Introduction to Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

October 6, 2019

Topics

1. Sampling Error and Sampling Distributions of Statistics
2. Sampling Distribution of the Sample Mean $\bar{X}$

Objectives

Objectives:

- Interpret a sampling error as the difference between an estimate (statistic) an a true value (population parameter).
- Interpret sampling distributions as probability distributions of statistics.
- State the two conditions under which the sample mean follows a normal distribution.
- Identify the mean and standard deviation (standard error) of the sampling distribution of the sample mean.
- Use the sampling distribution of the sample mean to obtain probabilities (proportions) involving a sample mean.

Sampling Error and Sampling Distributions of Statistics (7.1)

Statistics and Population Parameters

- Recall that a ***statistic*** is a numerical value computed from **random sample** data.

Sampling Error and Sampling Distributions of Statistics (7.1)

Statistics and Population Parameters

- Recall that a ***statistic*** is a numerical value computed from **random sample** data.

  A ***parameter*** is a numerical characteristic of a **population.**

Sampling Error and Sampling Distributions of Statistics (7.1)

Statistics and Population Parameters

- Recall that a ***statistic*** is a numerical value computed from **random sample** data.

  A ***parameter*** is a numerical characteristic of a **population.**

- The **value of a statistic** will exhibit ***chance variation*** from **one sample to the next.**

Sampling Error and Sampling Distributions of Statistics (7.1)

Statistics and Population Parameters

- Recall that a ***statistic*** is a numerical value computed from **random sample** data.

  A ***parameter*** is a numerical characteristic of a **population.**

- The **value of a statistic** will exhibit ***chance variation*** from **one sample to the next.**

Sampling Error and Sampling Distributions of Statistics (7.1)

Statistics and Population Parameters

- Recall that a ***statistic*** is a numerical value computed from **random sample** data.

  A ***parameter*** is a numerical characteristic of a **population.**

- The **value of a statistic** will exhibit ***chance variation*** from **one sample to the next.**

  The **value of a population parameter remains constant.**

Example

If we take a **random sample** from the **population** of U.S. adolescents and measure their **blood cholesterol** levels, then:

Example

If we take a **random sample** from the **population** of U.S. adolescents and measure their **blood cholesterol** levels, then:

- The **sample mean** blood cholesterol level $\bar{x}$ is a **statistic.**

Example

If we take a **random sample** from the **population** of U.S. adolescents and measure their **blood cholesterol** levels, then:

- The **sample mean** blood cholesterol level $\bar{x}$ is a **statistic.**
- The **population mean** blood cholesterol level μ is a **parameter.**

Using Statistics to Estimate Population Parameters

- **Statistics** are used to **estimate** the corresponding **population parameters**:

Statistics as Estimators of Population Parameters:

	Population Parameter	Statistic Used to Estimate the Parameter
Mean	μ	$\bar{x}$
Standard Deviation	σ	s

Example

The **sample mean** blood cholesterol level $\bar{x}$ in a **random sample** of U.S. adolescents is an **estimate** of the true (unknown) **population mean** level μ.

Sampling Error

- Because the value of a statistic is subject to **chance variation** from one sample to the next, there will be a slight **error** when it's used to **estimate** a population parameter.

Sampling Error

- Because the value of a statistic is subject to **chance variation** from one sample to the next, there will be a slight **error** when it's used to **estimate** a population parameter.

  We call this error the ***sampling error*** of the estimate.

Sampling Error

- Because the value of a statistic is subject to **chance variation** from one sample to the next, there will be a slight **error** when it's used to **estimate** a population parameter.

  We call this error the ***sampling error*** of the estimate.

Sampling Error of a Statistic:

$$\text{Sampling Error} = \underbrace{\text{Estimate}}_{\substack{\text{Sample}\\ \text{Statistic}}} - \underbrace{\text{True Value}}_{\substack{\text{Population}\\ \text{Parameter}}}$$

- When the **sample mean** $\bar{x}$ is used to estimate a **population mean** μ, the **sampling error** is:

Sampling Error of the Sample Mean:

$$\text{Sampling Error} \;=\; \bar{x} - \mu$$

Example

According to the Centers for Disease Control, the **mean** blood cholesterol level in the **population** of U.S. adolescents is $\mu = 160$ mg/dL.

Example

According to the Centers for Disease Control, the **mean** blood cholesterol level in the **population** of U.S. adolescents is $\boldsymbol{\mu = 160}$ mg/dL.

If a **random sample** of $n = 100$ adolescents has a **sample mean** $\boldsymbol{\bar{x} = 167}$, the **sampling error** of this estimate is

$$\begin{aligned} \text{Sampling Error} &= \bar{x} - \mu \\ &= 167 - 160 \\ &= \mathbf{7}. \end{aligned}$$

- The **sampling error** can be **positive** or **negative**, depending on whether $\bar{x}$ is an **overestimate** or an **underestimate** of μ.

