Probability and Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

April 8, 2019

Topics

Objectives

Objectives:

- Recognize independent random variables.
- Explain what is meant by the sampling distribution of a statistic.
- Use the sampling distribution of the sample mean to find probabilities.
- Obtain the expected value and variance of a linear combination of random variables.
- Obtain the distribution of a linear combination of normal random variables.

Independent Random Variables (5.1)

- Two **random variables** X and Y are said to be ***independent*** if for *every* two intervals

$$A = (a_0, a_1) \qquad \text{and} \qquad B = (b_0, b_1),$$

the events $X \in A$ and $Y \in B$ are independent **events**, i.e.

$$P(Y \in B \,|\, X \in A) = P(Y \in B)$$

Independent Random Variables (5.1)

- Two **random variables** X and Y are said to be ***independent*** if for *every* two intervals

$$A = (a_0, a_1) \qquad \text{and} \qquad B = (b_0, b_1),$$

the events $X \in A$ and $Y \in B$ are independent **events**, i.e.

$$P(Y \in B \,|\, X \in A) = P(Y \in B)$$

or equivalently

$$P\left((X \in A) \cap (Y \in B) \right) = P(X \in A) \times P(Y \in B).$$

Independent Random Variables (5.1)

- Two **random variables** X and Y are said to be ***independent*** if for *every* two intervals

$$A = (a_0, a_1) \qquad \text{and} \qquad B = (b_0, b_1),$$

  the events $X \in A$ and $Y \in B$ are independent **events**, i.e.

$$P(Y \in B \,|\, X \in A) = P(Y \in B)$$

  or equivalently

$$P\left((X \in A) \cap (Y \in B) \right) = P(X \in A) \times P(Y \in B).$$

- Intuitively, X and Y are **independent** if their values **don't influence each other.**

- We can extend the definition of **independence** to more than two random variables.

- We can extend the definition of **independence** to more than two random variables.

  $X_1, X_2, \ldots, X_n$ are said to be ***independent*** if for every k ($k = 2, 3, \ldots, n$), every set of indices $i_1, i_2, \ldots, i_k$, and every collection of intervals $A_1, A_2, \ldots, A_k$,

$$P\left((X_{i_1} \in A_1) \cap (X_{i_2} \in A_2) \cap \cdots \cap (X_{i_k} \in A_k) \right)$$

$$= P(X_{i_1} \in A_1) \times P(X_{i_2} \in A_2) \times \cdots \times P(X_{i_k} \in A_k)$$

- We can extend the definition of **independence** to more than two random variables.

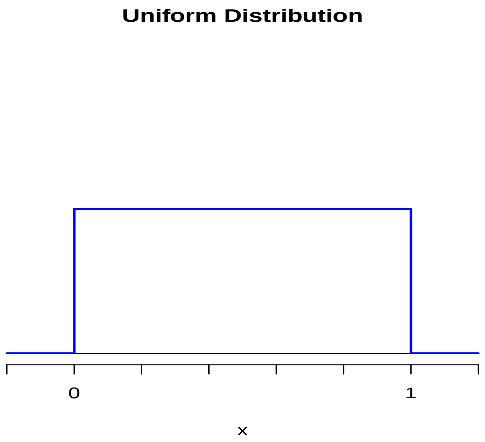
  $X_1, X_2, \ldots, X_n$ are said to be ***independent*** if for every k ($k = 2, 3, \ldots, n$), every set of indices $i_1, i_2, \ldots, i_k$, and every collection of intervals $A_1, A_2, \ldots, A_k$,

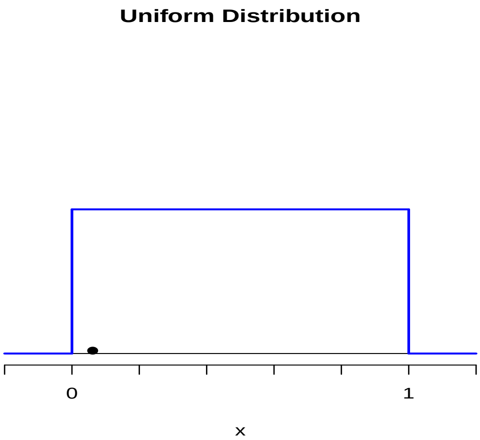
$$P\left((X_{i_1} \in A_1) \cap (X_{i_2} \in A_2) \cap \cdots \cap (X_{i_k} \in A_k) \right)$$
$$= P(X_{i_1} \in A_1) \times P(X_{i_2} \in A_2) \times \cdots \times P(X_{i_k} \in A_k)$$

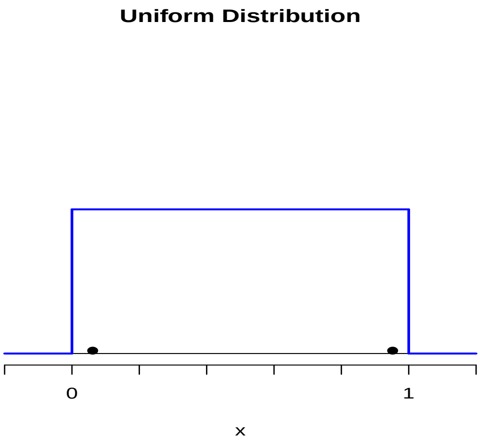
- Intuitively, $X_1, X_2, \ldots, X_n$ are **independent** if their values **don't influence each other.**

