Intro to Sampling

Prof. Wells

Math 209, 3/1/23

Outline

In this lecture, we will. . .

Outline

In this lecture, we will...

- Discuss random sampling: the heart of statistics!
- Perform a group sampling activity

Section 1

Sampling

Sampling Overview

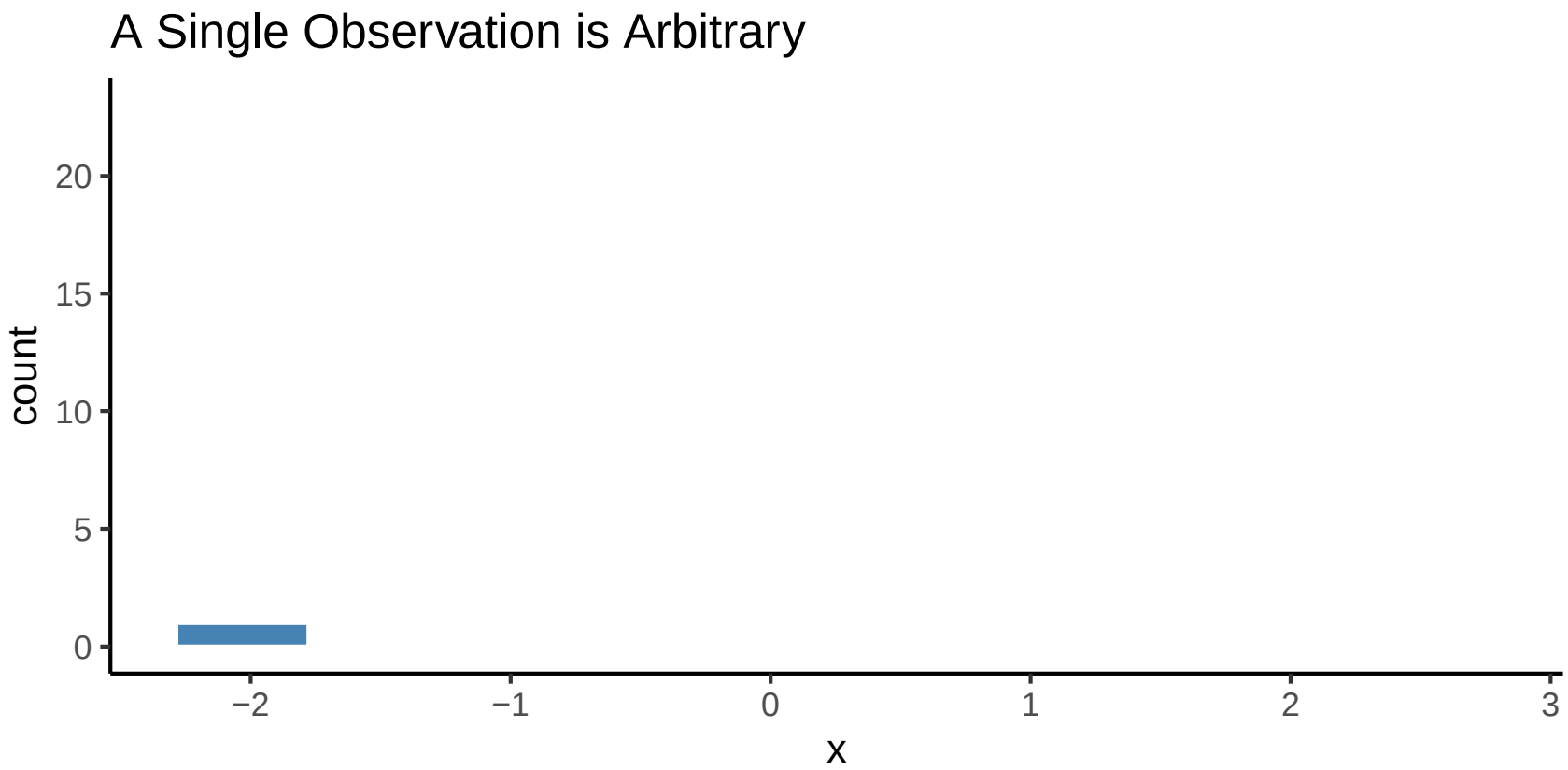
- The distribution of a data set allow us to quantify the shape, center, and spread of the data.

Sampling Overview

- The distribution of a data set allow us to quantify the shape, center, and spread of the data.
- While a single observation in a data set may appear arbitrary, repeated trials show that outcomes indeed follow prescribed patterns.

