Sampling distributions

ECON 3640–001

Marcio Santetti

Spring 2022

Motivation

Housekeeping

Notes based on `Keller (2009)`:

- Chapter **9**, section 9.1.

Motivation

We will spend the remaining lectures on the ***frequentist*** approach to statistical inference.

Recall that the frequentist **interpretation** of probability relies on it coming out of **repeated** experiments.

In this context, a fundamental element for understanding frequentist inference is **sampling distributions**.

Sampling distributions

Sampling distributions

There are **2** ways to approach sampling distributions.

The **first** is to repeatedly draw **samples of the same size** (n) from a **population** of interest (N), and calculate the statistic of interest.

- However, it is almost impossible to access data for an entire population.

The **second** is to use the laws of **Expected Value and Variance**, which we have already studied, to derive sampling distributions.

- More feasible!

Sampling distributions

Let us demonstrate the first approach, using the `AmesHousing` data set.

- It includes data on **all** residential home sales in Ames, Iowa, between 2006 and 2010.
- Thus, these data may serve as a **populational** reference.

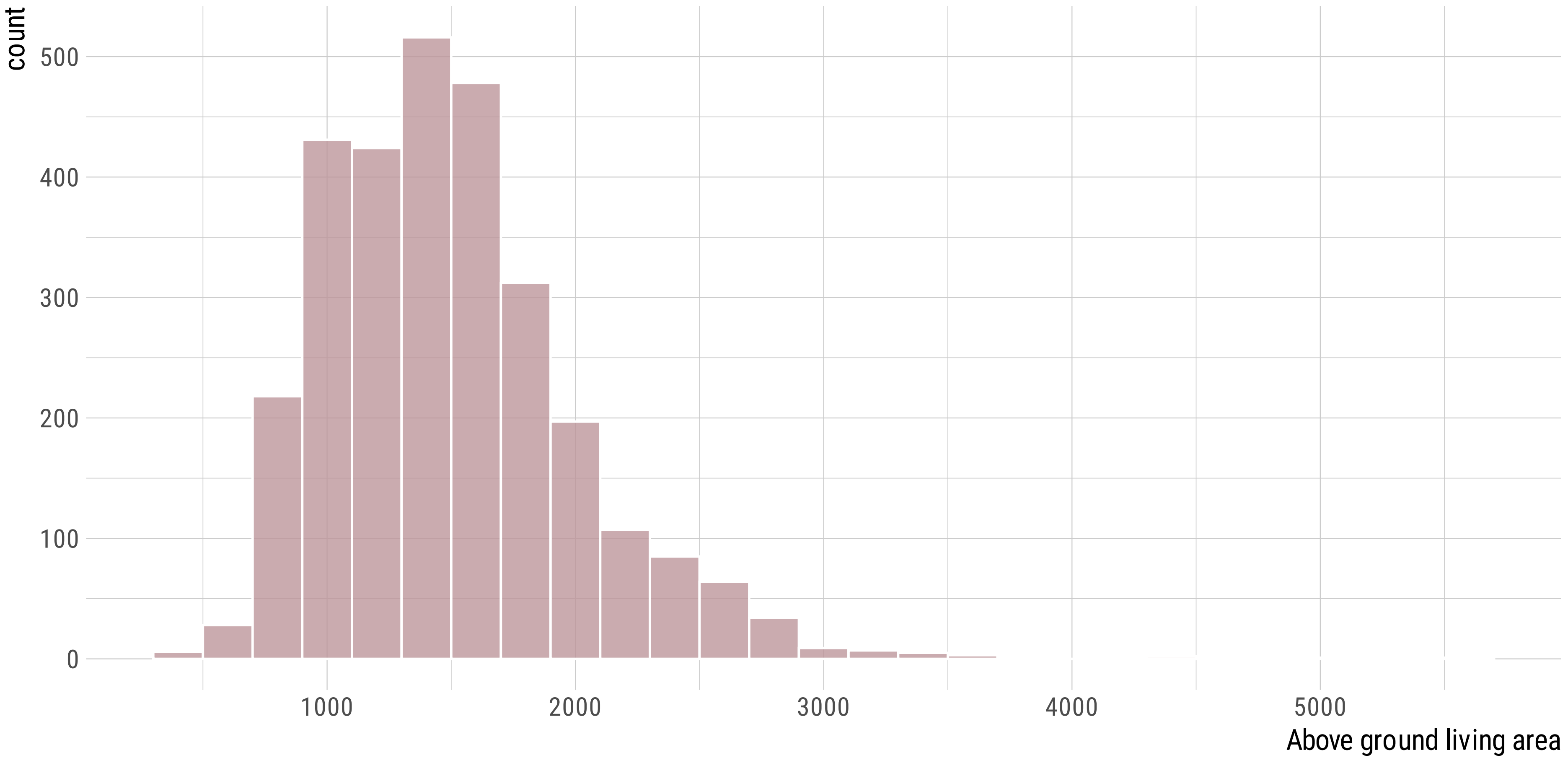
```
library(AmesHousing)    ## where the data come from
library(janitor)        ## package for data cleaning.

ames ← ames_raw         ## picking one of the package's data sets.


ames ← ames %>%
  clean_names()         ## using 'janitor' to clean the column names.
```

