5E: Statistics with Economics and Business Applications

Chapter 7 Sampling Distribution

7.3-7.6 Point Estimation, Sampling Distribution, Central Limit Theorem

Read 7.1, 7.2, Skip 7.7

Introduction

- **<u>Parameters</u>** are numerical descriptive measures for populations.
  - Two parameter for a binomial distribution: the number of trials ***n*** and the success probability of each trial ***p***.
  - Two parameters for a normal distribution: mean μ and standard deviation σ.
- Very often the values of parameters that specify the exact form of a distribution are unknown.
- We need to rely on the **sample** to learn about these parameters.

Sampling

Examples:

- A pollster is sure that the responses to his "agree/disagree" question will follow a binomial distribution, but p, the proportion of those who "agree" in the population, is unknown.
- Every year UCSB received thousands of applications from prospective students. The individual's SAT score is approximately normally distributed, but the mean μ and the standard deviation σ of the SAT score are unknown.

If you want the sample to provide reliable information about the population, you must select your sample in a certain way!

Simple Random Sampling

- The **sampling plan** or **experimental design** determines the amount of information you can extract, and often allows you to measure the **reliability of your inference**.

- **Simple random sampling** is a method of sampling that allows each possible sample of size *n* an equal probability of being selected.

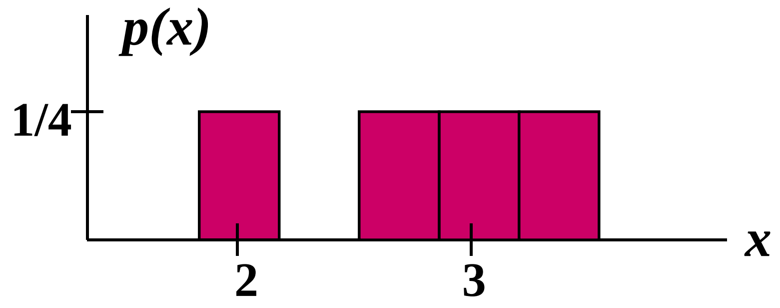
Illustrative Example: Simple Random Sampling

Population: 3, 5, 2, 1

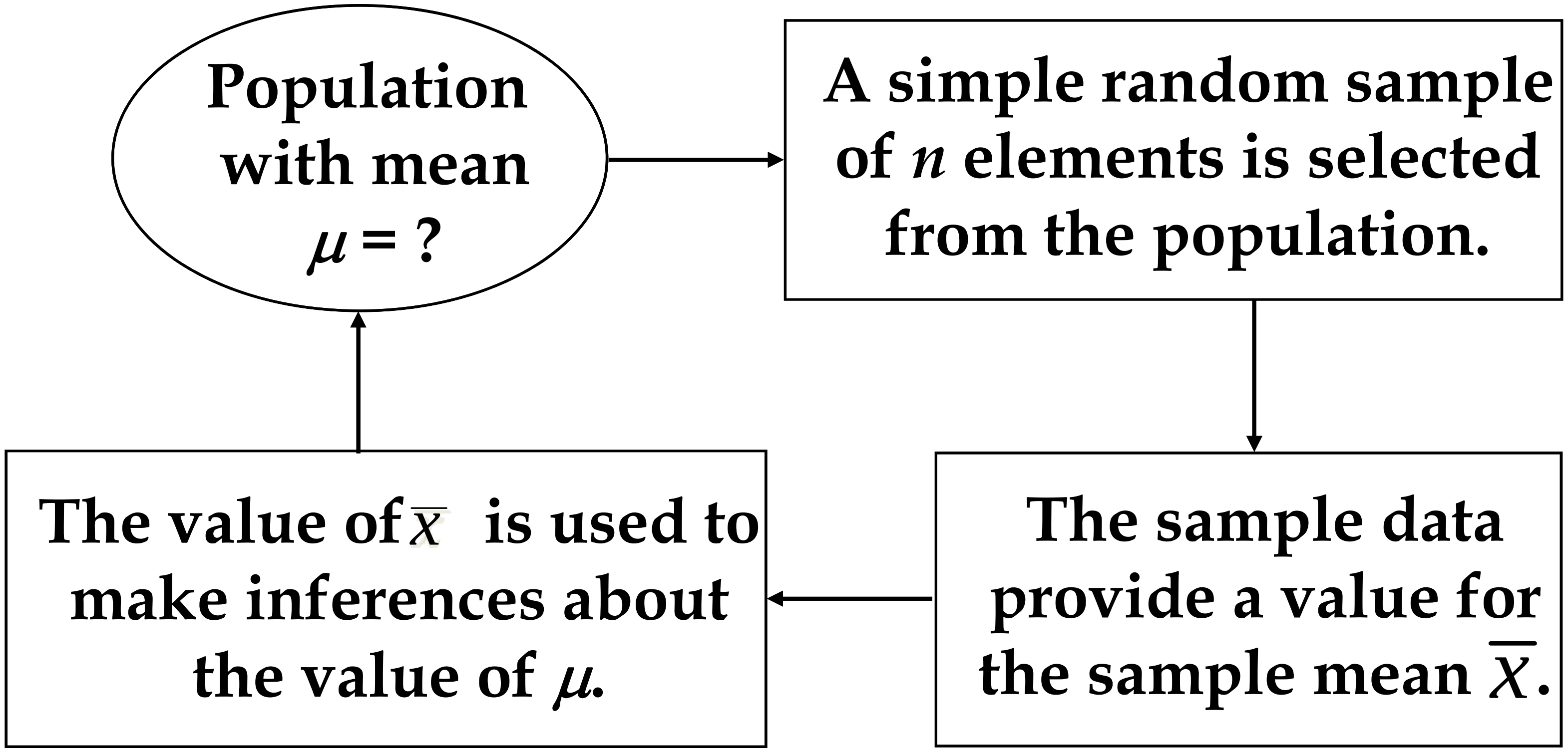
Draw samples of size n = 3 without replacement

