Ch7. Sampling Distributions

In the past several chapters, we studied *populations* and the *parameters* that describe them. These populations were either discrete or continuous, and we used *probability* as a tool for determining how likely certain sample outcomes might be. In this chapter, our focus changes as we begin to study *samples* and the *statistics* that describe them. These sample statistics are used to make inferences about the corresponding population parameters. This chapter involves sampling and sampling distributions, which describe the behavior of sample statistics in repeated sampling.

윤 연 옥

Introduction

- **Parameters** are numerical descriptive measures for populations.
  - For the normal distribution, the location and shape are described by μ and σ.
  - For a binomial distribution consisting of n trials, the location and shape are determined by p.
- Often the values of parameters that specify the exact form of a distribution are unknown.
- You must rely on the sample to learn about these parameters.

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.
- An agronomist believes that the yield per acre of a variety of wheat is approximately normally distributed, but the mean μ and the standard deviation σ of the yields 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.

Example

- suppose you want to select a sample of size n=2 from a population containing N= 4 objects.
- If the four objects are x_1, x_2, x_3, x_4, there are six distinct pairs could be selected.
- If the sample of n=2 observations is selected so that each of these six samples has the same chance of selection, given by 1/6

: resulting sample is called a simple random sample

(or a random sample.)

Sample	Observations in Sample
1	x_1, x_2
2	x_1, x_3
3	x_1, x_4
4	x_2, x_3
5	x_2, x_4
6	x_3, x_4

How to select random sample? Use random number(Table10 of Appendix I)

Example)

A computer database at a downtown law firm contains files for N=1000 clients.

The firm wants to select n=5 files for review.

Select a simple random sample of five files from this database.

Sol) Select first three digits from Table 10 of Appendix I

The random number 001 corresponds to file #1,

and the last file, #1000 corresponds to the random number 000.

For example, select files corresponding to #816, #309, #763, #78, #61

Table 10 Random Numbers

	Column										
Line	1	2	3	4	5	6	7	8	9	10	11
1	10480	15011	01536	02011	81647	91646	69179	14194	62590	36207	20969
2	22368	46573	25595	85393	30995	89198	27982	53402	93965	34095	52666
3	24130	48360	22527	97265	76393	64809	15179	24830	49340	32081	30680
4	42167	93093	06243	61680	07856	16376	39440	53537	71341	57004	00849
5	37570	39975	81837	16656	06121	91782	60468	81305	49684	60672	14110

Types of Samples

1. Observational studies: The data existed before you decided to study it.

✓ *Usually occur in sample survey*

❖ Frequently occurring problems

- ✓ **Nonresponse:** Are the responses biased because only opinionated people responded?
- ✓ **Undercoverage:** Are certain segments of the population systematically excluded? Ex) telephone survey
- ✓ **Wording bias:** The question may be too complicated or poorly worded.

2. Experimentation: The data are generated by imposing an experimental condition or treatment on the experimental units.

- ✓ **Hypothetical populations** can make random sampling difficult.
- ✓ Samples must sometimes be chosen so that the experimenter believes they are **representative** of the whole population.
- ✓ Selecting a simple random sample is more difficult.

Other Sampling Methods

- There are several other sampling plans that still involve **randomization**:

1. **Stratified random sample:** Divide the population into subpopulations or **strata** and select a simple random sample from each strata.

2. **Cluster sample:** Divide the population into subgroups called **clusters;** select a simple random sample of clusters and take a census of every element in the cluster.

3. **1-in-k systematic sample**: Randomly select one of the first k elements in an ordered population, and then select every k-th element thereafter.

◆ Example

- Divide California into counties and take a simple random sample within each county. : stratified
- Divide a city into city blocks, choose a simple random sample of 10 city blocks, and interview all who live there. : cluster
- Choose an entry at random from the phone book, and select every 50th number thereafter. : 1-in-50 systematic

Non-Random Sampling Methods

- There are several other sampling plans that do not involve **randomization**.
- They should **NOT** be used for statistical inference!

1. **Convenience sample:** A sample that can be taken easily without random selection.
   ex) People walking by on the street
2. **Judgment sample:** The sampler decides who will and won't be included in the sample.
   ex) teacher select a student who represent the class
3. **Quota sample**: The makeup of the sample must reflect the makeup of the population on some selected characteristic.
   ex) Race, ethnic origin, gender, etc.

