6.3 THE CENTRAL LIMIT THEOREM

In this section, we will consider one of the most remarkable results in probability — namely, the *central limit theorem*. Loosely speaking, this theorem asserts that the sum of a large number of independent random variables has a distribution that is approximately normal. Hence, it not only provides a simple method for computing approximate probabilities for sums of independent random variables, but it also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit a bell-shaped (that is, a normal) curve.

In its simplest form, the central limit theorem is as follows:

Theorem 6.3.1 The Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then for n large, the distribution of

$$X_1 + \cdots + X_n$$

is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

It follows from the central limit theorem that

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

is approximately a standard normal random variable; thus, for n large,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < x\right\} \approx P\{Z < x\}$$

where Z is a standard normal random variable.

EXAMPLE 6.3a An insurance company has 25,000 automobile policy holders. If the yearly claim of a policy holder is a random variable with mean 320 and standard deviation 540, approximate the probability that the total yearly claim exceeds 8.3 million.

SOLUTION Let X denote the total yearly claim. Number the policy holders, and let X_i denote the yearly claim of policy holder i. With $n = 25{,}000$, we have from the central limit theorem that $X = \sum_{i=1}^{n} X_i$ will have approximately a normal distribution with mean $320 \times 25{,}000 = 8 \times 10^6$ and standard deviation $540\sqrt{25{,}000} = 8.5381 \times 10^4$. Therefore,

$$P\{X > 8.3 \times 10^6\} = P\left\{\frac{X - 8 \times 10^6}{8.5381 \times 10^4} > \frac{8.3 \times 10^6 - 8 \times 10^6}{8.5381 \times 10^4}\right\}$$

$$= P\left\{\frac{X - 8 \times 10^6}{8.5381 \times 10^4} > \frac{.3 \times 10^6}{8.5381 \times 10^4}\right\}$$

$$\approx P\{Z > 3.51\} \qquad \text{where } Z \text{ is a standard normal}$$
$$\approx .00023$$

Thus, there are only 2.3 chances out of 10,000 that the total yearly claim will exceed 8.3 million. ■

EXAMPLE 6.3b Civil engineers believe that W, the amount of weight (in units of 1,000 pounds) that a certain span of a bridge can withstand without structural damage resulting, is normally distributed with mean 400 and standard deviation 40. Suppose that the weight (again, in units of 1,000 pounds) of a car is a random variable with mean 3 and standard deviation .3. How many cars would have to be on the bridge span for the probability of structural damage to exceed .1?

SOLUTION Let P_n denote the probability of structural damage when there are n cars on the bridge. That is,

$$\begin{aligned} P_n &= P\{X_1 + \cdots + X_n \geq W\} \\ &= P\{X_1 + \cdots + X_n - W \geq 0\} \end{aligned}$$

where X_i is the weight of the ith car, $i = 1, \ldots, n$. Now it follows from the central limit theorem that $\sum_{i=1}^{n} X_i$ is approximately normal with mean $3n$ and variance $.09n$. Hence, since W is independent of the $X_i, i = 1, \ldots, n$, and is also normal, it follows that $\sum_{i=1}^{n} X_i - W$ is approximately normal, with mean and variance given by

$$E\left[\sum_{1}^{n} X_i - W\right] = 3n - 400$$
$$\text{Var}\left(\sum_{1}^{n} X_i - W\right) = \text{Var}\left(\sum_{1}^{n} X_i\right) + \text{Var}(W) = .09n + 1{,}600$$

Therefore, if we let

$$Z = \frac{\sum_{i=1}^{n} X_i - W - (3n - 400)}{\sqrt{.09n + 1{,}600}}$$

then

$$P_n = P\left\{Z \geq \frac{-(3n - 400)}{\sqrt{.09n + 1{,}600}}\right\}$$

where Z is approximately a standard normal random variable. Now $P\{Z \geq 1.28\} \approx .1$, and so if the number of cars n is such that

