7 | THE CENTRAL LIMIT THEOREM

Figure 7.1 If you want to figure out the distribution of the change people carry in their pockets, using the central limit theorem and assuming your sample is large enough, you will find that the distribution is normal and bell-shaped. (credit: John Lodder)

Introduction

Chapter Objectives

By the end of this chapter, the student should be able to:

- Recognize central limit theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the central limit theorem for means.
- Apply and interpret the central limit theorem for sums.

Why are we so concerned with means? Two reasons are: they give us a middle ground for comparison, and they are easy to calculate. In this chapter, you will study means and the **central limit theorem**.

The **central limit theorem** (clt for short) is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing finite samples size n from a population with a known mean, μ, and a known standard deviation, σ. The first alternative says that if we collect samples of size n with

a "large enough *n*," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size *n* that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

In either case, it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the distribution of sample means and the sums tend to follow the normal distribution.

The size of the sample, *n*, that is required in order to be "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. **Sampling is done with replacement.**

Collaborative Exercise

Suppose eight of you roll one fair die ten times, seven of you roll two fair dice ten times, nine of you roll five fair dice ten times, and 11 of you roll ten fair dice ten times.

Each time a person rolls more than one die, he or she calculates the sample **mean** of the faces showing. For example, one person might roll five fair dice and get 2, 2, 3, 4, 6 on one roll.

The mean is $\frac{2+2+3+4+6}{5}$ = 3.4. The 3.4 is one mean when five fair dice are rolled. This same person would roll the five dice nine more times and calculate nine more means for a total of ten means.

Your instructor will pass out the dice to several people. Roll your dice ten times. For each roll, record the faces, and find the mean. Round to the nearest 0.5.

Your instructor (and possibly you) will produce one graph (it might be a histogram) for one die, one graph for two dice, one graph for five dice, and one graph for ten dice. Since the "mean" when you roll one die is just the face on the die, what distribution do these **means** appear to be representing?

Draw the graph for the means using two dice. Do the sample means show any kind of pattern?

Draw the graph for the means using five dice. Do you see any pattern emerging?

Finally, draw the graph for the means using ten dice. Do you see any pattern to the graph? What can you conclude as you increase the number of dice?

As the number of dice rolled increases from one to two to five to ten, the following is happening:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the central limit theorem (clt).

The central limit theorem tells you that as you increase the number of dice, **the sample means tend toward a normal distribution (the sampling distribution).**

7.1 | The Central Limit Theorem for Sample Means (Averages)

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose:

a. μ_X = the mean of X
b. σ_X = the standard deviation of X

If you draw random samples of size *n*, then as *n* increases, the random variable $\bar{X}$ which consists of sample means, tends to be **normally distributed** and

$\bar{X} \sim N\left(\mu_x, \frac{\sigma x}{\sqrt{n}}\right)$.

The **central limit theorem** for sample means says that if you keep drawing larger and larger samples (such as rolling one, two, five, and finally, ten dice) and **calculating their means,** the sample means form their own **normal distribution** (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by, the sample size. The variable n is the number of values that are averaged together, not the number of times the experiment is done.

To put it more formally, if you draw random samples of size n, the distribution of the random variable $\bar{X}$, which consists of sample means, is called the **sampling distribution of the mean**. The sampling distribution of the mean approaches a normal distribution as n, the **sample size**, increases.

The random variable $\bar{X}$ has a different z-score associated with it from that of the random variable X. The mean $\bar{x}$ is the value of $\bar{X}$ in one sample.

$$z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

μ_X is the average of both X and $\bar{X}$.

$\sigma\bar{x} = \frac{\sigma x}{\sqrt{n}}$ = standard deviation of $\bar{X}$ and is called the **standard error of the mean.**

Using the TI-83, 83+, 84, 84+ Calculator

To find probabilities for means on the calculator, follow these steps.

2nd DISTR
2:normalcdf

$$normalcdf\left(lower\ value\ of\ the\ area,\ upper\ value\ of\ the\ area,\ mean,\ \frac{standard\ \ deviation}{\sqrt{sample\ \ size}}\right)$$

where:

- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = n

Example 7.1

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size $n = 25$ are drawn randomly from the population.

a. Find the probability that the **sample mean** is between 85 and 92.

Solution 7.1

a. Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean**.

Let $\bar{X}$ = the mean of a sample of size 25. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 25$,

$\bar{X} \sim N\left(90, \frac{15}{\sqrt{25}}\right)$.

Find $P(85 < \bar{x} < 92)$. Draw a graph.

$P(85 < \bar{x} < 92) = 0.6997$

The probability that the sample mean is between 85 and 92 is 0.6997.

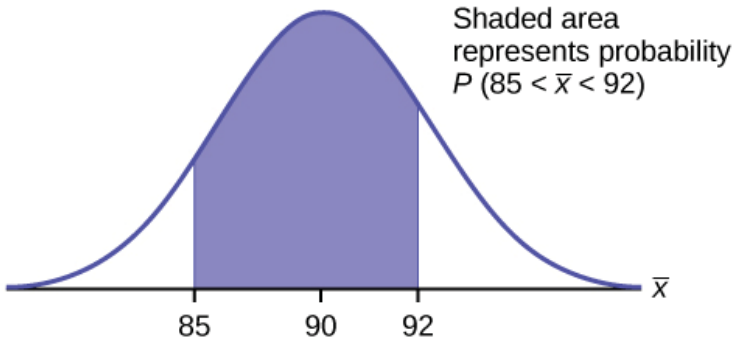

Figure 7.2

Using the TI-83, 83+, 84, 84+ Calculator

normalcdf(lower value, upper value, mean, standard error of the mean)

The parameter list is abbreviated (lower value, upper value, μ, $\frac{\sigma}{\sqrt{n}}$)

normalcdf(85,92,90, $\frac{15}{\sqrt{25}}$) = 0.6997

b. Find the value that is two standard deviations above the expected value, 90, of the sample mean.