Sampling Distributions

- Numerical descriptive measures calculated from the sample are called **statistics**.
- Statistics vary from sample to sample and hence are random variables.
- The probability distributions for statistics are called **sampling distributions**.
- In repeated sampling, they tell us what values of the statistics can occur and how often each value occurs.

Definition: The **sampling distribution of a statistic** is the probability distribution for the possible values of the statistic that results when random samples of size n are repeatedly drawn from the population.

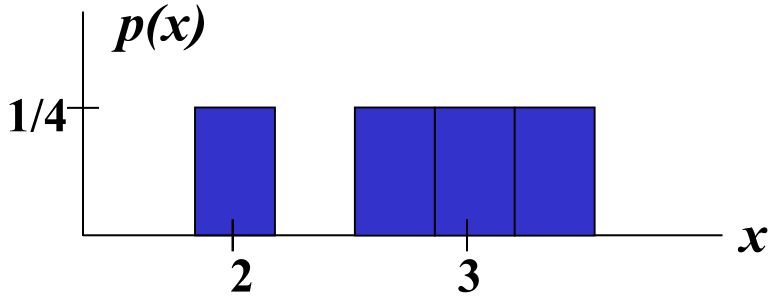
Example

Population: 3, 5, 2, 1

Draw samples of size $n = 3$ without replacement

Possible 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 $\bar{x}$ is equally likely, with probability 1/4

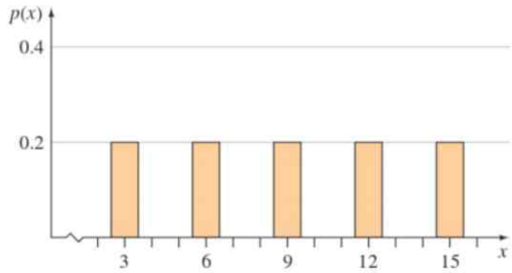


Example

- A population consists of N=5 numbers: 3, 6, 9, 12, 15. If a random sample of size n=3 is selected without replacement, find the sampling distributions for the sample mean $\bar{x}$ and the sample median m.

Sol) Population={3, 6, 9, 12, 15}

$$\mu = \frac{3+6+9+12+15}{5} = 9$$ and median M=9



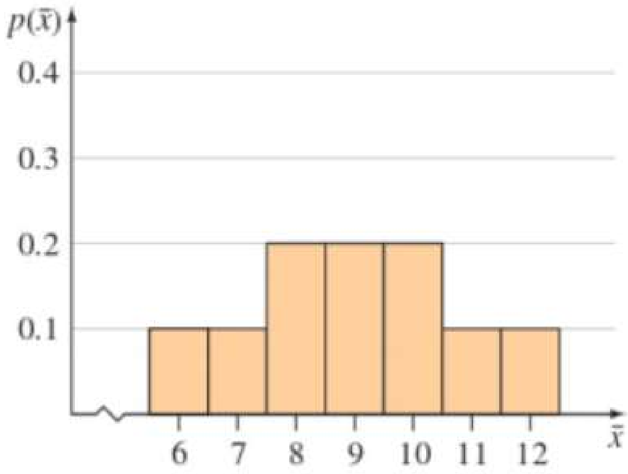
- $\binom{5}{3} = 10$ possible random samples of size n=3 and each is equally likely, with probability 1/10.

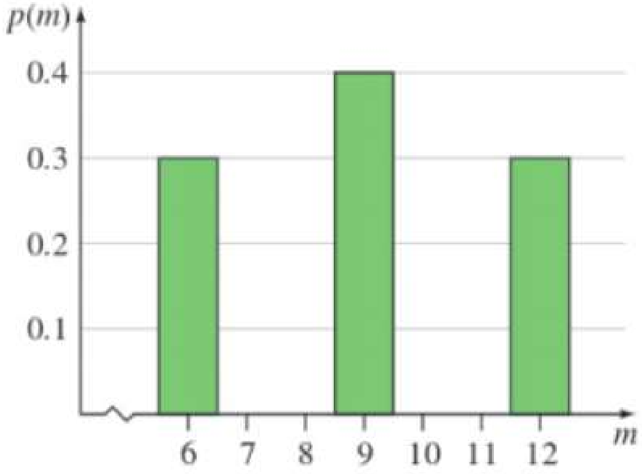
Sample	Sample Values	$\bar{x}$	m
1	3, 6, 9	6	6
2	3, 6, 12	7	6
3	3, 6, 15	8	6
4	3, 9, 12	8	9
5	3, 9, 15	9	9
6	3, 12, 15	10	12
7	6, 9, 12	9	9
8	6, 9, 15	10	9
9	6, 12, 15	11	12
10	9, 12, 15	12	12