Samples	$\bar{X}$
3, 5, 2	$10/3 = 3.33$
3, 5, 1	$9/3 = 3$
3, 2, 1	$6/3 = 2$
5, 2, 1	$8/3 = 2.67$

Each value of x-bar is equally likely, with probability 1/4



Recall: Process of Statistical Inference



$$\bar{X} \longrightarrow \mu$$

7. 3 Point Estimation

In <u>point estimation</u> we use the data from the sample to compute a value of a sample statistic that serves as an estimate of a population parameter.

We refer to $\bar{x}$ as the <u>point estimator</u> of the population mean μ.

s is the <u>point estimator</u> of the population standard deviation σ.

$\bar{p}$ is the <u>point estimator</u> of the population proportion p.

7.3 Point Estimation

Example: St. Andrew's College

Recall that St. Andrew's College received 900 applications from prospective students. The application form contains a variety of information including the individual's SAT score and whether or not the individual desires on-campus housing.

At a meeting in a few hours, the Director of Admissions would like to announce the average SAT score and the proportion of applicants that want to live on campus, for the population of 900 applicants.

Point Estimation

- **Example: St. Andrew's College**

However, the necessary data on the applicants have not yet been entered in the college's computerized database. So, the Director decides to estimate the values of the population parameters of interest based on sample statistics. The sample of 30 applicants can be selected at random from 900 populations.

Point Estimation

- **Draw 30 Samples from 900 population :**

SAT Score	On-Campus Housing
1207	No
1143	Yes
1091	Yes
1108	No
1227	Yes
982	Yes
1363	Yes
1108	No

Note: Rows 9-30 are not shown.

Point Estimation

- **$\bar{x}$ as Point Estimator of μ**

$$\bar{x} = \frac{\sum x_i}{30} = \frac{32,910}{30} = 1097$$

- ***s*** **as Point Estimator of σ**

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{29}} = \sqrt{\frac{163,996}{29}} = 75.2$$

- **$\bar{p}$ as Point Estimator of *p -- the proportion of "Yes"***

$$\bar{p} = 20/30 = .68$$

Note: **Different samples would result in different point estimates.**

Point Estimation

Once all the data for the 900 applicants were entered in the college's database, the values of the population parameters of interest were calculated.