Solution 7.1

b. To find the value that is two standard deviations above the expected value 90, use the formula:

value = μ_x + (#ofTSDEVs) $\left(\frac{\sigma_x}{\sqrt{n}}\right)$

value = 90 + 2 $\left(\frac{15}{\sqrt{25}}\right)$ = 96

The value that is two standard deviations above the expected value is 96.

The standard error of the mean is $\frac{\sigma x}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3$. Recall that the standard error of the mean is a description of how far (on average) that the sample mean will be from the population mean in repeated simple random samples of size n.

Try It

7.1 An unknown distribution has a mean of 45 and a standard deviation of eight. Samples of size $n = 30$ are drawn randomly from the population. Find the probability that the sample mean is between 42 and 50.

Example 7.2

The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of two hours** and a **standard deviation of 0.5 hours**. A **sample of size *n* = 50** is drawn randomly from the population. Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Solution 7.2

Let X = the time, in hours, it takes to play one soccer match.

The probability question asks you to find a probability for the **sample mean time, in hours**, it takes to play one soccer match.

Let $\bar{X}$ = the **mean** time, in hours, it takes to play one soccer match.

If μ_X = _____, σ_X = _____, and n = _____, then $X \sim N$(_____, _____) by the **central limit theorem for means**.

$\mu_X = 2$, $\sigma_X = 0.5$, $n = 50$, and $X \sim N\left(2, \frac{0.5}{\sqrt{50}}\right)$

Find $P(1.8 < \bar{x} < 2.3)$. Draw a graph.

$P(1.8 < \bar{x} < 2.3) = 0.9977$

normalcdf$\left(1.8,2.3,2,\frac{.5}{\sqrt{50}}\right) = 0.9977$

The probability that the mean time is between 1.8 hours and 2.3 hours is 0.9977.

Try It

7.2 The length of time taken on the SAT for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours. A sample size of $n = 60$ is drawn randomly from the population. Find the probability that the sample mean is between two hours and three hours.

Using the TI-83, 83+, 84, 84+ Calculator

To find percentiles for means on the calculator, follow these steps.

2nd DIStR
3:invNorm

k = invNorm$\left(\text{area to the left of } k, \text{ mean}, \frac{\textit{standard deviation}}{\sqrt{\textit{sample size}}}\right)$

where:

- k = the k^{th} percentile
- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = n

Example 7.3

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. Take a sample of size $n = 100$.

a. What are the mean and standard deviation for the sample mean ages of tablet users?

b. What does the distribution look like?

c. Find the probability that the sample mean age is more than 30 years (the reported mean age of tablet users in this particular study).

d. Find the 95th percentile for the sample mean age (to one decimal place).

Solution 7.3

a. Since the sample mean tends to target the population mean, we have $\mu_{\bar{X}} = \mu = 34$. The sample standard deviation is given by $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$

b. The central limit theorem states that for large sample sizes(n), the sampling distribution will be approximately normal.

c. The probability that the sample mean age is more than 30 is given by $P(X > 30) =$ `normalcdf`(30,E99,34,1.5) = 0.9962

d. Let k = the 95th percentile.

$k = \text{invNorm}\left(0.95,34,\frac{15}{\sqrt{100}}\right) = 36.5$

Try It

7.3 In an article on Flurry Blog, a gaming marketing gap for men between the ages of 30 and 40 is identified. You are researching a startup game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

Example 7.4

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of 60.

a. What are the mean and standard deviation for the sample mean number of app engagement by a tablet user?

b. What is the standard error of the mean?

c. Find the 90th percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.

d. Find the probability that the sample mean is between eight minutes and 8.5 minutes.

Solution 7.4

a. $\mu_{\bar{x}} = \mu = 8.2\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.13$

b. This allows us to calculate the probability of sample means of a particular distance from the mean, in repeated samples of size 60.

c. Let k = the 90th percentile

k = invNorm$\left(0.90,8.2,\frac{1}{\sqrt{60}}\right)$ = 8.37. This values indicates that 90 percent of the average app engagement time for table users is less than 8.37 minutes.

d. $P(8 < \bar{x} < 8.5)$ = normalcdf$\left(8,8.5,8.2,\frac{1}{\sqrt{60}}\right)$ = 0.9293

Try It

7.4 Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample are measured and the statistics are $n = 34$, $\bar{x} = 16.01$ ounces. If the cans are filled so that $\mu = 16.00$ ounces (as labeled) and $\sigma = 0.143$ ounces, find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

7.2 | The Central Limit Theorem for Sums

Suppose X is a random variable with a distribution that may be **known or unknown** (it can be any distribution) and suppose:

a. μ_X = the mean of X

b. σ_X = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable ΣX consisting of sums tends to be **normally distributed** and $\Sigma X \sim N((n)(\mu_X), (\sqrt{n})(\sigma_X))$.

The central limit theorem for sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution), which approaches a normal distribution as the sample size increases. **The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.**

The random variable ΣX has the following z-score associated with it:

a. Σx is one sum.

b. $z = \frac{\Sigma x - (n)(\mu_X)}{(\sqrt{n})(\sigma_X)}$

i. $(n)(\mu_X)$ = the mean of ΣX

ii. $(\sqrt{n})(\sigma_X)$ = standard deviation of ΣX

Using the TI-83, 83+, 84, 84+ Calculator

To find probabilities for sums on the calculator, follow these steps.

2nd DISTR
2:normalcdf
normalcdf(lower value of the area, upper value of the area, (n)(mean), $(\sqrt{n})$(standard deviation))

where:

- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- ***sample size*** = n