Sampling Distributions for (a) the Sample Mean and (b) the Sample Median

(a)	$\bar{x}$	$p(\bar{x})$	(b)	m	$p(m)$
	6	.1		6	.3
	7	.1		9	.4
	8	.2		12	.3
	9	.2			
	10	.2			
	11	.1			
	12	.1			

- Probability histograms for the sampling distributions of the sample mean, $\bar{x}$, and the sample median, m





Sampling Distributions

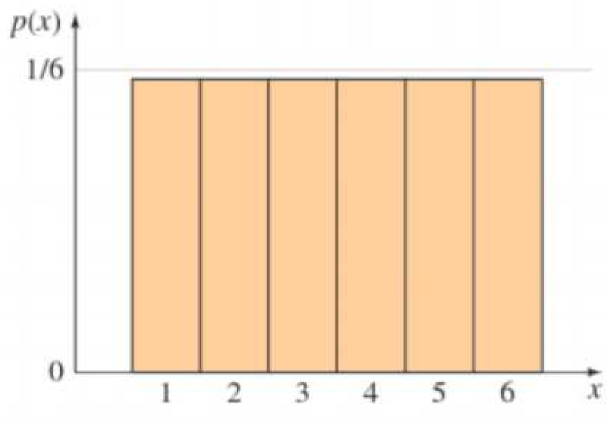
- Sampling distributions for statistics can be
  - ✓ Approximated with simulation techniques
  - ✓ Derived using mathematical theorems
    - The Central Limit Theorem is one such theorem.

Central Limit Theorem:
If random samples of n observations are drawn from a nonnormal population with finite μ and standard deviation σ , then, when n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normally distributed, with mean μ and standard deviation $\sigma/\sqrt{n}$. The approximation becomes more accurate as n becomes large.

Example

(1) Toss a fair die $n = 1$ time. The distribution of x the number on the upper face is flat or **uniform.**

$$\mu = \sum xp(x)$$
$$= 1(\frac{1}{6}) + 2(\frac{1}{6}) + \ldots + 6(\frac{1}{6}) = 3.5$$
$$\sigma = \sqrt{\sum (x-\mu)^2 p(x)} = 1.71$$



(2) Toss a fair die $n = 2$ times. The distribution of x the average number on the two upper faces is **mound-shaped.**

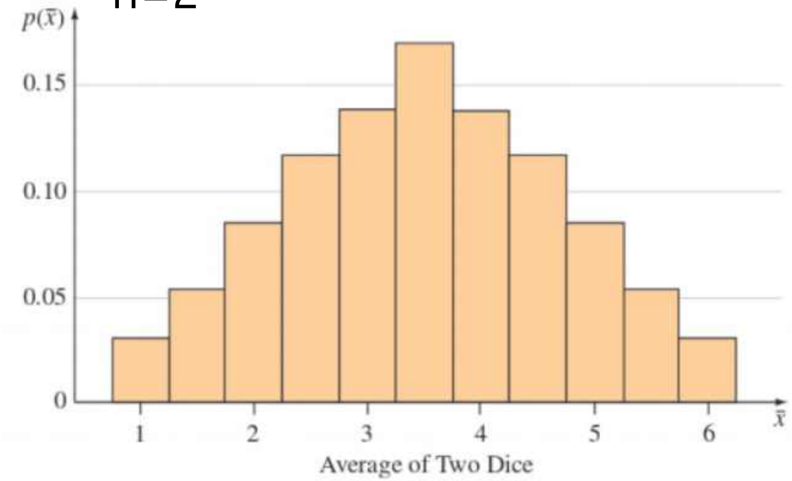
Sums of the Upper Faces of Two Dice

	First Die					
Second Die	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Sampling distribution of $\bar{x}$

