ECON 3818

Chapter 15

Kyle Butts

28 July 2021

Chapter 15: Parameters and Statistics

Parameters and Statistics

We have discussed using sample data to make inference about the population. In particular, we will use sample statistics to make inference about population **parameters**.

A **parameter** is a number that describes the population. In practice, parameters are unknown because we cannot examine the entire population.

A statistic is a number that can be calculated from sample data without using any unknown parameters. In practice, we use statistics to estimate parameters.

Greek Letters and Statistics

Latin Letters

- Latin letters like $\bar{x}$ and s^2 are calculations that represent guesses (estimates) at the population values.

Greek Letters

- Greek letters like μ and σ^2 represent the truth about the population.

The goal for the class is for the latin letters to be good guesses for the greek letters:

$$\text{Data} \longrightarrow \text{Calculation} \longrightarrow \text{Estimates} \longrightarrow^{hopefully!} \text{Truth}$$

For example,

$$X \longrightarrow \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \bar{x} \longrightarrow^{hopefullly!} \mu$$

Examples of Parameters

Some parameters of distributions we've encountered are

- n and p in $X \sim B(n, p)$ with probability mass function

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- a and b in $X \sim U(a, b)$ with probability density function

$$f(x) = \frac{1}{b - a}$$

- μ and σ^2 in $X \sim N(\mu, \sigma^2)$ with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2}$$

Mean and Variance

Two population parameters of particular interest are

- the mean, denoted μ, defined by $E(X)$
- the variance, denoted σ^2, defined by $E(X^2) - E(X)^2$

We **do not** observe these. Therefore, we guess using

- the sample mean, $\bar{X}$
- the sample variance, s^2

Why do we use these as our guess?

Getting the right sample

Before we talk about the properties of sample statistics, we need to make sure we have the right sample. We talked about good ways to generate a sample.

The right sample is the most important part of any data analysis.

A Simple Random Sample has no bias and has observations that are from the same population.

Identically Distributed

If every observation is from the same population, we say all of the observations in our sample are **identically distributed**. In math, this means for any two observations X_i and X_j,

$$Pr(X_i < x) = Pr(X_j < x)$$

Independent Observations

Does observing X_i impact our best guess of X_j? Sometimes yes (time series, spatial dependence), but hopefully not.

To simplify things, we need to assume **independent sample observations**, meaning

$$Pr(X_i = a \mid X_j = b) = Pr(X_i = a)$$

Intuitively, this means that *observing* one outcome doesn't help you *predict* any other outcome.

To summarize, we want an *i.i.d.* sample, i.e. sample observations that are **independent and identically distributed**.

Sample Statistics are Random Variables

For a sample $X_1, \ldots, X_n$ of the random variable X, any function of that sample, $\hat{\theta} = g(X_1, \ldots, X_n)$, is a **sample statistic**. For example,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Because $X_1, \ldots, X_n$ are random variables, any sample statistic $\hat{\theta} = g(X_1, \ldots, X_n)$ is itself a random variable!