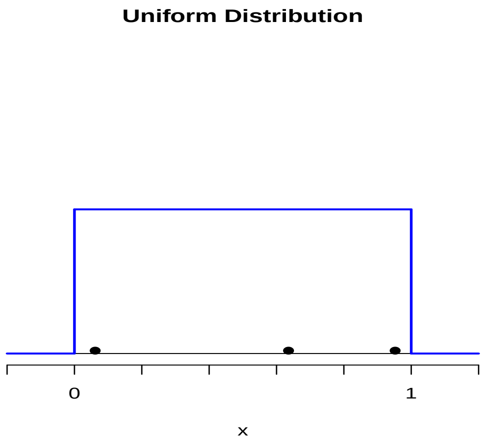
- Random variables $X_1, X_2, \ldots, X_n$ that are **independent** and all follow the **same probability distribution** are called a ***random sample*** from that distribution.

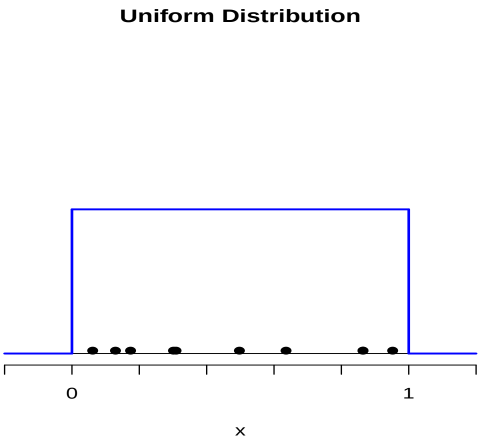
- Random variables $X_1, X_2, \ldots, X_n$ that are **independent** and all follow the **same probability distribution** are called a ***random sample*** from that distribution.
- Random samples are also sometimes called ***iid samples*** (for **independent** and **identically distributed**).