Sampling Distributions of Statistics

- Because the value of a statistic varies due to chance from one sample to the next, **a statistic** is a **random variable.**

Sampling Distributions of Statistics

- Because the value of a statistic varies due to chance from one sample to the next, **a statistic** is a **random variable.**
- The **probability distribution** of a **statistic** is called its ***sampling distribution***. It specifies two things:
  1. The values that are possible for the statistic.
  2. The probabilities of those values.

Sampling Distributions of Statistics

- Because the value of a statistic varies due to chance from one sample to the next, **a statistic** is a **random variable.**
- The **probability distribution** of a **statistic** is called its ***sampling distribution***. It specifies two things:
  1. The values that are possible for the statistic.
  2. The probabilities of those values.
- The **sampling distribution** of $\bar{x}$ can be used to gauge how large the **sampling error** of $\bar{x}$ might be when estimating μ.

- In the next example, we'll determine the **sampling distribution** of the **sample mean** $\bar{x}$ by listing every possible value of $\bar{x}$, then summarizing those values in a relative frequency distribution table.

Example

Suppose a *very small* **population** consists of six individuals named Ann, Bob, Cara, Dee, Earl, and Fran, and that their ages are as shown below.

Population

Individual	Age
Ann	10
Bob	20
Cara	30
Dee	40
Earl	50
Fran	60

$\mu = 35$

$\sigma = 17$

The population mean and standard deviation are $\mu = 35$ and $\sigma = 17$.

The population mean and standard deviation are $\mu = 35$ and $\sigma = 17$.

Consider taking a **random sample** of $n = 2$ individuals from the **population**.

The population mean and standard deviation are $\mu = 35$ and $\sigma = 17$.

Consider taking a **random sample** of $n = 2$ individuals from the **population**.

The table on the next slide shows all of the **samples** we might end up with along with their **sample mean** age values ($\bar{x}$ values).

Individuals in the Sample	Sample Values $x_1,\ x_2$	Value of $\bar{x}$
Ann, Bob	10, 20	15
Ann, Cara	10, 30	20
Ann, Dee	10, 40	25
Ann, Earl	10, 50	30
Ann, Fran	10, 60	35
Bob, Cara	20, 30	25
Bob, Dee	20, 40	30
Bob, Earl	20, 50	35
Bob, Fran	20, 60	40
Cara, Dee	30, 40	35
Cara, Earl	30, 50	40
Cara, Fran	30, 60	45
Dee, Earl	40, 50	45
Dee, Fran	40, 60	50
Earl, Fran	50, 60	55
		$\mu_{\bar{x}} = 35$
		$\sigma_{\bar{x}} \approx 12$

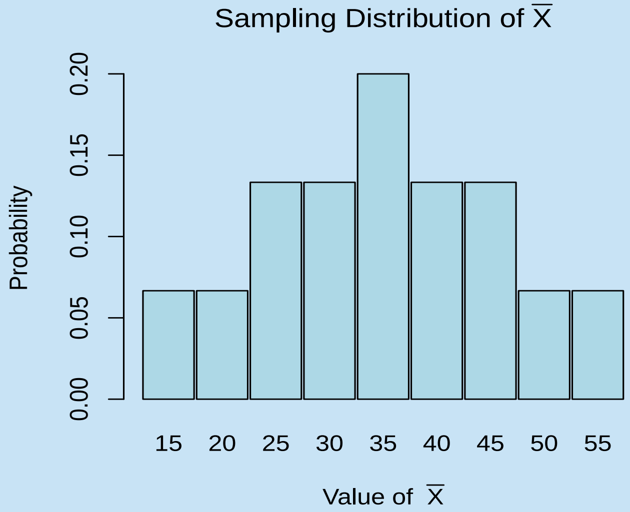
Note that some $\bar{x}$ values are duplicated.

Here's a summary of the $\bar{x}$ values in a **frequency distribution table**:

Distinct $\bar{X}$ Value	Frequency	Relative Frequency
15	1	1/15
20	1	1/15
25	2	2/15
30	2	2/15
35	3	3/15
40	2	2/15
45	2	2/15
50	1	1/15
55	1	1/15
	15	

If we interpret the *relative frequencies* as *probabilities*, we get the **sampling distribution of** $\bar{x}$ shown below (and graphed on the next slide).