Sampling distributions

```r
ames %>%
  select(gr_liv_area) %>%
  head(6)      ## above ground living area (in square feet).
```

```
#> # A tibble: 6 × 1
#>   gr_liv_area
#>         <int>
#> 1        1656
#> 2         896
#> 3        1329
#> 4        2110
#> 5        1629
#> 6        1604
```

Sampling distributions



Sampling distributions

Since we have the whole **population** data, we can compute population parameters, such as μ, σ^2, and σ:

```
ames %>%
  summarize(pop_mean = mean(gr_liv_area),
            pop_variance = var(gr_liv_area),
            pop_sd = sd(gr_liv_area))
```

```
#> # A tibble: 1 × 3
#>   pop_mean pop_variance pop_sd
#>      <dbl>        <dbl>  <dbl>
#> 1    1500.      255539.   506.
```

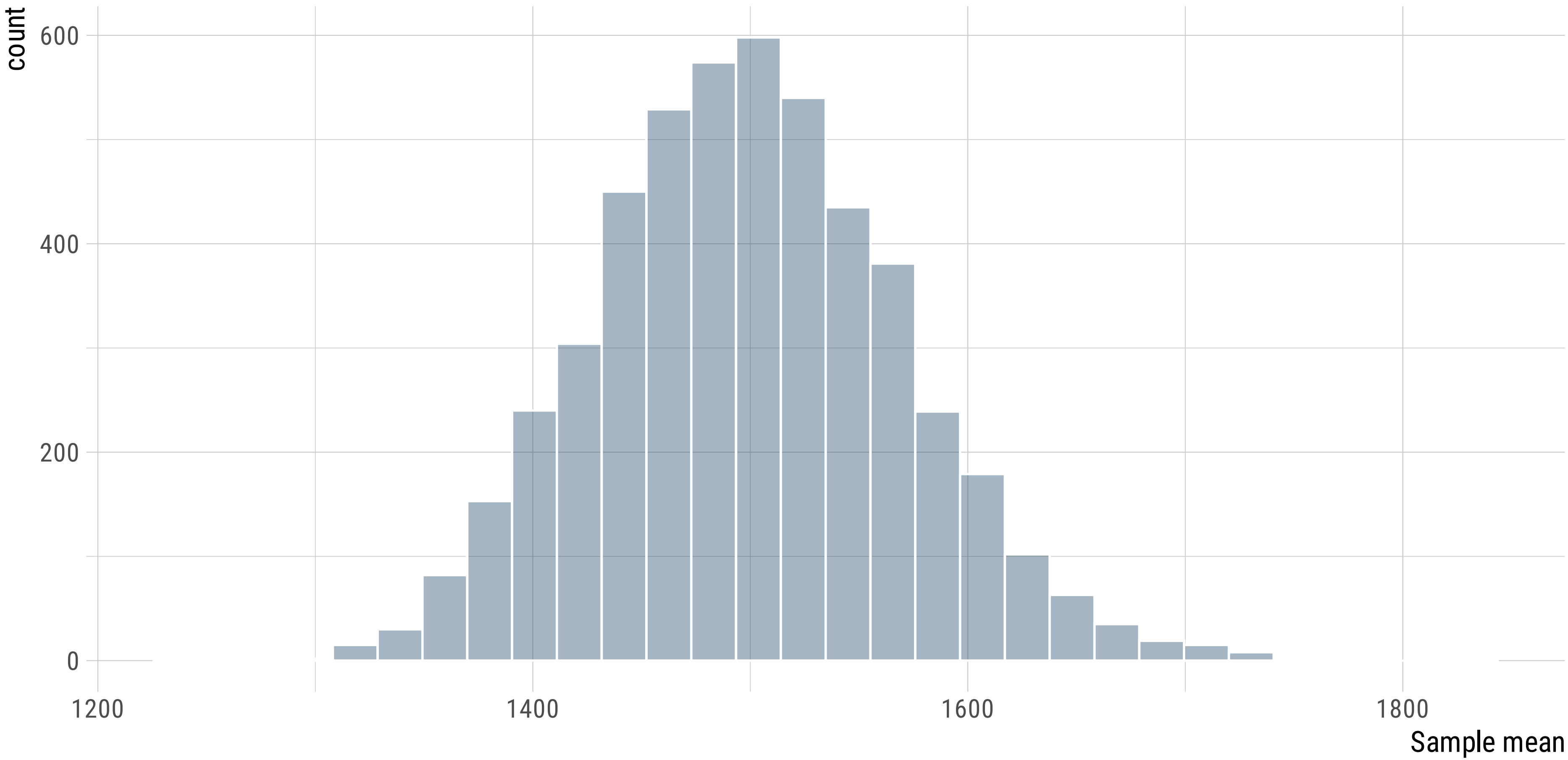
Now, let us repeatedly draw **samples of the same size** from this population, and see how the value of μ and σ^2 behave.

Sampling distributions

```r
area <- ames %>%
  pull(gr_liv_area)  ## pulling the values for the variable of interest.
```

```r
# A "for" loop:

sample_means50 <- rep(NA, 5000)  ## creating an empty vector of 5000 values.

for(i in 1:5000){                ## starting the loop (5,000 iterations).

  s50 <- sample(area, 50)        ## drawing samples of size n = 50

  sample_means50[i] <- mean(s50) ## filling the empty values with the sample means.

}
```

