ECN 102: Analysis of Economics Data

Chapter 3: The Sample Mean

Remy Beauregard

Random Variables

Overview

We previously discussed the sample mean, $\bar{x}$, but different samples of data may give us different sample means due to randomness in the data. How can we use the sample mean to say something substantive about the larger population?

Defining Random Variables

We call a variable a **random variable** (r.v.) if its outcome will be determined by some uncertain, *probabilistic*, or *stochastic* process.

A random variable X (big) could be whether a flipped coin lands heads or tails; x (little) is each potential outcome value X could take. For the coin:

$$X = \begin{cases} H & P[X = H] = 0.5 \\ T & P[X = T] = 0.5 \end{cases}$$

We say $P[X = x]$ for the chance our r.v. X takes value x.

Expected Value

We can summarize a random variable just like any other variable, to say something about its central tendency and spread. For a *probabilistic* r.v., its **expected value** is the probability-weighted average of all possible values x our variable X may take.

This expected value will be the *population mean* of our r.v. For a *discrete* r.v., we compute expected value using a sum[1]:

$$\mu \equiv E[X] = \sum_x x \cdot P[X = x]$$

As with all probabilistic outcome spaces, we must ensure that *our probabilities for all possible values of X sum to 1*.

[1] For a *continuous* r.v., we would use an integral to compute expected value.

Variance and Standard Deviation

We can also compute a measure of spread for a r.v., the expected value of squared deviations from our population mean. We denote this variance:

$$\sigma^2 = E[(X - \mu)^2]$$

and standard deviation:

$$\sigma = \sqrt{\sigma^2} = \sqrt{E[(X - \mu)^2]}$$

To compute this, we again use the probability $P[X = x]$ for each value of X:

$$\sigma^2 = \sum_x (x - \mu)^2 \cdot P[X = x]$$

To do this, we first need to compute μ.

Example: Computing Mean and Variance

An example: Given an X:

$$X = \begin{cases} 0 & 0.1 \\ 2 & 0.5 \\ 3 & 0.3 \\ 4 & 0.1 \end{cases}$$

we could compute μ, σ^2, and σ. What is this r.v. X in words?

Example: Computing Mean and Variance

An example: Given an X:

$$X = \begin{cases} 0 & 0.1 \\ 2 & 0.5 \\ 3 & 0.3 \\ 4 & 0.1 \end{cases}$$

we could compute μ, σ^2, and σ. What is this r.v. X in words?

X is a r.v. that takes value 0 with 10% chance, 2 with 50% chance, 3 with 30% chance, and 4 with 10% chance.

Properties of Expectation and Variance

We have some properties that help us evaluate the mean and variance of a r.v. X transformed with $a + bX$.

- $E[a + bX] = a + b \times E[X]$
- $V[a + bX] = b^2 \times V[X]$
- $V[X + Y] = V[X] + V[Y]$ if X and Y are independent

As we can see, adding or subtracting a constant a from a r.v. changes its measure of central tendency *but not its measure of spread*, while multiplying a r.v. by a constant b changes *both* measures. We will need these properties later for our proofs.

Random Samples

From Sample to Sample Mean

If we have a r.v. X, we can take a random *sample* of that variable of size n. This will contain n *realizations* of our r.v. X, and any summary statistics we compute for this sample will themselves be random variables.

We previously computed $\bar{x}$, the sample mean for a *given* sample of X. However, we could have drawn a different sample and computed a different sample mean, so our sample mean itself is a r.v. $\bar{X}$. $\bar{x}$ is a realization of $\bar{X}$.

Sample Variance as a Random Variable

Similarly, the variance (or standard deviation) we computed for our sample, s^2, is also dependent on the random sample of X we draw. Like the sample mean $\bar{x}$, each s^2 we compute from our data is a realization of the r.v. for the sample variance, S^2.