Sample Mean $\bar{x}$	15	20	25	30	35	40	45	50	55
Probability of $\bar{x}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{15}$


Sampling Distribution of X̄
Probability
0.00
0.05
0.10
0.15
0.20
15
20
25
30
35
40
45
50
55
Value of X̄

Sampling Distribution of the Sample Mean $\bar{X}$ (7.1, 7.2, 7.3)

Introduction

- The last example demonstrated the ***concept*** of the **sampling distribution of $\bar{x}$.**

Sampling Distribution of the Sample Mean $\bar{X}$ (7.1, 7.2, 7.3)

Introduction

- The last example demonstrated the ***concept*** of the **sampling distribution of $\bar{x}$.**

  But it was a bit unrealistic because:

  - The population was unrealistically small (six people).
  - The variable (age) was known already for everyone in the population (so there'd be no need to take a sample).

- A more realistic scenario is sampling from a *large* **population** that's described by a **normal distribution.**

- A more realistic scenario is sampling from a *large* **population** that's described by a **normal distribution.**
- In the slides ahead, we'll see that:

  1. When we sample from a **normal population**, the **sampling distribution of** $\bar{x}$ will be **normal** too.

- A more realistic scenario is sampling from a *large* **population** that's described by a **normal distribution.**
- In the slides ahead, we'll see that:
  1. When we sample from a **normal population**, the **sampling distribution of** $\bar{x}$ will be **normal** too.
  2. Furthermore, even if we sample from a **non-normal population** (e.g. a right skewed one), as long as the **sample size** n **is large**, the **sampling distribution of** $\bar{x}$ will be **approximately normal.**

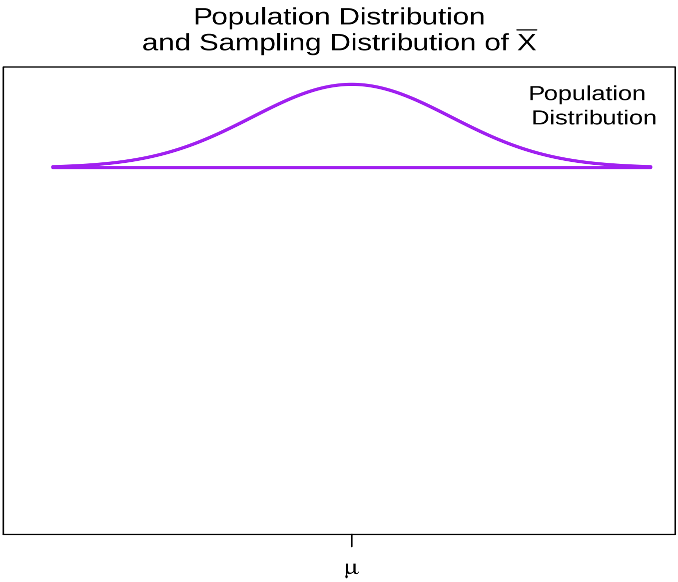
Normality of the Sampling Distribution of $\bar{X}$ When the Sample is from a Normal Population:

Normality of $\bar{X}$: If we take a **sample** of size n from a ***normal* population** whose mean is μ and whose standard deviation is σ, then:

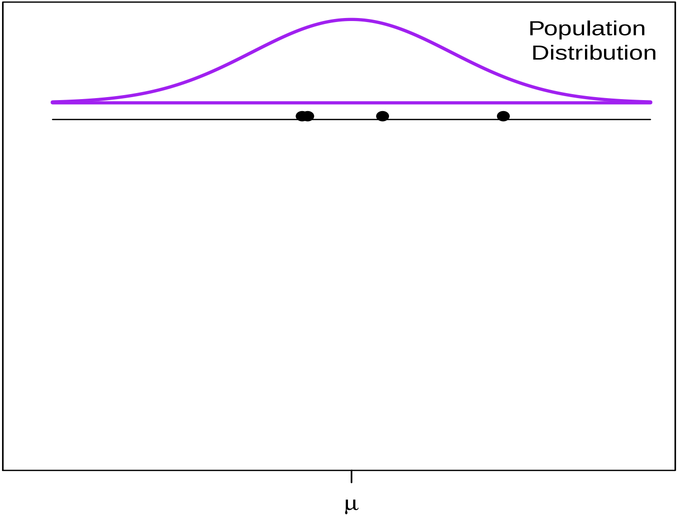
The $\bar{x}$ **distribution** will be ***normal*** with mean $\mu_{\bar{x}}$ and standard deviation $\sigma_{\bar{x}}$, where