$$\frac{400 - 3n}{\sqrt{.09n + 1{,}600}} \leq 1.28$$

or

$$n \geq 117$$

then there is at least 1 chance in 10 that structural damage will occur. ■

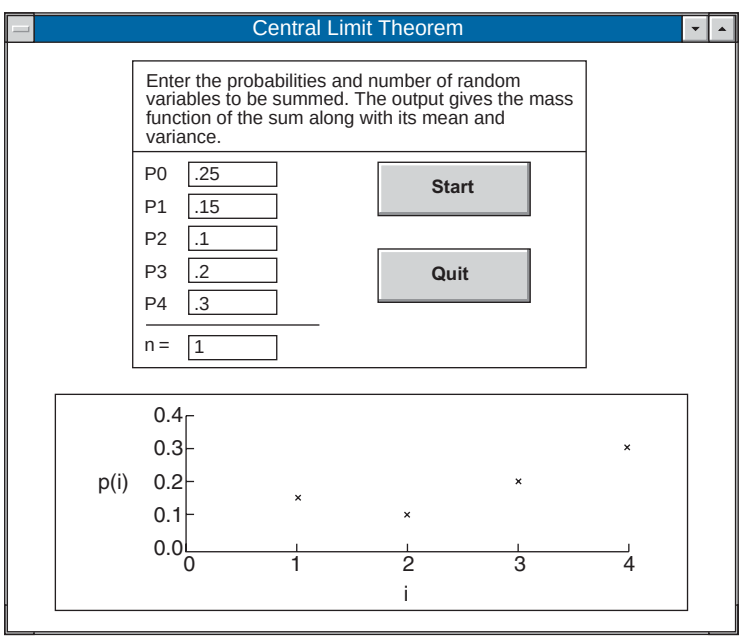
The central limit theorem is illustrated by Program 6.1 on the text disk. This program plots the probability mass function of the sum of n independent and identically distributed random variables that each take on one of the values 0, 1, 2, 3, 4. When using it, one enters the probabilities of these five values, and the desired value of n. Figures 6.2(a)–(f) give the resulting plot for a specified set of probabilities when $n = 1, 3, 5, 10, 25, 100$.

One of the most important applications of the central limit theorem is in regard to binomial random variables. Since such a random variable X having parameters (n, p) represents the number of successes in n independent trials when each trial is a success with probability p, we can express it as

$$X = X_1 + \cdots + X_n$$

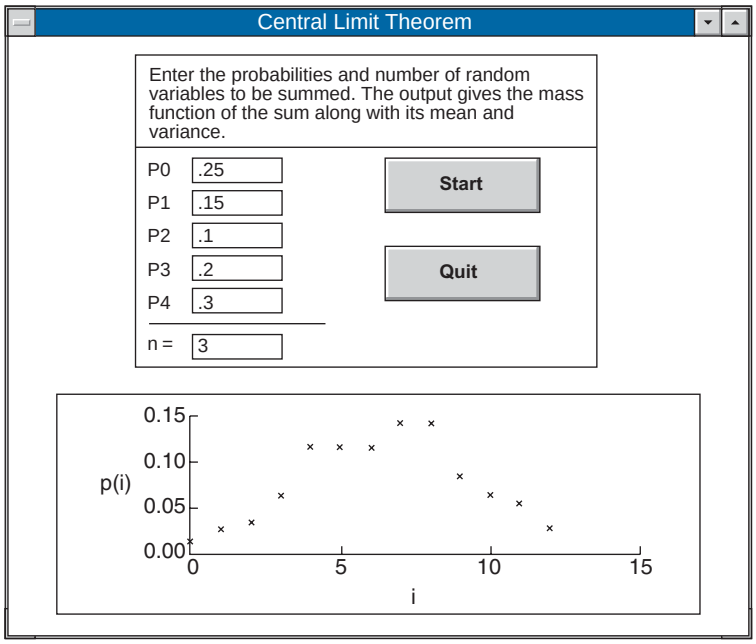
where

$$X_i = \begin{cases} 1 & \text{if the } i\text{th trial is a success} \\ 0 & \text{otherwise} \end{cases}$$