$\bar{x}$	$p(\bar{x})$	$\bar{x}$	$p(\bar{x})$
2/2 = 1	1/36	8/2 = 4	5/36
3/2 = 1.5	2/36	9/2 = 4.5	4/36
4/2 = 2	3/36	10/2 = 5	3/36
5/2 = 2.5	4/36	11/2 = 5.5	2/36
6/2 = 3	5/36	12/~~6~~ = 6	1/36
7/2 = 3.5	6/36		

n=2

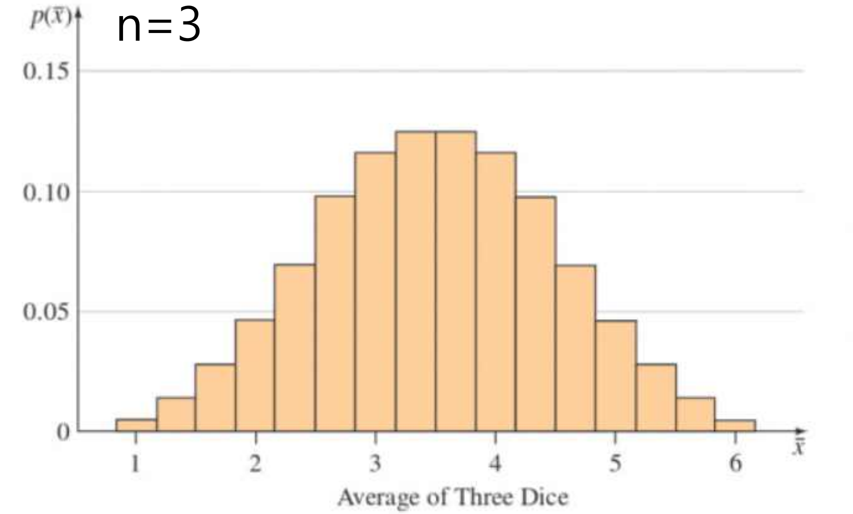


Mean : $\mu = 3.5$

Std Dev :

$\sigma/\sqrt{2} = 1.71/\sqrt{2} = 1.21$

n=3



Mean : $\mu = 3.5$

Std Dev :

$\sigma/\sqrt{3} = 1.71/\sqrt{3} = .987$

As n gets larger, the shape of $\bar{x}$ distribution is normally distributed

Why is this important?

- The **Central Limit Theorem** also implies that the sum of n measurements($\sum x_i$) is approximately normal with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.
- Many statistics that are used for statistical inference are **sums** or **averages** of sample measurements.
- When n is large, these statistics will have approximately **normal** distributions.
- This will allow us to describe their behavior and evaluate the **reliability** of our inferences.

How large is large?

- If the population is **normal**, then the sampling distribution of $\bar{x}$ will also be normal, no matter what the sample size
- When the sample population is approximately **symmetric**, the distribution becomes approximately normal for relatively small values of n.
- When the sample population is **skewed**, the sample size must be **at least 30** before the sampling distribution of $\bar{x}$ becomes approximately normal.

The Sampling Distribution of the Sample Mean

- A random sample of size n is selected from a population with mean μ and standard deviation σ.
- The sampling distribution of the sample mean $\bar{x}$ will have mean μ and standard deviation $\sigma/\sqrt{n}$.
- If the original population is **normal**, the sampling distribution will be normal for any sample size.
- If the original population is **nonnormal**, the sampling distribution will be normal when n is large.

✓ **The standard deviation of $\bar{x}$ is referred to as the standard error of the mean(SE($\bar{x}$)).**

Finding Probabilities for the Sample Mean

- If the sampling distribution of $\bar{x}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of

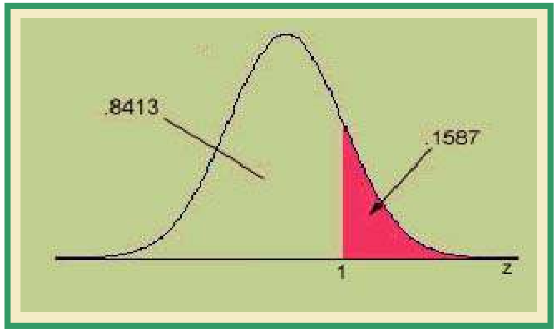
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- Find the appropriate area using Normal table (Table 3 in Appendix).