$$\mu_{\bar{x}} = \mu \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

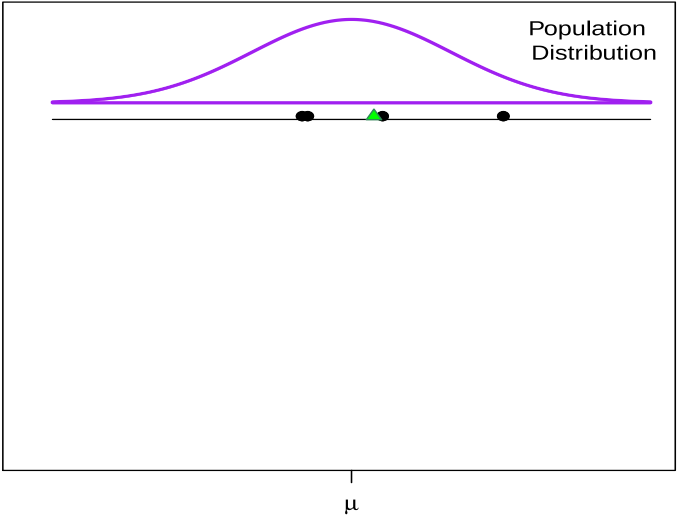
- The figures on the next slides illustrate.

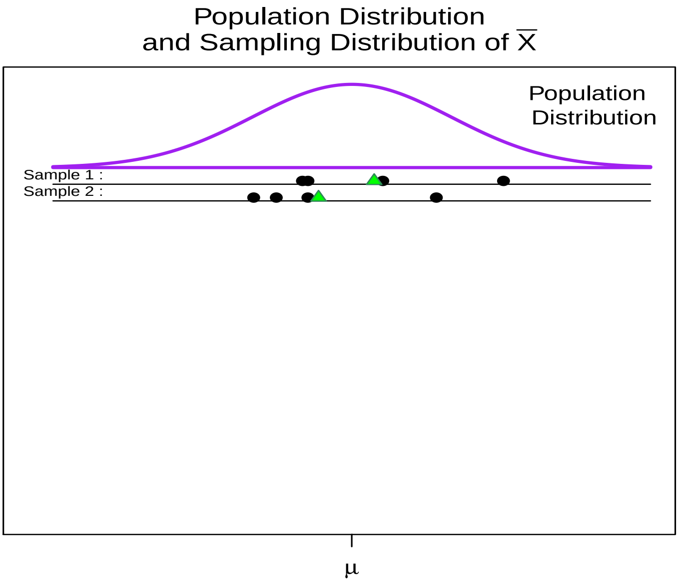

Population Distribution
and Sampling Distribution of X̄
Population
Distribution
μ

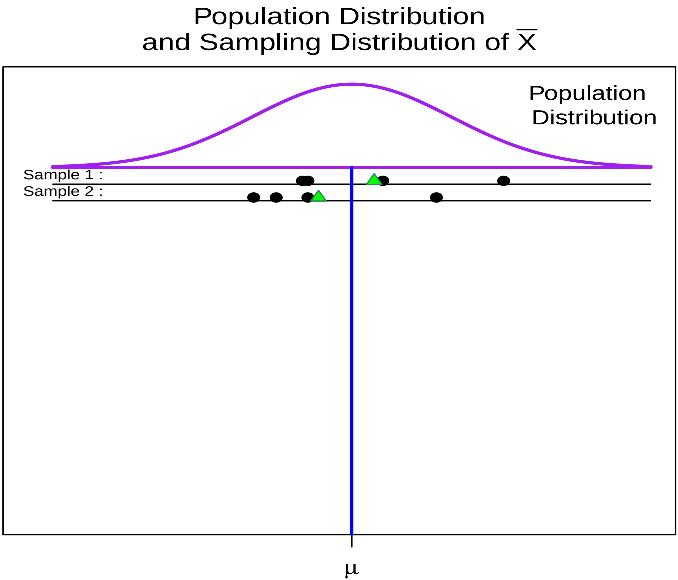
Population Distribution
and Sampling Distribution of $\overline{X}$



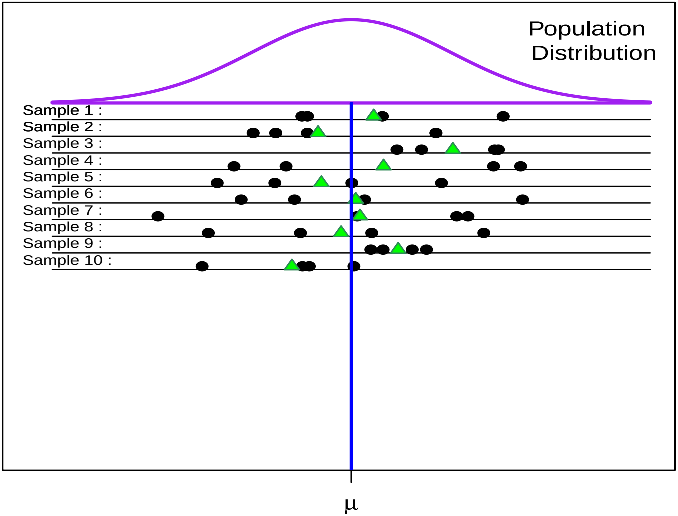
Population Distribution
and Sampling Distribution of $\overline{X}$