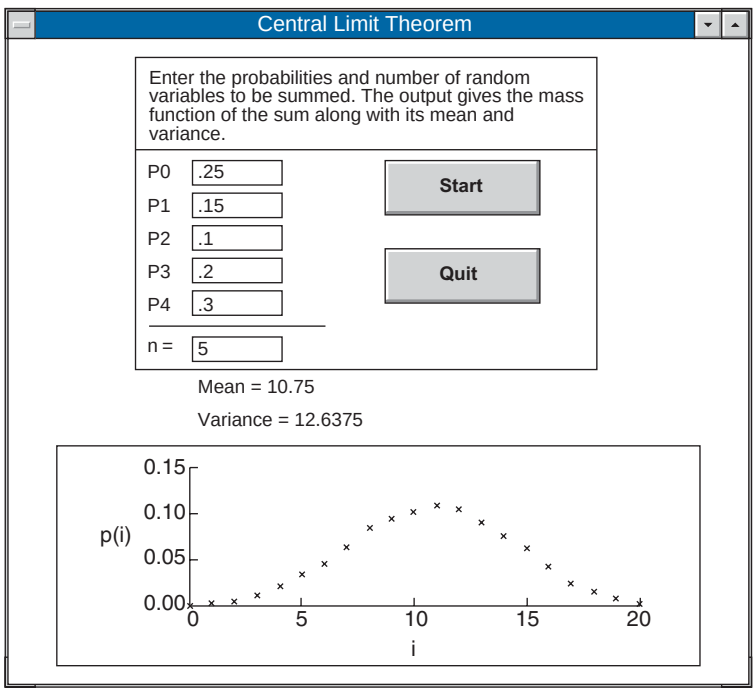


(a)

FIGURE 6.2 *(a)* $n = 1$, *(b)* $n = 3$, *(c)* $n = 5$, *(d)* $n = 10$, *(e)* $n = 25$, *(f)* $n = 100$.

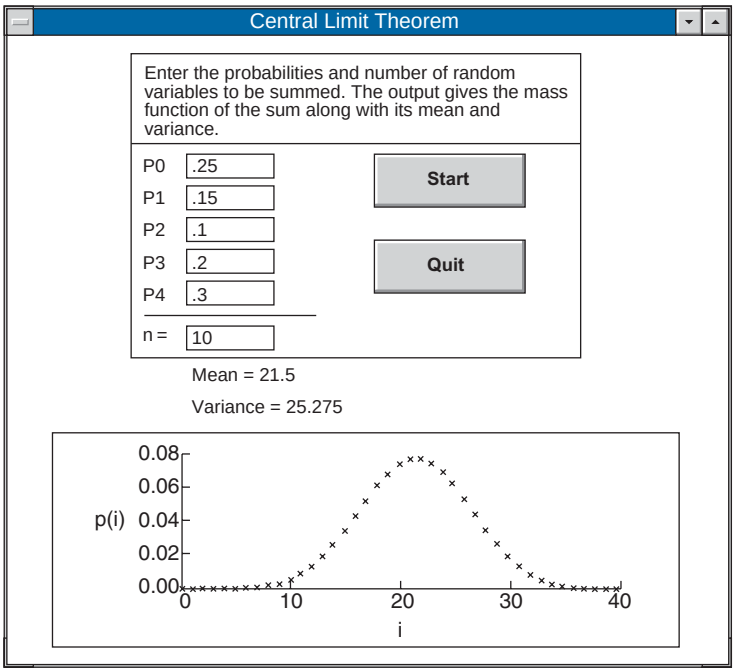


(b)