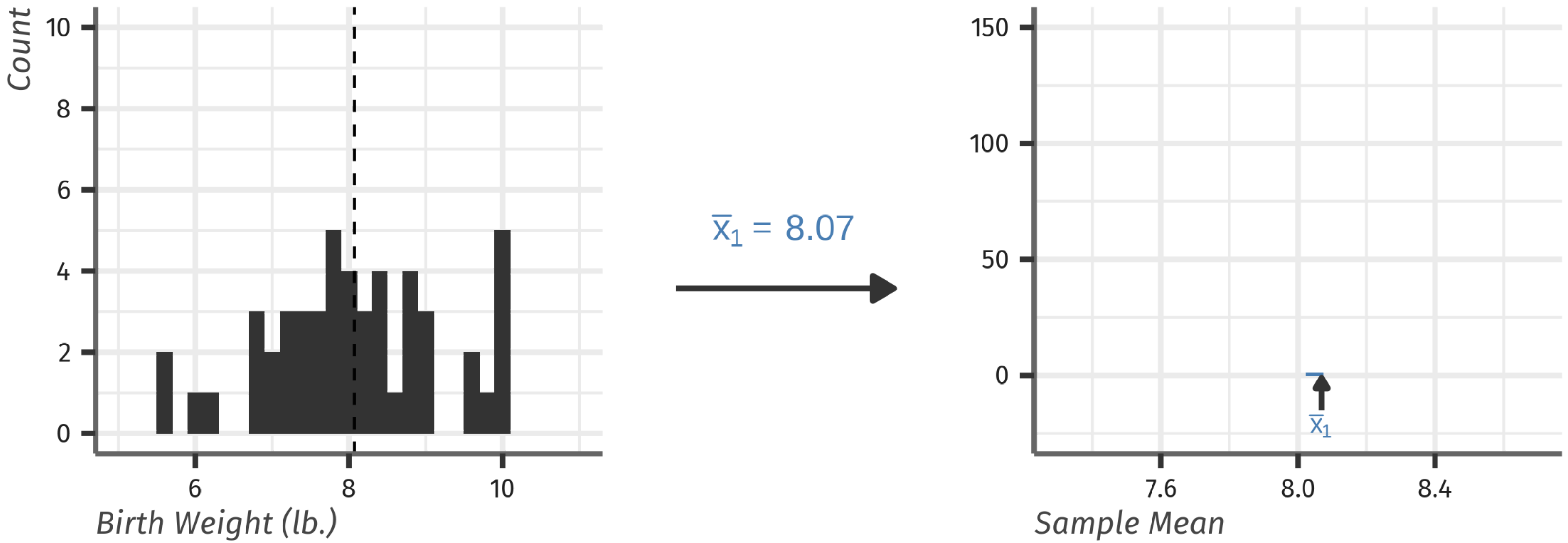
That means, there is some distribution for the values of $\hat{\theta}$

