Chapter 5 Sampling Distributions

5.1-3 The concept and properties of sampling distribution, and CLT for the means

- Establish that a sample statistic is a random variable with a probability distribution
- Define a *sampling distribution* as the probability distribution of a sample statistic
- Give two important properties of sampling distributions
- Learn that the sampling distribution of both the sample mean and sample proportion tends to be approximately normal

Recall:
A **parameter** is a numerical descriptive measure of a population. Because it is based on all the observations in the population, its value is almost always unknown.
A **sample statistic** is a numerical descriptive measure of a sample. It is calculated from the observations in the sample.

Common Statistics & Parameters

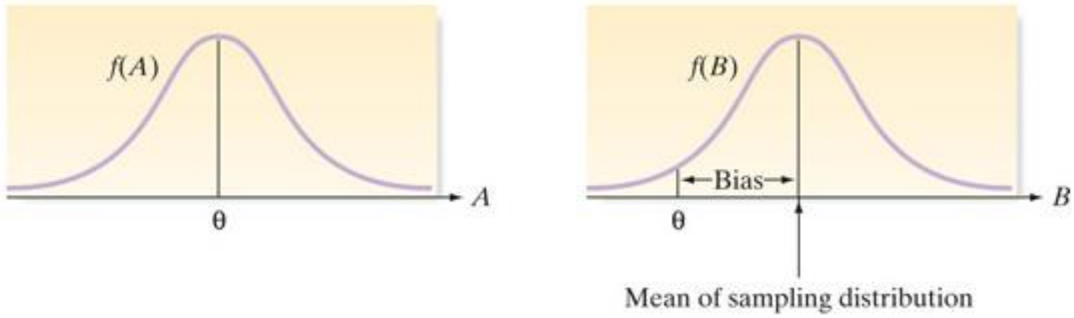
	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Standard Deviation	s	σ
Variance	s^2	σ^2
Binomial Proportion	$\hat{p}$	p

A **point estimator** of a population parameter is a rule or formula that tells us how to use the sample data to calculate a single number that can be used as an *estimate of* the population parameter.

The concept of sampling distribution: suppose, many different samples size n are selected from given population. In each sample a statistic (like sample mean, sample proportion or variance) was calculated (which itself is random variable, because its values vary for different samples).
Probability distribution of such sample statistics is called a sampling distribution of those statistics.

If the sampling distribution of a sample statistic has a mean **equal to the population parameter** the statistic is intended to estimate, the statistic is said to be an **unbiased estimate of the parameter.**

If the mean of sampling distribution is **not equal to the parameter**, the statistic is said to be a **biased estimate** of the parameter.



a. Unbiased sample statistic for the parameter θ

b. Biased sample statistic for the parameter θ

Example of sampling distribution: distribution of the mean numbers of dots rolled in repeated tosses of n dice

http://www.albany.edu/~jr853689/CentralLimitTheoremForDice.htm

or

https://en.wikipedia.org/wiki/File:Dice_sum_central_limit_theorem.svg

Sample of the size n=1 (one die toss)

n=1 outcomes: 1, 2, 3, 4, 5, 6; P(x)=1/6; uniform distribution, and **μ=…3.5**

Variance: $((1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(4-3.5)^2+(5-3.5)^2+(6-3.5)^2)/6=2.916667$

Standard deviation: 1.71

n=2 outcomes: 2/2, 3/2, 4/2, …, 11/2, 12/2; triangular distribution, and μ=…3.5

Standard deviation: …(=1.21)

n=5 outcomes: 5/5, 6/5, …, 30/5, ……(shape?)…….. distribution; μ=…3.5

…

Sample size 1: the rolls have **uniform** distribution. Population mean is the same, or almost the same as the average of all sample means when we repeat rolling a die many times.

Sample size 2: the rolls of two dice have **triangular** distribution if we repeat tossing two dice many times.

The more dice in a toss (the bigger n), the more normal distribution of sample means becomes.

Conclusion: as sample size increases, distribution of sample means (called sampling distribution) approaches NORMAL distribution.