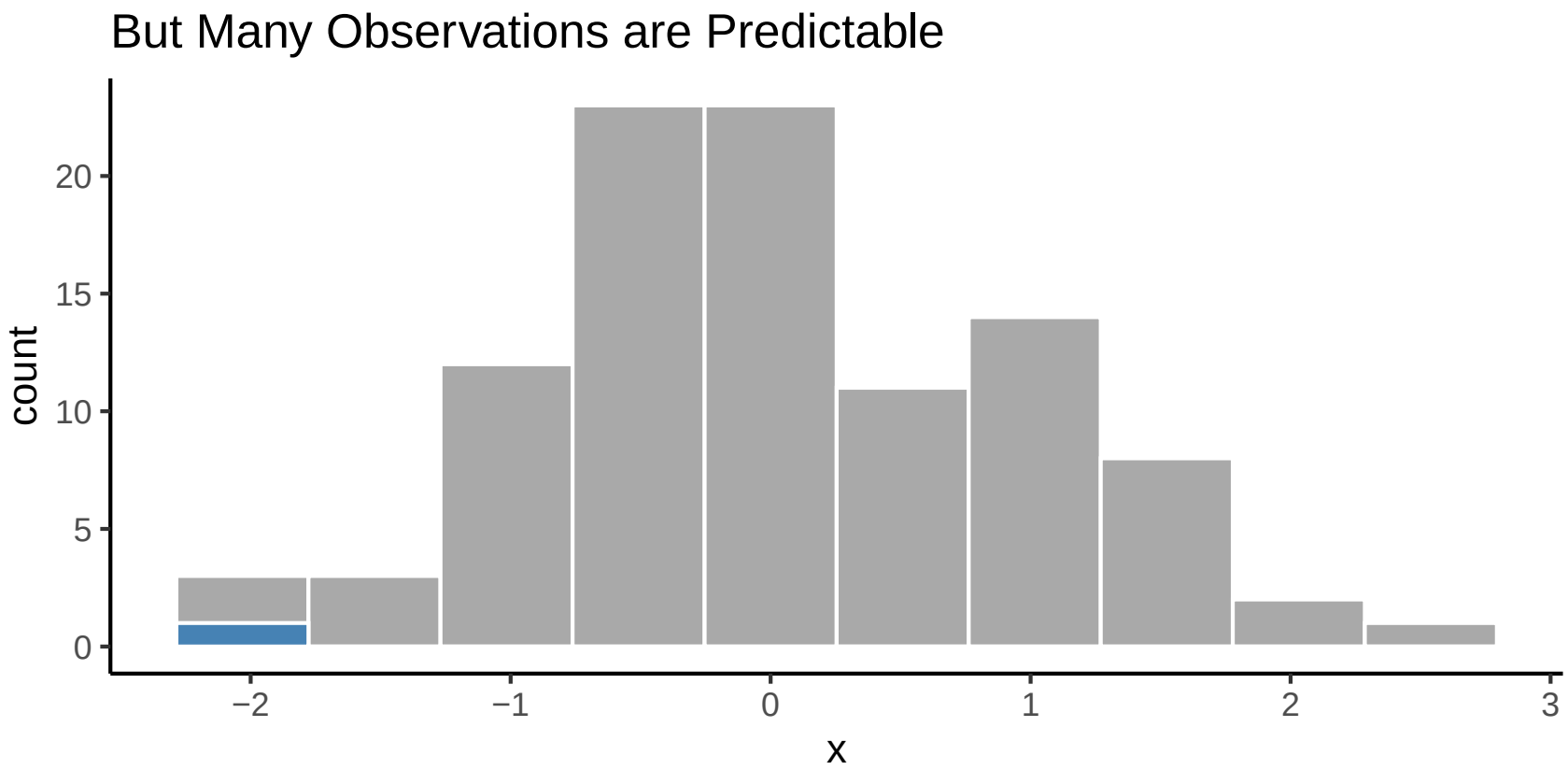
Sampling Overview

- The distribution of a data set allow us to quantify the shape, center, and spread of the data.
- While a single observation in a data set may appear arbitrary, repeated trials show that outcomes indeed follow prescribed patterns.



Sampling Overview

- The distribution of a data set allow us to quantify the shape, center, and spread of the data.
- While a single observation in a data set may appear arbitrary, repeated trials show that outcomes indeed follow prescribed patterns.



Sampling Overview

- A **parameter** is a numeric summary of a variable in a *population*. Researchers are often interested in learning the value of parameters. Ex:

Sampling Overview

- A **parameter** is a numeric summary of a variable in a *population*. Researchers are often interested in learning the value of parameters. Ex:
  - The *proportion* of U.S. voters who plan to vote for a particular presidential candidate.
  - The *mean* time until failure for a computer harddrive.

Sampling Overview

- A **parameter** is a numeric summary of a variable in a *population*. Researchers are often interested in learning the value of parameters. Ex:
  - The *proportion* of U.S. voters who plan to vote for a particular presidential candidate.
  - The *mean* time until failure for a computer harddrive.
- But it is often prohibitively expensive, impractical, time-consuming, or impossible to perform a census to collect complete information on the population.

Sampling Overview

- A **parameter** is a numeric summary of a variable in a *population*. Researchers are often interested in learning the value of parameters. Ex:
  - The *proportion* of U.S. voters who plan to vote for a particular presidential candidate.
  - The *mean* time until failure for a computer harddrive.
- But it is often prohibitively expensive, impractical, time-consuming, or impossible to perform a census to collect complete information on the population.
- A **statistic** is a numeric summary of a variable in a *sample*. Researchers collect a sample and measure the value of the statistic

Sampling Overview

- A **parameter** is a numeric summary of a variable in a *population*. Researchers are often interested in learning the value of parameters. Ex:
  - The *proportion* of U.S. voters who plan to vote for a particular presidential candidate.
  - The *mean* time until failure for a computer harddrive.
- But it is often prohibitively expensive, impractical, time-consuming, or impossible to perform a census to collect complete information on the population.
- A **statistic** is a numeric summary of a variable in a *sample*. Researchers collect a sample and measure the value of the statistic
  - The *proportion* of voters in a sample of size 10 who plan to vote for the candidate.
  - The *mean* time until failure for 10 randomly selected computer harddrives.

Sampling Overview

- A **parameter** is a numeric summary of a variable in a *population*. Researchers are often interested in learning the value of parameters. Ex:
  - The *proportion* of U.S. voters who plan to vote for a particular presidential candidate.
  - The *mean* time until failure for a computer harddrive.
- But it is often prohibitively expensive, impractical, time-consuming, or impossible to perform a census to collect complete information on the population.
- A **statistic** is a numeric summary of a variable in a *sample*. Researchers collect a sample and measure the value of the statistic
  - The *proportion* of voters in a sample of size 10 who plan to vote for the candidate.
  - The *mean* time until failure for 10 randomly selected computer harddrives.
- Sample statistics serve as our best estimators for the corresponding population parameters.

Sampling Overview

- A **parameter** is a numeric summary of a variable in a *population*. Researchers are often interested in learning the value of parameters. Ex:
  - The *proportion* of U.S. voters who plan to vote for a particular presidential candidate.
  - The *mean* time until failure for a computer harddrive.
- But it is often prohibitively expensive, impractical, time-consuming, or impossible to perform a census to collect complete information on the population.
- A **statistic** is a numeric summary of a variable in a *sample*. Researchers collect a sample and measure the value of the statistic
  - The *proportion* of voters in a sample of size 10 who plan to vote for the candidate.
  - The *mean* time until failure for 10 randomly selected computer harddrives.
- Sample statistics serve as our best estimators for the corresponding population parameters.
- Parameters are fixed values, but usually unknown.