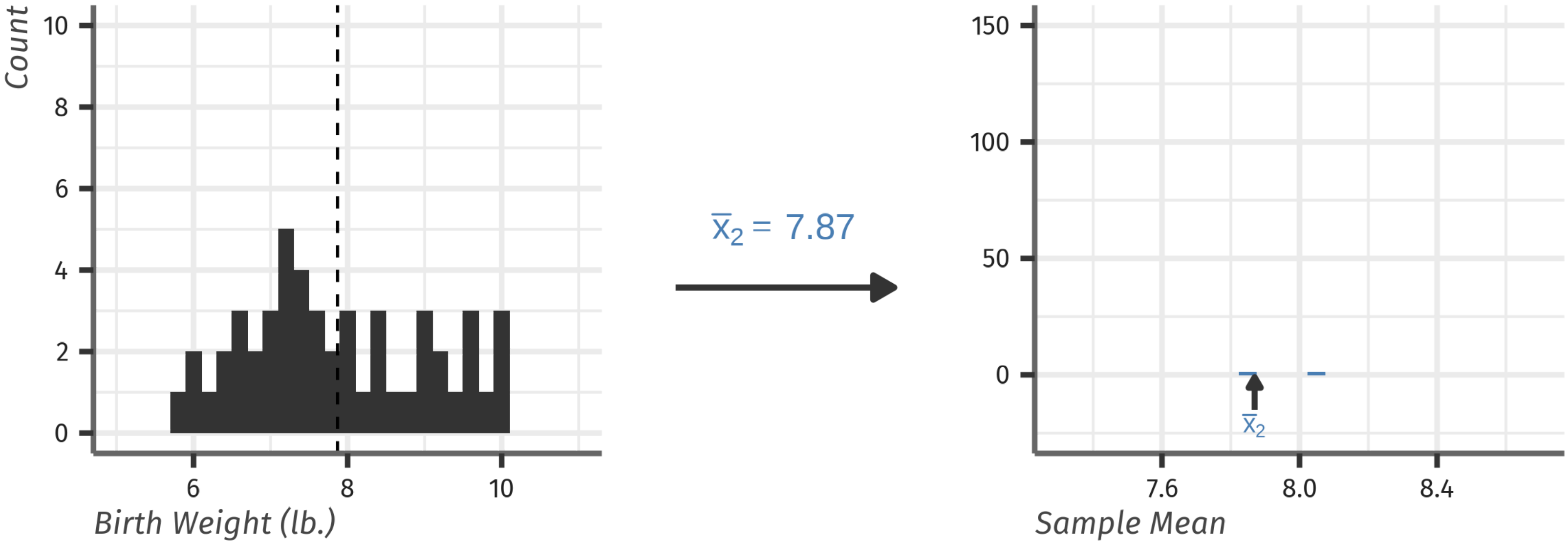
Sampling Distributions

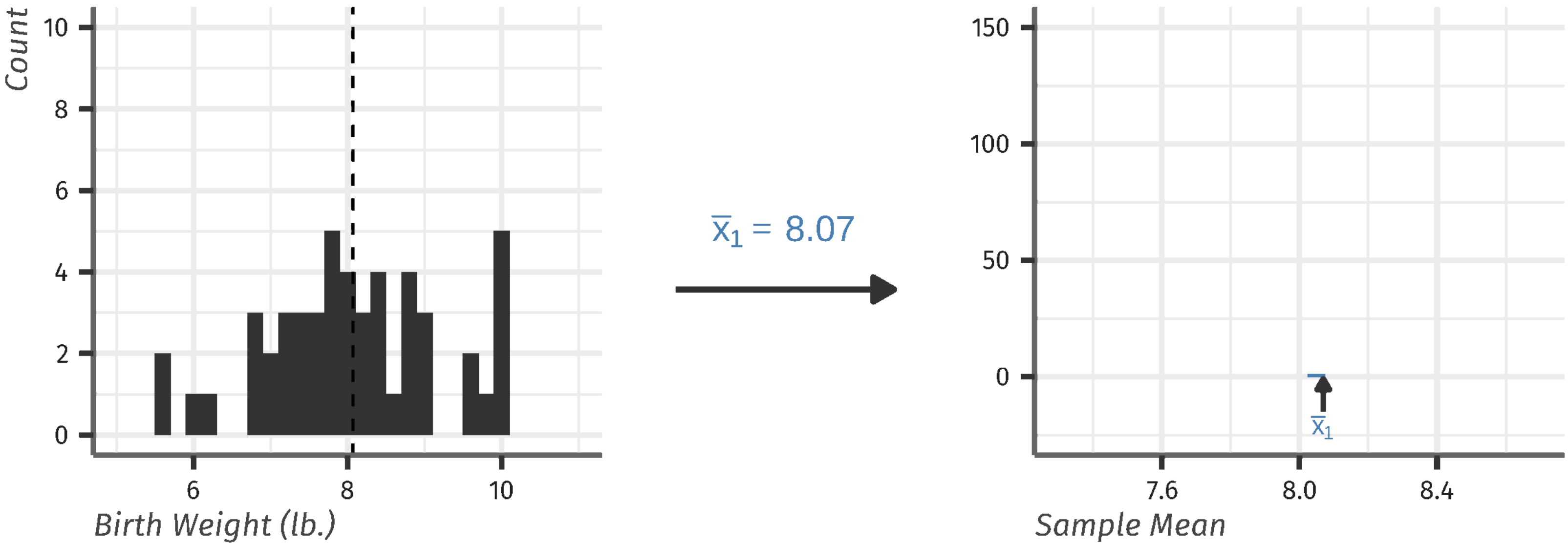
This is one of the most important concepts in the course. One **trial** would consist of the following:

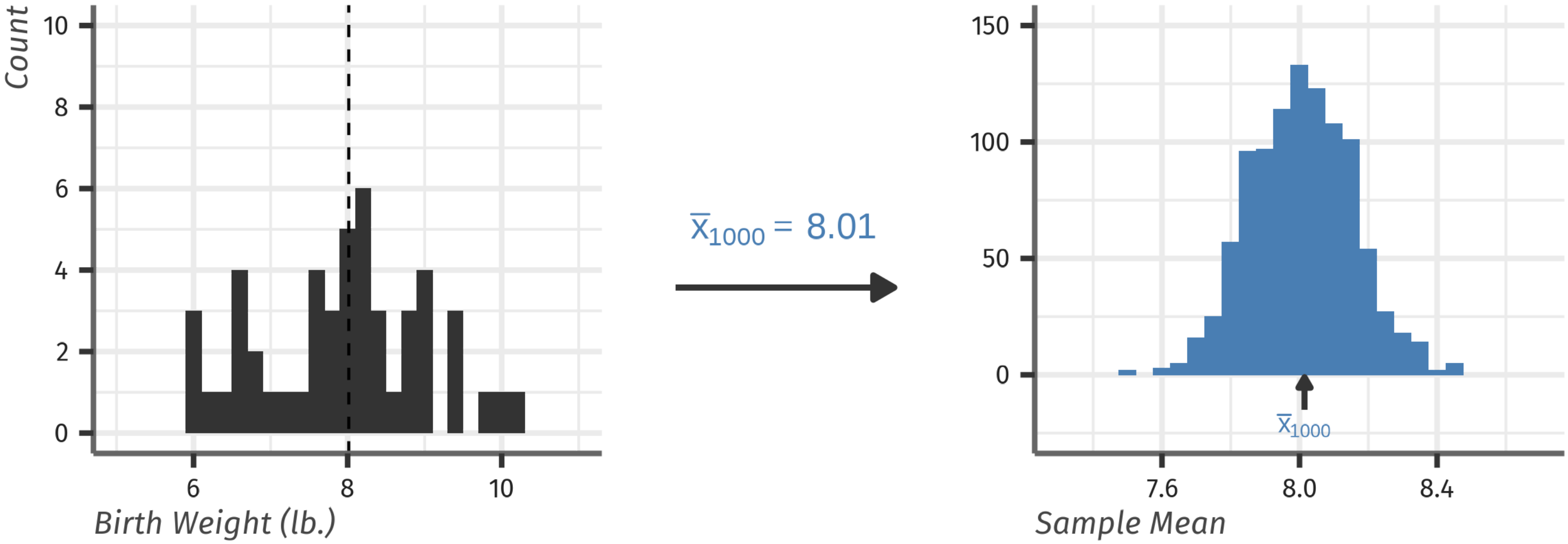
- Random Sample - Grab a group of observations from the population
- Sample Statistic - Take your particular random sample and calculate a sample statistic (e.g. sample mean)

Sampling Distribution - Imagine repeatedly grabbing a different group of observations from the population and calculating the sample mean. This is performing many **trials**. The sample means themselves will have a distribution.


Count
10
8
6
4
2
0
6
8
10
Birth Weight (lb.)
$\bar{x}_1 = 8.07$
150
100
50
0
7.6
8.0
8.4
$\bar{x}_1$
Sample Mean