Sampling distributions

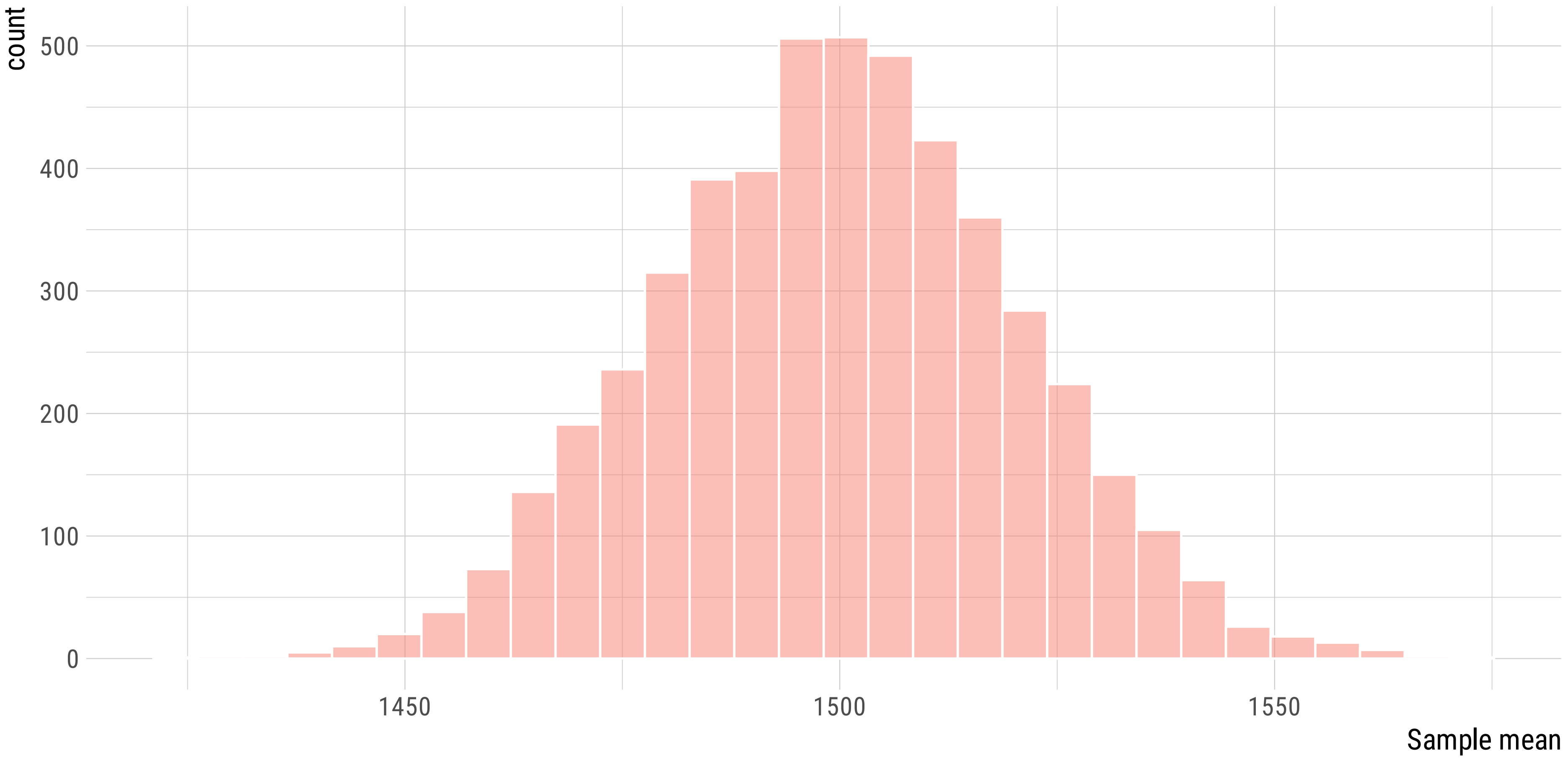


Sampling distributions

Now, instead of samples of size $n = 50$, what about $n = 500$?