Sampling Overview

- A **parameter** is a numeric summary of a variable in a *population*. Researchers are often interested in learning the value of parameters. Ex:
  - The *proportion* of U.S. voters who plan to vote for a particular presidential candidate.
  - The *mean* time until failure for a computer harddrive.
- But it is often prohibitively expensive, impractical, time-consuming, or impossible to perform a census to collect complete information on the population.
- A **statistic** is a numeric summary of a variable in a *sample*. Researchers collect a sample and measure the value of the statistic
  - The *proportion* of voters in a sample of size 10 who plan to vote for the candidate.
  - The *mean* time until failure for 10 randomly selected computer harddrives.
- Sample statistics serve as our best estimators for the corresponding population parameters.
- Parameters are fixed values, but usually unknown.
- Statistics are known, but vary from sample to sample.

Sampling Example

- Consider a collection of 10 novels on a bookshelf (*the population*)

```
##                          Title Pages
## 1            A Game of Thrones   694
## 2                         Dune   412
## 3  The Fellowship of the Ring   479
## 4               The Two Towers   352
## 5       The Return of the King   416
## 6                   Good Omens   288
## 7         The Name of the Wind   662
## 8                American Gods   465
## 9                   Foundation   255
## 10                    Hyperion   482
```

Sampling Example

- Consider a collection of 10 novels on a bookshelf (*the population*)

```
##                          Title Pages
## 1            A Game of Thrones   694
## 2                         Dune   412
## 3  The Fellowship of the Ring   479
## 4               The Two Towers   352
## 5      The Return of the King   416
## 6                   Good Omens   288
## 7        The Name of the Wind   662
## 8                American Gods   465
## 9                   Foundation   255
## 10                    Hyperion   482
```

- The books on the shelf have an average page count (*a parameter*)

```
##   avg_Pages
## 1     450.5
```

Sampling Example

- We could randomly choose 3 books from the shelf (*a sample*):

```
##                   Title Pages
## 1           Foundation   255
## 2       The Two Towers   352
## 3 The Name of the Wind   662
```

Sampling Example

- We could randomly choose 3 books from the shelf (*a sample*):

```
##                   Title Pages
## 1           Foundation   255
## 2       The Two Towers   352
## 3 The Name of the Wind   662
```

- Which have their own mean page count (*a statistic*):

```
##   avg_Pages
## 1       423
```

Sampling Example

- We could randomly choose 3 books from the shelf (*a sample*):

```
##                   Title Pages
## 1           Foundation   255
## 2       The Two Towers   352
## 3 The Name of the Wind   662
```

- Which have their own mean page count (*a statistic*):

```
##   avg_Pages
## 1       423
```

- Or we could choose 3 other books (*a sample*)

```
##                        Title Pages
## 1                 Good Omens   288
## 2               American Gods   465
## 3 The Fellowship of the Ring   479
```

Sampling Example

- We could randomly choose 3 books from the shelf (*a sample*):

```
##                   Title Pages
## 1           Foundation   255
## 2       The Two Towers   352
## 3 The Name of the Wind   662
```

- Which have their own mean page count (*a statistic*):

```
##   avg_Pages
## 1       423
```

- Or we could choose 3 other books (*a sample*)

```
##                        Title Pages
## 1                 Good Omens   288
## 2               American Gods   465
## 3 The Fellowship of the Ring   479
```

- Which will have a new mean page count (*a statistic*):

```
##   avg_Pages
## 1  410.6667
```

Sampling Example

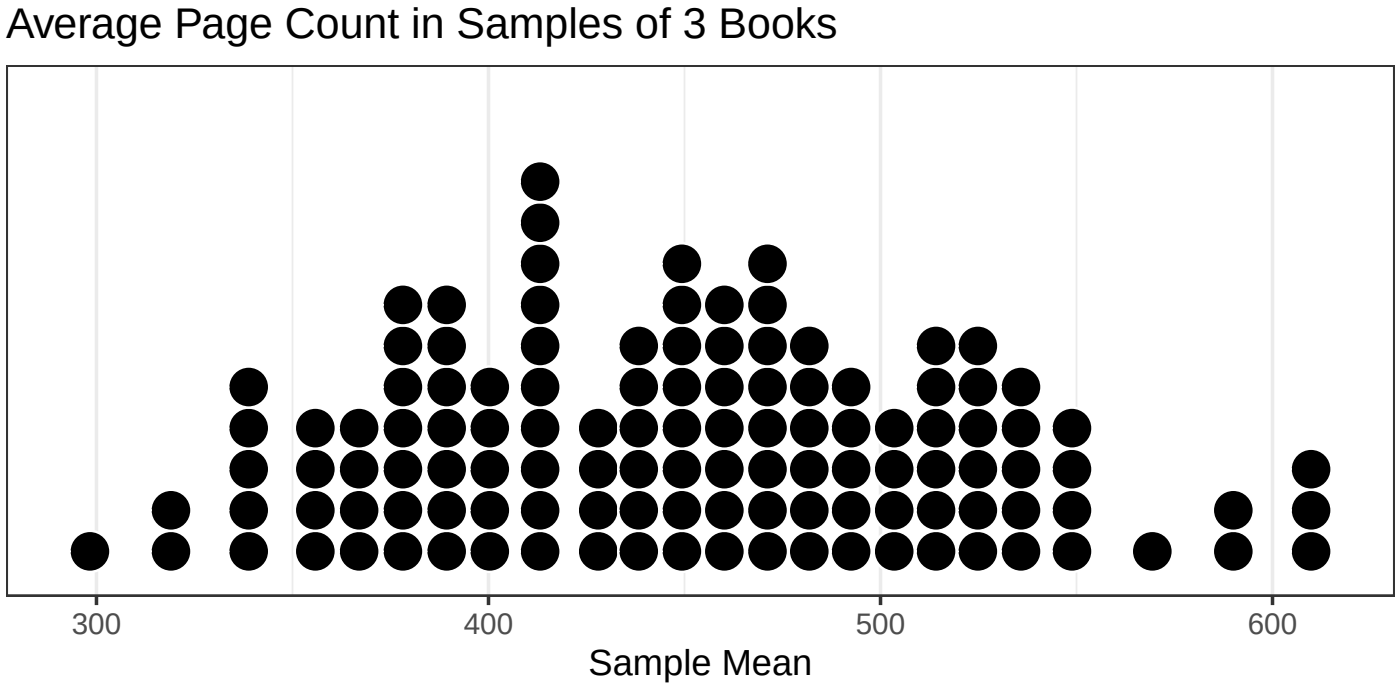
- There are 120 different samples of size 3 we could choose from a set of 10 books.

