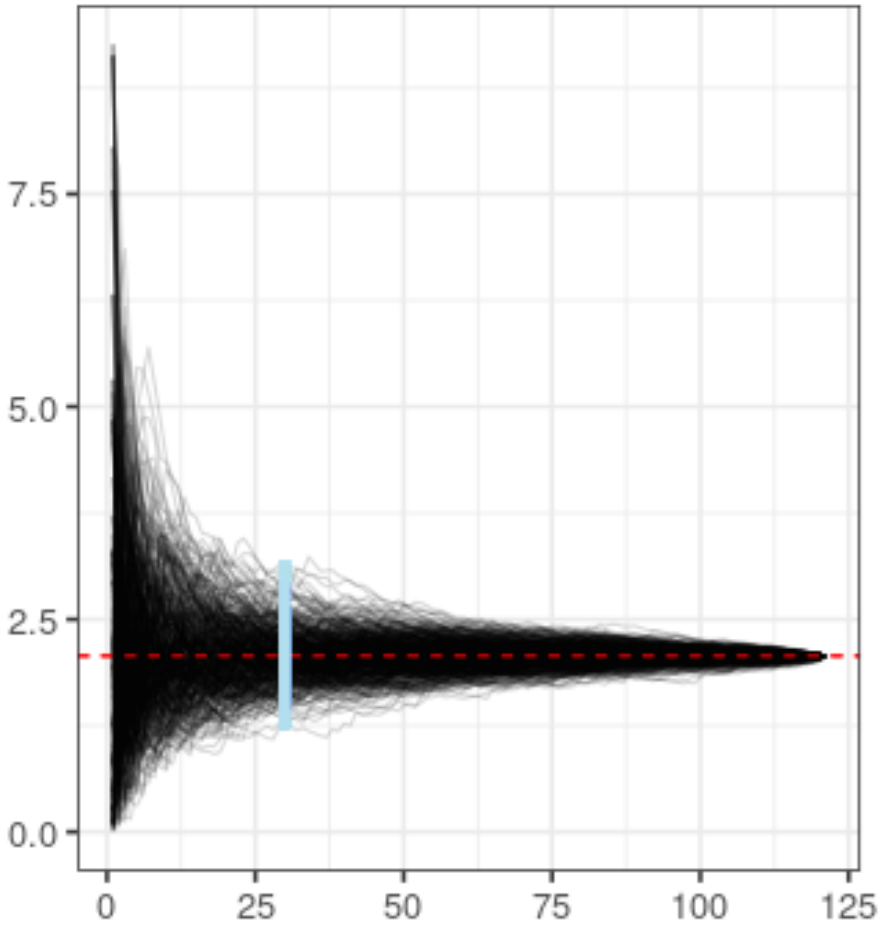
For illustration, we can treat this as our population. From this, we can determine that the true average monthly rainful for this period is $\mu = 2.07$ inches, with a standard deviation of $\sigma = 0.36$

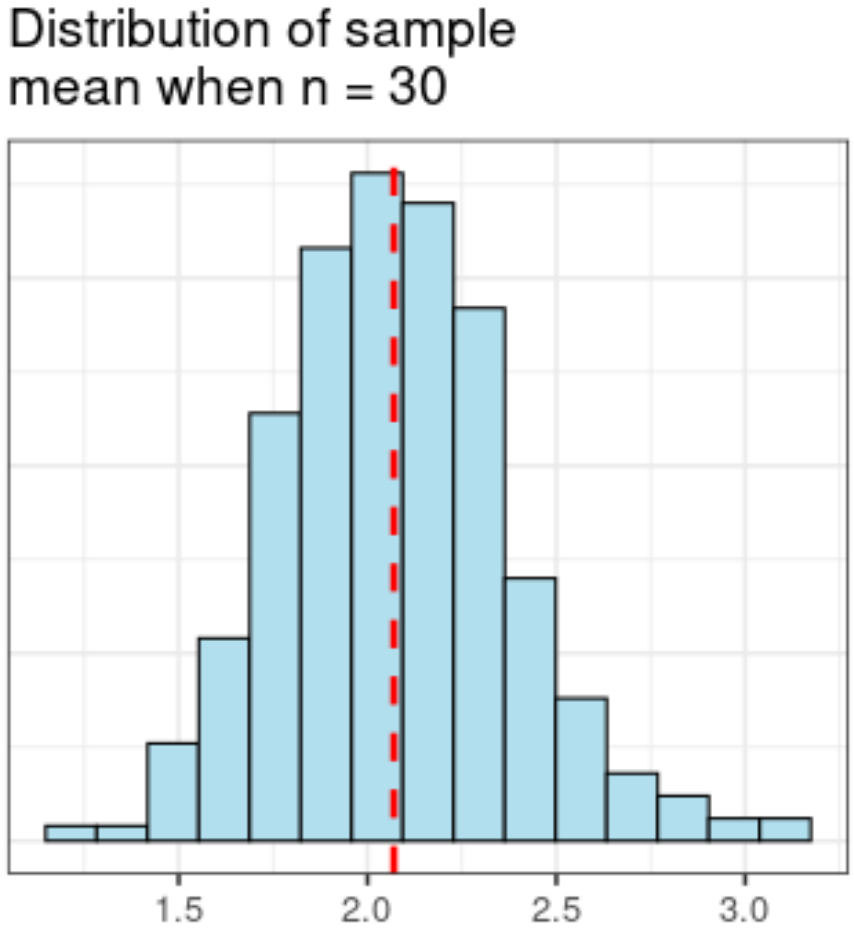

Grinnell Monthly Rain 2014-2024
Total Months
40
30
20
10
0
0.0
2.5
5.0
7.5
10.0
Precipitation (inches)


