

# 6

# Central Limit Theorem

**(Chs 6.4, 6.5)**

# Motivating Example

In the next few weeks, we will be focusing on making “statistical inference” about the true mean of a population by using sample datasets.

Examples?

# Random Samples

The r.v.'s  $X_1, X_2, \dots, X_n$  are said to form a (simple) **random sample** of size  $n$  if

1. The  $X_i$ 's are independent r.v.'s.
2. Every  $X_i$  has the same probability distribution.

We say that these  $X_i$ 's are independent and identically distributed (***iid***).

# Estimators and Their Distributions

We use **estimators** to summarize our i.i.d. sample.

Examples?

# Estimators and Their Distributions

We use **estimators** to summarize our i.i.d. sample.

Any estimator, including  $\bar{X}$ , is a random variable (since it is based on a random sample).

This means that  $\bar{X}$  has a distribution of its own, which is referred to as **sampling distribution of the sample mean**.

This sampling distribution depends on:

- 1) The population distribution (normal, uniform, etc.)
- 2) The sample size  $n$
- 3) The method of sampling

# Estimators and Their Distributions

Any estimator, including  $\bar{X}$ , is a random variable (since it is based on a random sample).

This means that  $\bar{X}$  has a distribution of its own, which is referred to as **sampling distribution of the sample mean**.

This sampling distribution depends on:

- 1) The population distribution (normal, uniform, etc.)
- 2) The sample size  $n$
- 3) The method of sampling

The standard deviation of this distribution is called **the standard error of the estimator**.

# Example

A certain brand of MP3 player comes in three models:

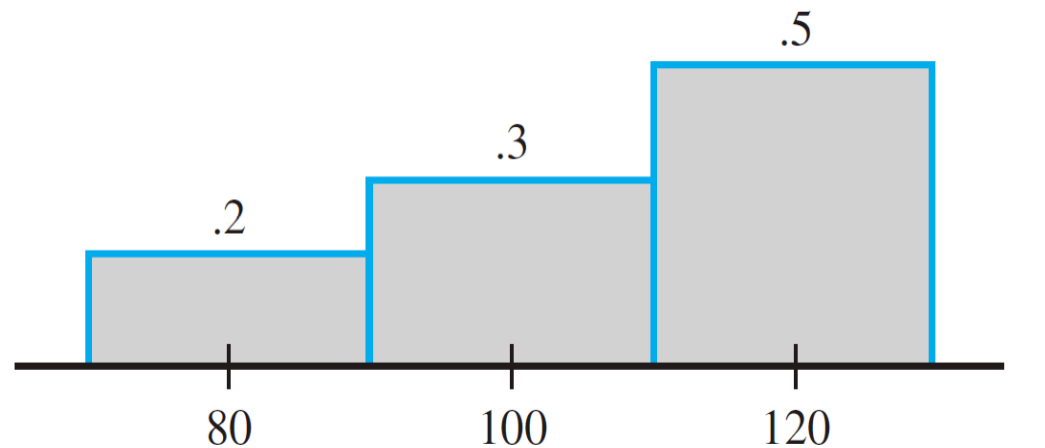
- 2 GB model, priced \$80,
- 4 GB model priced at \$100,
- 8 GB model priced \$120.

Suppose the probability distribution of the cost  $X$  of a single randomly selected MP3 player purchase is given by

$x$	80	100	120
$p(x)$	.2	.3	.5

# Example

We can use this pdf to calculate  $\mu = 106$ ,  $\sigma^2 = 244$ . This means that the average amount spent is \$106, and the standard deviation is \$15.60. A graph of this pdf is:



Original distribution:

$$\mu = 106, \sigma^2 = 244$$



# Example

Suppose on a particular day only two MP3 players are sold. Let  $X_1$  = the revenue from the first sale and  $X_2$  = revenue from the second.  $X_1$  and  $X_2$  are independent, and have the previously shown probability distribution.

In other words,  $X_1$  and  $X_2$  constitute a random sample from that distribution.