Sampling Example

- There are 120 different samples of size 3 we could choose from a set of 10 books.
- The collection of all 120 sample means form its own data set:

Sampling Example

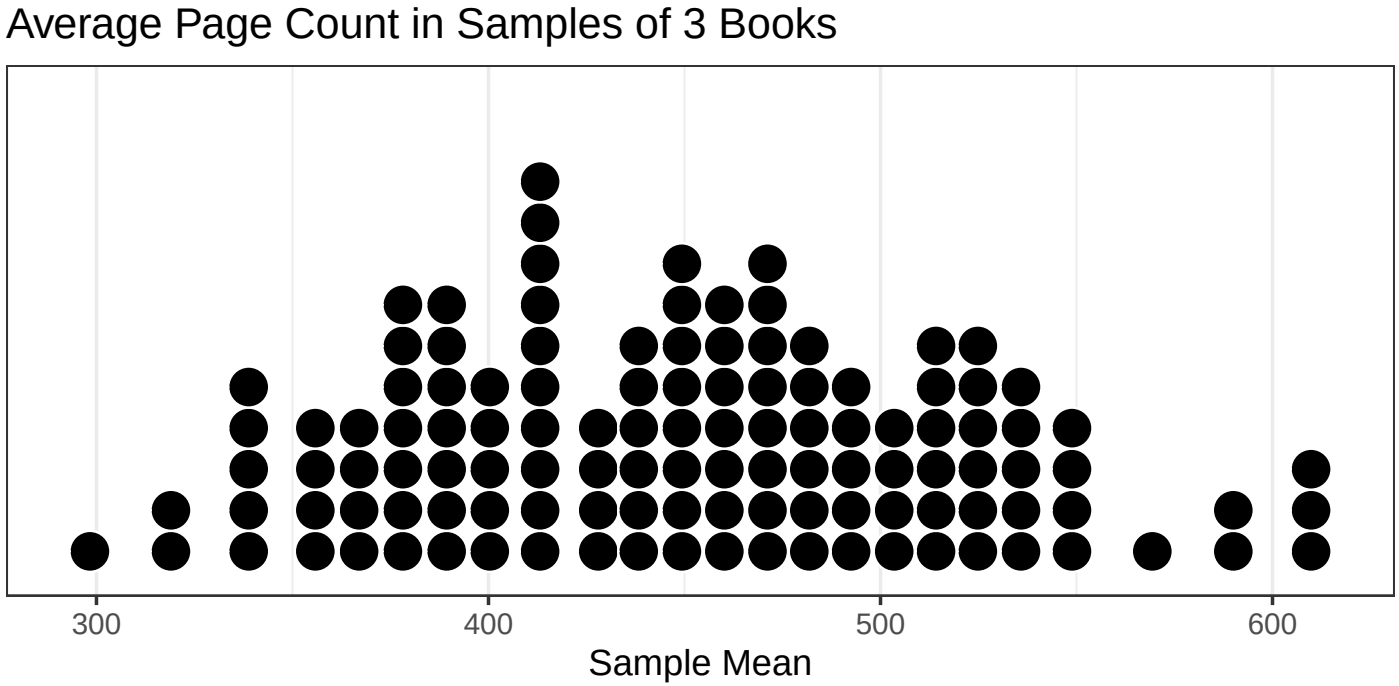
- There are 120 different samples of size 3 we could choose from a set of 10 books.
- The collection of all 120 sample means form its own data set:



- Each dot represents the mean page count in a sample of 3 books.