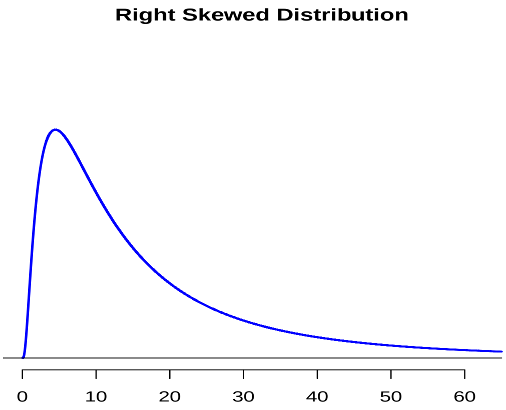

Uniform Distribution
0
1
x

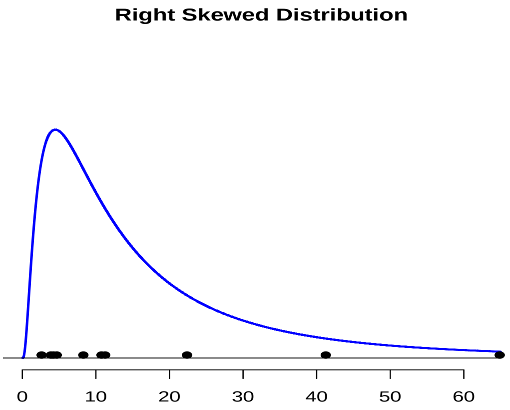

Uniform Distribution
0
1
x

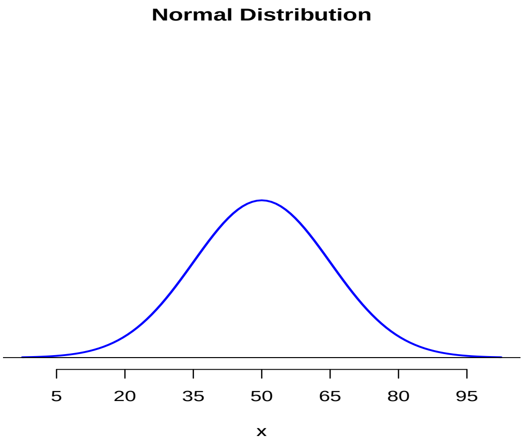

Uniform Distribution
0
1
x

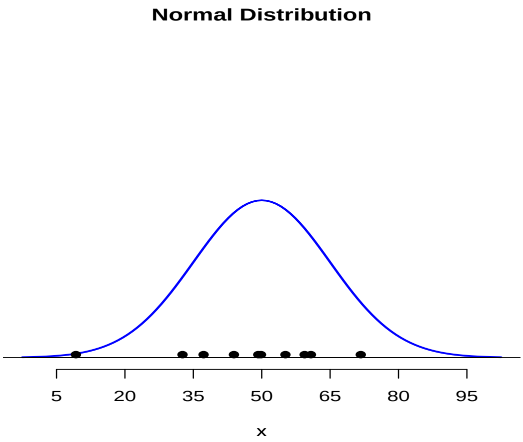

Uniform Distribution
0
1
x


Uniform Distribution
0
1
x


Right Skewed Distribution
0
10
20
30
40
50
60


Right Skewed Distribution
0
10
20
30
40
50
60


Normal Distribution
5
20
35
50
65
80
95
x


Normal Distribution
5
20
35
50
65
80
95
x

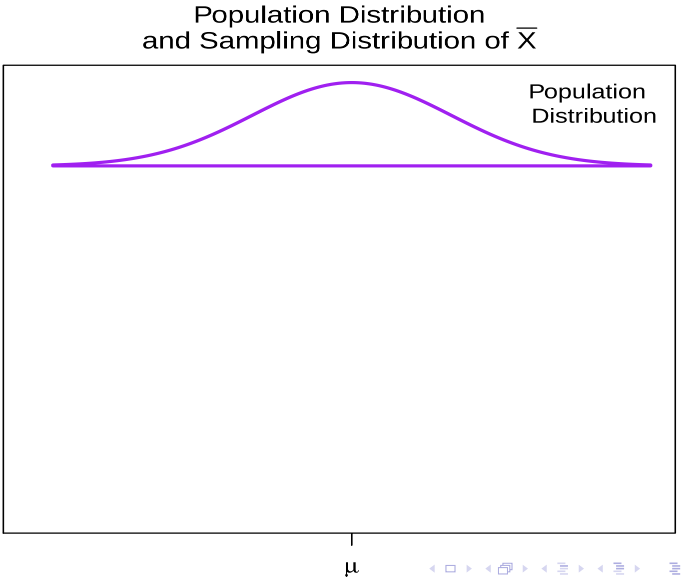
Sampling Distributions of Statistics (5.3)

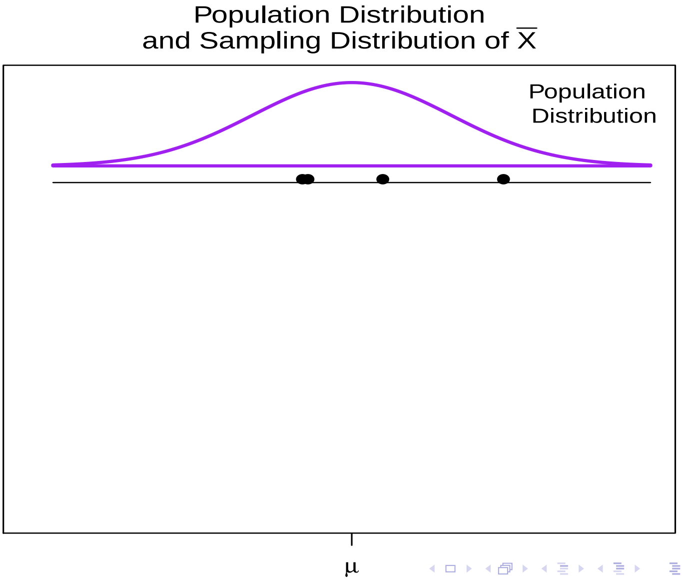
- A ***statistic*** is any numerical value computed from a set of **random sample** data. Therefore a statistic is a **random variable.**

Example

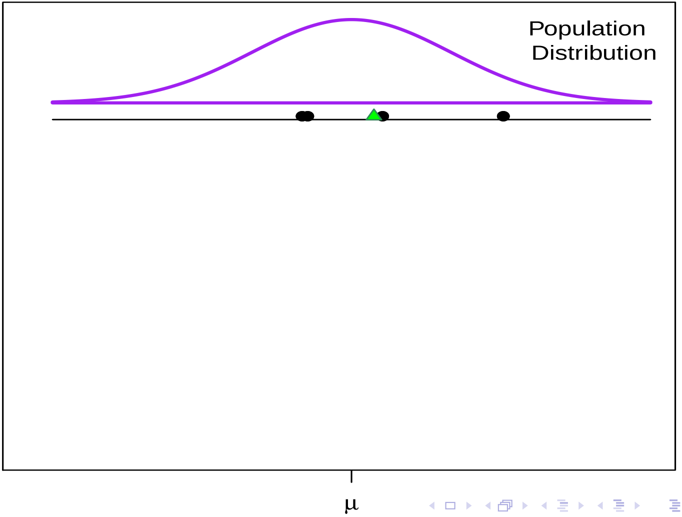
The sample mean $\bar{X}$, median $\tilde{X}$, and standard deviation S are all **statistics.**

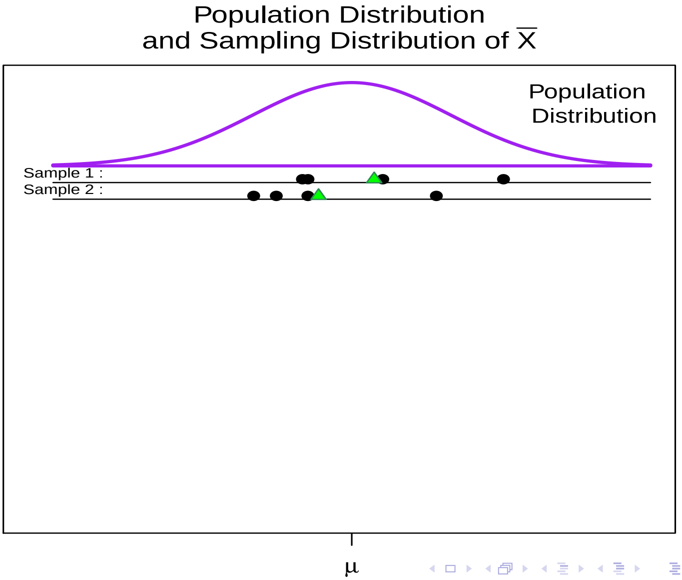
- The **probability distribution** of a statistic is called its ***sampling distribution***.

