Chapter 6.4_6.5
Sampling Distributions and Central Limit Theorem

Ram C Kafle, Ph.D.
Assistant Professor of Statistics
Department of Mathematics and Statistics
Sam Houston State University

Definitions:

A numerical feature of a population is called a **parameter**.

A **statistic** is a numerical valued function of the sample observations.

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

Mean and Variance of Sampling Distribution ($\bar{X}$):

Mean and Standard Deviation of $\overline{X}$

The distribution of the sample mean, based on a random sample of size n, has

$$E(\overline{X}) = \mu \qquad (= \text{Population mean})$$

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \qquad \left(= \frac{\text{Population variance}}{\text{Sample size}}\right)$$

$$\mathrm{sd}(\overline{X}) = \frac{\sigma}{\sqrt{n}} \qquad \left(= \frac{\text{Population standard deviation}}{\sqrt{\text{Sample size}}}\right)$$

MEAN AND STANDARD DEVIATION of $\bar{x}$

Mean of the Sampling Distribution of $\bar{x}$

The ***mean of the sampling distribution of*** $\bar{x}$ is always equal to the mean of the population. Thus,

$$\mu_{\bar{x}} = \mu$$

MEAN AND STANDARD DEVIATION OF $\bar{x}$

Standard Deviation of the Sampling Distribution of $\bar{x}$

The ***standard deviation of the sampling distribution of*** $\bar{x}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population and n is the sample size. This formula is used when $n/N \leq .05$, where N is the population size.

Find the mean and variance of sampling distribution ($\bar{X}$) of housing example

TABLE 4 Mean and Variance of $\bar{X} = (X_1 + X_2)/2$

Population Distribution			
x	$f(x)$	$xf(x)$	$x^2f(x)$
2	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{4}{3}$
3	$\frac{1}{3}$	$\frac{3}{3}$	$\frac{9}{3}$
4	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{16}{3}$
Total	1	3	$\frac{29}{3}$

$$\mu = 3$$

$$\sigma^2 = \frac{29}{3} - (3)^2 = \frac{2}{3}$$

Distribution of $\bar{X} = (X_1 + X_2)/2$			
$\bar{x}$	$f(\bar{x})$	$\bar{x}f(\bar{x})$	$\bar{x}^2f(\bar{x})$
2	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{4}{9}$
2.5	$\frac{2}{9}$	$\frac{5}{9}$	$\frac{12.5}{9}$
3	$\frac{3}{9}$	$\frac{9}{9}$	$\frac{27}{9}$
3.5	$\frac{2}{9}$	$\frac{7}{9}$	$\frac{24.5}{9}$
4	$\frac{1}{9}$	$\frac{4}{9}$	$\frac{16}{9}$
Total	1	3	$\frac{84}{9}$

$$E(\bar{X}) = 3 = \mu$$

$$\text{Var}(\bar{X}) = \frac{84}{9} - (3)^2 = \frac{1}{3}$$

Read Example 11 (Page 315 of your textbook)

Central Limit Theorem

$\overline{X}$ Is Normal When Sampling from a Normal Population

In random sampling from a **normal** population with mean μ and standard deviation σ, the sample mean $\overline{X}$ has the normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