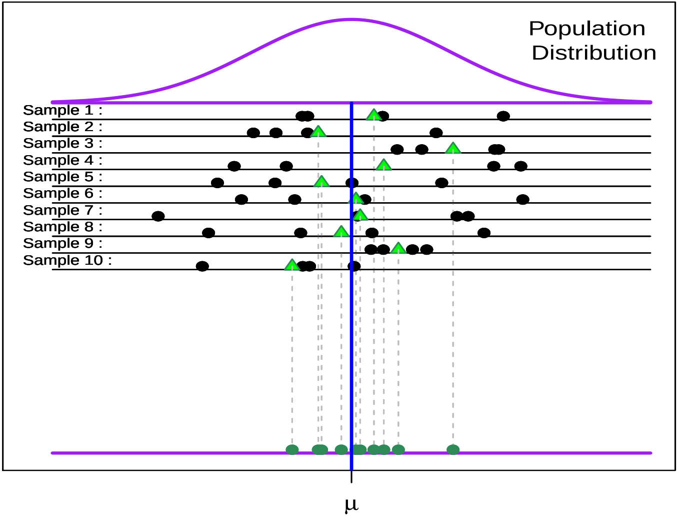

Population Distribution
and Sampling Distribution of X̄
Population
Distribution
Sample 1 :
Sample 2 :
μ

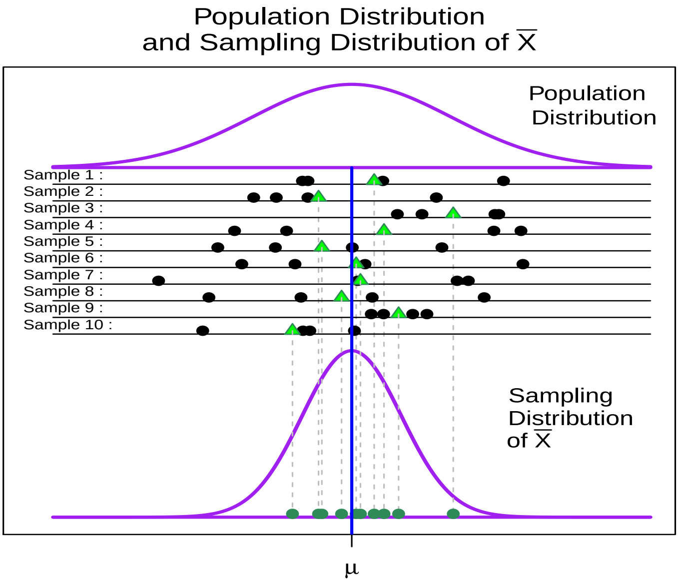

Population Distribution
and Sampling Distribution of X̄
Population
Distribution
Sample 1 :
Sample 2 :
μ

Population Distribution and Sampling Distribution of $\overline{X}$



Population Distribution and Sampling Distribution of $\overline{X}$




Population Distribution
and Sampling Distribution of X̄
Population
Distribution
Sample 1 :
Sample 2 :
Sample 3 :
Sample 4 :
Sample 5 :
Sample 6 :
Sample 7 :
Sample 8 :
Sample 9 :
Sample 10 :
Sampling
Distribution
of X̄
μ

- Because $\bar{x}$ follows a **normal** distribution, the **standardized** version of $\bar{x}$,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}},$$

follows a **standard normal** distribution.

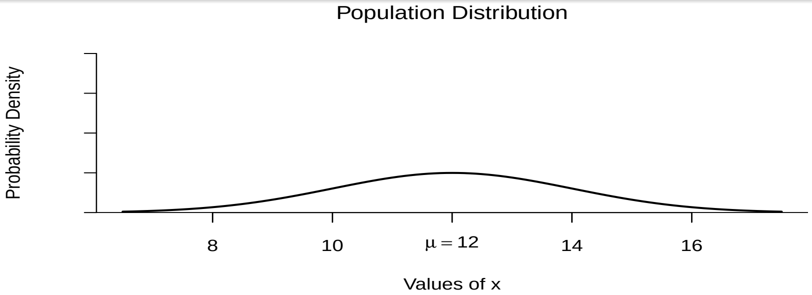
- **Interpretation of $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$:**
  - $\mu_{\bar{x}}$ is the value that $\bar{x}$ takes, **on average**. Thus, because $\mu_{\bar{x}} = \mu$, **on average** the **sample mean** equals the **population mean**.
  - $\sigma_{\bar{x}}$ represents a **typical deviation** of $\bar{x}$ away from μ, i.e. a typical **sampling error**. Thus, because $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, the size of a **typical sampling error** is $\sigma/\sqrt{n}$.