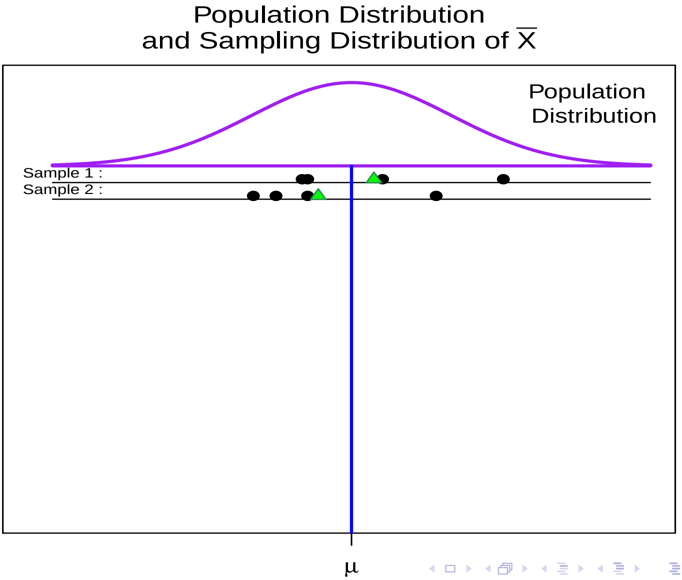

Population Distribution
and Sampling Distribution of $\overline{X}$
Population
Distribution
μ


Population Distribution
and Sampling Distribution of X̄
Population
Distribution
μ

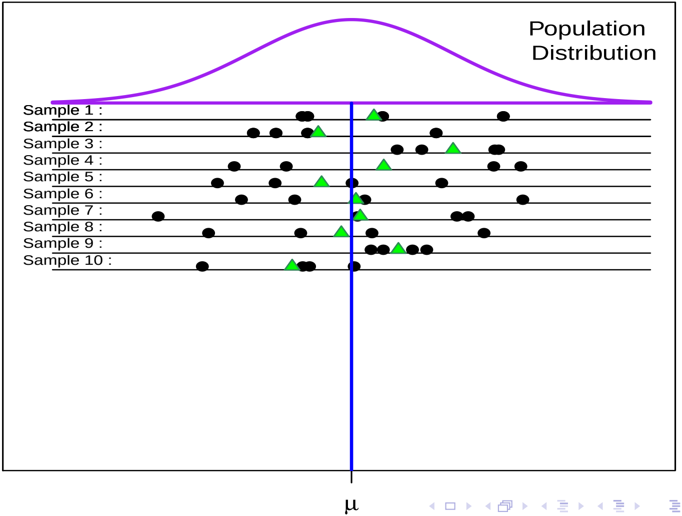
Population Distribution
and Sampling Distribution of $\overline{X}$



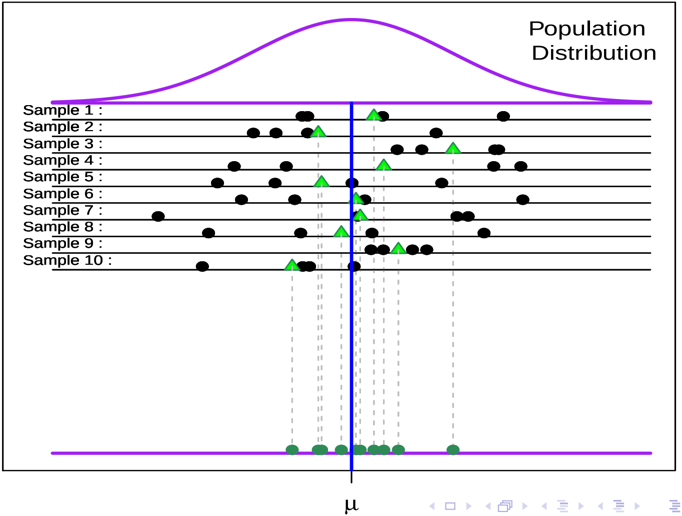

Population Distribution
and Sampling Distribution of X̄
Population
Distribution
Sample 1 :
Sample 2 :
μ

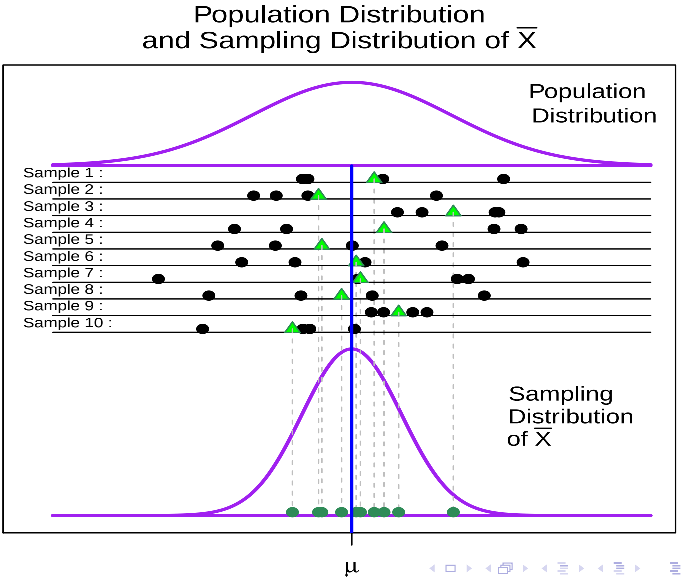

Population Distribution
and Sampling Distribution of $\overline{X}$
Population
Distribution
Sample 1 :
Sample 2 :
μ

