BIOS:4120 – Introduction to Biostatistics
Unit 7: Sampling Distribution of the Mean

Knute D. Carter

Department of Biostatistics
The University of Iowa

September 30, 2025

Learning Objectives

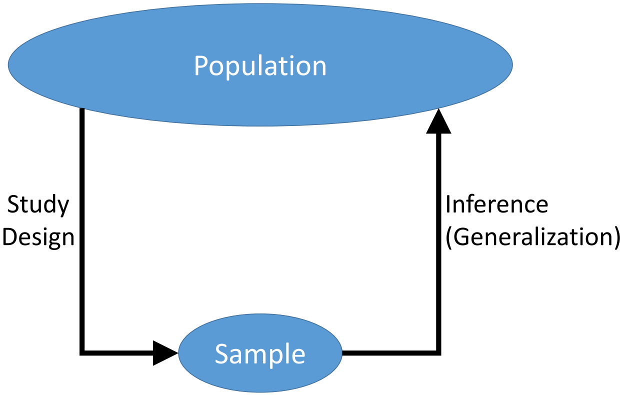
At the end of this session, you should be able to:

- Describe the sampling distribution of the mean in the context of repeated sampling.
- Calculate the expected value and standard error of the sample mean.
- Describe the central limit theorem and the conditions when it is valid.
- Perform probability calculations on the sample mean.

Overview

- Sampling Distributions
- The Central Limit Theorem
- Applications of the Central Limit Theorem

Sampling Distributions

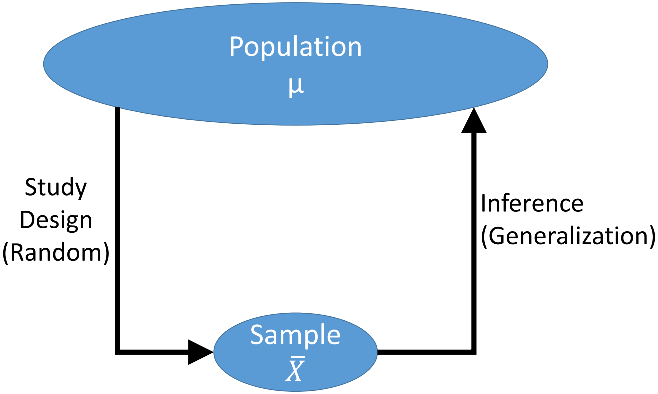


Sampling Distributions

- *Inferential Statistics* deals with methods for making generalizations about a population based on information contained in the sample.
- A *Statistic* is a numerical value describing a sample characteristic.
- A *Parameter* is a numerical value describing a population characteristic.

Sampling Distributions

- Perhaps we draw a random sample from a population, and use the statistic $\bar{X}$ (the sample mean), to estimate the parameter μ (the population mean).



Sampling Distributions

- One can view $\bar{X}$ as a numeric variable that assumes a value based on the outcome of a random experiment: i.e., the process of drawing a sample at random from the population.
- $\bar{X}$ is therefore, by definition, a random variable. It will vary from sample to sample.
- By the same reasoning, so is any statistic.
- The probability distribution of a statistic is referred to as a *Sampling Distribution*.
- The sampling distribution reflects which values of the statistic are likely and which values are improbable.

Sampling Distributions

- The mean of a statistic is often called the *Expected Value*, for example, $E(\bar{X})$.
- The standard deviation of a statistic is often called the *Standard Error*, e.g., the square root of $Var(\bar{X})$.
- The sampling distribution of a statistic is used in developing inferential procedures.

Sampling Distributions: Example 1

- Let X be the number of medications a person is taking.
- Suppose we have a population of five people to sample from, and the number of medications they take is as follows:

Person	# Medications
A	1
B	4
C	4
D	7
E	10

- So the population mean, $E[X] = \mu = 5.2$.

Sampling Distributions: Example 1

- How many different ways could we sample three people from this population?