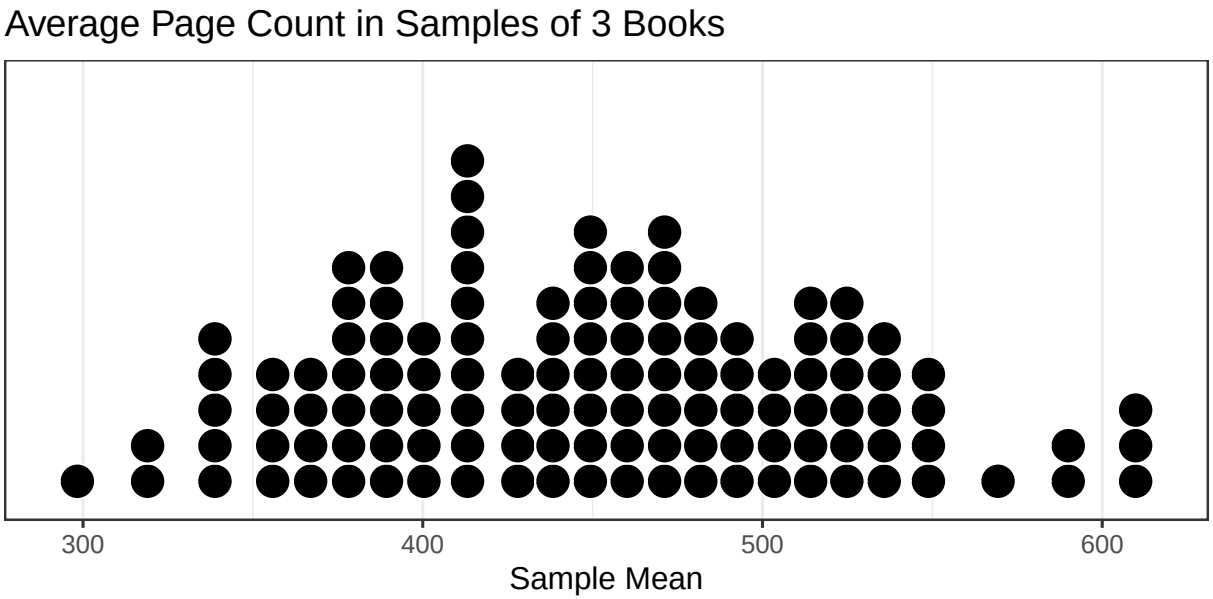
Sampling Example

- There are 120 different samples of size 3 we could choose from a set of 10 books.
- The collection of all 120 sample means form its own data set:



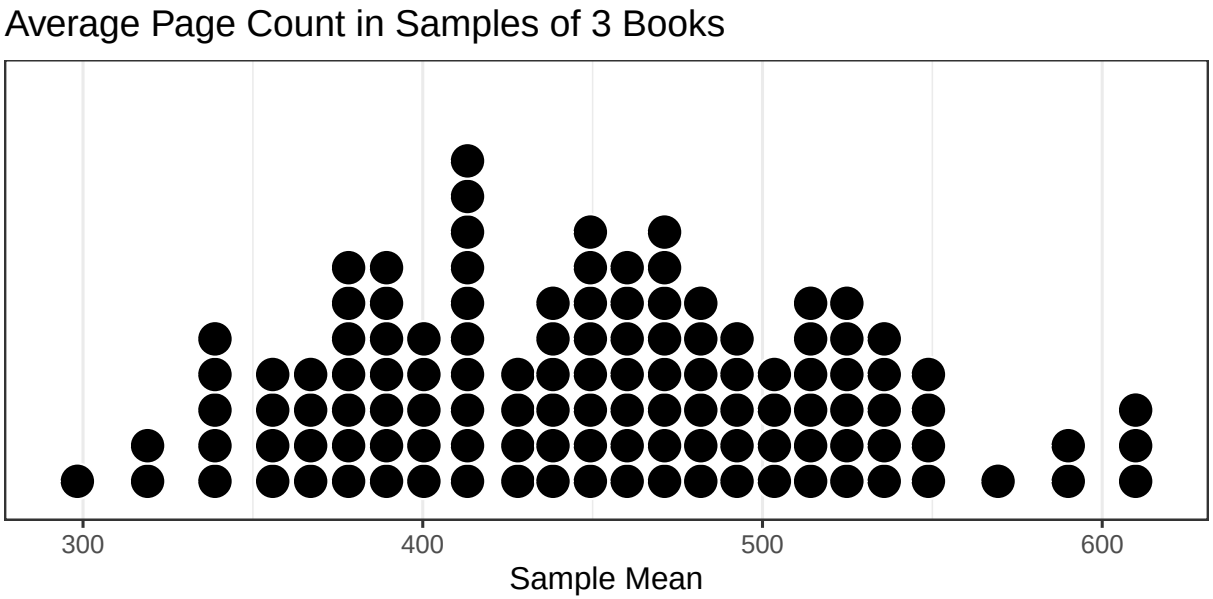
- Each dot represents the mean page count in a sample of 3 books.
  - Most samples have means close to the true mean (450 pages)
  - But other samples are more extreme (averages < 350 or > 550)

The Sampling Distribution



- As a data set, the collection of sample means themselves have a *mean*, *standard deviation* and a *5-number summary*

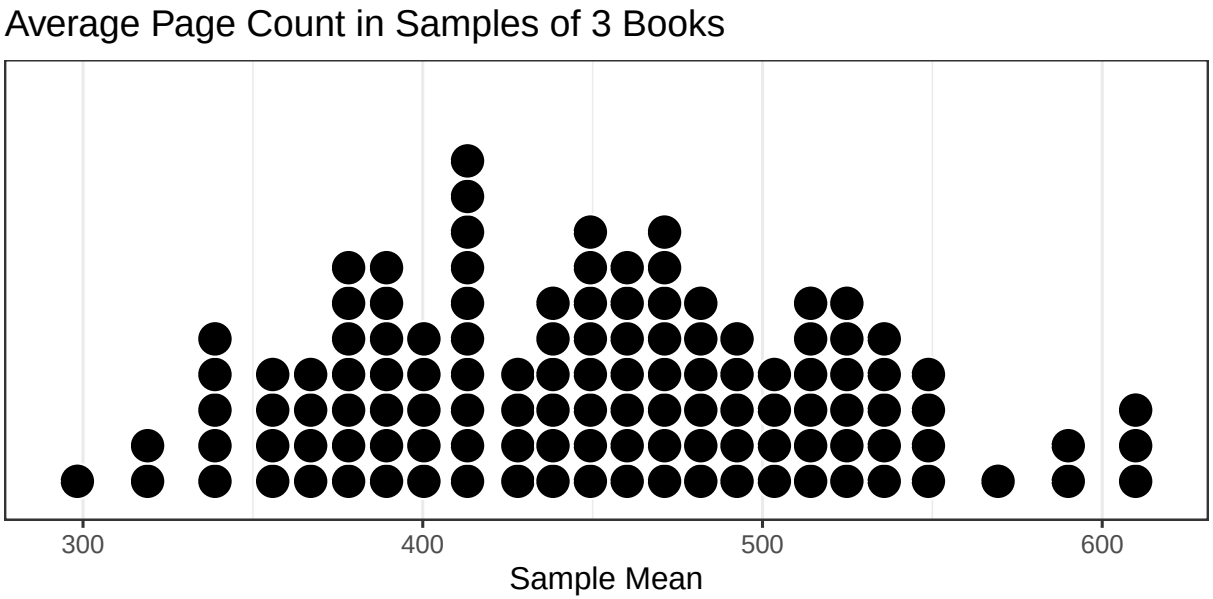
The Sampling Distribution



- As a data set, the collection of sample means themselves have a *mean*, *standard deviation* and a *5-number summary*

```
##   mean sd min  q1 median  q3 max
## 1  450 69 298 395    454 500 613
```