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



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




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

Sampling Distribution of the Sample Mean $\bar{X}$ (5.4)

Proposition

Mean and Variance of $\bar{X}$: Suppose $X_1, X_2, \ldots, X_n$ are a **random sample** from a population whose mean and standard deviation are μ and σ. Then the **sampling distribution of $\bar{X}$** has mean $\mu_{\bar{x}}$ given by

$$\mu_{\bar{x}} = E(\bar{X}) = \mu$$

and variance $\sigma^2_{\bar{x}}$ given by

$$\sigma^2_{\bar{x}} = V(\bar{X}) = \frac{\sigma^2}{n}.$$

- Recall that the sample mean $\bar{X}$ is an ***estimator*** of the population mean μ.

- Recall that the sample mean $\bar{X}$ is an ***estimator*** of the population mean μ.
- Because $E(\bar{X}) = \mu$, it's called an ***unbiased*** estimator of μ.

- Recall that the sample mean $\bar{X}$ is an ***estimator*** of the population mean μ.
- Because $E(\bar{X}) = \mu$, it's called an ***unbiased*** estimator of μ.
- The standard deviation

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

is sometimes called the ***standard error of*** $\bar{X}$. It represents a **typical deviation** of $\bar{X}$ away from μ.

- Recall that the sample mean $\bar{X}$ is an ***estimator*** of the population mean μ.
- Because $E(\bar{X}) = \mu$, it's called an ***unbiased*** estimator of μ.
- The standard deviation

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

  is sometimes called the ***standard error of $\bar{X}$***. It represents a **typical deviation** of $\bar{X}$ away from μ.
- The **standard error** will be **small** if either:
  1. The population standard deviation σ is **small**, or
  2. The sample size n is **large**

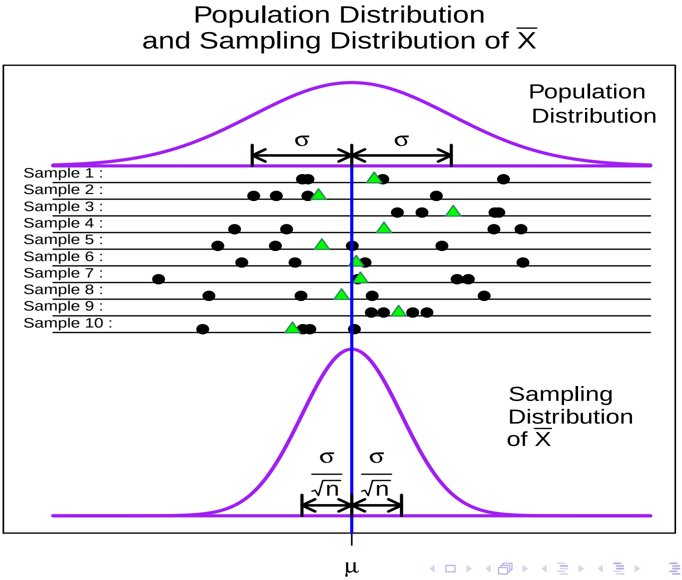
Proposition

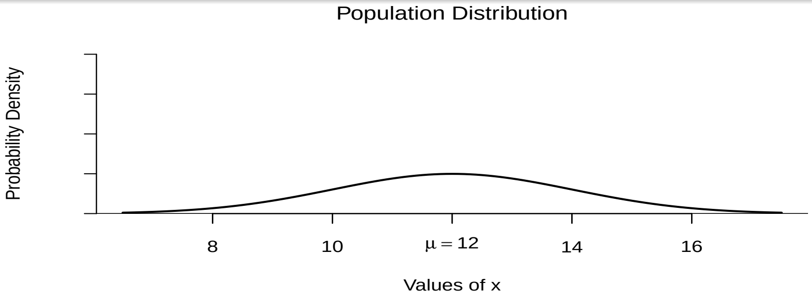
Normality of $\bar{X}$: Suppose $X_1, X_2, \ldots, X_n$ are a random sample from a **normal** population whose mean and standard deviation are μ and σ. Then

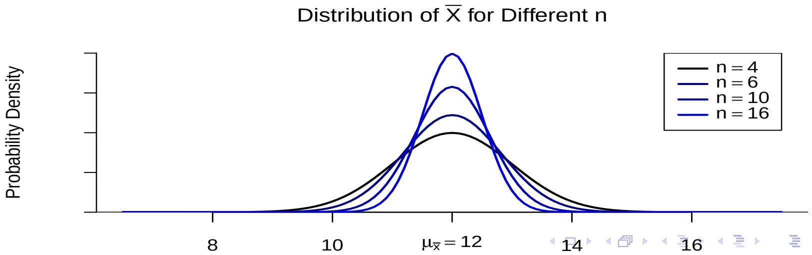
$$\bar{X} \sim \mathsf{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Thus the **standardized** version of $\bar{X}$ follows a **standard normal** distribution, i.e.