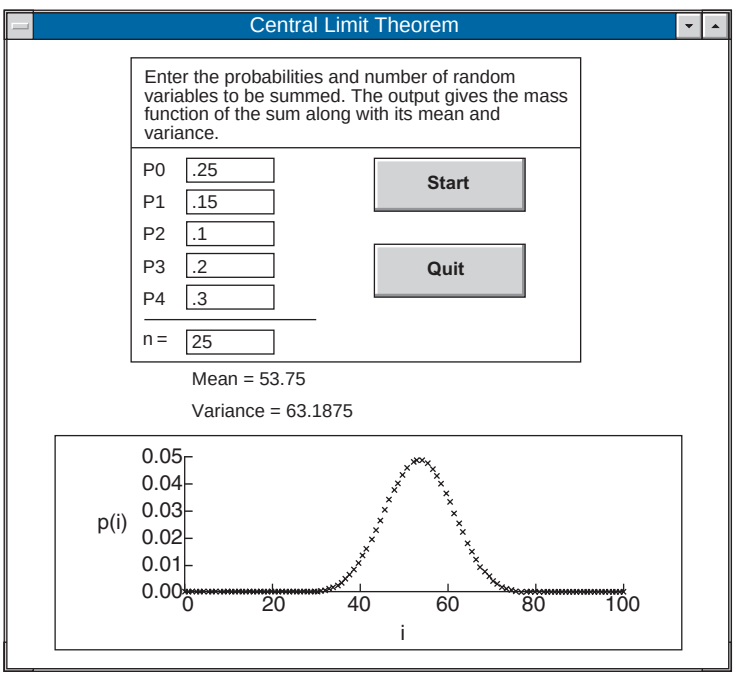


(c)

FIGURE 6.2 *(continued)*

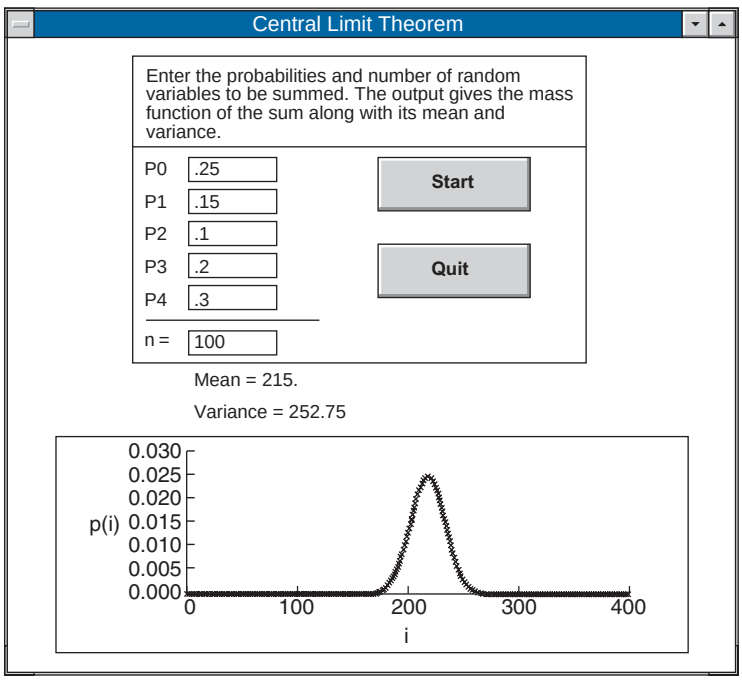


(d)



(e)

FIGURE 6.2 *(continued)*



(f)

FIGURE 6.2 *(continued)*

Because

$$E[X_i] = p, \qquad \text{Var}(X_i) = p(1-p)$$

it follows from the central limit theorem that for n large

$$\frac{X - np}{\sqrt{np(1-p)}}$$

will approximately be a standard normal random variable [see Figure 6.3, which graphically illustrates how the probability mass function of a binomial (n, p) random variable becomes more and more "normal" as n becomes larger and larger].

EXAMPLE 6.3c The ideal size of a first-year class at a particular college is 150 students. The college, knowing from past experience that, on the average, only 30 percent of those accepted for admission will actually attend, uses a policy of approving the applications of 450 students. Compute the probability that more than 150 first-year students attend this college.

SOLUTION Let X denote the number of students that attend; then assuming that each accepted applicant will independently attend, it follows that X is a binomial random

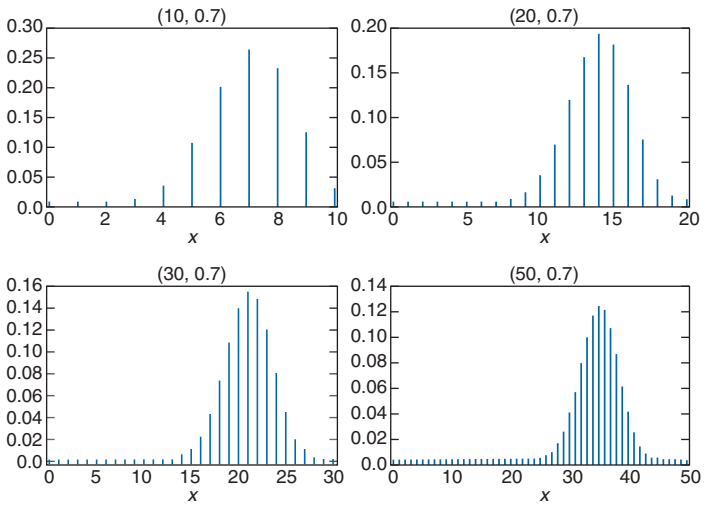


FIGURE 6.3 *Binomial probability mass functions converging to the normal density.*