Sampling Distributions: Example 1

- Possible samples of size 3 that we could select from this population and the resulting sample mean are:

Sample (person)	Sample (values)	$\bar{x}$
(C,D,E), (B,D,E)	(4,7,10), (4,7,10)	7
(B,C,E), (A,D,E)	(4,4,10), (1,7,10)	6
(B,C,D), (A,C,E), (A,B,E)	(4,4,7), (1,4,10), (1,4,10)	5
(A,C,D), (A,B,D)	(1,4,7), (1,4,7)	4
(A,B,C)	(1,4,4)	3

- Note that *none* of the possible sample means are equal to the population mean.

Sampling Distributions: Example 1

- The sampling distribution of $\bar{X}$ for a sample of size three from these data is given by:

	Probability
$P(\bar{X} = 7)$	0.2
$P(\bar{X} = 6)$	0.2
$P(\bar{X} = 5)$	0.3
$P(\bar{X} = 4)$	0.2
$P(\bar{X} = 3)$	0.1

- So what is $E[\bar{X}]$, the expected value of $\bar{X}$?

Sampling Distributions: Example 1

- If we took samples over and over again, we would end up with 20% with a mean of 7; 20% with a mean of 6; 30% with a mean of 5; and so on.

- So the average (expected) value we would observe if we took repeated samples would be:

$$20\% \times 7 + 20\% \times 6 + 30\% \times 5 + 20\% \times 4 + 10\% \times 3$$

  which is

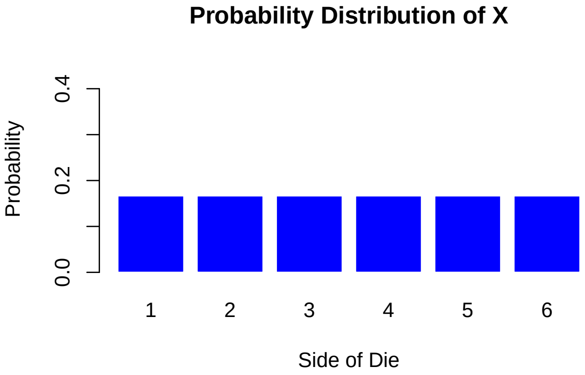
$$0.2 \times 7 + 0.2 \times 6 + 0.3 \times 5 + 0.2 \times 4 + 0.1 \times 3 = 5.2$$

- That is, $E[\bar{X}] = \mu = 5.2$.

- Because of this property, we say that $\bar{X}$ is an *unbiased estimator* of μ.

Sampling Distributions: Example 2

- Let X = the outcome of the toss of a fair (unbiased) six-sided die.
- Sample space for $X : \mathcal{S} = \{1, 2, 3, 4, 5, 6\}$.



Sampling Distributions: Example 2

- Consider measuring X twice, and averaging the two observed values of X to obtain a sample mean, $\bar{X}$.
- We can view the two measurements on X as a sample, and the sample mean $\bar{X}$ as a statistic.
- We can view the population mean μ as the mean of the random variable X. Thus, $\mu = 3.5$.
- What would be the sampling distribution of $\bar{X}$?

Sampling Distributions: Example 2

To determine the sampling distribution, we would need to determine the following:

i) every possible sample mean,

ii) the samples of size two corresponding to every possible sample mean,

iii) the probability corresponding to every possible sample mean.

Sampling Distributions: Example 2

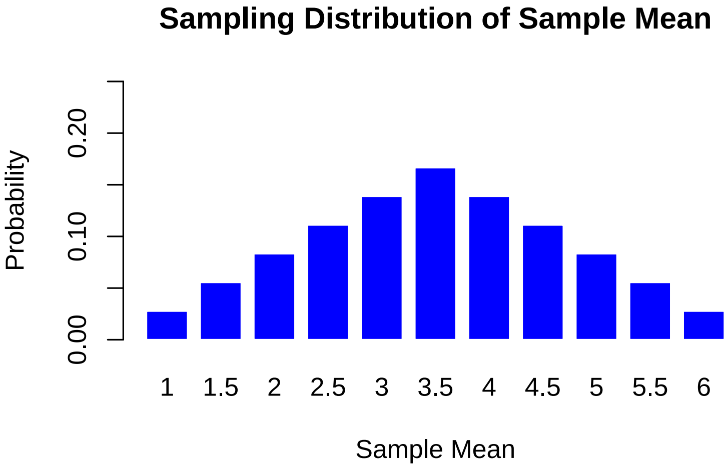
- Sampling distribution of mean of two throws of a six-sided die.