Sample Mean Grinnell Rain
7.5
5.0
2.5
0.0
$\bar{X}_n$
0
25
50
75
100
125
Sample Size (n)

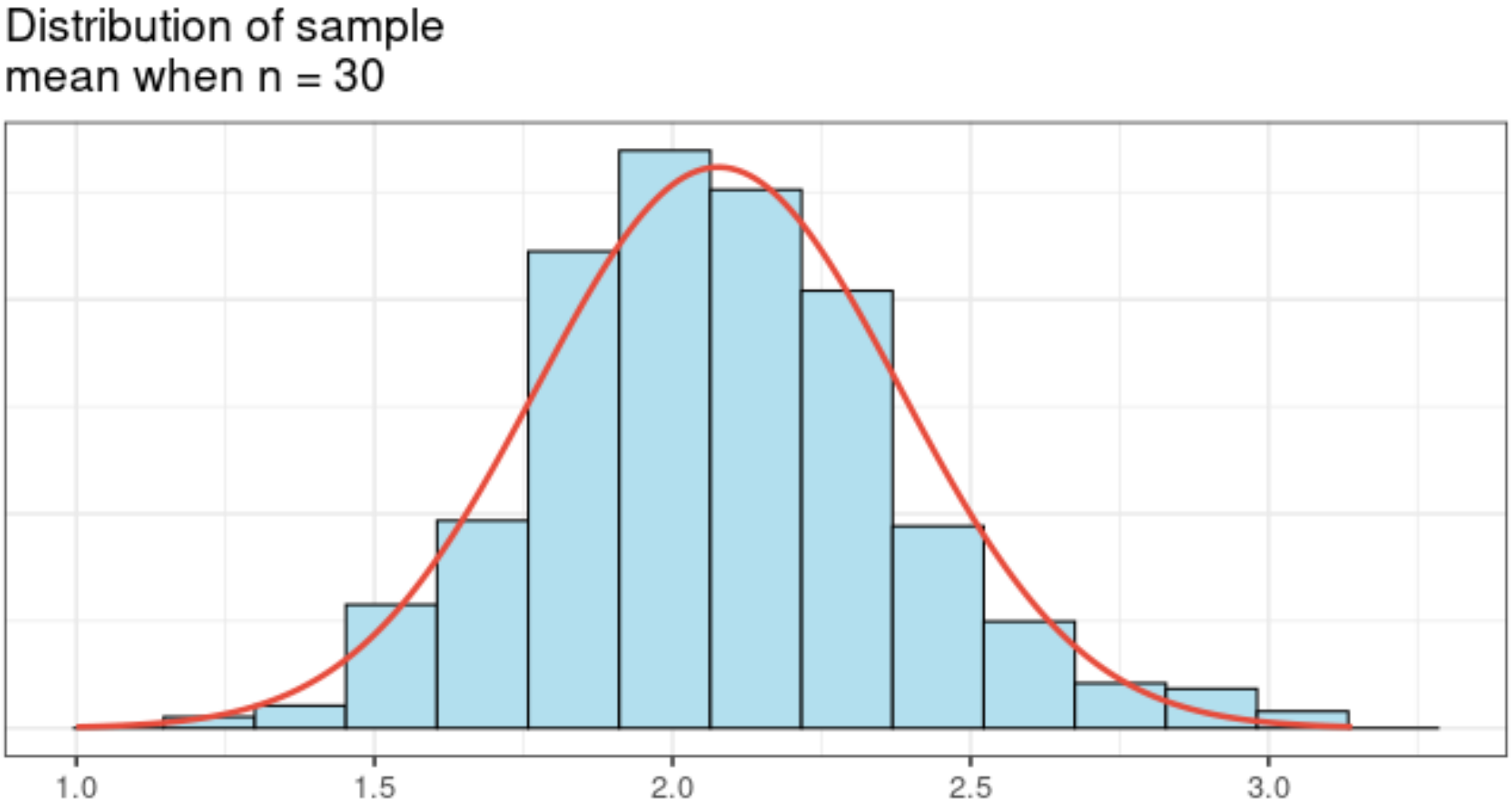
$\overline{x} = 2.07$, $\hat{\sigma} = 0.31$, $n = 30$

$\mu = 2.07$, $\sigma/\sqrt{n} = 0.36$





We can compare this to what the CLT tells us the distribution should be



Review

- Collecting a sample is an example of a random process
- The randomness present in a sample will be present in the computed statistics
- A **sampling distribution** tells us about the distribution of a sample statistic
- The **Central Limit Theorem** tells us that the sampling distribution of the mean or a proportion will be normally distributed