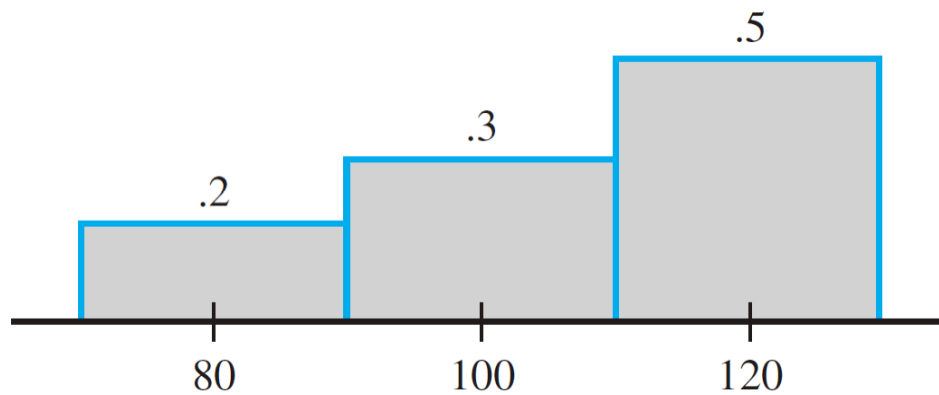
How do we find the mean and variance of this random sample?

# Example

cont' d

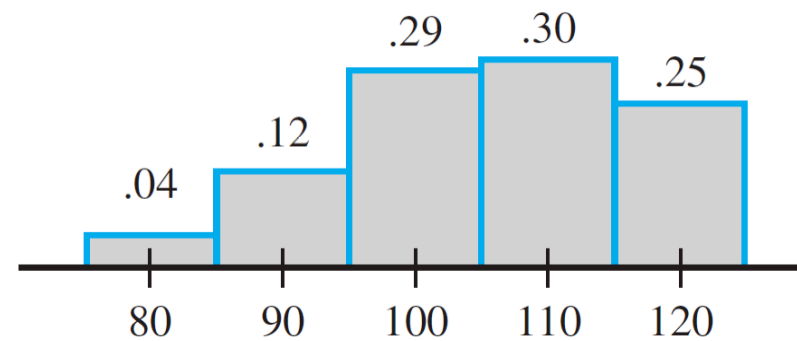
The complete sampling distribution of this  $\bar{X}$  is :

$\bar{x}$	80	90	100	110	120
$p_{\bar{X}}(\bar{x})$	.04	.12	.29	.30	.25



Original distribution:

$$\mu = 106, \sigma^2 = 244$$



$\bar{X}$  's distribution

# Example

cont' d

What do you think the mean and variance would be if we had four samples instead of 2?

# Example

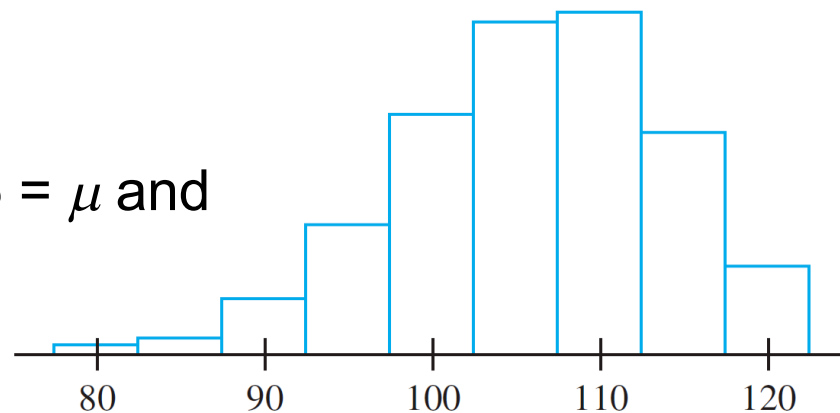
cont' d

If there had been four purchases on the day of interest, the sample average revenue  $\bar{X}$  would be based on a random sample of four  $X_i$ 's, each having the same distribution.