- **Population Mean SAT Score**

$$\mu = \frac{\sum x_i}{900} = 1090$$

- **Population Standard Deviation for SAT Score**

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{900}} = 80$$

- **Population Proportion Wanting On-Campus Housing**

$$p = \frac{648}{900} = .72$$

Summary of Point Estimates Obtained from a Simple Random Sample

Population Parameter	Parameter Value	Point Estimator	Point Estimate
μ = Population mean SAT score	1090	$\overline{x}$ = Sample mean SAT score	1097
σ = Population std. deviation for SAT score	80	s = Sample std. deviation for SAT score	75.2
p = Population pro- portion wanting campus housing	0.72	$\overline{p}$ = Sample pro- portion wanting campus housing	0.68

__Note:__ **Different samples will result in different values of estimate. Those estimators are random variables!**

Sampling Distributions

- **Parameters** are numerical descriptive measures for populations.
- **Statistics** are numerical descriptive measures calculated from the samples. ($\overline{x}$, s, $\overline{p}$, etc.)

In point estimation we use the data from the sample to compute a value of a sample statistic that serves as an estimate of a population parameter.

Example:

A simple random sample of 8 employees of a corporation provided the following information.

Employee :	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Age:	**25**	**36**	**28**	**29**	**26**	**34**	**43**	**39**
Gender:	**M**	**M**	**M**	**M**	**F**	**M**	**F**	**M**

a. Determine the point estimate for the average age of all emloyees.
b. What is the point estimate for the standard deviation of the population?
c. Determine a point estimate for the proportion of all employees who are female.

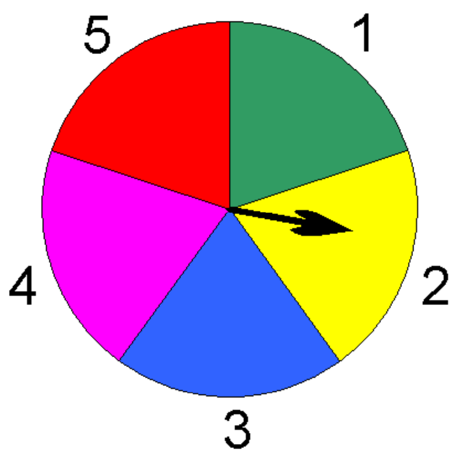
$\overline{X}$ is a random variable !

- $\overline{X}$'s vary from sample to sample and hence are random variables, which have the probability distributions called sampling distributions.

- Repeated sampling for many times enable us to observe what values of $\overline{X}$ can occur and how often each value occurs.

Example

Consider a population that consists of the numbers 1, 2, 3, 4 and 5 generated in a manner that the probability of each of those values is 0.2 no matter what the previous selections were. This population could be described as the outcome associated with a spinner such as given below with the distribution next to it.



x	p(x)
1	0.2
2	0.2
3	0.2
4	0.2
5	0.2

μ=?

Example

For n=2, we repeat doing sampling 25 times and analyze the distribution for those sample means:

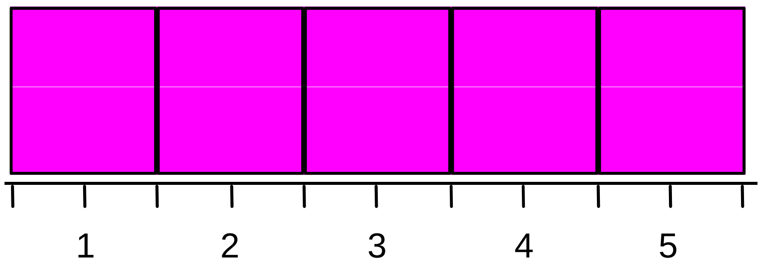
Sample	$\bar{X}$
1, 1	1
1, 2	1.5
1, 3	2
1, 4	2.5
1, 5	3
2, 1	1.5
2, 2	2
2, 3	2.5
2, 4	3
2, 5	3.5
3, 1	2
3, 2	2.5
3. 3	3

Sample	$\bar{X}$
3, 4	3.5
3, 5	4
4, 1	2.5
4, 2	3
4, 3	3.5
4, 4	4
4, 5	4.5
5, 1	3
5, 2	3.5
5, 3	4
5, 4	4.5
5, 5	5