Central Limit Theorem

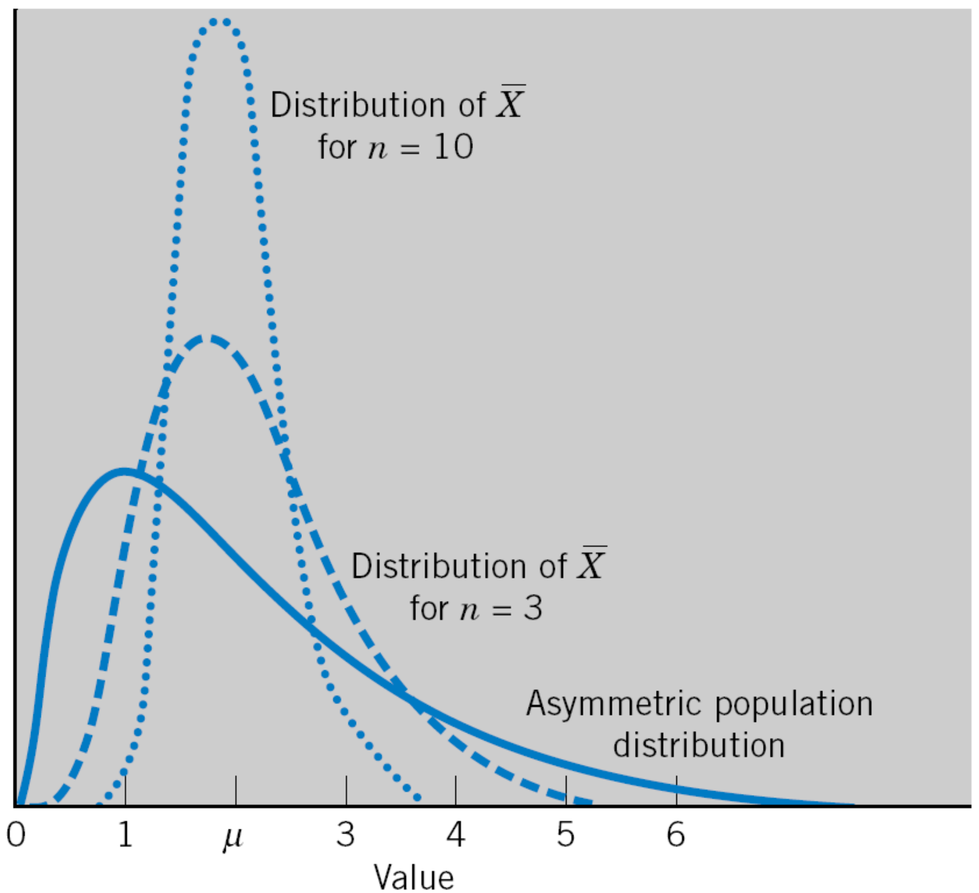
Whatever the population, the distribution of $\overline{X}$ is approximately normal when n is large.

In random sampling from an arbitrary population with mean μ and standard deviation σ, when n is large, the distribution of $\overline{X}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$. Consequently,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{is approximately } N(0, 1)$$

Graphic Example of Central Limit Theorem

The population distribution is represented by solid curve which is a continuous asymmetric distribution with mean mu= 2 and standard deviation sigma = 1.41. Distributions of $\overline{x}$ for n = 3 and n = 10 in sampling from an asymmetric population



Notes:

- If the distribution of population (X) is normal, then regardless of sample size its sampling distribution ($\overline{x}$) is always normal.

- If the sample size is large (n≥30), then regardless of the distribution of the population (X), the distribution of its sampling size ($\overline{x}$) is approximately normal. (CLT)

- If the population distribution (X) is unknown (or not normal), and the sample size is small (n<30), then the sampling distribution is unknown.

1. Determine whether the sampling distribution of $\bar{x}$ is normal, approximately normal, or unknown.
   (a) Heights of American women aged 18 to 24 are distributed with mean 65.5 inches and standard deviation 2.5 inches and the distribution is unknown. A random sample of 40 is selected.

   (b) Average salaries for full-time associate professors in United States doctoral departments of psychology are normally distributed with a mean of $71955 and a standard deviation of $13513. A sample of size 20 is taken.

   (c) A study involving the economic burden of congestive heart failure found that the lengths of hospital stays for patients are not normally distributed with a mean of 7.8 days and a standard deviation of 9.1 days. A random sample of size 27 is taken.

   (d) According to the Census Bureau's 2002 American Community Survey, the average travel time to work of workers 16 years and over living in Boston, MA, who did not work at home, was 28.2 minutes with a standard deviation of 0.79 minutes. A sample of 35 commute times is taken.

Suppose the weights of the contents of cans of mixed nuts have a normal distribution with mean 32.4 ounces and standard deviation .4 ounce.

(a) If every can is labeled 32 ounces, what proportion of the cans have contents that weigh less than the labeled amount?

ANSWER

(b) If two packages are randomly selected, specify the mean, standard deviation, and distribution of the average weight of the contents.

(c) If two packages are randomly selected, what is the probability that the average weight is less than 32 ounces?

ANSWER

SOLUTION

Denote $X = \text{weight of a package}$. We are given that X is normal with mean 32.4 and standard deviation 0.4.

(a) We convert to the standard normal to obtain

$$P[X < 32] = P\left[\frac{X - 32.4}{0.4} < \frac{32 - 32.4}{0.4}\right] = P[Z < -1] = 0.1587.$$

Hence, about 16% of the packages weigh less than the labeled amount.