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


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


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
μ_X̄ = 12
14
16

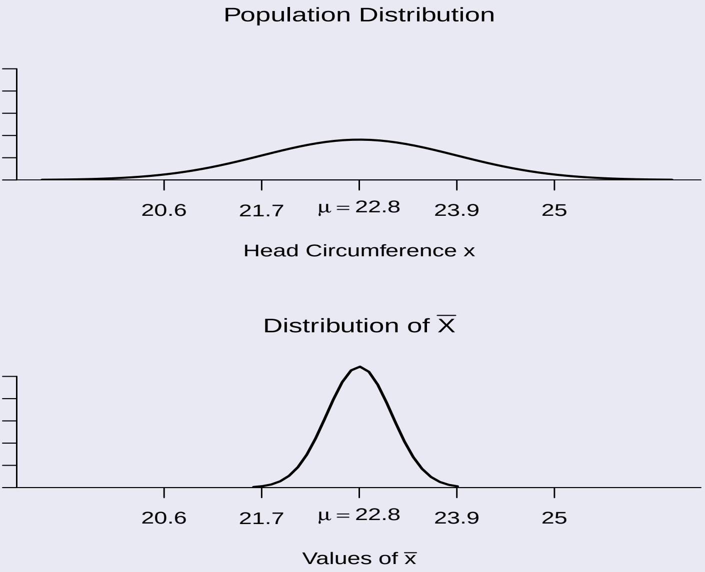
Example

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

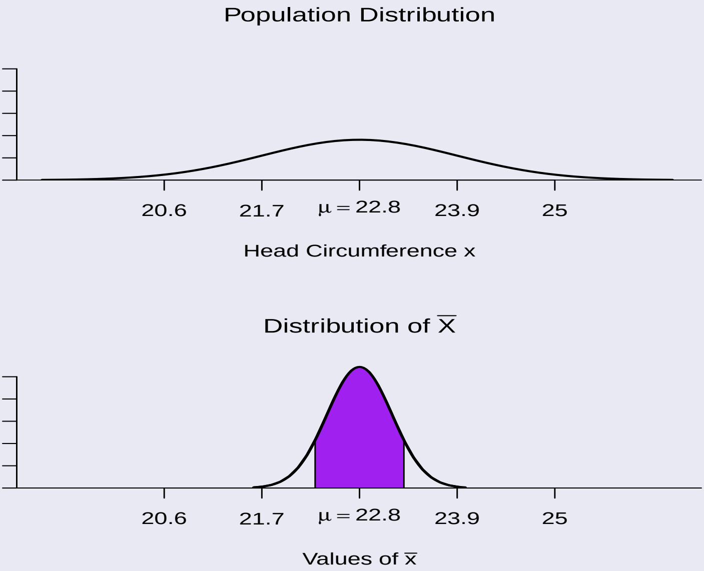
A random sample of $n = 9$ soldiers is to be taken.

The **sampling distribution of** $\bar{X}$ is:

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$


Population Distribution
20.6
21.7
μ = 22.8
23.9
25
Head Circumference x
Distribution of $\overline{X}$
20.6
21.7
μ = 22.8
23.9
25
Values of $\overline{x}$

We'll find the probability $P(22.3 \leq \bar{X} \leq 23.3)$ that $\bar{X}$ will be **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 probability is

The probability is

$$P(22.3 \le \bar{X} \le 23.3) = P\left(\frac{22.3-\mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{23.3-\mu}{\sigma/\sqrt{n}}\right)$$

The probability is

$$\begin{aligned} P(22.3 \leq \bar{X} \leq 23.3) &= P\left(\frac{22.3-\mu}{\sigma/\sqrt{n}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{23.3-\mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(\frac{22.3-22.8}{1.1/\sqrt{9}} \leq Z \leq \frac{23.3-22.8}{1.1/\sqrt{9}}\right) \end{aligned}$$

The probability is

$$
\begin{aligned}
P(22.3 \leq \bar{X} \leq 23.3) &= P\left(\tfrac{22.3-\mu}{\sigma/\sqrt{n}} \leq \tfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \tfrac{23.3-\mu}{\sigma/\sqrt{n}}\right) \\
&= P\left(\tfrac{22.3-22.8}{1.1/\sqrt{9}} \leq Z \leq \tfrac{23.3-22.8}{1.1/\sqrt{9}}\right) \\
&= P(-1.36 \leq Z \leq 1.36)
\end{aligned}
$$

The probability is

$$\begin{aligned} P(22.3 \leq \bar{X} \leq 23.3) &= P\left(\frac{22.3-\mu}{\sigma/\sqrt{n}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{23.3-\mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(\frac{22.3-22.8}{1.1/\sqrt{9}} \leq Z \leq \frac{23.3-22.8}{1.1/\sqrt{9}}\right) \\ &= P(-1.36 \leq Z \leq 1.36) \\ &= \phi(1.36) - \phi(-1.36) \end{aligned}$$