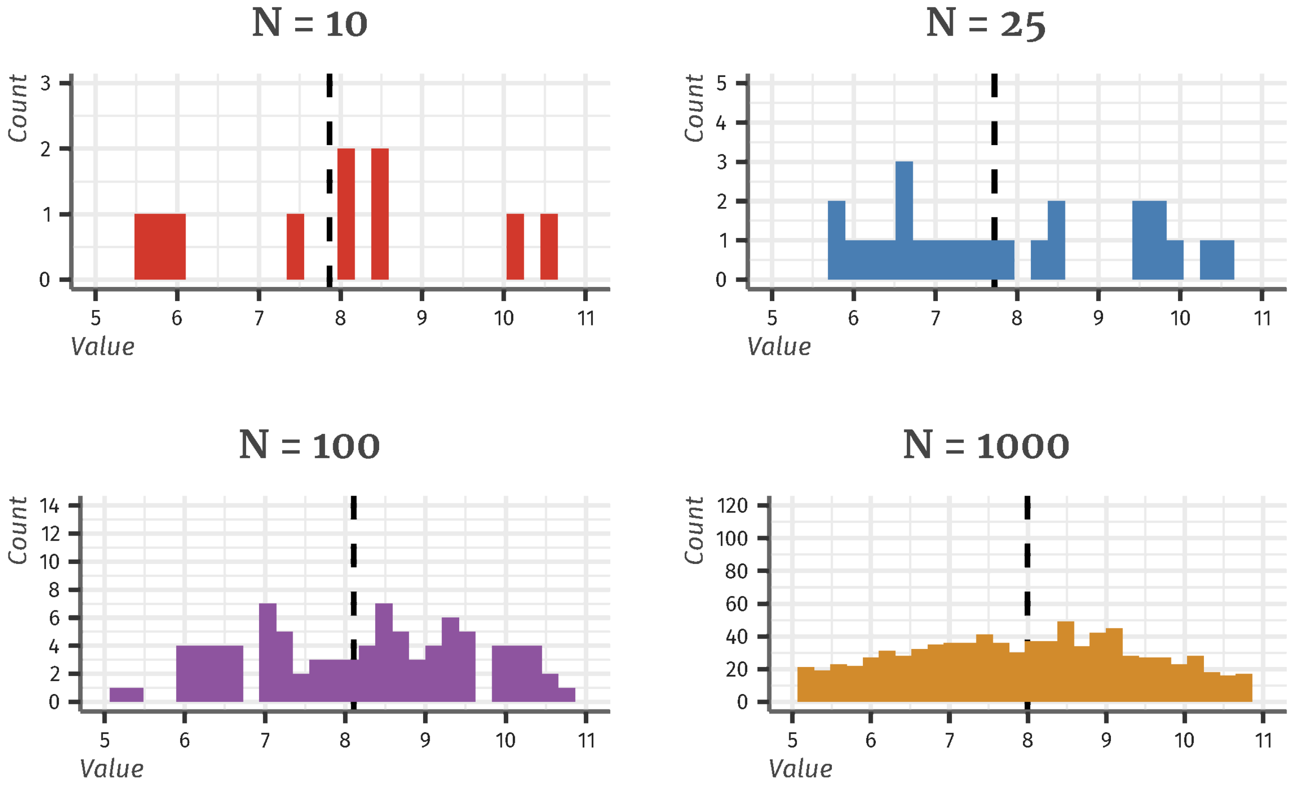

Count
10
8
6
4
2
0
6
8
10
Birth Weight (lb.)
$\bar{x}_2 = 7.87$
150
100
50
0
7.6
8.0
8.4
Sample Mean
$\bar{x}_2$


Count
10
8
6
4
2
0
6
8
10
Birth Weight (lb.)
$\bar{x}_1 = 8.07$
150
100
50
0
$\bar{x}_1$
7.6
8.0
8.4
Sample Mean


Count
10
8
6
4
2
0
6
8
10
Birth Weight (lb.)
$\overline{x}_{1000}$ = 8.01
150
100
50
0
$\overline{x}_{1000}$
7.6
8.0
8.4
Sample Mean

Sample Size

The variance of the *sampling distribution* depends on the sample size. As n gets larger, each individual **trial** gives a better guess at the mean. Hence, the sampling distribution gets more narrow