(b) Let X_1 and X_2 denote the weight of two randomly chosen packages. Observe that:

$$\begin{aligned} E(\overline{X}) &= 32.4 \\ \text{sd}(\overline{X}) &= \tfrac{0.4}{\sqrt{2}} = 0.2828 \end{aligned}$$

Hence, $\overline{X} = \dfrac{X_1 + X_2}{2}$ is normal with mean 32.4 and standard deviation 0.2828.

(c) Again, we convert to the standard normal (using part (b)) to obtain

$$P[\overline{X} < 32] = P\left[\frac{\overline{X} - 32.4}{0.2828} < \frac{32 - 32.4}{0.2828}\right] = P[Z < -1.414] = 0.0786.$$

Hence, there is about an 8% chance that the average weight of two packages will be less than the labeled amount of 32 ounces.

7.27 The lengths of the trout fry in a pond at the fish hatchery are approximately normally distributed with mean 3.4 inches and standard deviation .8 inch. Three dozen fry are netted and their lengths measured.

(a) What is the probability that the sample mean length of the 36 netted trout fry is less than 3.2 inches?

(b) Why might the fish in the net not represent a random sample of trout fry in the pond?

SOLUTION

The population of fry has mean $\mu = 3.4$ and standard deviation $\sigma = 0.8$, so that

$$E(\overline{X}) = \mu = 3.4 \text{ and sd}(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.8}{\sqrt{36}} = 0.1333$$

and the standardized variable is $Z = \dfrac{\overline{X} - 3.4}{0.1333}$.

(a) $P[\overline{X} < 3.2] = P\left[Z < \dfrac{3.2 - 3.4}{0.1333}\right] = P[Z < -1.5] = 0.0668$

(b) Those caught in the net may be slower, less active fish, or even the less healthy ones. Consequently, they may tend to be on the smaller side of the distribution.

Determining Probabilities Concerning $\overline{X}$ —Normal Populations

The weight of a pepperoni and cheese pizza from a local provider is a random variable whose distribution is normal with mean 16 ounces and standard deviation 1 ounce. You intend to purchase four pepperoni and cheese pizzas. What is the probability that:

(a) The average weight of the four pizzas will be greater than 17.1 ounces?

(b) The total weight of the four pizzas will not exceed 61.0 ounces?

Because the population is normal, the distribution of the sample mean $\overline{X} = (X_1 + X_2 + X_3 + X_4)/4$ is exactly normal with mean 16 ounces and standard deviation $1/\sqrt{4} = .5$ ounce.

(a) Since $\overline{X}$ is $N\,(16, .5)$

$$\begin{aligned} P[\overline{X} > 17.1] &= P\left[\frac{\overline{X}-16}{.5} > \frac{17.1-16}{.5}\right] \\ &= P[Z > 2.20] = 1 - .9861 = .0139 \end{aligned}$$

Only rarely, just over one time in a hundred purchases of four pizzas, would the average weight exceed 17.1 ounces.

(b) The event that the total weight $X_1 + X_2 + X_3 + X_4 = 4\overline{X}$ does not exceed 61.0 ounces is the same event that the average weight $\overline{X}$ is less than or equal to $61.0/4 = 15.25$. Consequently

$$\begin{aligned} P[X_1 + X_2 + X_3 + X_4 \le 61.0] &= P[\overline{X} \le 15.25] \\ &= P\left[\frac{\overline{X}-16}{.5} \le \frac{15.25-16}{.5}\right] \\ &= P[Z \le -1.50] = .0668 \end{aligned}$$

Only about seven times in one hundred purchases would the total weight be less than 61.0 ounces.

Example 7-6

According to Moebs Services Inc., an individual checking account at major U.S. banks costs the banks between \$350 and \$450 per year (*Time*, November 21, 2011). Suppose that the current average cost of all checking accounts at major U.S. banks is \$400 per year with a standard deviation of \$30. Let $\bar{X}$ be the current average annual cost of a random sample of 225 individual checking account at major banks in America.

(a) What is the probability that the average annual cost of the checking accounts in this sample is within \$4 of the population mean?

(b) What is the probability that the average annual cost of the checking accounts in this sample is less than the population mean by \$2.70 or more?