The probability is

$$\begin{aligned} P(22.3 \le \bar{X} \le 23.3) &= P\left(\frac{22.3-\mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{23.3-\mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(\frac{22.3-22.8}{1.1/\sqrt{9}} \le Z \le \frac{23.3-22.8}{1.1/\sqrt{9}}\right) \\ &= P(-1.36 \le Z \le 1.36) \\ &= \phi(1.36) - \phi(-1.36) \\ &= 0.9131 - 0.0869 \end{aligned}$$

The probability is

$$\begin{aligned} P(22.3 \le \bar{X} \le 23.3) &= P\left(\frac{22.3-\mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{23.3-\mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(\frac{22.3-22.8}{1.1/\sqrt{9}} \le Z \le \frac{23.3-22.8}{1.1/\sqrt{9}}\right) \\ &= P(-1.36 \le Z \le 1.36) \\ &= \phi(1.36) - \phi(-1.36) \\ &= 0.9131 - 0.0869 \\ &= \mathbf{0.8262}. \end{aligned}$$

Proposition

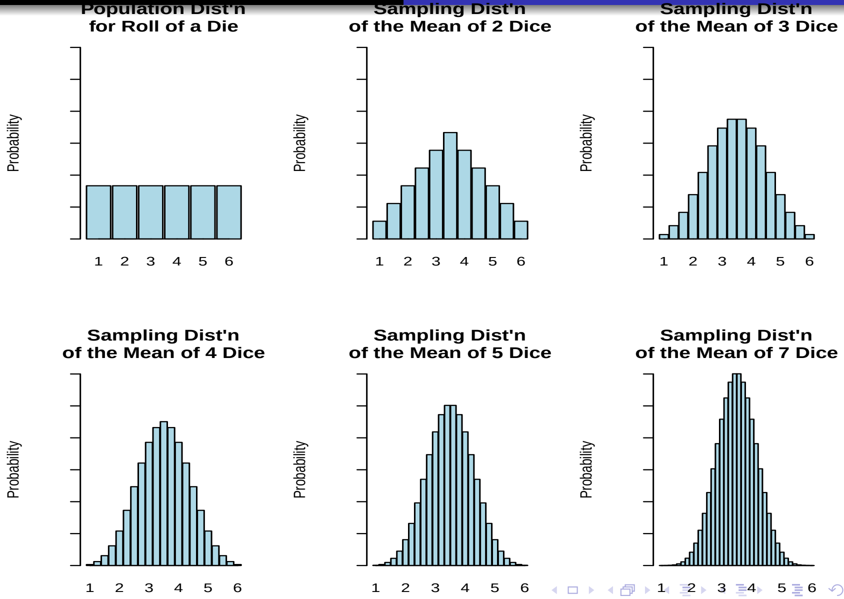
The Central Limit Theorem: Suppose $X_1, X_2, \ldots, X_n$ are a random sample from **any** population whose mean is μ and whose standard deviation is σ (with $\sigma < \infty$). Then **if n is large**,

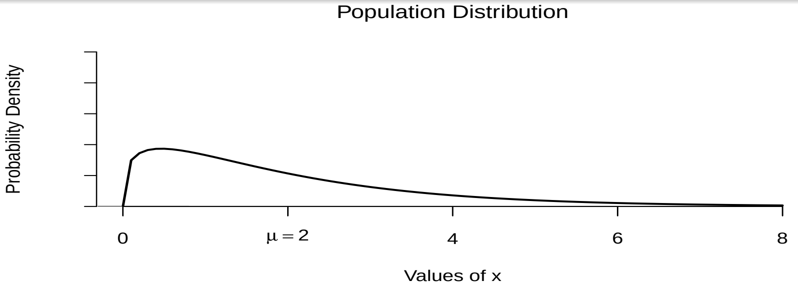
$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad \text{(approximately)}$$

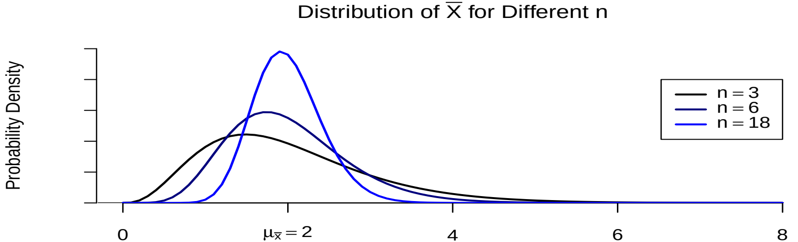
The larger n is, the closer to a normal distribution the $\bar{X}$ distribution will be.

Furthermore, if $T_o = X_1 + X_2 + \cdots + X_n$, then

$$T_o \sim \mathrm{N}\left(n\mu, \sqrt{n}\sigma\right) \qquad \text{(approximately)}$$


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


Population Distribution
Probability Density
0
μ = 2
4
6
8
Values of x


Distribution of X̄ for Different n
Probability Density
n = 3
n = 6
n = 18
0
μ_x̄ = 2
4
6
8

- In practice, n is (usually) **large enough** for the **Central Limit Theorem** to apply as long as $\boldsymbol{n \geq 30}$.