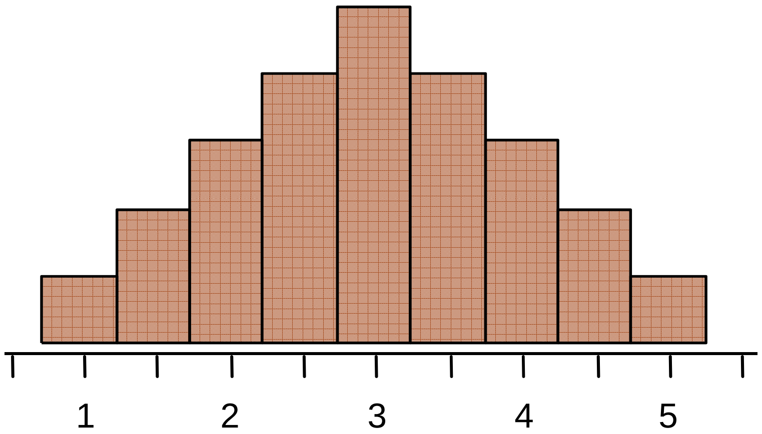
	frequency	p(x)
1	1	0.04
1.5	2	0.08
2	3	0.12
2.5	4	0.16
3	5	0.20
3.5	4	0.16
4	3	0.12
4.5	2	0.08
5	1	0.04
	25	

Example

The original distribution and the sampling distribution of means of samples with n=2 are given below.



Original distribution

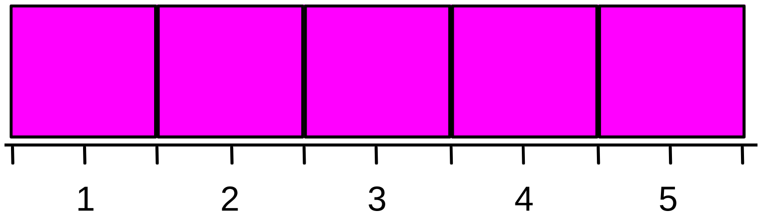


Sampling distribution

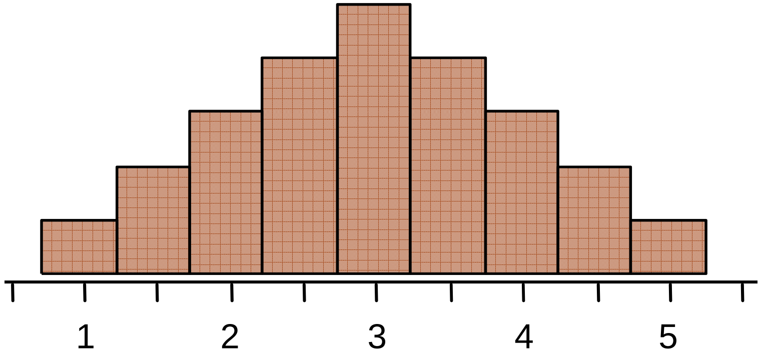
n = 2

Example

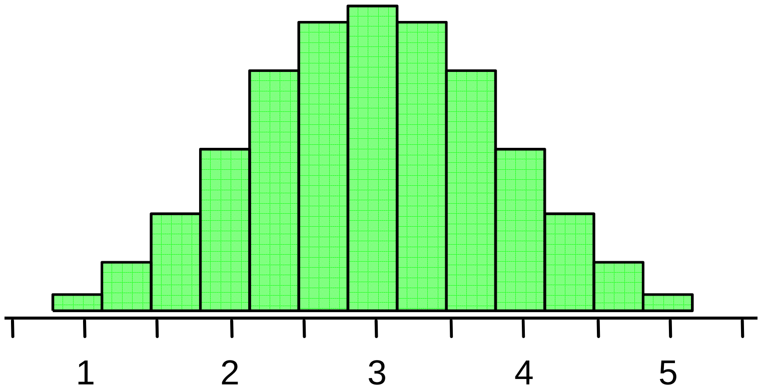
Sampling distributions of sample means for n=3 and n=4 were calculated and are illustrated below. The shape is getting closer and closer to the normal distribution.