Example 7-6: Solution

μ = \$400 and σ = \$30. The shape of the probability distribution of the population is unknown. However, the sampling distribution of $\bar{x}$ is approximately normal because the sample is large (n > 30).

$$\mu_{\bar{x}} = \mu = \mathbf{\$400}$$

$$\sigma_{\bar{x}} = \sigma / \sqrt{n} = \mathbf{30} / \sqrt{\mathbf{225}} = \mathbf{\$2}$$

Example 7-6: Solution

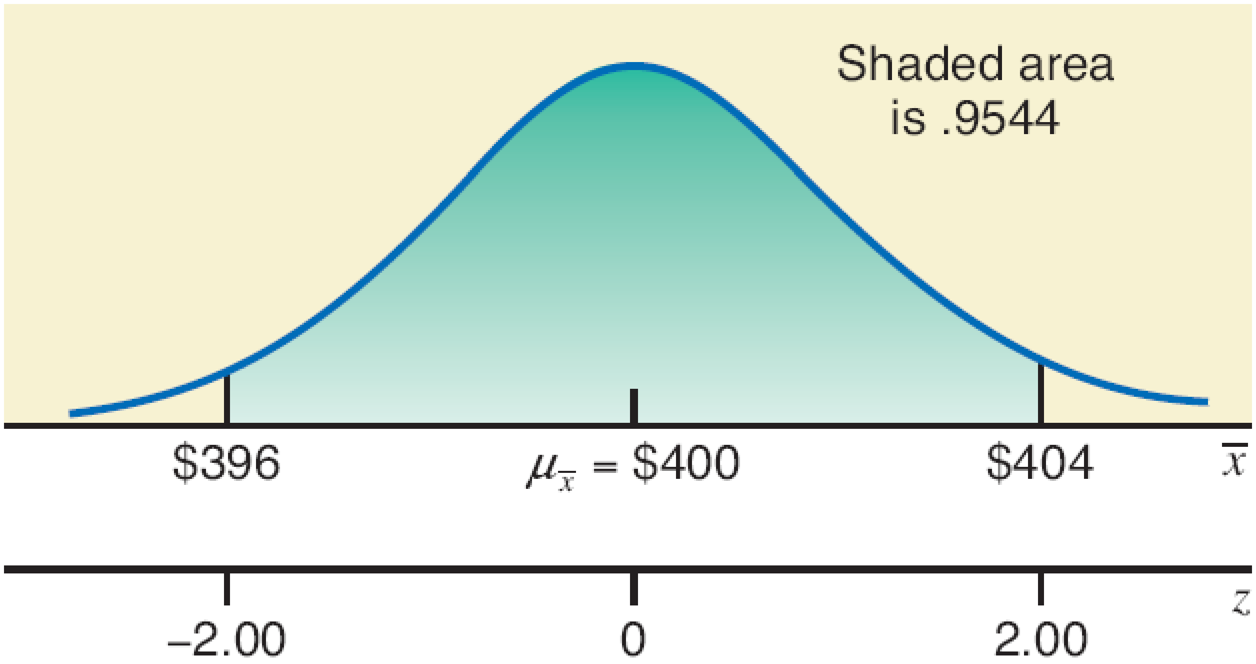
(a) Apply
$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$P(\$396 \le \bar{X} \le \$404) = P(-2.00 \le z \le 2.00)$

$= .9772 - .0228$

$= \mathbf{.9544}$

Figure 7.13



(a) Therefore, the probability that the average annual cost of the 225 checking accounts in this sample is within $4 of the population mean is .9544.