Linear Combinations of Random Variables (5.5)

- Given a collection of random variables $X_1, X_2, \ldots, X_n$ and constants $\boldsymbol{a_1, a_2, \ldots, a_n}$, we call

$$Y \;=\; a_1X_1 + a_2X_2 + \cdots + a_nX_n \;=\; \sum_{i=1}^{n} a_iX_i$$

a ***linear combination*** the X_i's.

Linear Combinations of Random Variables (5.5)

- Given a collection of random variables $X_1, X_2, \ldots, X_n$ and constants $\boldsymbol{a_1, a_2, \ldots, a_n}$, we call

$$Y \;=\; a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \;=\; \sum_{i=1}^{n} a_i X_i$$

  a ***linear combination*** the X_i's.

- A linear combination of random variables is itself a random variable.

Linear Combinations of Random Variables (5.5)

- Given a collection of random variables $X_1, X_2, \ldots, X_n$ and constants $\boldsymbol{a_1, a_2, \ldots, a_n}$, we call

  $$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n = \sum_{i=1}^{n} a_iX_i$$

  a ***linear combination*** the X_i's.

- A linear combination of random variables is itself a random variable.

- The sample mean $\bar{X}$ is a linear combination of the sample $X_1, X_2, \ldots, X_n$ (with $a_i = 1/n$ for all i).

Example

Consider someone who owns **100** shares of stock **A**, **200** shares of stock **B**, and **500** shares of stock **C**. Denote the share prices of these three stocks by $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, and $\boldsymbol{X_3}$, respectively.

Then the value of this individual's stock holdings is the **linear combination**

$$Y = 100\,X_1 + 200\,X_2 + 500\,X_3.$$

Proposition

Mean and Variance of a Linear Combination: Suppose $X_1, X_2, \ldots, X_n$ have means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. Let $a_1, a_2, \ldots, a_n$ be any constants. Then

1. Regardless of whether or not the X_i's are independent,

$$\begin{aligned} &E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\ &\quad = a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n) \\ &\quad = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n\,. \end{aligned}$$

1. If the X_i's are **independent**, then

$$\begin{aligned} & V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\ &= a_1^2V(X_1) + a_2^2V(X_2) + \cdots + a_n^2V(X_n) \\ &= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 \end{aligned}$$

and thus

$$\begin{aligned} & SD(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\ &= \sqrt{a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2}. \end{aligned}$$

Example

A town has two industrial plants: a cement production plant and a steel mill.

The daily SO_2 emissions (lbs) X_1 from the cement plant varies from day to day, with

$$X_1 \sim \mathbf{N}(8800,\ 340).$$

Likewise, the daily emissions X_2 from the steel mill varies, with

$$X_2 \sim \mathbf{N}(410,\ 75).$$

The total **combined** emissions from these two sources, $\boldsymbol{X_1 + X_2}$, is a random variable, with

$$E(X_1 + X_2) \;=\; 8,800 + 410 \;=\; \mathbf{9,210} \quad \text{lb.}$$

and

$$SD(X_1 + X_2) \;=\; \sqrt{340^2 + 75^2} \;=\; \mathbf{348} \quad \text{lb.}$$

- An important special case of the last proposition is the **difference** between two random variables $X_1 - X_2$.

Corollary

Mean and Variance of a Difference:

1. Regardless of whether or not X_1 and X_2 are independent,

$$E(X_1 - X_2) \;=\; E(X_1) \;-\; E(X_2) \;=\; \mu_1 \;-\; \mu_2$$

2. If X_1 and X_2 are **independent**, then

$$V(X_1 - X_2) \;=\; V(X_1) + V(X_2) \;=\; \sigma_1^2 + \sigma_2^2$$

and thus

$$SD(X_1 - X_2) \;=\; \sqrt{\sigma_1^2 + \sigma_2^2}$$

Proposition

Distribution of a Linear Combination: Suppose $X_1, X_2, \ldots,$ X_n are independent **normal** random variables (with possibly different means and/or variances).

Then any **linear combination** of the X_i's also follows a **normal** distribution (with mean and variance as in the last proposition).

Example

Consider again the daily SO_2 emissions (lbs) from a town's cement plant, X_1, and from its its steel mill, X_2, with with

$$X_1 \sim \mathbf{N}(8800,\ 340) \quad \text{and} \quad X_2 \sim \mathbf{N}(410,\ 75).$$

Then the total **combined** emissions from these two sources, $X_1 + X_2$, is a random variable, with

$$X_1 + X_2 \sim \mathrm{N}(9210,\ 348).$$

Then the total **combined** emissions from these two sources, $\boldsymbol{X_1 + X_2}$, is a random variable, with

$$\boldsymbol{X_1 + X_2} \sim \mathrm{N}(\mathbf{9210,\ 348}).$$

- An important special case of the last proposition is the **difference $X_1 - X_2$** between two **normal** random variables.

Corollary

Distribution of a Difference: Suppose $\boldsymbol{X_1 \sim}$ **N**$\boldsymbol{(\mu_1, \sigma_1)}$ and $\boldsymbol{X_2 \sim}$ **N**$\boldsymbol{(\mu_2, \sigma_2)}$, and X_1 and X_2 are **independent.** Then

$$X_1 - X_2 \;\sim\; \mathrm{N}\left(\mu_1 - \mu_2,\; \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$