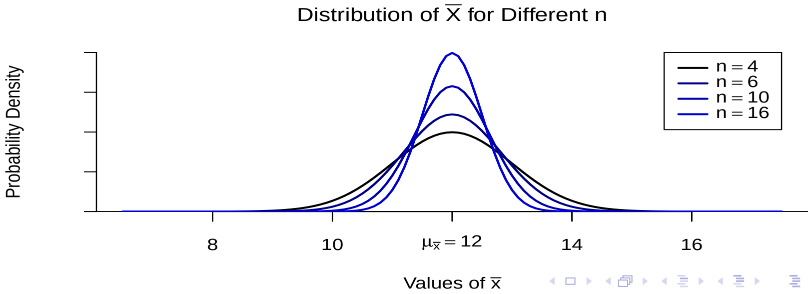
- **Interpretation of $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$:**
  - $\mu_{\bar{x}}$ is the value that $\bar{x}$ takes, **on average**. Thus, because $\mu_{\bar{x}} = \mu$, **on average** the **sample mean** equals the **population mean**.
  - $\sigma_{\bar{x}}$ represents a **typical deviation** of $\bar{x}$ away from μ, i.e. a typical **sampling error**. Thus, because $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, the size of a **typical sampling error** is $\sigma/\sqrt{n}$.
- $\sigma/\sqrt{n}$ is often called the ***standard error*** of $\bar{x}$.

- The **standard error** of $\bar{x}$ **will be small** if either:
  1. The population standard deviation $\boldsymbol{\sigma}$ **is small** (i.e. the population is fairly ***homogeneous***).
  2. The sample size n **is large**.

  Under either of these conditions, $\bar{x}$ will be a **precise estimator** of μ.

- The figure on the next slide shows the **standard error** of $\bar{x}$ becoming **smaller** as the **sample size** n gets **bigger.**


Population Distribution
Probability Density
8
10
μ = 12
14
16
Values of x


Distribution of X̄ for Different n
Probability Density
n = 4
n = 6
n = 10
n = 16
8
10
μx̄ = 12
14
16
Values of x̄

- As mentioned earlier, even if the sample comes from a **non-normal population**, as long as n **is large**, $\bar{x}$ will still follow a **normal** distribution approximately.

Normality of $\bar{X}$ When the Population *Isn't* Normal but n is Large

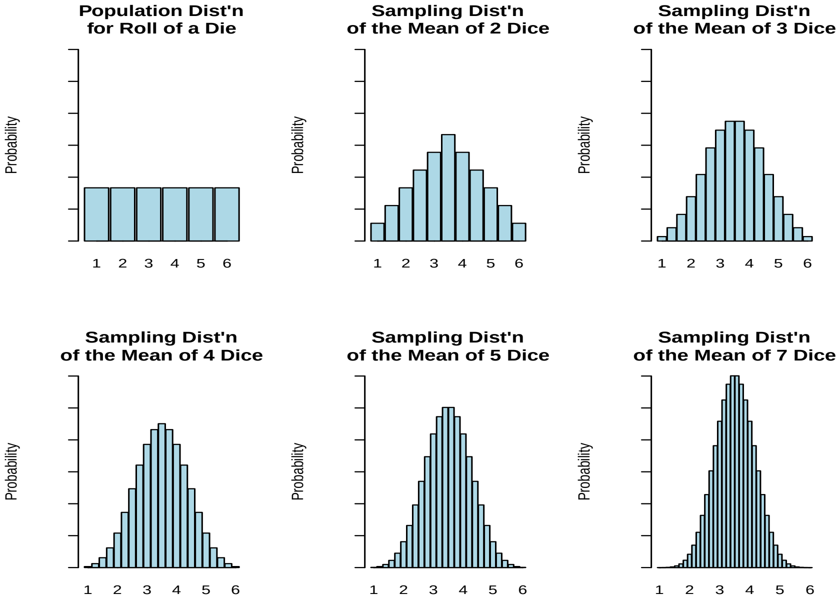
Normality of $\bar{X}$: If we take a **sample** of size n from a ***non-normal* population** whose mean is μ and whose standard deviation is σ, then as long as the **sample size** n is ***large***:

> The $\bar{x}$ **distribution** will be (at least approximately) ***normal*** with mean $\mu_{\bar{x}}$ and standard deviation $\sigma_{\bar{x}}$, where
>
> $$\mu_{\bar{x}} = \mu \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

This is known as the ***Central Limit Theorem***. The larger n is, the closer the $\bar{x}$ distribution gets to a normal distribution.

Usually $\boldsymbol{n \geq 30}$ is large enough.