variable with parameters $n = 450$ and $p = .3$. Since the binomial is a discrete and the normal a continuous distribution, it is best to compute $P\{X = i\}$ as $P\{i-.5 < X < i+.5\}$ when applying the normal approximation (this is called the continuity correction). This yields the approximation

$$P\{X > 150.5\} = P\left\{\frac{X - (450)(.3)}{\sqrt{450(.3)(.7)}} \geq \frac{150.5 - (450)(.3)}{\sqrt{450(.3)(.7)}}\right\}$$
$$\approx P\{Z > 1.59\} = .06$$

Hence, only 6 percent of the time do more than 150 of the first 450 accepted actually attend. ■

It should be noted that we now have two possible approximations to binomial probabilities: The Poisson approximation, which yields a good approximation when n is large and p small, and the normal approximation, which can be shown to be quite good when $np(1 - p)$ is large. [The normal approximation will, in general, be quite good for values of n satisfying $np(1 - p) \geq 10$.]

6.3.1 Approximate Distribution of the Sample Mean

Let $X_1, \ldots, X_n$ be a sample from a population having mean μ and variance σ^2. The central limit theorem can be used to approximate the distribution of the sample mean

$$\overline{X} = \sum_{i=1}^{n} X_i/n$$

Since a constant multiple of a normal random variable is also normal, it follows from the central limit theorem that $\overline{X}$ will be approximately normal when the sample size n is large. Since the sample mean has expected value μ and standard deviation $\sigma/\sqrt{n}$, it then follows that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

has approximately a standard normal distribution.

EXAMPLE 6.3d The weights of a population of workers have mean 167 and standard deviation 27.

(a) If a sample of 36 workers is chosen, approximate the probability that the sample mean of their weights lies between 163 and 170.

(b) Repeat part (a) when the sample is of size 144.

SOLUTION Let Z be a standard normal random variable.

(a) It follows from the central limit theorem that $\overline{X}$ is approximately normal with mean 167 and standard deviation $27/\sqrt{36} = 4.5$. Therefore,

$$\begin{aligned} P\{163 < \overline{X} < 170\} &= P\left\{\frac{163 - 167}{4.5} < \frac{\overline{X} - 167}{4.5} < \frac{170 - 167}{4.5}\right\} \\ &= P\left\{-.8889 < \frac{\overline{X} - 167}{4.5} < .8889\right\} \\ &\approx 2P\{Z < .8889\} - 1 \\ &\approx .6259 \end{aligned}$$

(b) For a sample of size 144, the sample mean will be approximately normal with mean 167 and standard deviation $27/\sqrt{144} = 2.25$. Therefore,

$$\begin{aligned} P\{163 < \overline{X} < 170\} &= P\left\{\frac{163 - 167}{2.25} < \frac{\overline{X} - 167}{2.25} < \frac{170 - 167}{2.25}\right\} \\ &= P\left\{-1.7778 < \frac{\overline{X} - 167}{4.5} < 1.7778\right\} \\ &\approx 2P\{Z < 1.7778\} - 1 \\ &\approx .9246 \end{aligned}$$

Thus increasing the sample size from 36 to 144 increases the probability from .6259 to .9246. ■

EXAMPLE 6.3e An astronomer wants to measure the distance from her observatory to a distant star. However, due to atmospheric disturbances, any measurement will not yield the exact distance d. As a result, the astronomer has decided to make a series of measurements and then use their average value as an estimate of the actual distance. If the astronomer believes that the values of the successive measurements are independent random variables with a mean of d light years and a standard deviation of 2 light years, how many measurements need she make to be at least 95 percent certain that her estimate is accurate to within $\pm .5$ light years?