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

$$P(\bar{x} > 12) = P\left(z > \frac{12-10}{\frac{8}{\sqrt{16}}}\right)$$

$$= P(z > 1) = 1 - .8413 = .1587$$

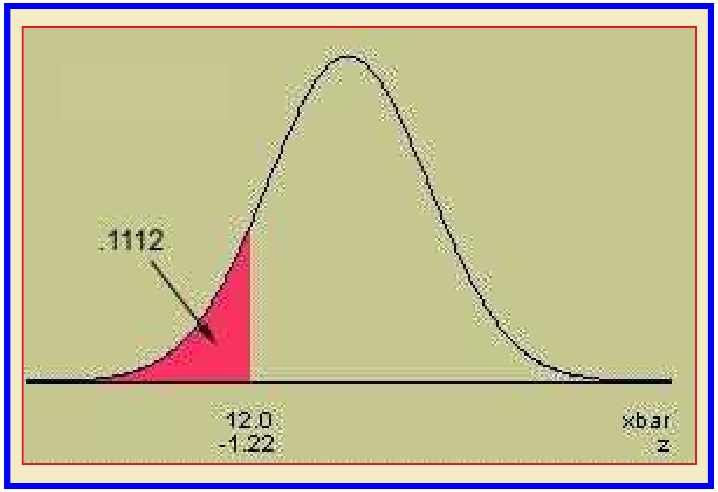


Example

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

$$P(\bar{x} < 12) = P\left(z < \frac{12 - 12.1}{.2/\sqrt{6}}\right)$$

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



The Sampling Distribution of the Sample Proportion

- The **Central Limit Theorem** can be used to conclude that the binomial random variable *x* is approximately normal when *n* is large, with mean ***np*** and standard deviation $\sqrt{npq}$.

- The sample proportion, $\hat{p} = \frac{x}{n}$ is simply a *rescaling* of the binomial random variable *x*, dividing it by *n*.

- From the Central Limit Theorem, the sampling distribution of $\hat{p}$ will also be **approximately normal,** with a *rescaled* mean and standard deviation.

The Sampling Distribution of the **Sample Proportion**

- A random sample of size ***n*** is selected from a binomial population with parameter ***p***.

- The sampling distribution of the sample proportion, $\hat{p} = \frac{x}{n}$

  will have mean ***p*** and standard deviation $\sqrt{\frac{pq}{n}}$

- If *n* is large, and *p* is not too close to zero or one, the sampling distribution of $\hat{p}$ will be **approximately normal.**

✓ **The standard deviation of $\hat{p}$ is referred as the Standard Error (SE) of $\hat{p}$.**

Finding Probabilities for the Sample Proportion

- If the sampling distribution of $\hat{p}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

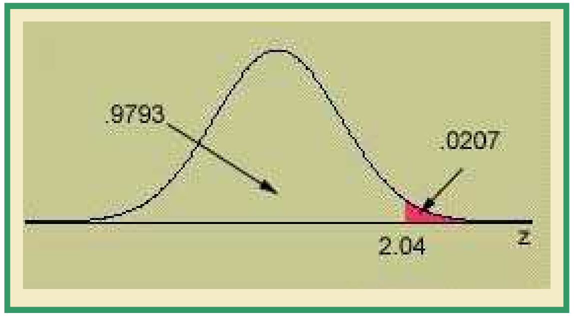
- Find the appropriate area using Normal Table(Table3 in Appendix).

✓ Check np> 5 and nq> 5 for normal approximation!!

Example: A random sample of size n = 100 from a binomial population with p = .4.

$$P(\hat{p} > .5) \approx P\left(z > \frac{.5 - .4}{\sqrt{\frac{.4(.6)}{100}}}\right)$$

$$= P(z > 2.04) = 1 - .9793 = .0207$$



Example

The soda bottler in the previous example claims 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?

Sol) $n = 200$, Success: underfilled can, $p = P(S) = .05$, $q = .95$

$np = 10$ $nq = 190$: Ok to use the normal approximation.

$$P(\hat{p} > .10) \approx P\left(z > \frac{.10 - .05}{\sqrt{\frac{.05(.95)}{200}}}\right) = P(z > 3.24) = 1 - .9994 = .0006$$

✓ *This would be very unusual, if indeed p=.05*

Key Concepts

I. Sampling Plans and Experimental Designs

1. Simple random sampling
   a. Each possible sample is equally likely to occur.
   b. Use a computer or a table of random numbers.
   c. Problems are nonresponse, undercoverage, and wording bias.
2. Other sampling plans involving randomization
   a. Stratified random sampling
   b. Cluster sampling
   c. Systematic 1-in-k sampling

3. Nonrandom sampling

a. Convenience sampling

b. Judgment sampling

c. Quota sampling

II. Statistics and Sampling Distributions

1. Sampling distributions describe the possible values of a statistic and how often they occur in repeated sampling.
2. Sampling distributions can be derived mathematically, approximated empirically, or found using statistical theorems.
3. The **Central Limit Theorem** states that sums and averages of measurements from a nonnormal population with finite mean μ and standard deviation σ have approximately normal distributions for large samples of 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 $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

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 $\hat{p}$ will have an approximately normal distribution with mean p and variance pq/n as long as $np > 5$ and $nq > 5$.

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

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$