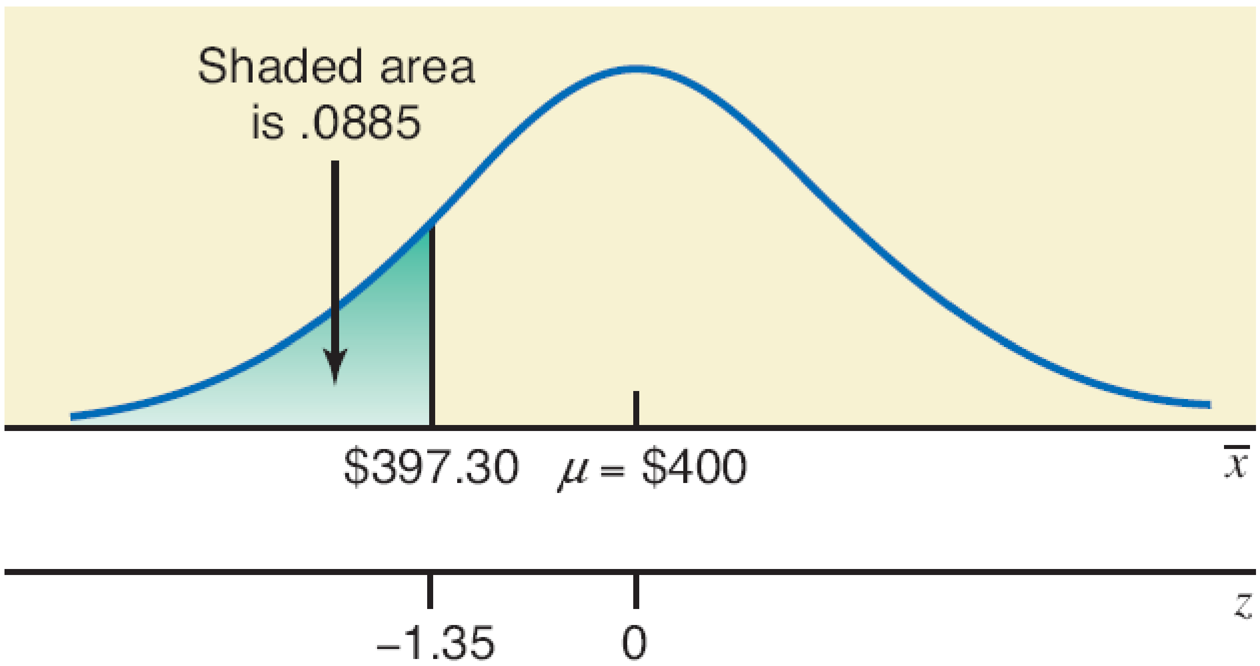
Example 7-6: Solution

(b) $$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = (397.30 - 400) / 2$$

$P(\bar{x} \leq \$397.50) = P(z \leq -1.35) =$ **.0885**

Figure 7.14

$P(\bar{x} \leq \$397.30)$



(b) Thus, the probability that the average annual cost of the checking accounts in this sample is less than the population mean by $2.70 or more is .0885.

Example 13 – *Central Limit Theorem*

- A certain strain of bacteria occurs in all raw milk. Let *x* be milliliter of milk. The health department has found that if the milk is not contaminated, then *x* has a distribution that is more or less mound-shaped and symmetrical.

- The mean of the *x* distribution is μ = 2500, and the standard deviation is σ = 300.In a large commercial dairy, the health inspector takes 42 random samples of the milk produced each day.

-

  At the end of the day, the bacteria count in each of the 42 samples is averaged to obtain the sample mean bacteria count $\overline{x}$.

Example 13(a) – *Central Limit Theorem*

- Assuming the milk is not contaminated, what is the distribution of $\overline{x}$?

- Solution:
  The sample size is *n* = 42.Since this value exceeds 30, the central limit theorem applies, and we know that $\overline{x}$ will be approximately normal, with mean and standard deviation

$$\mu_{\overline{x}} = \mu = 2500$$

$$\sigma_{\overline{x}} = \sigma/\sqrt{n} = 300/\sqrt{42} \approx 46.3$$

Example 13(b) – *Central Limit Theorem*

- Assuming the milk is not contaminated, what is the probability that the average bacteria count $\overline{x}$ for one day is between 2350 and 2650 bacteria per milliliter?

- Solution:
  We convert the interval

  $2350 \leq \overline{x} \leq 2650$

- to a corresponding interval on the standard *z* axis.

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\overline{x} - 2500}{46.3}$$

Example 13(b) – *Solution* cont'd

$$\bar{x} = 2350 \quad \text{converts to} \quad z = \frac{2350 - 2500}{46.3} \approx -3.24$$

$$\bar{x} = 2650 \quad \text{converts to} \quad z = \frac{2650 - 2500}{46.3} \approx 3.24$$

Therefore,

$P(2350 \le \bar{x} \le 2650) = P(-3.24 \le z \le 3.24)$

$= 0.9994 - 0.0006$

$= 0.9988$

The probability is 0.9988 that $\bar{x}$ is between 2350 and 2650.

Example 13(c) – *Central Limit Theorem*

- *Interpretation* At the end of each day, the inspector must decide to accept or reject the accumulated milk that has been held in cold storage awaiting shipment.

- Suppose the 42 samples taken by the inspector have a mean bacteria count $\bar{x}$ that is *not* between 2350 and 2650.

- If you were the inspector, what would be your comment on this situation?