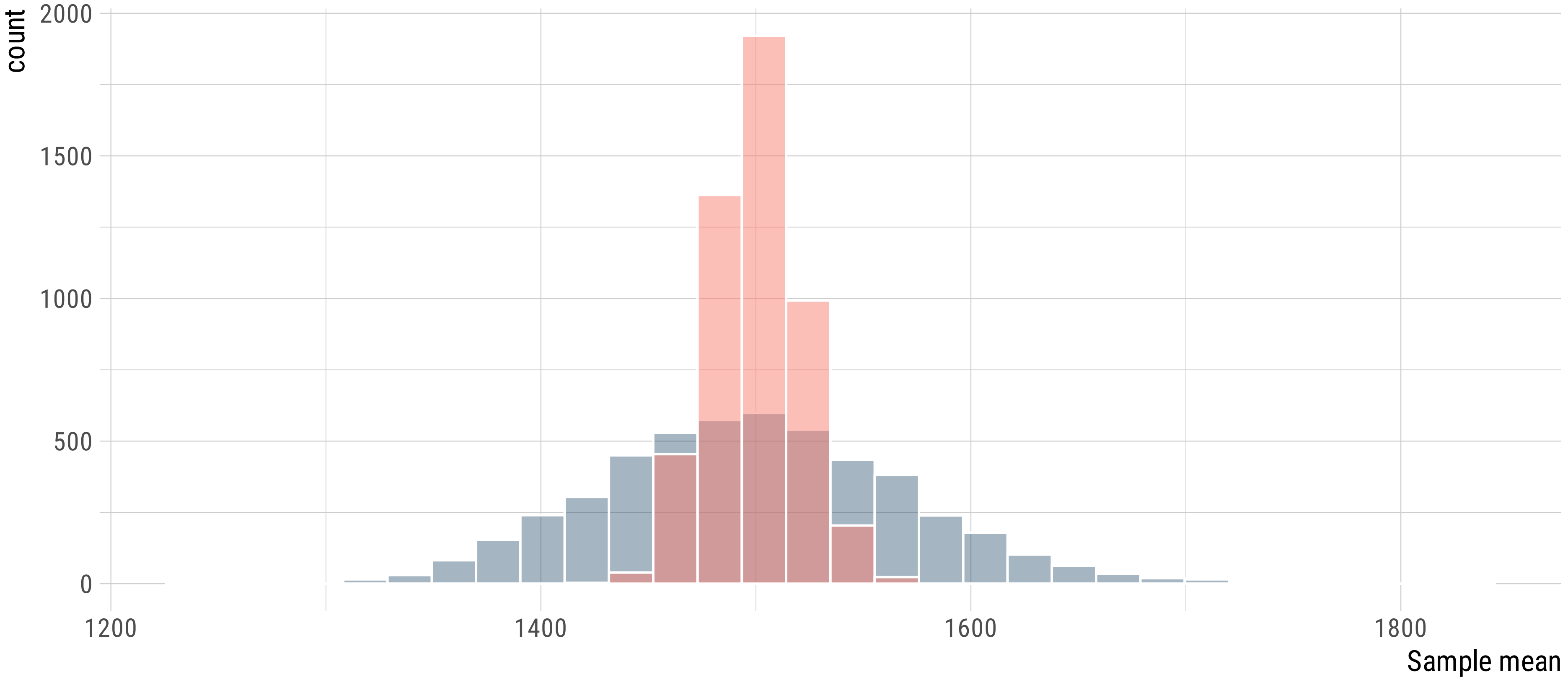
```
sample_means500 <- rep(NA, 5000)

for(i in 1:5000){

  s500 <- sample(area, 500)

  sample_means500[i] <- mean(s500)

}
```

Sampling distributions



Sampling distributions

Now, the two together...



Sampling distributions

Having access to the whole population, we may draw samples of the same size and **repeatedly** compute *sample statistics* from these samples.

And as the sample size **increases**, the *variance* (and standard deviation) is reduced.

- More precision!