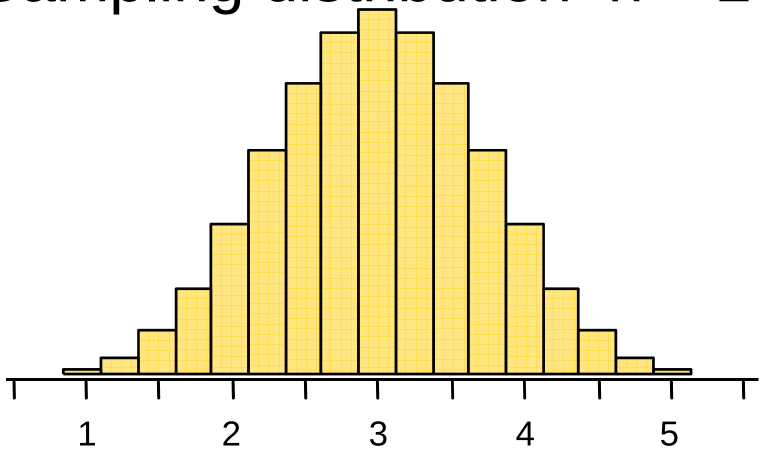
Original distribution



Sampling distribution n = 2

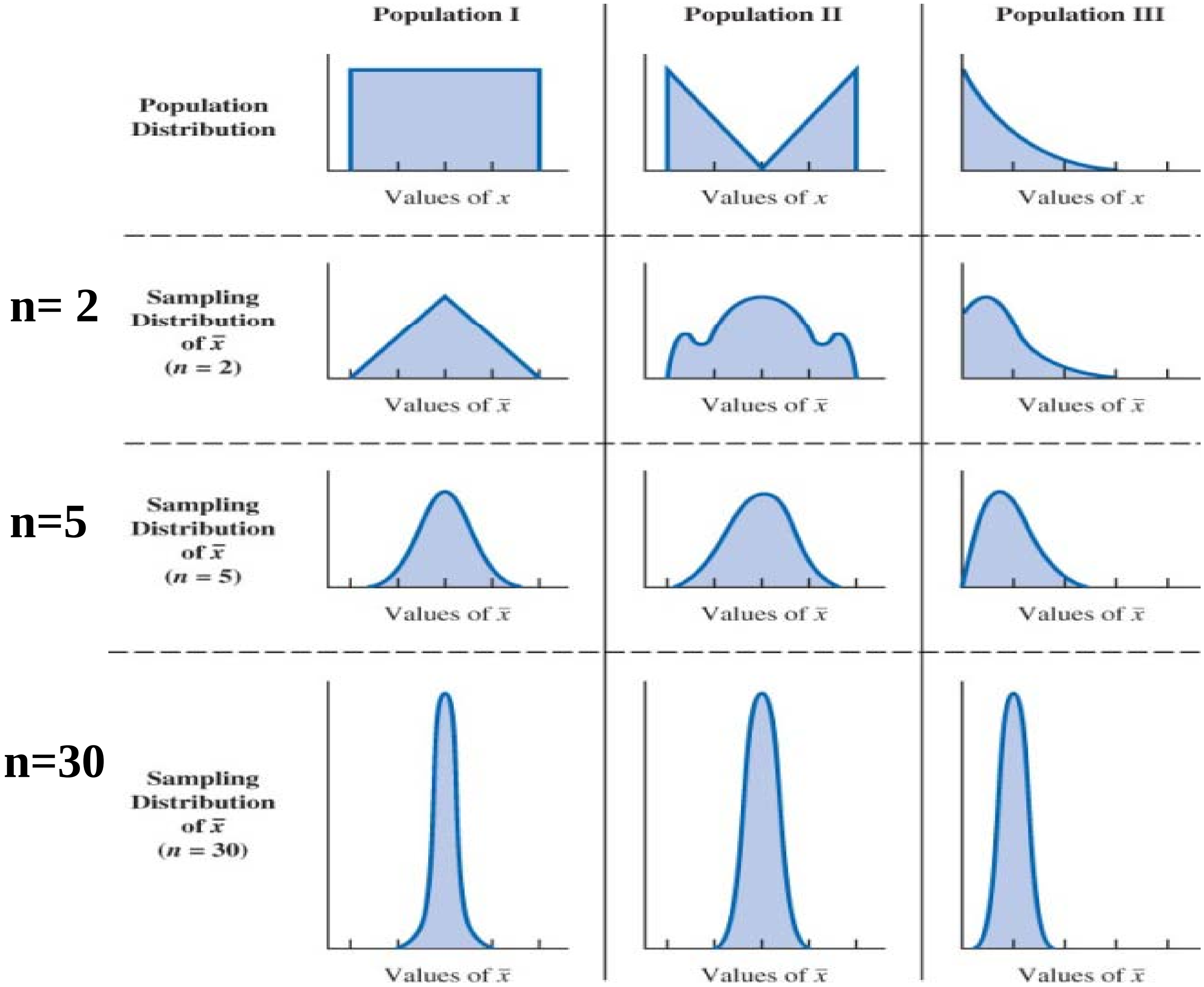


Sampling distribution n = 3



Sampling distribution n = 4

Illustration of Central Limit Theorem for three population



7.5 Sampling Distribution of $\overline{X}$

$$E(\overline{X}) = \mu$$

$$\sigma_{\overline{x}} = \sigma / \sqrt{n}$$

where μ is the population mean

σ is the standard deviation

n is the sample size.

To further emphasize the difference between σ and $\sigma_{\overline{x}}$ we refer to $\sigma_{\overline{x}}$ as the standard error of the mean.

Central Limit Theorem (CLT) :

Central Limit Theorem: The sampling distribution of $\overline{X}$ can be approximated by a normal distribution with mean μ and SD $\sigma / \sqrt{n}$ as n becomes large.

Corollary: CLT also implies that the sum of *n* sample values (Σxi) is approximately normal with mean $n\mu$ and SD $\sigma\sqrt{n}$.

- The sampling distribution of $\overline{X}$ can be used to provide *probability information* about how close the sample mean is to the population mean μ.

How large is the sample size ?

1. When samples of size *n* are drawn from a normal population, the sample mean $\overline{X}$ has a normal distribution with mean μ and standard deviation $\sigma / \sqrt{n}$ for any n!

2. When samples of size *n* are drawn from a nonnormal population, CLT ensures that the sample mean $\overline{X}$ will have an approximately normal distribution with mean μ and standard deviation $\sigma / \sqrt{n}$ when *n* is large ($n \geq 30$).

Finding Probabilities for the Sample Mean

If the sampling distribution of $\bar{x}$ is normal or approximately normal, ***standardize*** *it by*

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

Find the appropriate area using Table 1.