SOLUTION If the astronomer makes n measurements, then $\overline{X}$, the sample mean of these measurements, will be approximately a normal random variable with mean d and standard deviation $2/\sqrt{n}$. Thus, the probability that it will lie between $d \pm .5$ is obtained as follows:

$$\begin{aligned} P\{-.5 < \overline{X} - d < .5\} &= P\left\{\frac{-.5}{2/\sqrt{n}} < \frac{\overline{X} - d}{2/\sqrt{n}} < \frac{.5}{2/\sqrt{n}}\right\} \\ &\approx P\{-\sqrt{n}/4 < Z < \sqrt{n}/4\} \\ &= 2P\{Z < \sqrt{n}/4\} - 1 \end{aligned}$$

where Z is a standard normal random variable.

Thus, the astronomer should make n measurements, where n is such that

$$2P\{Z < \sqrt{n}/4\} - 1 \geq .95$$

or, equivalently,

$$P\{Z < \sqrt{n}/4\} \geq .975$$

Since $P\{Z < 1.96\} = .975$, it follows that n should be chosen so that

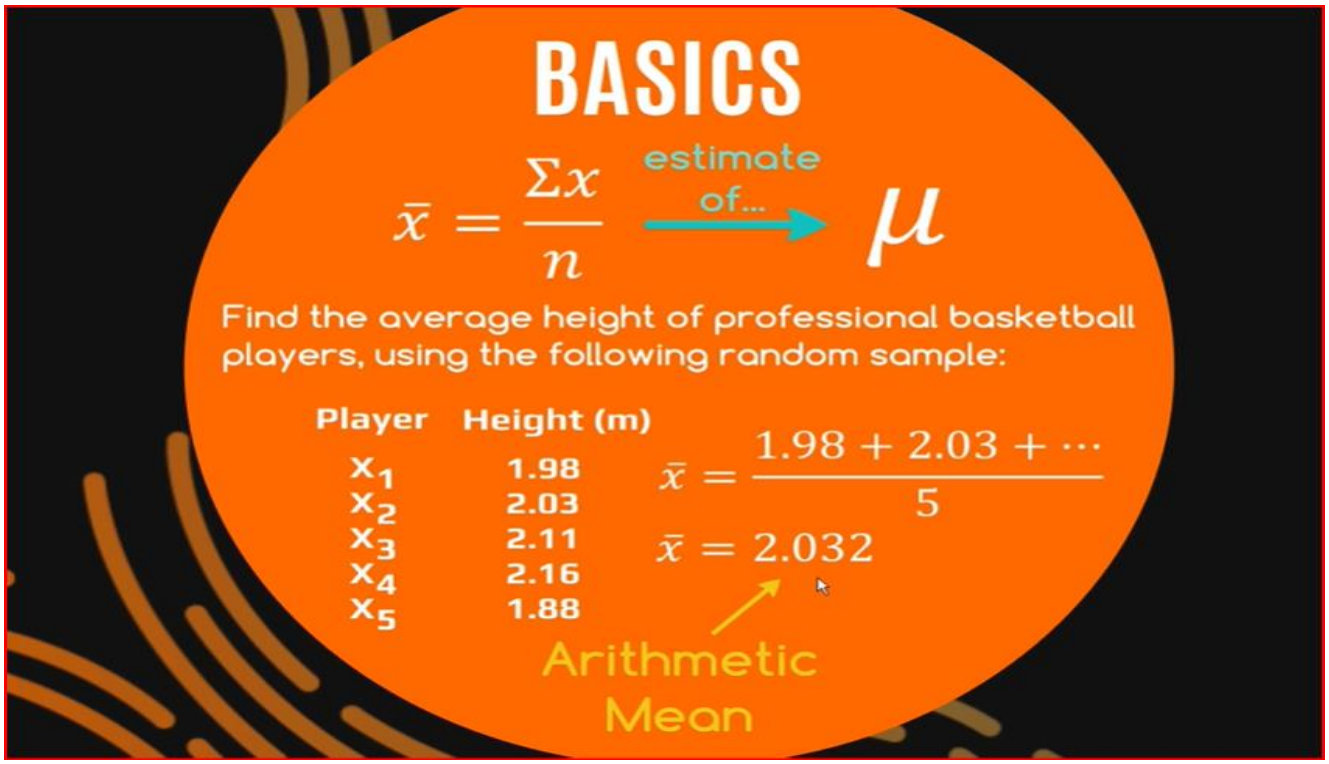
$$\sqrt{n}/4 \geq 1.96$$

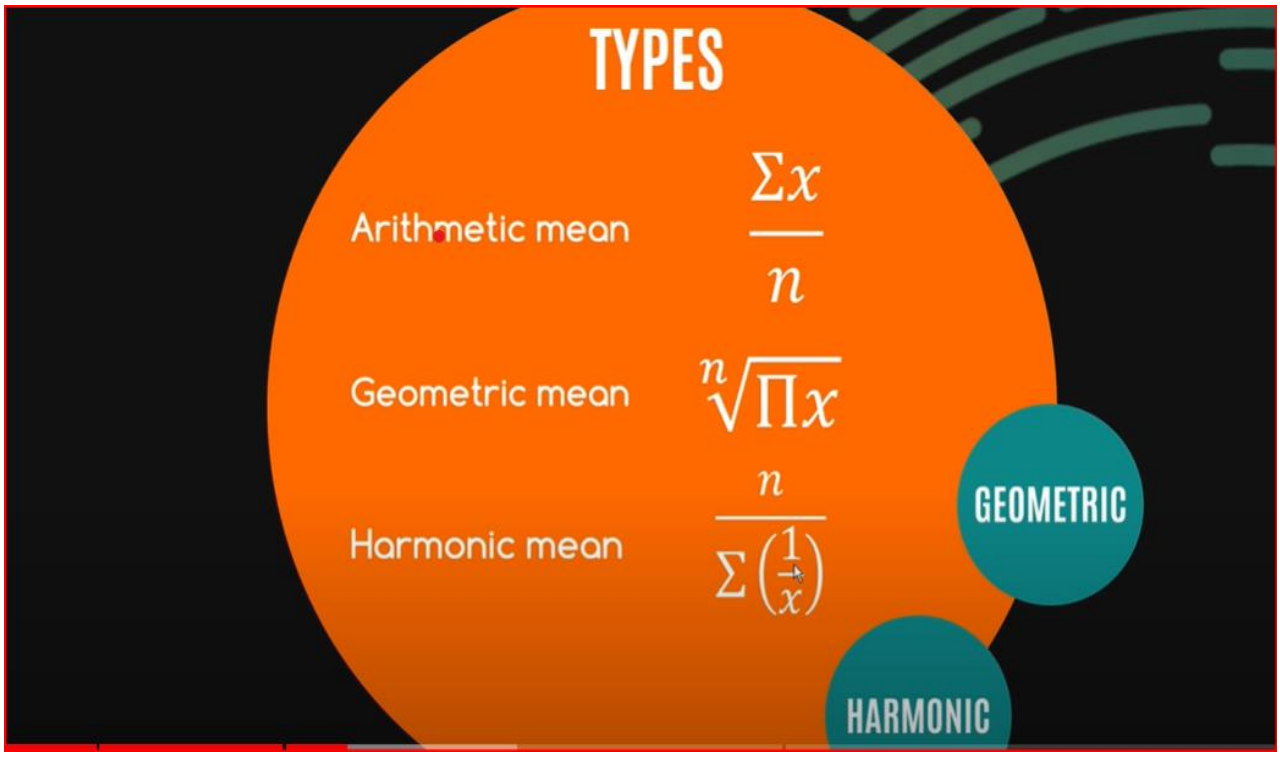
That is, at least 62 observations are necessary. ■

6.3.2 How Large a Sample Is Needed?

The central limit theorem leaves open the question of how large the sample size n needs to be for the normal approximation to be valid, and indeed the answer depends on the population distribution of the sample data. For instance, if the underlying population distribution is normal, then the sample mean $\overline{X}$ will also be normal regardless of the sample size. A general rule of thumb is that one can be confident of the normal approximation whenever the sample size n is at least 30. That is, practically speaking, no matter how nonnormal the underlying population distribution is, the sample mean of a sample of size at least 30 will be approximately normal. In most cases, the normal approximation is valid for much

Basis idea about Arithmetic mean, Geometric mean, Harmonic mean





The arithmetic mean, often simply called the average, is the sum of a set of numbers divided by the count of those numbers. It is the most commonly used measure of central tendency.