- The figure on the next slide illustrates the **Central Limit Theorem.**


Population Dist'n
for Roll of a Die
Probability
1 2 3 4 5 6
Sampling Dist'n
of the Mean of 2 Dice
Probability
1 2 3 4 5 6
Sampling Dist'n
of the Mean of 3 Dice
Probability
1 2 3 4 5 6
Sampling Dist'n
of the Mean of 4 Dice
Probability
1 2 3 4 5 6
Sampling Dist'n
of the Mean of 5 Dice
Probability
1 2 3 4 5 6
Sampling Dist'n
of the Mean of 7 Dice
Probability
1 2 3 4 5 6

- The next few exercises show how to use the sampling distribution of $\bar{x}$ to compute **probabilities involving $\bar{x}$.**

Example

The U.S. army reports that head circumferences among the **population** of male soldiers follow a **normal** distribution with **mean** $\mu = 22.8$ inches and **standard deviation** $\sigma = 1.1$ inches.

Example

The U.S. army reports that head circumferences among the **population** of male soldiers follow a **normal** distribution with **mean** $\mu = 22.8$ inches and **standard deviation** $\sigma = 1.1$ inches.

A random sample of $n = 9$ soldiers is to be taken.

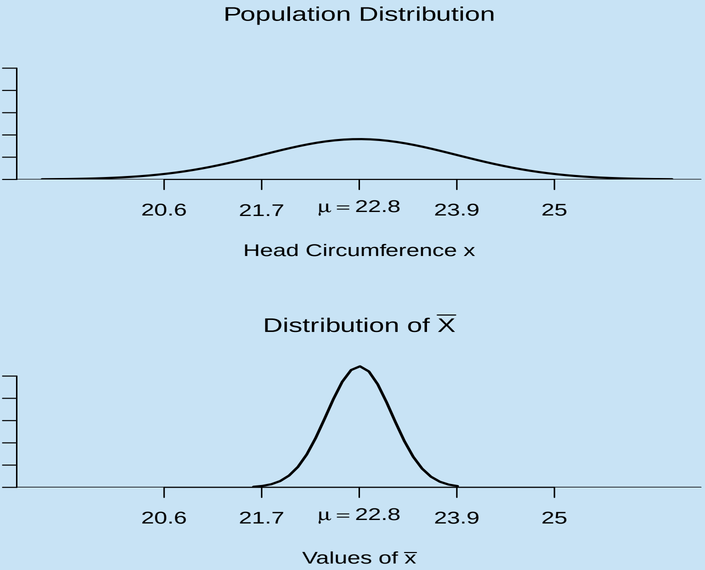
Example

The U.S. army reports that head circumferences among the **population** of male soldiers follow a **normal** distribution with **mean** $\mu = 22.8$ inches and **standard deviation** $\sigma = 1.1$ inches.

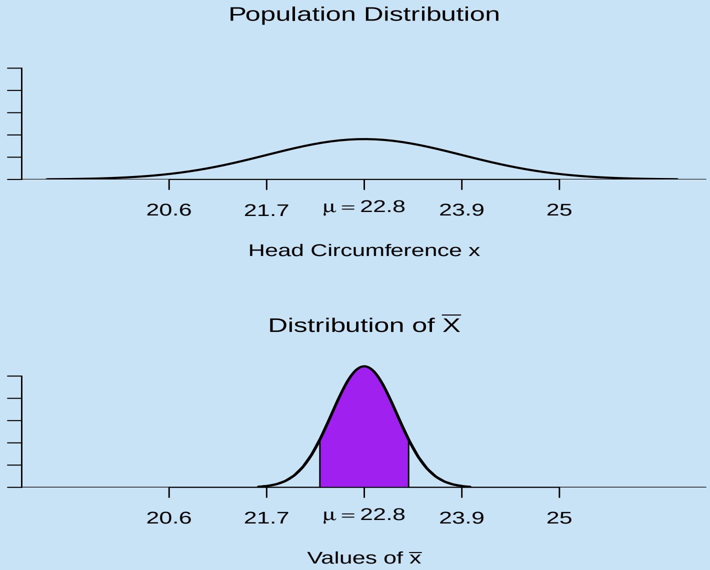
A random sample of $n = 9$ soldiers is to be taken.

The **sampling distribution** of $\bar{x}$ is **normal** with **mean** and **standard error**

$$\boldsymbol{\mu_{\bar{x}}} = \mu = \mathbf{22.8} \qquad \text{and} \qquad \boldsymbol{\sigma_{\bar{x}}} = \frac{\sigma}{\sqrt{n}} = \frac{\mathbf{1.1}}{\sqrt{\mathbf{9}}} = \mathbf{0.37}.$$


Population Distribution
20.6
21.7
μ = 22.8
23.9
25
Head Circumference x
Distribution of X̄
20.6
21.7
μ = 22.8
23.9
25
Values of x̄

We'll find the **proportion** of times (i.e. the **probability**) that a sample of size $n = 9$ would produce a sample mean $\bar{x}$ **between 22.3** and **23.3** inches (i.e. **within 0.5** of an inch **of the population mean** μ (22.8 inches)).