The Sampling Distribution

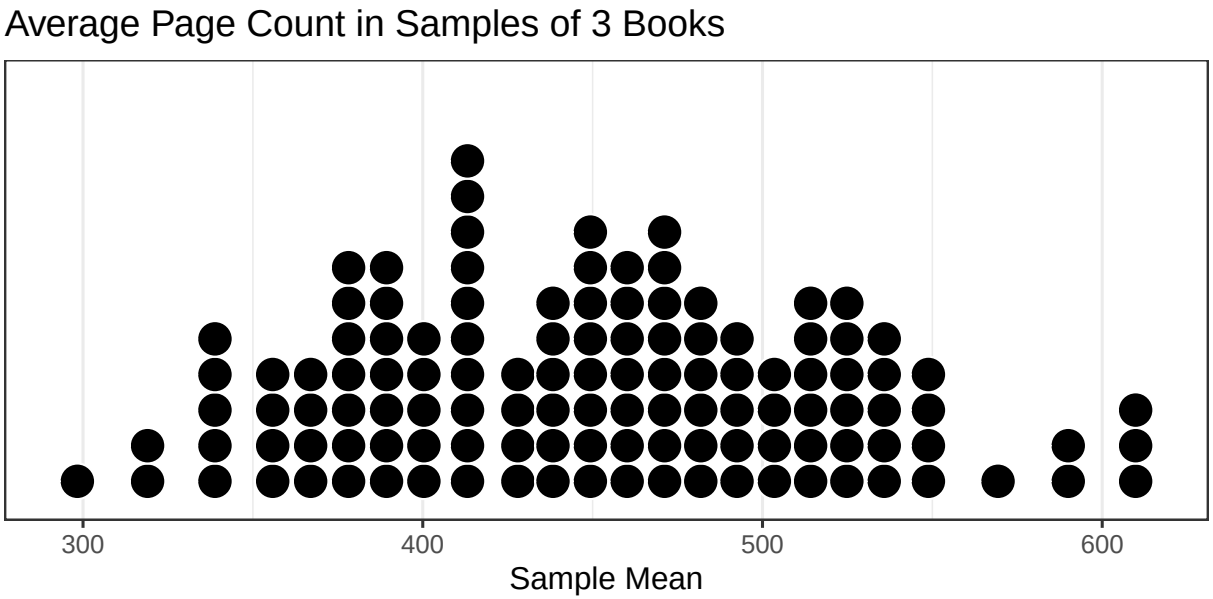


- As a data set, the collection of sample means themselves have a *mean*, *standard deviation* and a *5-number summary*

```
##   mean sd min  q1 median  q3 max
## 1  450 69 298 395    454 500 613
```

- The mean tells us the typical value of the statistic in random sample.

The Sampling Distribution



- As a data set, the collection of sample means themselves have a *mean*, *standard deviation* and a *5-number summary*

```
##   mean sd min  q1 median  q3 max
## 1  450 69 298 395    454 500 613
```

- The mean tells us the typical value of the statistic in random sample.
- The standard deviation indicates how the statistic varies from sample to sample.

Sampling Activity

- Consider a standard deck of 52 playing cards, consisting of...
  - 4 cards each of Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King

Sampling Activity

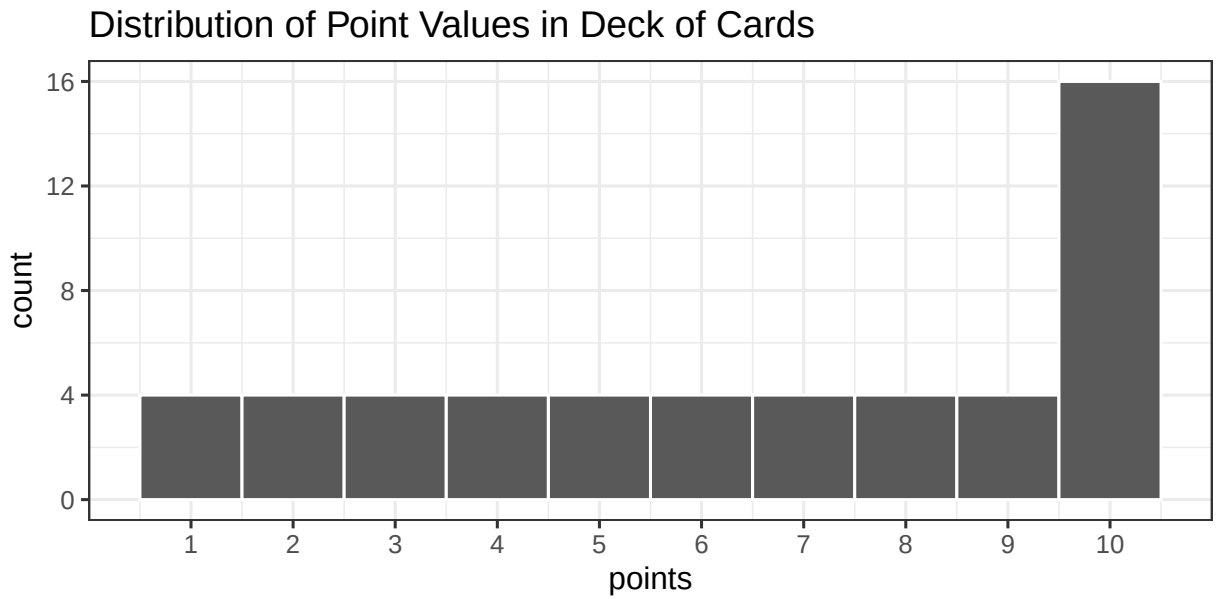
- Consider a standard deck of 52 playing cards, consisting of...
  - 4 cards each of Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King

- Assign a point value to each card:
  - Face Cards (Jack, Queen, King) are worth 10
  - Ace is worth 1
  - All numeric cards are worth their printed number.