When we compute s^2, we cannot do so directly. First, we must compute $\bar{x}$ from our data and then include that computed statistic in our formula $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Since we use **one** statistic already computed from our data in the formula for s^2, we must subtract **one degree of freedom** from our number of observations:

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

Distribution of Sample Means

Our r.v. X may not be (and often is not) *normally distributed*. However, something quite magical happens to the distribution of *sample means* when our sample is big enough.

Note: we are not talking about the individual realizations of X in each sample, but instead the measure of central tendency $\bar{x}$ for each sample.

Let's simulate 1000 draws of increasingly large samples of fair coin tosses, with a 50% chance of heads or tails.

Sample Means with n = 10

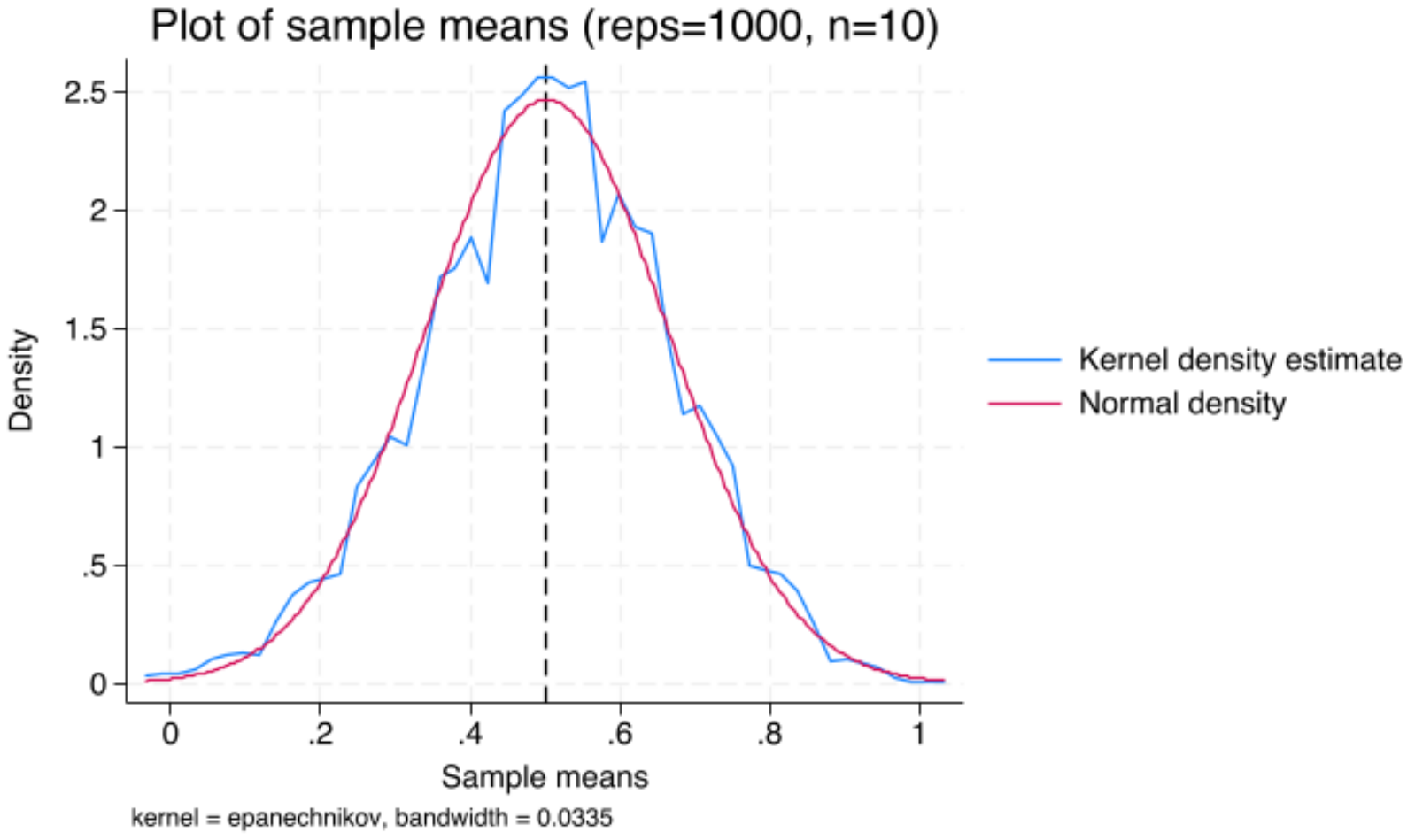

Figure 1: Kernel density of 1000 sample means from n = 10 coin tosses.