Population Distribution
20.6
21.7
μ = 22.8
23.9
25
Head Circumference x
Distribution of X̄
20.6
21.7
μ = 22.8
23.9
25
Values of x̄

The z**-score** for a ***sample mean*** of **23.3** inches is

$$z = \frac{23.3 - 22.8}{1.1/\sqrt{9}} = \mathbf{1.36},$$

The z**-score** for a ***sample mean*** of **23.3** inches is

$$z = \frac{23.3 - 22.8}{1.1/\sqrt{9}} = \mathbf{1.36},$$

and the z**-score** for a ***sample mean*** of **22.3** inches is

$$z = \frac{22.3 - 22.8}{1.1/\sqrt{9}} = \mathbf{-1.36},$$

The z**-score** for a ***sample mean*** of **23.3** inches is

$$z = \frac{23.3 - 22.8}{1.1/\sqrt{9}} = \mathbf{1.36},$$

and the z**-score** for a ***sample mean*** of **22.3** inches is

$$z = \frac{22.3 - 22.8}{1.1/\sqrt{9}} = \mathbf{-1.36},$$

From **Table II**, the **proportion** of z-scores *below* **1.36** is **0.9131**, and the **proportion** *below* **-1.36** is **0.0869**.

The z**-score** for a ***sample mean*** of **23.3** inches is

$$z = \frac{23.3 - 22.8}{1.1/\sqrt{9}} = \mathbf{1.36},$$

and the z**-score** for a ***sample mean*** of **22.3** inches is

$$z = \frac{22.3 - 22.8}{1.1/\sqrt{9}} = \mathbf{-1.36},$$

From **Table II**, the **proportion** of z-scores *below* **1.36** is **0.9131**, and the **proportion** *below* **-1.36** is **0.0869.**

Thus, the **proportion** ***between*** **1.36** and **-1.36** is

$$0.9131 - 0.0869 = \mathbf{0.8262}.$$

In other words, the *sample mean* will fall **between 22.3** and **23.3** inches in **82.62%** of all samples of size $n = 9$.

Exercise

Recall that head circumferences among the **population** of male soldiers follow a **normal** distribution with **mean** $\mu = 22.8$ inches and **standard deviation** $\sigma = 1.1$ inches.

a) In the last example, we found that for a sample of size $n = 9$, there's an **82.62%** chance that $\bar{x}$ will fall **within 0.5** inch of μ.

Exercise

Recall that head circumferences among the **population** of male soldiers follow a **normal** distribution with **mean** $\mu = 22.8$ inches and **standard deviation** $\sigma = 1.1$ inches.

a) In the last example, we found that for a sample of size $n = 9$, there's an **82.62%** chance that $\bar{x}$ will fall **within 0.5** inch of μ.

   If a **larger sample**, of size $n = 16$, is to be taken, do you think $\bar{x}$ will be **more likely** or **less likely** to fall within **0.5** of μ?

Exercise

b) Recalculate the **proportion** of times $\bar{x}$ would fall **between** **22.3** and **23.3** inches, but this time using $n = 16$. Compare the result to **0.8262** (from when n was 9).

Exercise

For Canadians, systolic blood pressure readings have a distribution whose **mean** is $\mu = 121$ whose **standard deviation** is $\sigma = 16$.

Exercise

For Canadians, systolic blood pressure readings have a distribution whose **mean** is $\mu = 121$ whose **standard deviation** is $\sigma = 16$.

a) A random sample of $n = 80$ Canadians is to be taken and their blood pressures measured.

Exercise

For Canadians, systolic blood pressure readings have a distribution whose **mean** is $\mu = 121$ whose **standard deviation** is $\sigma = 16$.

a) A random sample of $n = 80$ Canadians is to be taken and their blood pressures measured.

   Sketch the **sampling distribution** $\bar{x}$, with the values of its mean $\mu_{\bar{x}}$ and standard error $\sigma_{\bar{x}}$ marked on the horizontal axis.

b) What **proportion** of times would a sample of size $n = 80$ produce a *sample mean* $\bar{x}$ that's **between 119** and **123** (i.e. **within 2.0** units of the **population mean** μ (121))?

b) What **proportion** of times would a sample of size $n = 80$ produce a *sample mean* $\bar{x}$ that's **between 119** and **123** (i.e. **within 2.0** units of the **population mean** μ (121))?

c) If blood pressures in the **population** followed a **non-normal**, **right skewed** distribution, would the $\bar{x}$ distribution be (approximately) **normal** nonetheless? Explain.