More calculation eventually yields the pmf of  $\bar{X}$  for  $n = 4$  as

$\bar{x}$	80	85	90	95	100	105	110	115	120
$p_{\bar{X}}(\bar{x})$	.0016	.0096	.0376	.0936	.1761	.2340	.2350	.1500	.0625

From this,  $\mu_{\bar{X}} = 106 = \mu$  and  
 $\sigma_{\bar{X}}^2 = 61 = \sigma^2/4$ .



# Distribution of the Sample Mean

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with mean value  $\mu$  and standard deviation  $\sigma$ . Then

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

2.  $V(\bar{X}) = \sigma_{\bar{X}}^2 = \sigma^2/n$  and  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$

The standard deviation  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$  is also called the *standard error of the mean*

# Distribution of the Sample Mean

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with mean value  $\mu$  and standard deviation  $\sigma$ . Then

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

2.  $V(\bar{X}) = \sigma_{\bar{X}}^2 = \sigma^2/n$  and  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$

The standard deviation  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$  is also called the *standard error of the mean*

Great, but what is the \*distribution\* of the sample mean?

# A Normal Population Distribution

## Proposition:

Let  $X_1, X_2, \dots, X_n$  be a random sample from a *normal* distribution with mean  $\mu$  and standard deviation  $\sigma$ . Then for *any*  $n$ ,  $\bar{X}$  is normally distributed (with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ )

We know everything there is to know about the distribution of the sample mean when the population distribution is normal.

# A Normal Population Distribution

## Proposition:

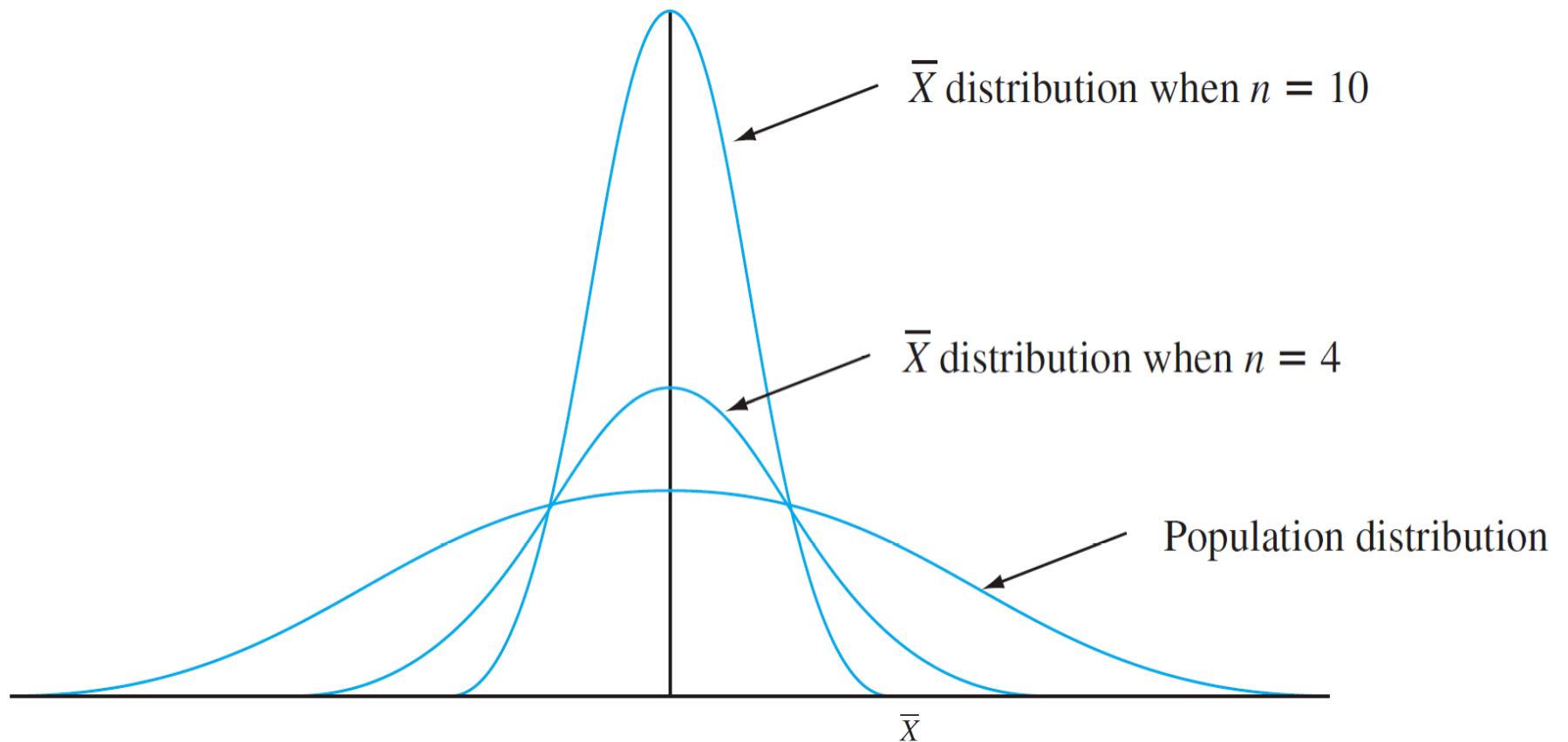
Let  $X_1, X_2, \dots, X_n$  be a random sample from a *normal* distribution with mean  $\mu$  and standard deviation  $\sigma$ . Then for *any*  $n$ ,  $\bar{X}$  is normally distributed (with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ )


We know everything there is to know about the distribution of the sample mean when the population distribution is normal.

In particular, probabilities such as  $P(a \leq \bar{X} \leq b)$  can be obtained simply by standardizing.



# A Normal Population Distribution





**But what if the underlying  
distribution of  $X_i$ 's is not  
normal?**

## **The Central Limit Theorem**

# The Central Limit Theorem (CLT)

When the  $X_i$ 's are normally distributed, so is  $\bar{X}$  for every sample size  $n$ .

Even when the population distribution is highly non-normal, averaging produces a distribution more bell-shaped than the one being sampled.

A reasonable conjecture is that if  $n$  is large, a suitable normal curve will approximate the actual distribution of  $\bar{X}$ .

The formal statement of this result is one of the most important theorems in probability: CLT

# The Central Limit Theorem

## Theorem

### The Central Limit Theorem (CLT)

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with mean  $\mu$  and variance  $\sigma^2$ .

Then if  $n$  is “**large enough**”,  $\bar{X}$  has approximately a normal distribution with  $\mu_{\bar{X}} = \mu$  and  $\sigma_{\bar{X}}^2 = \sigma^2/n$ ,

The larger the value of  $n$ , the better the approximation.

# The Central Limit Theorem

## Theorem

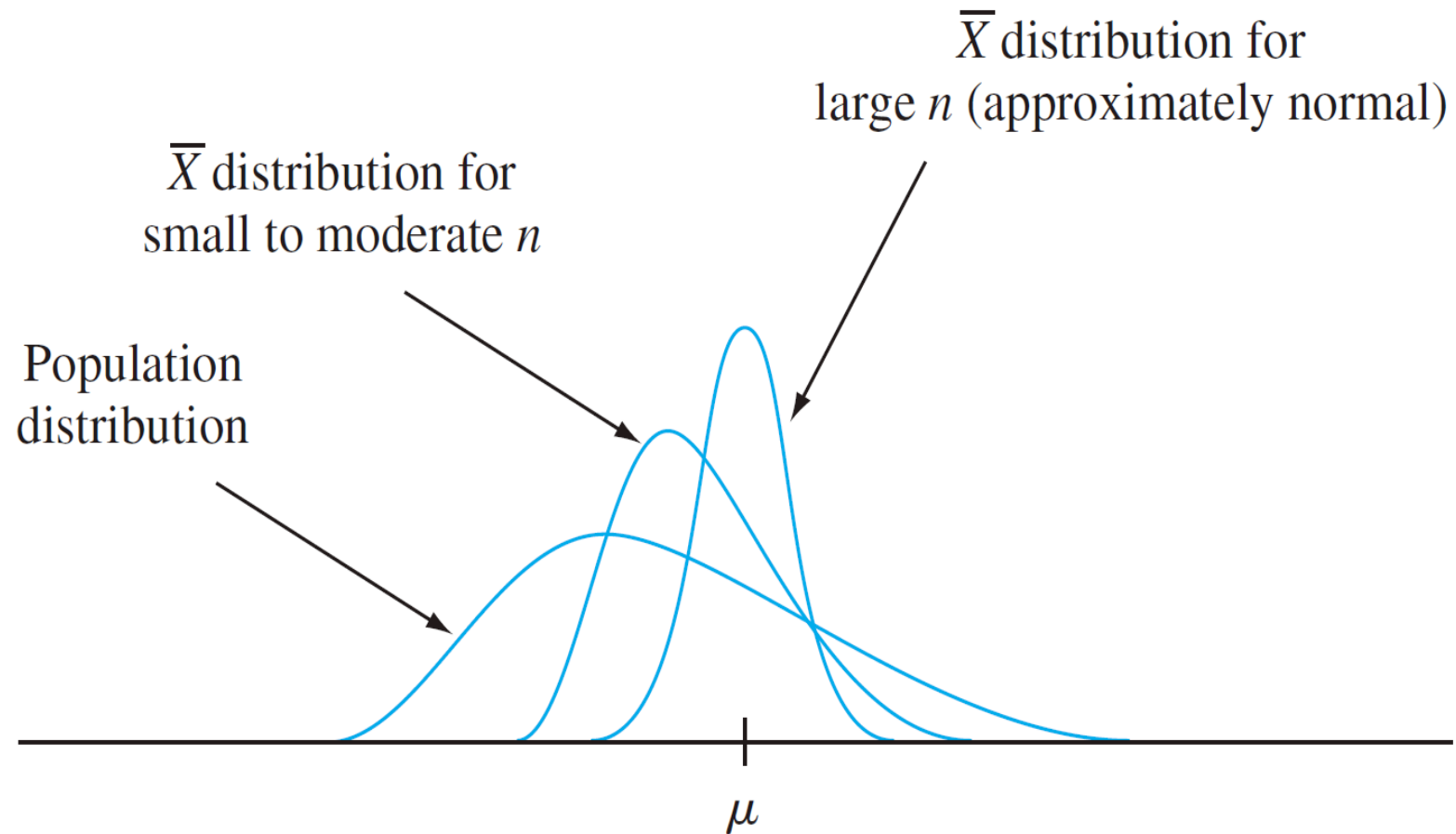
### The Central Limit Theorem (CLT)

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with mean  $\mu$  and variance  $\sigma^2$ .

Then if  $n$  is “**large enough**”,  $\bar{X}$  has approximately a normal distribution with  $\mu_{\bar{X}} = \mu$  and  $\sigma_{\bar{X}}^2 = \sigma^2/n$ ,

What is “**large enough**”? It depends. In this class, we'll say  $n \geq 30$  is large enough.

# The Central Limit Theorem



The Central Limit Theorem illustrated

# Example

The amount of impurity in a batch of a chemical product is a random variable with mean value 4.0 g and standard deviation 1.5 g. (unknown distribution)

If 50 batches are independently prepared, what is the (approximate) probability that the average amount of impurity in these 50 batches is between 3.5 and 3.8 g?

# The Central Limit Theorem

The CLT provides insight into why many random variables have probability distributions that are approximately normal.

For example, the measurement error in a scientific experiment can be thought of as the sum of a number of underlying perturbations and errors of small magnitude.

A practical difficulty in applying the CLT is in knowing when  $n$  is sufficiently large. The problem is that the accuracy of the approximation for a particular  $n$  depends on the shape of the original underlying distribution being sampled.



# R CODE

Normal random variables in R.  
The CLT.