Sample Means with n = 25

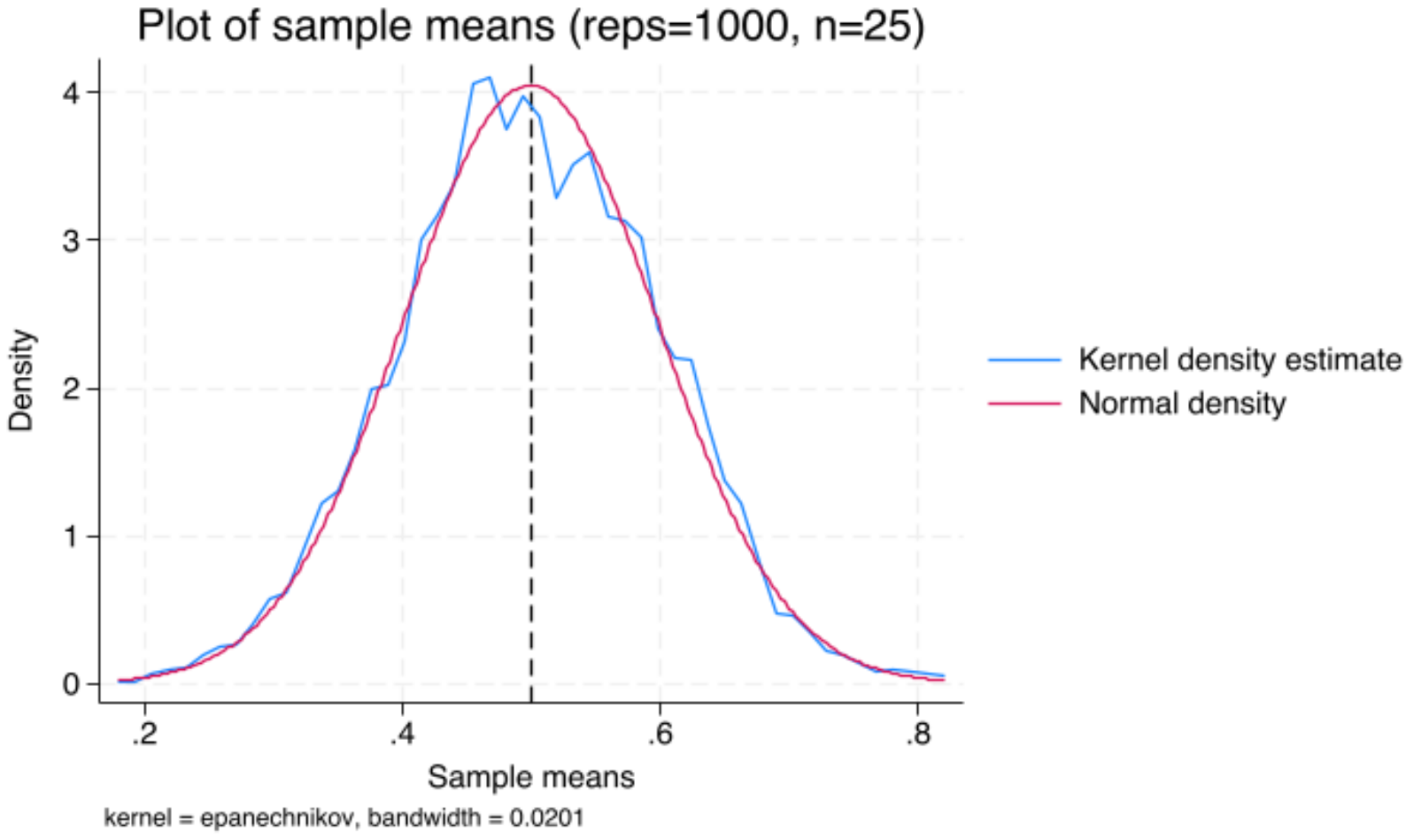

Figure 2: Kernel density of 1000 sample means from n = 25 coin tosses.