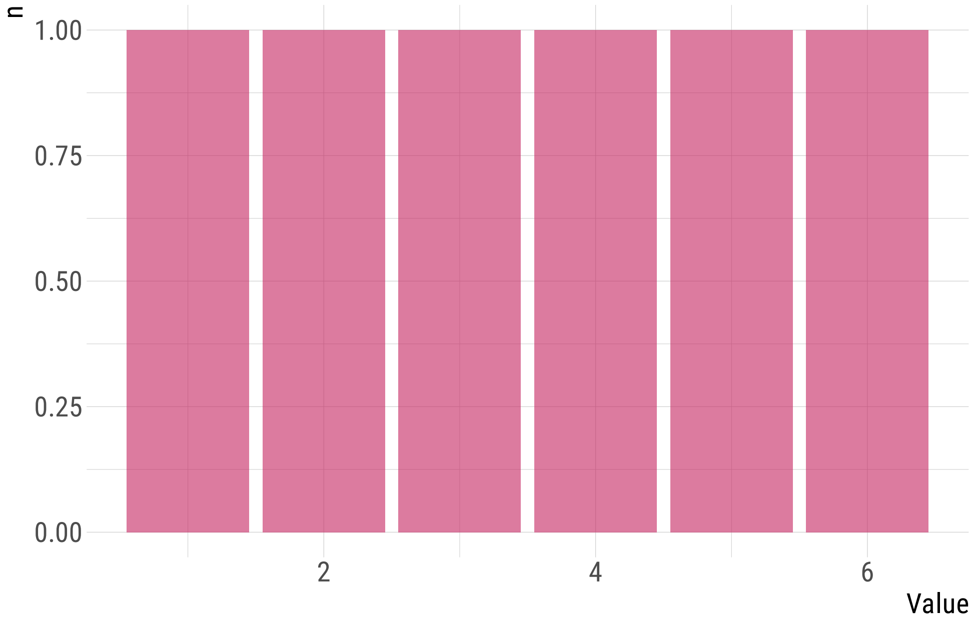
But when we do not have the luxury of accessing the whole population, we may appeal to the laws of *Expected Value and Variance* we've already studied.

Sampling distributions

Let us start with a single **die roll**.

The population is created by throwing a fair die *infinitely* many times, with the random variable X being the number of spots showing on any one throw.

What is the probability of each specific value of X, $P(x)$?



Sampling distributions

As we all know, the probability of a **1** is the same as the probability of a **6** from this single die roll.

$$\mu = \sum_{all\ x} xP(x) = 1(1/6) + 2(1/6) + \ldots + 6(1/6) = 3.5$$

$$\sigma^2 = \sum_{all\ x} (x - \mu)^2 P(x) = (1 - 3.5)^2(1/6) + (2 - 3.5)^2(1/6) + \ldots + (6 - 3.5)^2(1/6) = 2.92$$

$$\sigma = \sqrt{2.92} = 1.71$$

Sampling distributions

Now, what if we draw samples of size $n = 2$?