Note: it is not important what shape the distribution of one single object is.

We don't care for the shape of the original population. All we need is r**andomization in selecting samples, and large enough samples**.

Goal: We use sampling distribution of a statistic to estimate the value of an unknown population parameter.

The Central Limit Theorem for sampling distribution of sample means

This fact was proven in 1810 by Pierre-Simeon La Place.

Consider a random sample of *n* observations selected from a population (***any* probability distribution)** with mean μ and standard deviation σ. Then, when *n* is sufficiently large, the sampling distribution of sample means x-bar will be **approximately normal:**

1. Mean of the sampling distribution equals mean of sampled population, that is,

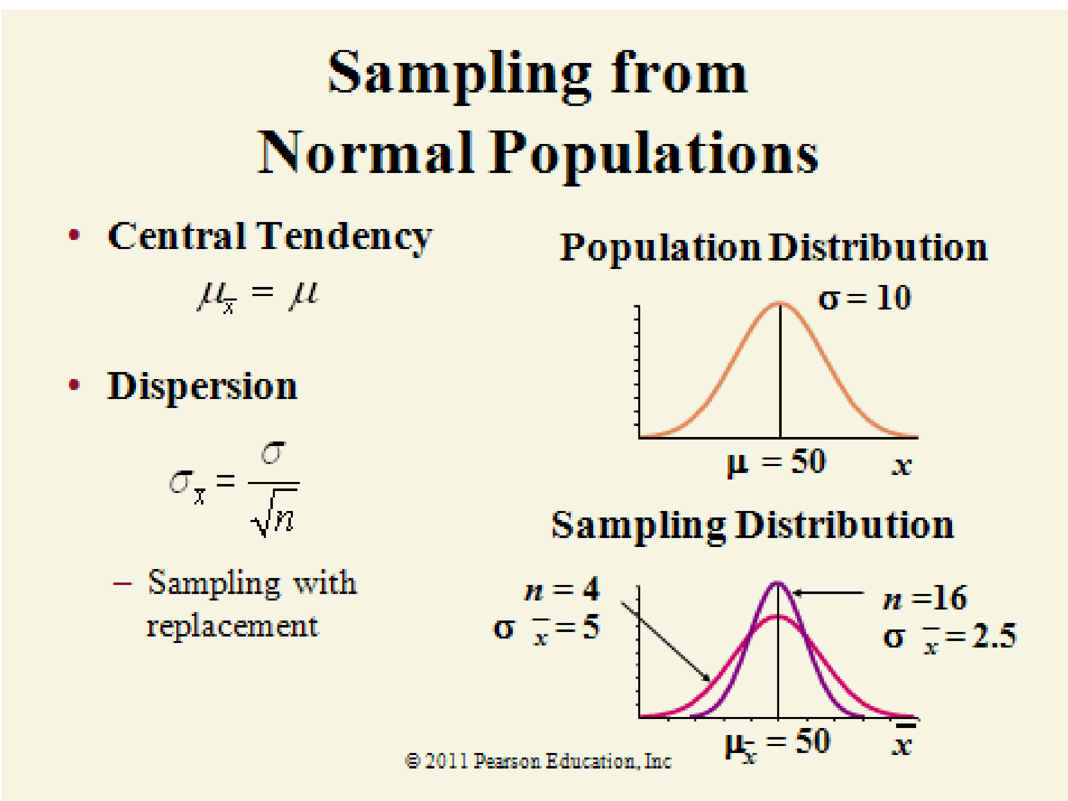
$$\mu_{\bar{x}} = E\left(\bar{x}\right) = \mu.$$

2. Standard deviation of the sampling distribution equals

$$\frac{\text{Standard deviation of sampled population}}{\text{Square root of sample size}} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

The bigger sample size (n), the better normal approximation.

The standard deviation $\sigma_{\bar{x}}$ is often referred to as the **standard error of the mean**.