$\bar{x}$	Sample	Probability
1.0	$(1, 1)$	$1/36$
1.5	$(1, 2), (2, 1)$	$2/36$
2.0	$(1, 3), (3, 1), (2, 2)$	$3/36$
2.5	$(1, 4), (4, 1), (2, 3), (3, 2)$	$4/36$
3.0	$(1, 5), (5, 1), (2, 4), (4, 2), (3, 3)$	$5/36$
3.5	$(1, 6), (6, 1), (3, 4), (4, 3), (2, 5), (5, 2)$	$6/36$
4.0	$(2, 6), (6, 2), (3, 5), (5, 3), (4, 4)$	$5/36$
4.5	$(3, 6), (6, 3), (4, 5), (5, 4)$	$4/36$
5.0	$(4, 6), (6, 4), (5, 5)$	$3/36$
5.5	$(5, 6), (6, 5)$	$2/36$
6.0	$(6, 6)$	$1/36$

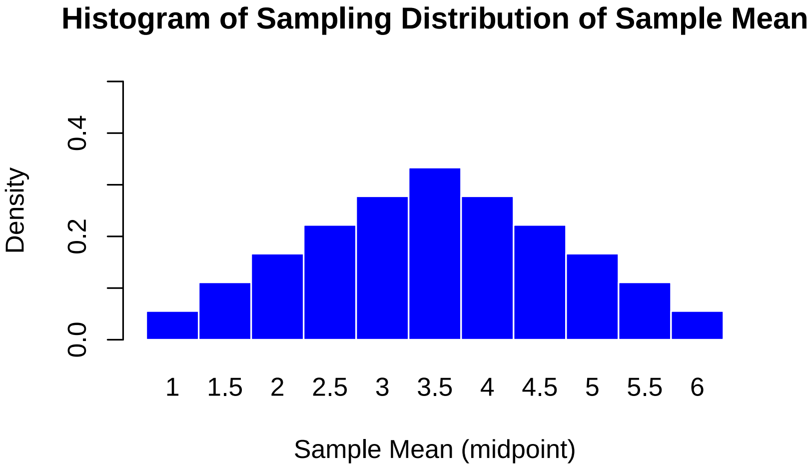
Sampling Distributions: Example 2

- The sampling distribution reflects which values of $\bar{X}$ are likely and which values are improbable.
- The most likely value of $\bar{X}$ is $\mu = 3.5$. This is meaningful, since we are viewing $\bar{X}$ as an estimator of μ.
- Values of $\bar{X}$ near μ are more likely than values more distant from μ.

Sampling Distributions: Example 2



Sampling Distributions: Example 2



Sampling Distributions

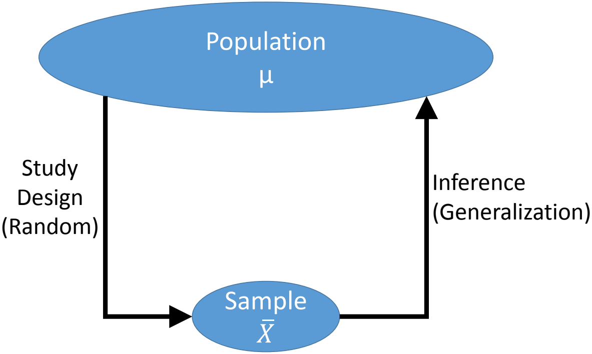
- How do you think the sampling distribution of $\bar{X}$ would change if we based $\bar{X}$ on a sample of size 3? A sample of size 30? A sample of size 1000?
- The sampling distribution reflects the shape of a histogram that would result if a long set of measurements on $\bar{X}$ was compiled, and a histogram was constructed based on the relative frequency distribution of these measurements.

Sampling Distributions

- *Repeated Sampling* refers to the process of repeatedly drawing samples from a population, and repeatedly computing a statistic (such as $\bar{X}$) based on these samples.
- Repeated sampling is not done in practice. However, understanding the behavior of a statistic in repeated sampling (i.e., understanding the sampling distribution of a statistic) helps us to characterize the properties of the statistic as an estimator of a parameter.