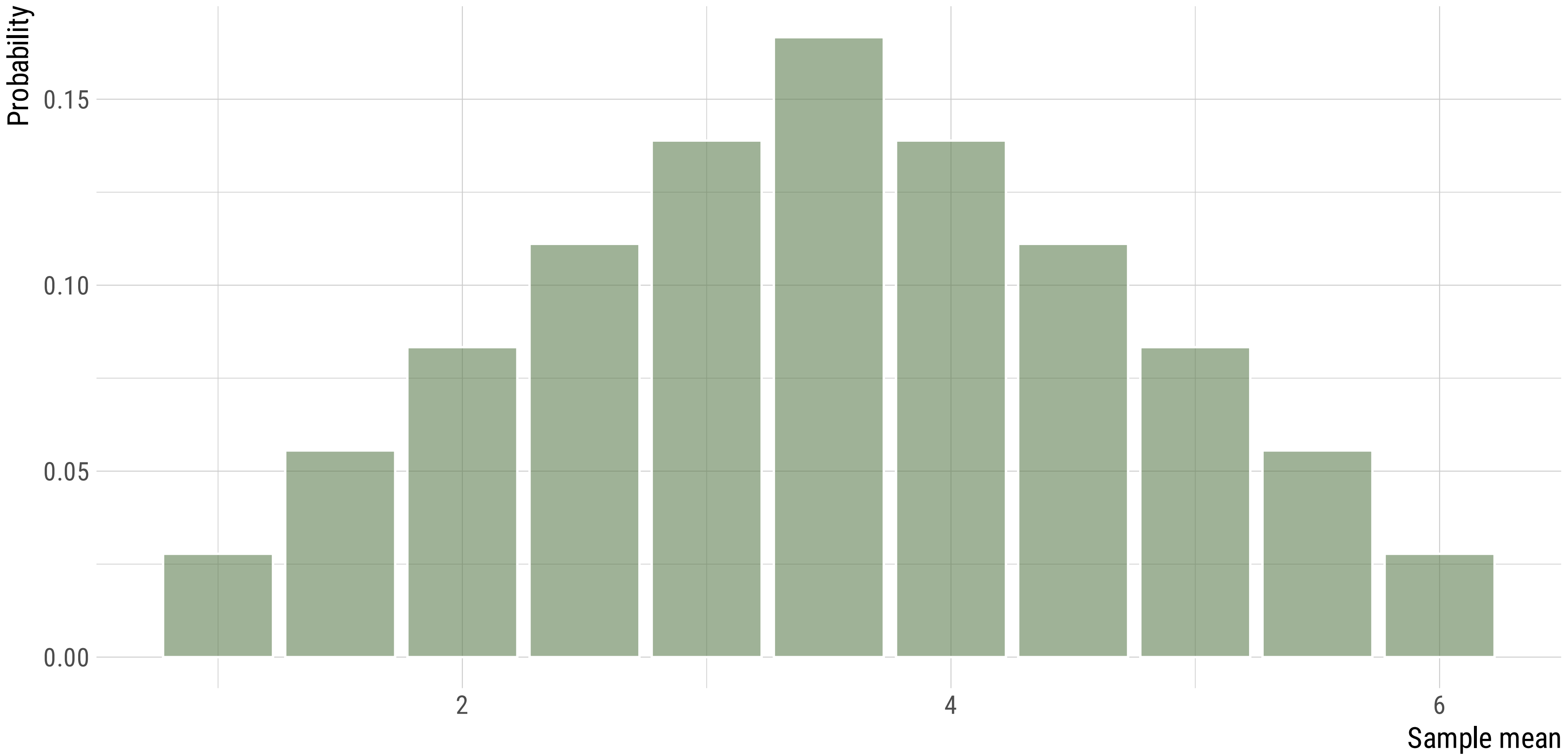
In other words, we throw **2 dice**, and study the *mean* and *variance* from these throws.

By tossing two dice, we have **36** different possible samples of size 2.

Each of these 36 possible pairs will have **different means**.

Therefore, the means are **not** the same as the the ones in the probability distribution from rolling a single die.

Sampling distributions



Sampling distributions

The Expected Value of the sample mean is the **same** as with 1 dice roll.

The **variance**, however, is different:

$$\sigma_{\bar{x}}^2 = \sum_{all\ \bar{x}} (\bar{x} - \mu_{\bar{x}})^2 P(\bar{x}) = (1 - 3.5)^2(1/36) + (1.5 - 3.5)^2(2/36) + \ldots + (6 - 3.5)^2(1/6) = 1.46$$

But they are related!

- $\sigma_{\bar{x}}^2 = \sigma/2$

If we repeat the same sampling process, but now *increasing* the sample size to, say, 5, 10, or 25 dice rolls, we will **still** observe the same sampling mean of 3.5.

Sampling distributions

The **variance** of the sampling distribution of the sample mean will be the variance of *X*, divided by the sample size, *n*.

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

Not surprisingly, the **standard deviation** will be

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Moreover, as the **sample size increases**, that is, as the number of dice rolls increases, the sampling distribution of $\bar{x}$ becomes *increasingly bell-shaped*.

In other words, its bell curve becomes **narrower** as the sample size is increased.

Sampling distributions

The latter phenomenon is summarized by the **Central Limit Theorem** (CLT).

> The sampling distribution of the mean of a random sample drawn from any population is **approximately Normal** for a sufficiently large sample size. The larger the sample size, the more closely the sampling distribution of $\bar{X}$ will resemble a Normal distribution.

In many practical situations, a sample size of **30** may be sufficiently large to allow us to use the Normal distribution as an approximation for the sampling distribution of $\bar{X}$.

Next time: Properties of sampling means; Confidence intervals