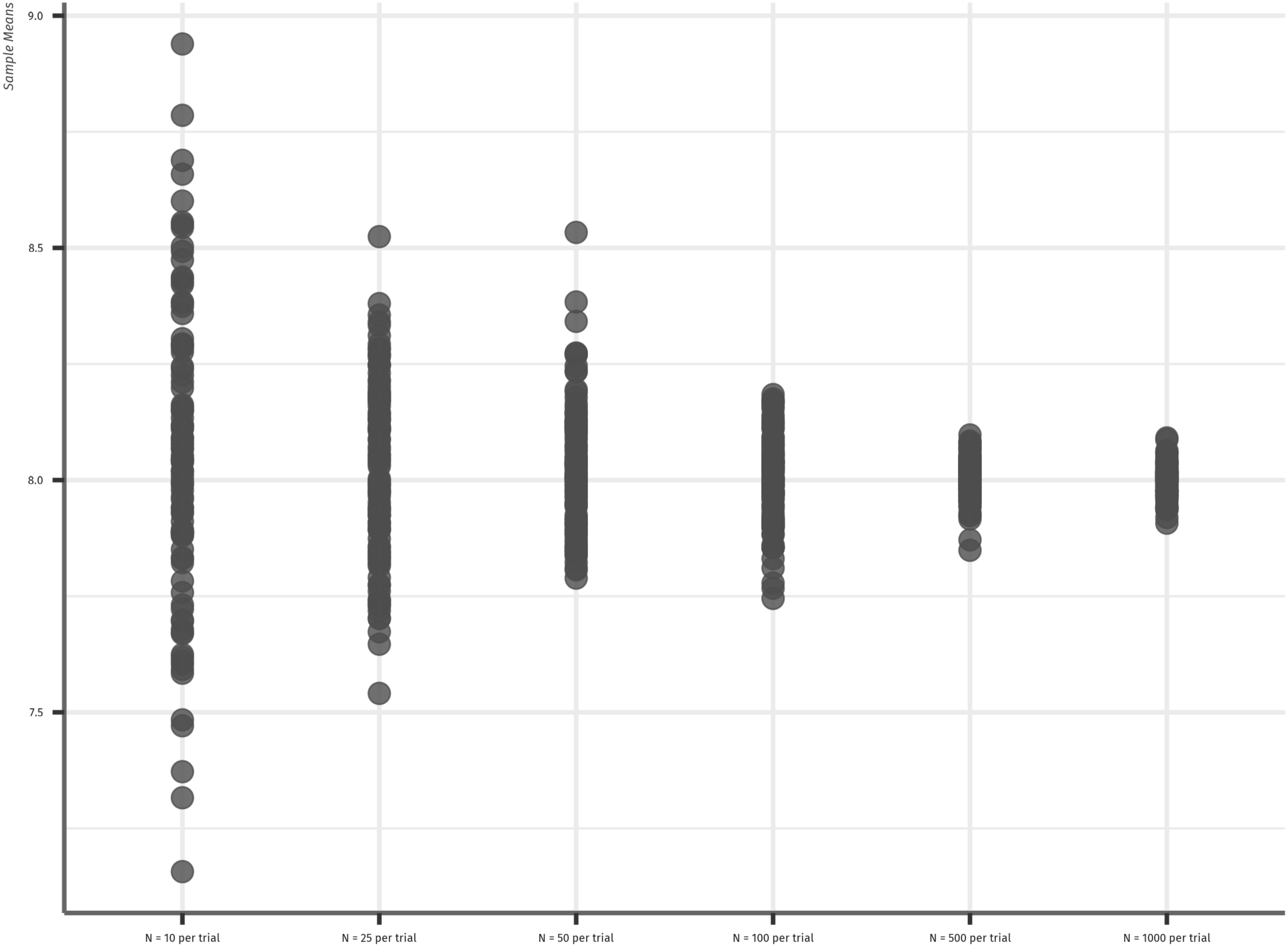


N = 10

N = 25

N = 100

N = 1000


Sample Means
9.0
8.5
8.0
7.5
N = 10 per trial
N = 25 per trial
N = 50 per trial
N = 100 per trial
N = 500 per trial
N = 1000 per trial

Sampling Distributions

We will only observe 1 sample in the world though.

How does the concept of sampling distribution help us?

- Since we don't know the true population parameter, Our sample statistic will be our best guess at the possible true value.
- If we know the sampling distribution, then we can consider uncertainty about our sample statistic.

Law of Large Numbers

Is $\bar{X}$ actually a good guess for μ? Under certain conditions, we can use the **Law of Large Numbers (LLN)** to guarantee that $\bar{X}$ approaches μ as the sample size grows large.