The Central Limit Theorem

- We will now consider the specific problem of using the statistic $\bar{X}$ (the sample mean) to estimate the parameter μ (the population mean).



The Central Limit Theorem

Properties of the sampling distribution of $\bar{X}$:

- The mean or expected value of the statistic $\bar{X}$ is the same as the mean of the sampled population, μ.
- The standard deviation or standard error of the statistic $\bar{X}$ is given by $\sigma/\sqrt{n}$, where σ is the standard deviation of the sampled population, and n is the sample size.

The Central Limit Theorem

The preceding are general properties of the sampling distribution of $\bar{X}$ that hold for any sample size n.

- The quantity $\sigma/\sqrt{n}$ is often called the *Standard Error of the Mean*.
- The magnitude of this quantity reflects the accuracy of the sample mean as an estimator of the population mean.

The Central Limit Theorem

- **Central Limit Theorem**: If the sample size is ‘large,’ the sampling distribution of $\bar{X}$ is approximately normal, regardless of the characteristics of the underlying population.
- A ‘large’ sample is generally considered to be one where $n \geq 30$.
- Suppose the random variable being measured to collect the sample data has a normal distribution. Then the sampling distribution of $\bar{X}$ is normal for *any* sample size.

The Central Limit Theorem

Note:

- If $\bar{X}$ is approximately $N(\mu, \sigma^2/n)$, then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

  is approximately standard normal.

- Observations must be independently drawn from and be representative of the population.

- The central limit theorem applies to the sampling distribution of the mean, not necessarily to the sampling distribution of other statistics.

Central Limit Theorem: Application 1

- In Norway, the distribution of birth weights for infants whose gestational age is 40 weeks is approximately normal with mean $\mu = 3500$ grams and standard deviation $\sigma = 430$ grams.
- For parts a) and b), consider a randomly selected newborn whose gestational age is 40 weeks. Let X denote the birth weight of this infant.

a) Illustrate the probability distribution of X.

Central Limit Theorem: Application 1

b) What is the probability that the infant's birth weight is less than 3400 grams?

- For the remaining parts, consider a random sample of 25 newborns whose gestational age is 40 weeks. Let $\bar{X}$ denote the mean birth weight of these 25 infants.

c) What is the mean of $\bar{X}$? (That is, in repeated sampling, what would be the average value of $\bar{X}$?)

Central Limit Theorem: Application 1

d) What is the standard deviation (i.e., standard error) of $\bar{X}$?

e) Illustrate the sampling distribution of $\bar{X}$.

Central Limit Theorem: Application 1

f) What is the probability that $\bar{X}$ will be less than 3400 grams? (That is, in repeated sampling, what proportion of sample means will be less than 3400 grams?)

g) What is the probability that $\bar{X}$ will be between 3400 and 3600 grams? (That is, in repeated sampling, what proportion of sample means will be between 3400 and 3600 grams?)

Central Limit Theorem: Application 2

- In the Netherlands, healthy males between the ages of 65 and 79 have a distribution of serum uric acid levels that is approximately normal with mean $\mu = 341$ μmol/L and standard deviation $\sigma = 79$ μmol/L.

a) What proportion of the males have a serum uric acid level between 300 and 400 μmol/L?

Central Limit Theorem: Application 2

b) What proportion of samples of size 5 have a mean serum uric acid level between 300 and 400 μmol/L?

Central Limit Theorem: Application 2

c) What proportion of samples of size 10 have a mean serum uric acid level between 300 and 400 μmol/L?

Central Limit Theorem: Application 2

d) Construct an interval that encloses 95% of the means of samples of size 10.

Central Limit Theorem: Application 2

- How does this compare to $Pr(300 < \bar{Y}_{10} < 400)$?
- How do the lengths of the two intervals compare?
- Does it make sense? Why or why not?

Learning Objectives

At the end of this session, you should be able to:

- Describe the sampling distribution of the mean in the context of repeated sampling.
- Calculate the expected value and standard error of the sample mean.
- Describe the central limit theorem and the conditions when it is valid.
- Perform probability calculations on the sample mean.