Example: A random sample of size $n = 16$ from a normal distribution with $\mu = 10$ and $\sigma = 8$.

$$P(\bar{x} > 12) = P\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < \frac{12 - 10}{8 / \sqrt{16}}\right)$$

$$= P(z > 1) = .5 - .3413 = .1587$$

Example 1

A soda filling machine is supposed to fill cans of soda with 12 fluid ounces. Suppose that the fills are actually normally distributed with a mean of 12.1 oz and a standard deviation of 0.2 oz.

(1) The probability of one can less than 12 oz is

$$P(X < 12) = P(\frac{X-\mu}{\sigma} < \frac{12-12.1}{.2}) = P(Z < -.5) = 0.3085$$

(2) What is the probability that the average fill for a 6-pack of soda is less than 12 oz?

$$P(\bar{x} < 12) =$$

$$P(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < \frac{12-12.1}{.2/\sqrt{6}}) =$$

$$P(z < -1.22) = .1112$$

What is the sample size? Is it large enough to guarantee the normal distribution of $\bar{X}$? Why?

- Example 2

A population has mean of 200 and a standard deviation of 50. Suppose a simple random sample of size n is selected and $\overline{x}$ is used to estimate μ.

(1) n=100, What is the probability that the sample mean will be in ± 5 of the population mean? (2) n=400?

Note:

1. The standard error decreases as n becomes large.
2. A larger sample size will provide a higher probability that the sample mean falls within a specified distance of the population mean.

7.6 Sampling Distribution of $\bar{p}$

$$\bar{p} = \frac{x}{n}$$

$x =$ **# of elements in the sample that possess the characteristic of interest**

$n =$ **# of sample size**

Expected Value of $\bar{p}$ **:**

$$E(\bar{p}) = p$$

Standard Deviation of $\bar{p}$ **:**

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

where p = the population proportion

- $\sigma_{\bar{p}}$ is referred to as the standard error of the proportion.