Mathematical formula:

$$\text{Arithmetic Mean} = \frac{x_1+x_2+\ldots+x_n}{n}$$

Where:

- $x_1, x_2, \ldots, x_n$ are the individual numbers in the set.
- n is the count of numbers in the set.

Example:

Consider the set of numbers: 2, 5, 8, 11, 14. The arithmetic mean is calculated as follows:

$$\text{Arithmetic Mean} = \frac{2+5+8+11+14}{5} = \frac{40}{5} = 8$$

Geometric Mean:

The geometric mean is the nth root of the product of n numbers. It is used when dealing with quantities that are proportional to each other.

Mathematical formula:

$$\text{Geometric Mean} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Where:

- $x_1, x_2, \ldots, x_n$ are the individual numbers in the set.
- n is the count of numbers in the set.

Example:

Consider the set of numbers: 2, 4, 8, 16. The geometric mean is calculated as follows:

$$\text{Geometric Mean} = \sqrt[4]{2 \cdot 4 \cdot 8 \cdot 16} = \sqrt[4]{1024} = 8$$

Harmonic Mean:

The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of a set of numbers. It is often used in situations where rates or ratios are involved.

Mathematical formula:

$$\text{Harmonic Mean} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}}$$

Where:

- $x_1, x_2, \ldots, x_n$ are the individual numbers in the set.
- n is the count of numbers in the set.

Example:

Consider the set of numbers: 2, 4, 8, 16. The harmonic mean is calculated as follows:

$$\text{Harmonic Mean} = \frac{4}{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}} = \frac{4}{\frac{8+4+2+1}{16}} = \frac{4}{\frac{15}{16}} = \frac{64}{15}$$

In summary AM$\geq GM \geq HM$

In summary:

Arithmetic mean is suitable for situations where the data is evenly distributed.

Geometric mean is useful for situations involving growth rates, ratios, or proportions.

Harmonic mean is appropriate when dealing with rates, such as speed or efficiency.

Finding Mode and Median for Grouped Data

Mode:

The mode is the value or values that occur most frequently. For grouped data, you can find the mode using the formula:

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times c$$

Where:

- L is the lower class boundary of the modal class,
- f_1 is the frequency of the modal class,
- f_0 is the frequency of the class before the modal class,
- f_2 is the frequency of the class after the modal class,
- c is the width of the class interval.

Median:

For grouped data, the median can be found using the formula:

$$\text{Median} = L + \left(\frac{\frac{N}{2} - F}{f}\right) \times c$$

Where:

- L is the lower class boundary of the median class,
- N is the total number of observations,
- F is the cumulative frequency of the class before the median class,
- f is the frequency of the median class,
- c is the width of the class interval.

Let's go through an example to illustrate these concepts.

Consider the following grouped data:

Class Interval	Frequency
$10-20$	5
$20-30$	8
$30-40$	12
$40-50$	6
$50-60$	9

1. **Mode:**
   - Modal class: 30-40
   - $L = 30$ (lower boundary of the modal class)
   - $f_1 = 12$ (frequency of the modal class)
   - $f_0 = 8$ (frequency of the class before the modal class)
   - $f_2 = 6$ (frequency of the class after the modal class)
   - $c = 10$ (width of the class interval)

   Substituting these values into the mode formula:

   $$\text{Mode} = 30 + \left(\frac{12-8}{2\times12-8-6}\right) \times 10$$

2. **Median:**
   - Total number of observations (N) = Sum of frequencies = $5 + 8 + 12 + 6 + 9 = 40$
   - Median class: 30-40
   - $L = 30$ (lower boundary of the median class)
   - $F = 5 + 8 = 13$ (cumulative frequency of the class before the median class)
   - $f = 12$ (frequency of the median class)
   - $c = 10$ (width of the class interval)

   Substituting these values into the median formula:

   $$\text{Median} = 30 + \left(\frac{\frac{40}{2}-13}{12}\right) \times 10$$

Now, you can calculate these values to find the mode and median for the given data.