Sample Means with n = 125

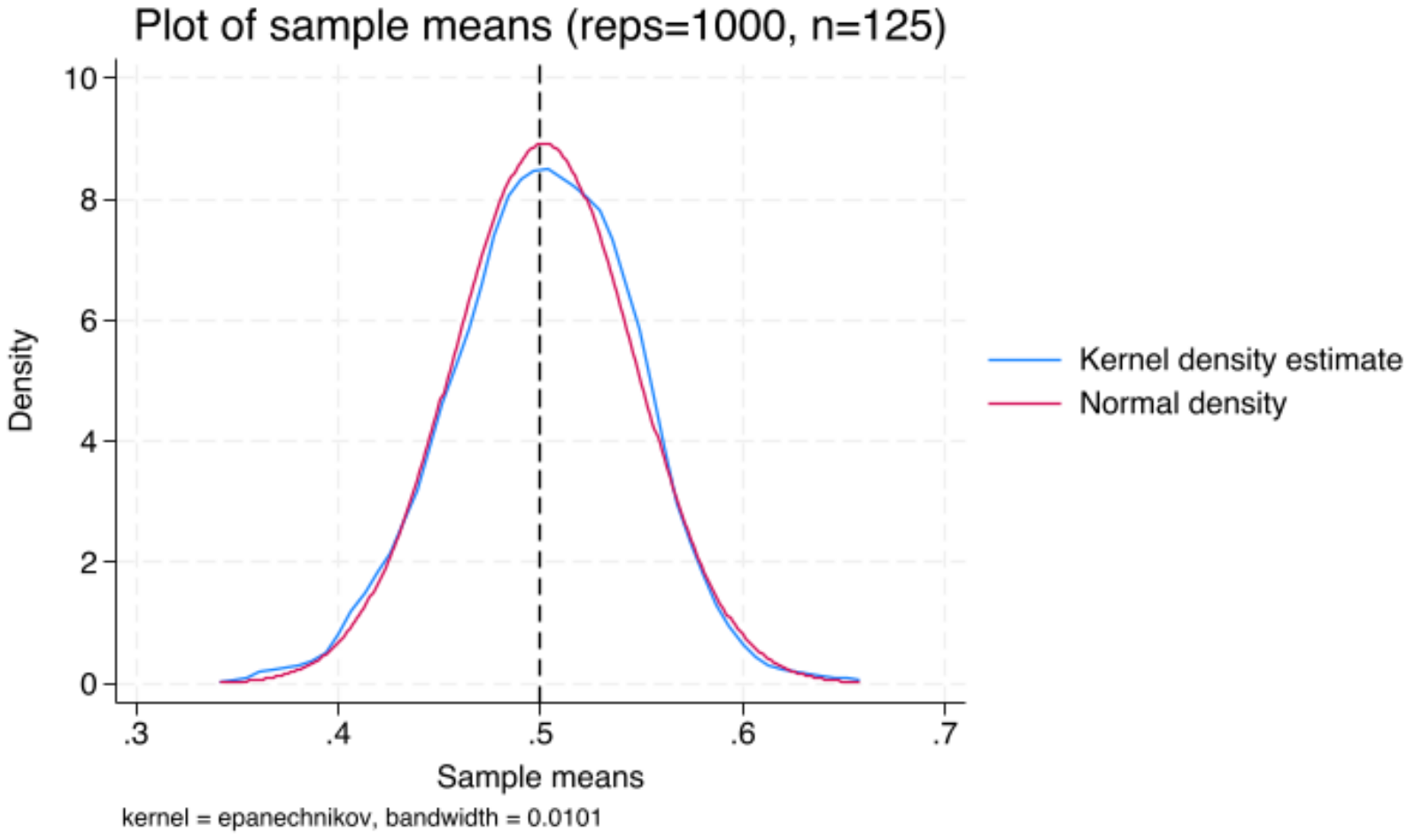

Figure 3: Kernel density of 1000 sample means from n = 125 coin tosses.

Central Limit Theorem: Intuition

What is happening?

$\Rightarrow$ Our distribution of *sample means* is looking more and more normal as we simulate larger samples of random coin tosses. Our coin toss r.v. is definitely not normally distributed, but the distribution of its sample means appears to be as $n \to \infty$.

Note: n here refers to the **sample size** (10, 25, 125), NOT the number of repetitions (always 1000).

$\Rightarrow$ In addition, the average of our sample means ($\mu_{\bar{X}}$) appears to be roughly equal to the average of our coin toss r.v., $\mu = 0.5$[2].

[2] $\mu \equiv \sum_x x \cdot P[X = x] = 1 \cdot 0.5 + 0 \cdot 0.5 = 0.5$

Assumptions: Simple Random Samples

Let us place some formal structure on our discussion of our r.v. X:

A) X_i has a common mean $\mu : E[X_i] = \mu$ for all i

B) X_i has a common variance $\sigma^2 : V[X_i] = \sigma^2$ for all i

C) Different realizations of X do not influence each other; X_i is statistically independent of X_j for all $i \neq j$

Together, these imply that $X \sim (\mu, \sigma^2)$, where $\sim$ is read as "is distributed with"; *X is a random variable distributed with population mean μ and population variance σ^2.* We call samples with these properties **(simple) random samples**.

Mean of the Sample Mean

The **population mean of the sample mean** $\bar{X}$ is:

$$\mu_{\bar{X}} \equiv E[\bar{X}] = \mu$$

Proof:

Mean of the Sample Mean

The **population mean of the sample mean** $\bar{X}$ is:

$$\mu_{\bar{X}} \equiv E[\bar{X}] = \mu$$

Proof:

$$\mu_{\bar{X}} = E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$$

Variance of the Sample Mean

The **population variance of the sample mean** $\bar{X}$ is:

$$\sigma^2_{\bar{X}} = V[\bar{X}] \equiv E[(\bar{X} - \mu_{\bar{X}})^2] = \frac{\sigma^2}{n}$$

Proof:

Variance of the Sample Mean

The **population variance of the sample mean** $\bar{X}$ is:

$$\sigma^2_{\bar{X}} = V[\bar{X}] \equiv E[(\bar{X} - \mu_{\bar{X}})^2] = \frac{\sigma^2}{n}$$

Proof:

$$\sigma^2_{\bar{X}} = V[\bar{X}] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Standard Deviation of the Sample Mean

Thus, the **standard deviation of the sample mean** is given by:

$$\sigma_{\bar{X}} \equiv \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

This quantity represents the *spread of our sample means* for a given sample size n, our *precision* in estimating our sample mean.

We can see that *larger samples mechanically lead to greater precision in estimating* μ: $\sigma_{\bar{X}} \to 0$ as $n \to \infty$.

Characterizing the Sample Mean

So far, we have shown that $\bar{X} \sim (\mu, \frac{\sigma^2}{n})$ or that:

"X-bar is a random variable distributed with population mean μ and population variance σ^2/n".

Characterizing the Sample Mean

So far, we have shown that $\bar{X} \sim (\mu, \frac{\sigma^2}{n})$ or that:

"X-bar is a random variable distributed with population mean μ and population variance σ^2/n".