Form of the Sampling Distribution of $\bar{p}$

The sampling distribution of $\bar{p}$ can be approximated by a **normal distribution** whenever the sample size is large.

The sample size is considered large whenever these two conditions are satisfied:

$np \geq 5$ and $n(1 - p) \geq 5$

Form of the Sampling Distribution of $\bar{p}$

For values of p near 0.5, sample sizes as small as 10 permit a normal approximation.

With very small (approaching 0) or very large (approaching 1) values of p, much larger sample size n is needed.

In practice, the sample size n is almost always large enough to permit the use of normal approximation.

Finding Probabilities for $\overline{p}$

✓If the sampling distribution of $\overline{p}$ is approximately normal, *standardize* $\overline{p}$ *by*

$$Z = \frac{\overline{p} - p}{\sigma_{\overline{p}}} = \frac{\overline{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

✓Find the appropriate area using Table 1.

Example

In the previous soda filling machine example, it is claimed that only 5% of the soda cans are underfilled. A quality control technician randomly samples 200 cans of soda. What is the probability that more than 10% of the cans are underfilled?

***n* = 200**

S: underfilled can

***p* = P(S) = .05**

***1-p* = .95**

***np* = 10 *n(1-p)* = 190**

OK to use the normal approximation

$$P(\bar{p} > .10)$$

$$= P\left(z > \frac{.10 - .05}{\sqrt{\frac{.05(1-.05)}{200}}}\right) = P(z > 3.24)$$

$$< 1 - 0.999 = 0.001$$

This would be very unusual, if indeed $p = .05$!

Example

Suppose 3% of the people contacted by phone are receptive to a certain sales pitch and buy your product. If your sales staff contacts 2000 people, what is the probability that more than 100 of the people contacted will purchase your product?

n=2000, p= 0.03, np=60, n(1-p)=1940,

OK to use the normal approximation

$$P(\overline{p} > 100/2000) = P\left(z > \frac{.05 - .03}{\sqrt{\frac{.03(.97)}{2000}}}\right) = P(z > 5.24) \approx 0$$

Key Concepts

I. **Simple random sampling**: Each possible sample is equally likely to occur.

II. **Statistics and Sampling Distributions**

1. Sampling distributions describe the possible values of a statistic and how often they occur when repeating sampling.
2. The **Central Limit Theorem** states that the sample mean can be approximated by a normal distributions for large sample size *n*.

Key Concepts

III. Sampling Distribution of the Sample Mean

1. When samples of size *n* are drawn from a normal population with mean μ and variance σ^2, the sample mean $\bar{x}$ has a normal distribution with mean μ and variance σ^2/n.

2. When samples of size *n* are drawn from a nonnormal population with mean μ and variance σ^2, the Central Limit Theorem ensures that the sample mean $\bar{x}$ will have an approximately normal distribution with mean μ and variance σ^2/n when *n* is large ($n \geq 30$).

3. Probabilities involving the sample mean μ can be calculated by standardizing the value of $\bar{x}$ using

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

Key Concepts

IV. Sampling Distribution of the Sample Proportion

1. When samples of size *n* are drawn from a binomial population with parameter *p*, the sample proportion $\bar{p}$ will have an approximately normal distribution with mean *p* and variance *p(1-p)*/*n* as long as *np* > 5 and *n(1-p)* > 5.

2. Probabilities involving the sample proportion can be calculated by standardizing the value $\bar{p}$ using

$$Z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Exercise: A population of 1,000 students spends an average of $10 a day on dinner. The standard deviation of the expenditure is $3. A simple random sample of 64 students is taken.

- a.What are the expected value, standard deviation, and shape of the sampling distribution of the sample mean?
- b.What is the probability that these 64 students will spend a combined total of more than $700?
- c.What is the probability that these 64 students will spend a combined total between $650 and $700?

Exercise: There 500 employees in a firm, 45% are female. A sample of 60 employees is selected randomly.

- a. Determine the standard error of the proportion.
- b. What is the probability that the sample proportion of females is between 0.40 and 0.55?