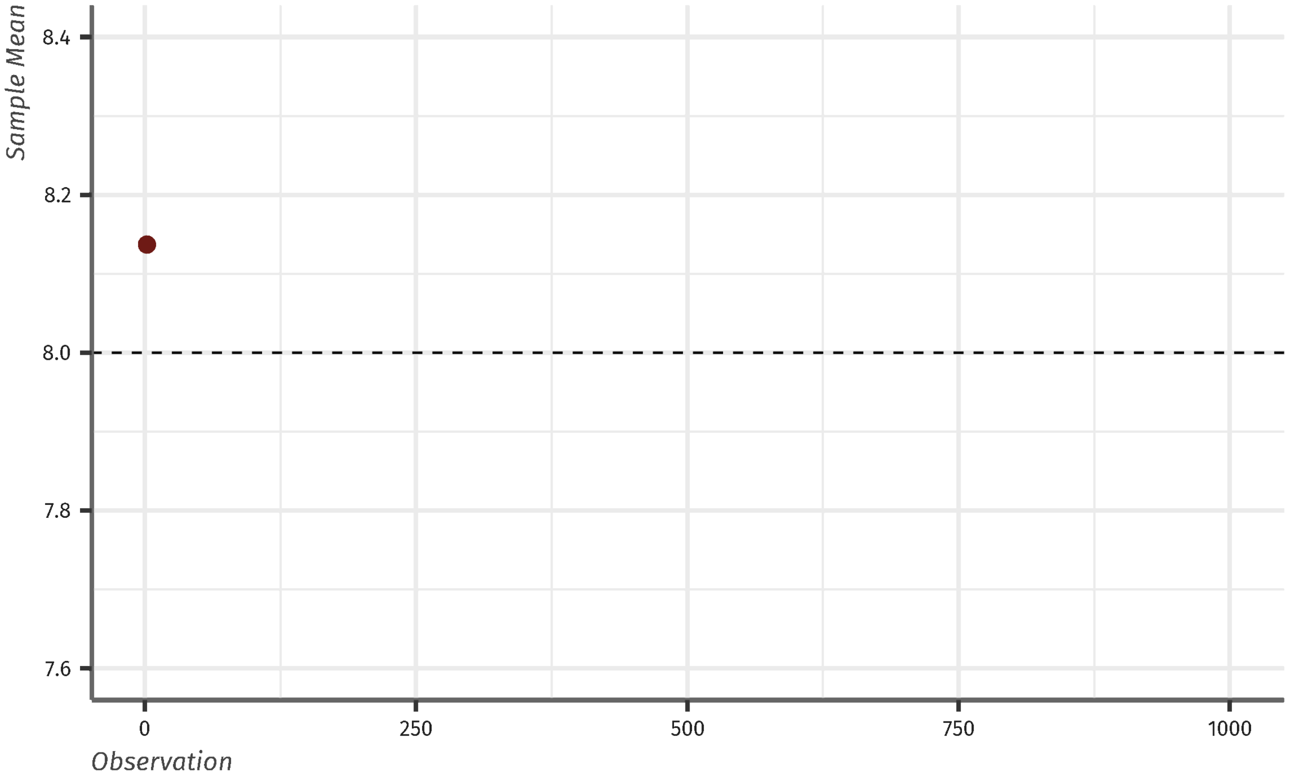
Theorem: Let $X_1, X_2, \ldots, X_n$ be an i.i.d. set of observations with $E(X_i) = \mu$.

Define the sample mean of size n as $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\bar{X}_n \to \mu \quad \text{as} \quad n \to \infty.$$

Intuitively, as we observe a larger and larger sample, we average over randomness and our sample mean approaches the true population mean.

Law of Large Numbers



Law of Large Numbers

Properties of the sample mean

Theorem: Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Then

$$E(\bar{X}_n) = \mu$$

$$Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

Intuitively, we grab many samples from a population. The variance of our sample averages shrinks as we observe more observations per sample.

Clicker Question

Suppose we sample 100 observations from a distribution with $\mu = 15$ and $\sigma^2 = 25$. What are $E(\bar{X}_{100})$ and $Var(\bar{X}_{100})$?

a. $E(\bar{X}_{100}) = 15, Var(\bar{X}_{100}) = 25$
b. $E(\bar{X}_{100}) = 0.15, Var(\bar{X}_{100}) = 0.25$
c. $E(\bar{X}_{100}) = 15, Var(\bar{X}_{100}) = 5$
d. $E(\bar{X}_{100}) = 15, Var(\bar{X}_{100}) = 0.25$

When is the sample mean Normally Distributed?

Although we know the mean and variance of $\bar{X}$, we generally don't know its distribution function.

Theorem: Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample with $X_i \sim N(\mu, \sigma^2)$ for $i = 1, 2, \ldots, n$.

Then

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}).$$

Intuitively, if all the observations come from the same normal distribution then the sample average is also normally distributed and centered at the true mean (but much more narrow).

Central Limit Theorem

What if X_i are not normally distributed?

If the number of observation, n, per sample is large (we will discuss this more later), then the distribution of X_i doesn't matter. We will always have

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}).$$