We need one more result to complete our characterization of $\bar{X}$.

The Central Limit Theorem

Using the **central limit theorem**, we can say that $\bar{X}$ will be *approximately normally distributed* whenever $n > 30$. Note, this is the sample size, *not the number of resamplings we do*.

With the CLT, we now have that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

We say that $\bar{X}$ is **asymptotically normally distributed** since its distribution becomes normal as $n \to \infty$.

Standardizing the Sample Mean

Finally, we can *standardize* $\bar{X}$ to turn it into a z-score. Since $\bar{X}$ is normally distributed by the CLT when $n > 30$, this will be a standard *normal* distribution.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- "Standard" means we have mean 0 and variance 1
- "Normal" means our distribution is perfectly symmetric (skew=0) and has kurtosis of exactly 3
- The standard normal distribution is well-understood, making it easy to conduct statistical tests

Estimating Population Variance

For a given sample, however, we are unlikely to know our population variance σ^2 from a single sample. Instead, we can replace σ^2 with s^2, the sample estimate of variance.

This is given by:

$$s^2_{\bar{X}} = \frac{s^2}{n} = \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}{n}$$

The Standard Error

Similarly, we can take:

$$s_{\bar{X}} = \frac{\sqrt{s^2}}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}}{\sqrt{n}}$$

Since this is no longer the standard deviation of the sample mean, we give this quantity a new name: the **standard error** of $\bar{X}$. We are able to calculate this quantity using only *sample statistics* from our given dataset (no μ or σ^2 here).

Revisiting the Assumptions

All of the above identities rely on the three previous assumptions:

1. Common mean
2. Common variance
3. Independence of observations

Later, we will relax these assumptions.

Sample Representativeness and Weights

So far, we have implicitly assumed that our samples are **representative** of our population — they are simple random samples. However, we may encounter instances where our sample is not representative of the broader population. In these cases, sample statistics derived from biased samples will lead to improper estimations of population parameters.

Sample Representativeness and Weights

So far, we have implicitly assumed that our samples are **representative** of our population — they are simple random samples. However, we may encounter instances where our sample is not representative of the broader population. In these cases, sample statistics derived from biased samples will lead to improper estimations of population parameters.

$\Rightarrow$ What can be done? One common solution is to use **sample weights**, quantities to increase or decrease how represented a given value is in a sample when estimating a population. For example, we could weight all observations in the sample by the inverse of the probability p that such a value appears in the sample.

Estimators

What Is an Estimator?

In the cases above, we have been *estimating* the population mean μ with the sample mean $\bar{X}$, as we do not know the true value of the former (we also estimated σ with s). We want to think about properties of a good estimator that will lead us to a proper conjecture about the true (unknown) population parameter.

Unbiasedness

One such property is **unbiasedness**: the expected value of an estimator is equal to the population parameter.

We have shown that $\bar{X}$ is an unbiased estimator for μ since:

$$E[\bar{X}] = \mu$$

Visualizing Bias

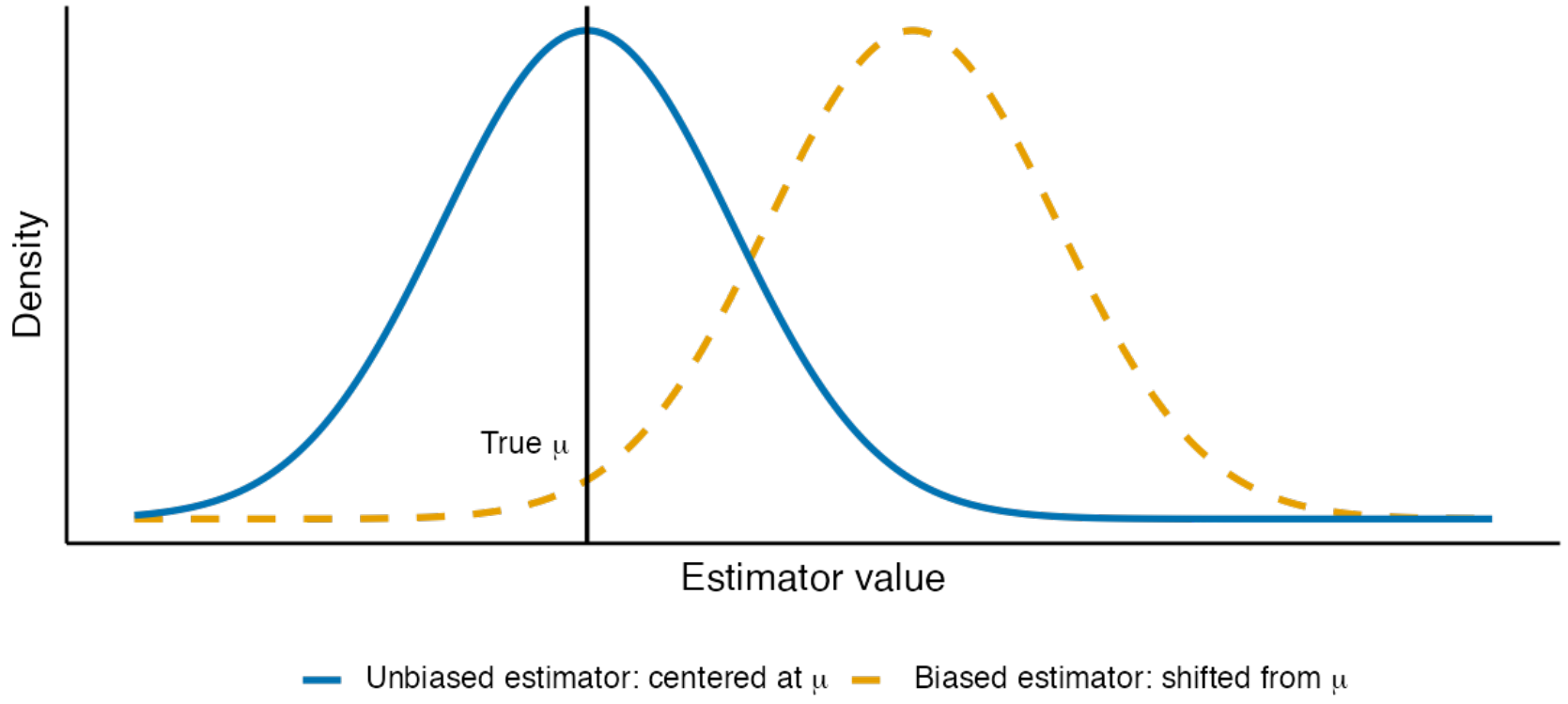

Figure 4: Sampling distributions of a biased and an unbiased estimator for a population mean μ.

Consistency

Another property is **consistency**: the estimator gets closer and closer to the true parameter value as $n \to \infty$ and the variance of the estimator shrinks to zero.

We have also shown that $\bar{X} \to \mu$ and $\frac{\sigma^2}{n} \to 0$ as $n \to \infty$.

Thus, our estimator $\bar{X}$ for μ is both *unbiased* and *consistent*.

Minimum Variance

We may have many unbiased estimators to choose from. A way to further narrow down this list is to designate the **best estimator** as that with the *minimum variance*. We have shown what the population variance is for $\bar{X}$, but we have not shown it to be the minimum variance for all possible estimators of μ.

End of Lecture Material

Knowledge Check 3

Suppose $Y \sim (10, 25)$ with $\bar{Y}$ as the sample mean for $n = 55$.

- Describe in words how Y is distributed.
- What do you expect the mean of $\bar{Y}$ to be?
- What do you expect the variance of $\bar{Y}$ to be?
- How do you expect $\bar{Y}$ to be distributed? Explain.
- Complete the identity $\bar{Y} \sim \_\_\_(\_\_\_, \_\_\_)$