Practical Rules:

1. Large samples: For samples of size *n* **at least 30**
   - the distribution of sample means can be approximated quite well by normal distribution. *(Although we'll learn later that if population standard deviation is unknown, then T-distribution will be more accurate.)*
   - Original shape of distribution is not important.
2. Small samples (size n less than 30):
   - **If the original population is itself normally distributed**, then the sample means are normally distributed for any sample size *n* (even as small as n=2).
   - **If the sample does not seem to come from normal distribution,** then we **cannot use this tool (CLT)**.

Example: You're an operations analyst for AT&T. Long-distance telephone calls are normally distributed with $\boldsymbol{\mu}$ **= 8** min. and $\boldsymbol{\sigma}$ **= 2** min. If you select random samples of **25** calls, what percentage of the **sample mean calls** would be between **7.2** & **8.8** minutes?
Exercises: p. 283 #5.17 a, c; #5.18, 5.19 d, e, 5.20, 5.26

5.17 Suppose a random sample of *n* measurements is selected from a population with mean μ=100 and variance σ^2=100. For each of the following values of *n*, give the mean and standard deviation of the sampling distribution of the sample mean $\bar{x}$.

a. n=4

b. n=25

c. n=100

5.18 A random sample of n=64 observations is drawn from a population with a mean equal to 20 and a standard deviation equal to 16.

a. Give the mean and standard deviation of the (repeated) sampling distribution of $\bar{x}$.

b. Describe the shape of the sampling distribution of $\bar{x}$. Does your answer depend on the sample size?

c. Calculate the standard normal *z*-score corresponding to a value of $\bar{x}$=15.5.

d. Calculate the standard normal *z*-score corresponding to $\bar{x}$=23.

5.19 Refer to Exercise 5.18. Find the probability that

d. $\bar{x}$ falls between 16 and 22.

e. $\bar{x}$ is less than 14.

5.20 A random sample of $n=900$ observations is selected from a population with $\mu=100$ and $\sigma=10$.

a. What are the largest and smallest values of $\bar{x}$ that you would expect to see?

b. How far, at the most, would you expect $\bar{x}$ to deviate from μ?

c. Did you have to know μ to answer part **b**? Explain.

5.28 Requests to a Web server. In Exercise 4.139 (p. 247) you learned that Brighton Webs LTD modeled the arrival time of requests to a Web server within each hour using a uniform distribution. Specifically, the number of seconds *x* from the start of the hour that the request is made is uniformly distributed between 0 and 3,600 seconds. In a random sample of $n=60$ Web server requests, let $\bar{x}$ represent the sample mean number of seconds from the start of the hour that the request is made.

a. Find $E(\bar{x})$ and interpret its value.

b. Find $Var(\bar{x})$.

c. Describe the shape of the sampling distribution of $\bar{x}$.

d. Find the probability that $\bar{x}$ is between 1,700 and 1,900 seconds.

e. Find the probability that $\bar{x}$ exceeds 2,000 seconds.

5.4 The Sampling Distribution of Sample Proportions

Target parameter: p = unknown proportion of elements of some type (say S) in a population (population proportion)

Point estimator: sample proportion $\hat{p}$

$$\hat{p} = \frac{number\ of\ elements\ of\ type\ S\ in\ a\ sample}{sample\ size} = \frac{x}{n}$$

where x = number of elements of type S in a sample, and n = sample size

Notation:

p = population proportion, $q = 1\text{-}p$,

$\hat{p}$ = sample proportion and $\hat{q} = 1 - \hat{p}$

Example: If 313 out of 1250 randomly selected consumers were found to prefer brand X cereal, then the <u>sample proportion</u> of those who prefer brand X in this sample is

$$\hat{p} = \frac{313}{1250} = 0.2504 \quad or \quad 25.04\%$$

This is our point estimate of the proportion of **all** potential cereal consumers who prefer brand X.

The sample proportion is a good estimator of the population proportion.