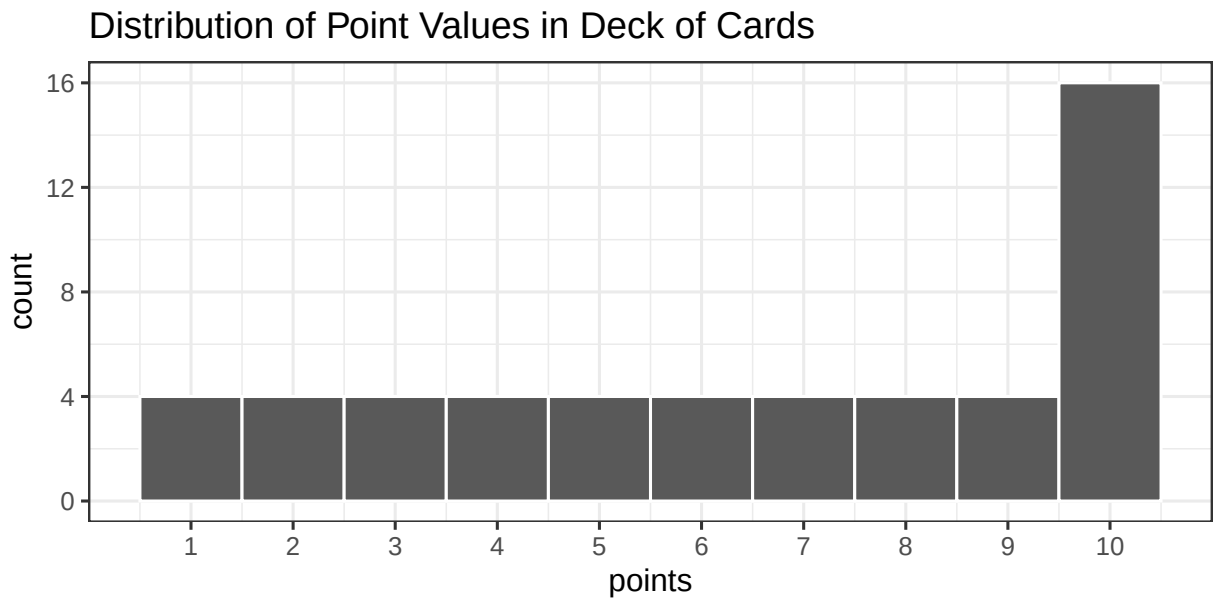
Sampling Activity

- Consider a standard deck of 52 playing cards, consisting of...
  - 4 cards each of Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King

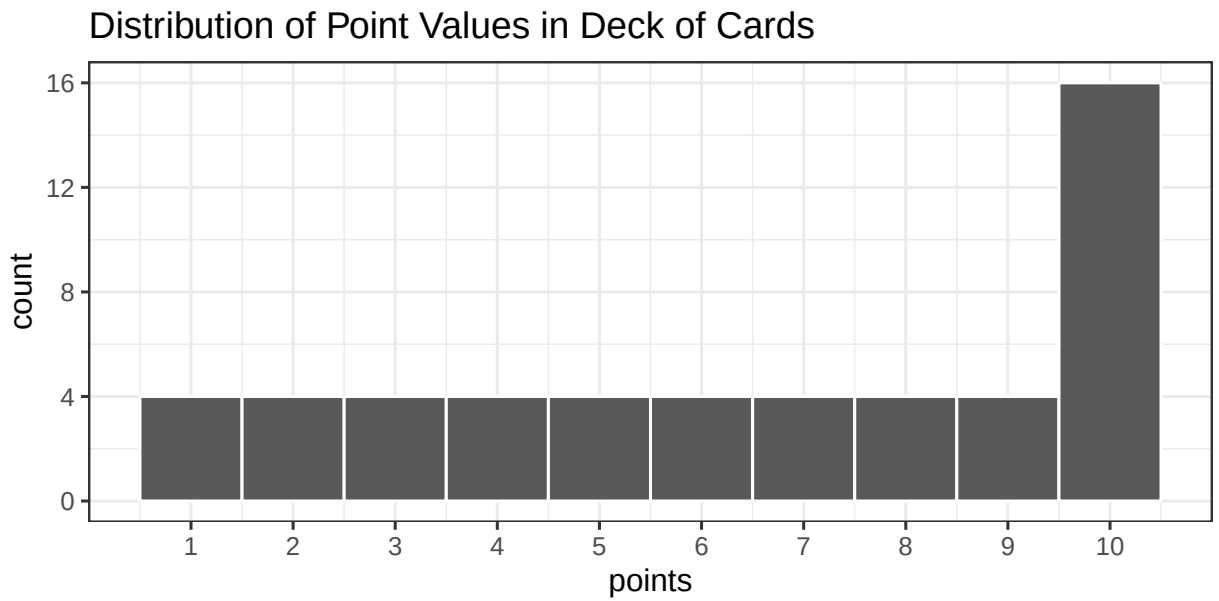
- Assign a point value to each card:
  - Face Cards (Jack, Queen, King) are worth 10
  - Ace is worth 1
  - All numeric cards are worth their printed number.