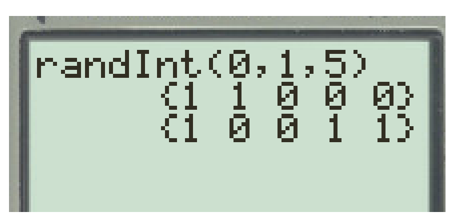
<u>Example</u>: How are proportions of heads in ten-coin-tosses distributed?

We'll toss ten coins (of one coin 10 times). The size of our sample is n=10, and x=observed number of heads. The sample proportion of heads p-hat=x/n.

Find your p-hat: …………..

Next we'll collect the results obtained by each person. We expect …….% of heads. Find the shape and statistics of the distribution of all p-hats collected so far:

(No coins? No problem: simulate on the calculator. Go to MATH, PRB, 5:randInt and type 0,1,5 ENTER. The calculator will produce five randomly selected zero-one digits at a time. Click ENTER again to get the list of ten numbers altogether. Count 1 as heads; summarize the ones and divide the result by 10)



0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
(Our results are collected in this dot—plot)
Discuss the shape, mean and standard deviation of this distribution of the sample proportions.

…………………………………………………..

Recall the distribution of the number of heads in a single toss of ten coins (or tossing a coin 10 times).

The population parameters: the mean and standard deviation for a single experiment is easy to compute, because it is a binomial model: n=10 trials, x=number of heads, p=probability of success is constant (0.5), the trials are independent, and there are only two outcomes per trial. This is a binomial distribution.

The mean, the expected number of heads in a 10 fair coins toss is np=5, and standard deviation is $\sqrt{npq} = \sqrt{10 * .5 * .5} = 1.58$

Distribution of the numbers of heads in 10-coins toss is binomial.
What is the distribution of the proportions of heads?

Sampling Distribution for Proportions

Central Limit Theorem for Proportions:

For large *n* (*np>15* and *nq>15*) the sampling distribution of $\hat{p}$ is approximately normal:

$N\left(p, \sqrt{\dfrac{pq}{n}}\right)$ that is, $\hat{p}$ is normal with mean $E(\hat{p}) = p$ and

standard deviation $SD(\hat{p}) = \sigma(\hat{p}) = \sqrt{\dfrac{pq}{n}}$

Before you use the theorem to solve a problem, always check the conditions.

Exercises: Page 288, #36a, 37a, 40, 46, 48

5.36 Suppose a random sample of *n* measurements is selected from a binomial population with probability of success p=.2. For each of the following values of *n*, give the mean and standard deviation of the sampling distribution of the sample proportion, p^.

a. n=50

5.37 Suppose a random sample of n=500 measurements is selected from a binomial population with probability of success *p.* For each of the following values of *p*, give the mean and standard deviation of the sampling distribution of the sample proportion, p^.

a. p=.1

5.40 A random sample of n=1,500 measurements is drawn from a binomial population with probability of success .4. What are the smallest and largest values of p^ you would expect to see?

5.46 Stock market participation and IQ. Refer to *The Journal of Finance* (December 2011) study of whether the decision to invest in the stock market is dependent on IQ, Exercise 3.42 (p. 153). The researchers found that the probability of a Finnish citizen investing in the stock market differed depending on IQ score. For those with a high IQ score, the probability is .44; for those with an average IQ score, the probability is .26; and for those with a low IQ score, the probability is .14.

a. In a random sample of 500 Finnish citizens with high IQ scores, what is the probability that more than 150 invest in the stock market?

b. In a random sample of 500 Finnish citizens with average IQ scores, what is the probability that more than 150 invest in the stock market?

c. In a random sample of 500 Finnish citizens with low IQ scores, what is the probability that more than 150 invest in the stock market?

5.48 Hotel guest satisfaction. Refer to the results of the 2009 North American Hotel Guest Satisfaction Index Study referenced in Exercise 4.48 (p. 210). Recall that 66% of hotel guests were aware of the hotel's “green” conservation program; of these guests, 72% actually participated in the program by reusing towels and bed linens. In a random sample of 100 hotel guests, find the probability that fewer than 42 were aware of and participated in the hotel's conservation efforts.