Sampling Activity



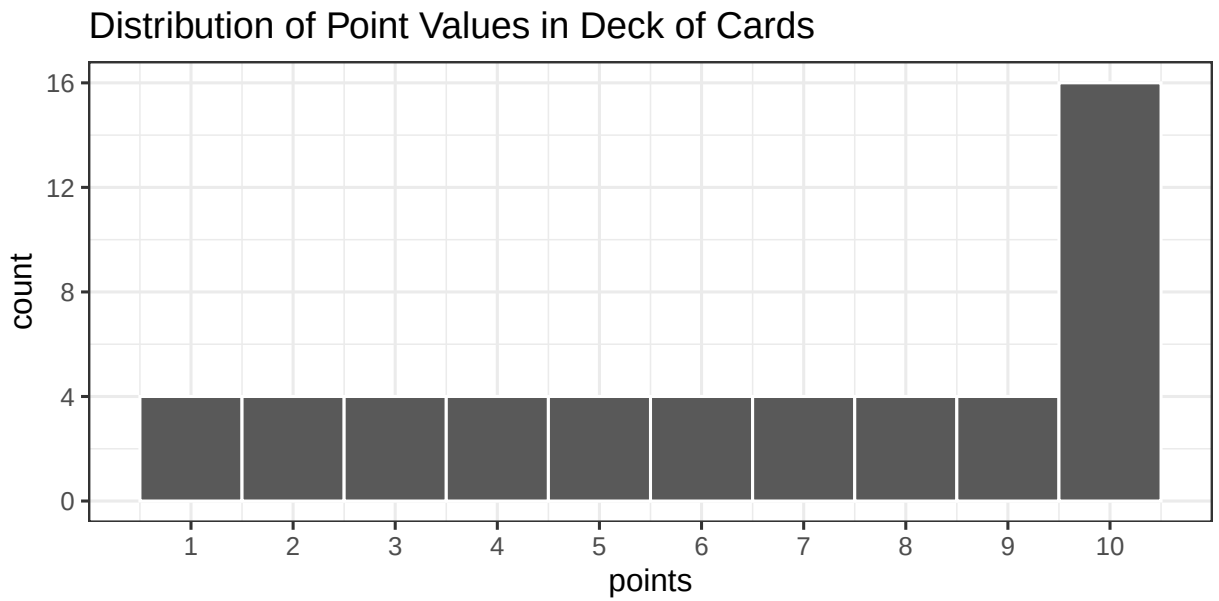
Sampling Activity



- The mean, standard deviation and median of point values in the deck are:

```
##   mean  sd median
## 1  6.5 3.2      7
```

Sampling Activity

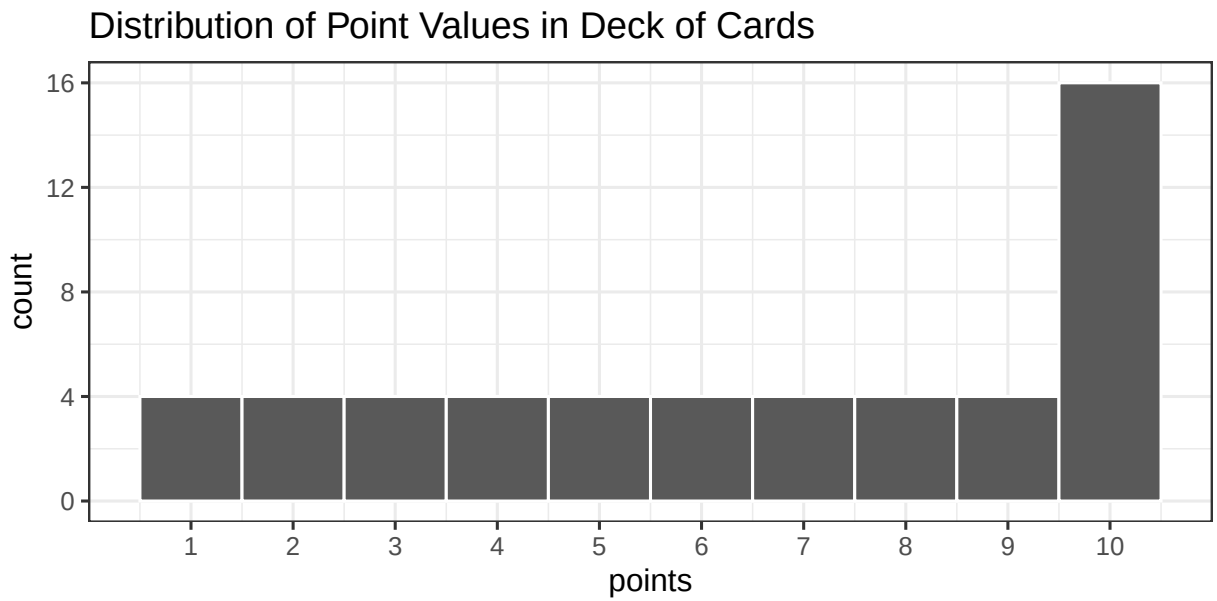


- The mean, standard deviation and median of point values in the deck are:

```
##   mean  sd median
## 1  6.5 3.2      7
```

- Note that the distribution of point values is slightly left-skewed (there are a large number of high point cards)

Sampling Activity



- The mean, standard deviation and median of point values in the deck are:

```
##   mean  sd median
## 1  6.5 3.2      7
```

- Note that the distribution of point values is slightly left-skewed (there are a large number of high point cards)

- Suppose we draw a hand of 10 cards and compute the average point value of the hand.
  - What is the distribution of average point values across all possible samples?

Sampling Activity

Activity

1. Thoroughly shuffle one of your group's deck of cards.
2. Draw 10 cards from the deck (without replacement) to form a sample.
3. Compute the mean point value of your hand.
4. Write the value of the mean on a sticky note and add to whiteboard.
5. Repeat steps 1 - 4 four additional times, per person.

Discussion

Answer the following questions in your group:

- What appears to be the average value of the sample mean?
- How does this compare to the average point value in the deck of cards?
- Which distribution appears to have more variability: the distribution of sample means or the distribution of point values in the deck?
- How do the shapes of the two distributions compare?
- What are the approximate 1st and 3rd quartiles for the distribution of sample means?
- What does this suggest about the value of *most* sample means?

Theoretical Sampling Distribution