Keys to the Central Limit Theorem

- One of the most important theorems in statistics and probability theory is the Central Limit Theorem .
- It is used almost everywhere we apply statistics.
- It’s usefulness lies in its simple definition.
- The central limit theorem states that if some certain conditions are satisfied, then the distribution of the arithmetic mean of a number of independent random variables approaches a normal distribution as the number of variables approaches infinity.
- In other words, there is no need to know very much about the actual distribution of the variables, as long as there are enough instances of them - their sum can be treated as normally distributed.

Keys to the Central Limit Theorem

The bottom line is:

“The beauty of the theorem thus lies in its simplicity.”

Equations for x-bar distributions….
(when we want to know information about our sample drawn from a population)

$$\mu_{\overline{X}} = \mu$$

Mean of X-bar distribution

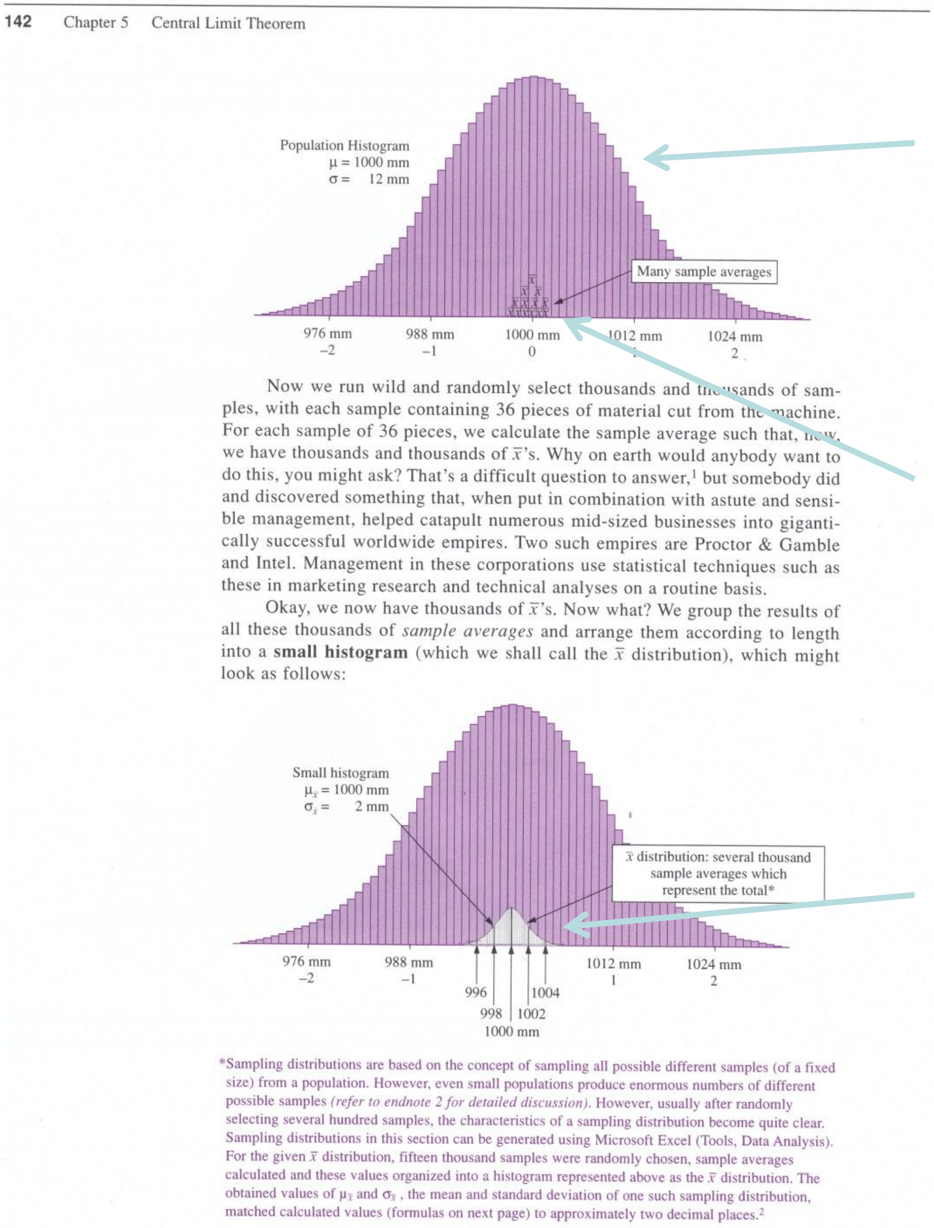
$$\sigma_{\overline{X}} = \sigma / \sqrt{n}$$

Standard Deviation of X-bar distribution
(in SPSS called “Standard Error”)

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

Z- score (in this case, you must use standard deviation of X-Bar)

With these equations, you can use the normal table to predict SAMPLE DISTRIBUTION VALUES.


Population Histogram
μ = 1000 mm
σ = 12 mm
Many sample averages
976 mm
–2
988 mm
–1
1000 mm
0
1012 mm
1
1024 mm
2

142 Chapter 5 Central Limit Theorem

Now we run wild and randomly select thousands and thousands of samples, with each sample containing 36 pieces of material cut from the machine. For each sample of 36 pieces, we calculate the sample average such that, now, we have thousands and thousands of $\overline{x}$'s. Why on earth would anybody want to do this, you might ask? That's a difficult question to answer,[1] but somebody did and discovered something that, when put in combination with astute and sensible management, helped catapult numerous mid-sized businesses into gigantically successful worldwide empires. Two such empires are Proctor & Gamble and Intel. Management in these corporations use statistical techniques such as these in marketing research and technical analyses on a routine basis.

Okay, we now have thousands of $\overline{x}$'s. Now what? We group the results of all these thousands of *sample averages* and arrange them according to length into a **small histogram** (which we shall call the $\overline{x}$ distribution), which might look as follows:

Small histogram
$\mu_{\overline{x}}$ = 1000 mm
$\sigma_{\overline{x}}$ = 2 mm
$\overline{x}$ distribution: several thousand sample averages which represent the total*
976 mm
–2
988 mm
–1
1012 mm
1
1024 mm
2
996
998
1000 mm
1002
1004

*Sampling distributions are based on the concept of sampling all possible different samples (of a fixed size) from a population. However, even small populations produce enormous numbers of different possible samples *(refer to endnote 2 for detailed discussion)*. However, usually after randomly selecting several hundred samples, the characteristics of a sampling distribution become quite clear. Sampling distributions in this section can be generated using Microsoft Excel (Tools, Data Analysis). For the given $\overline{x}$ distribution, fifteen thousand samples were randomly chosen, sample averages calculated and these values organized into a histogram represented above as the $\overline{x}$ distribution. The obtained values of $\mu_{\overline{x}}$ and $\sigma_{\overline{x}}$, the mean and standard deviation of one such sampling distribution, matched calculated values (formulas on next page) to approximately two decimal places.[2]

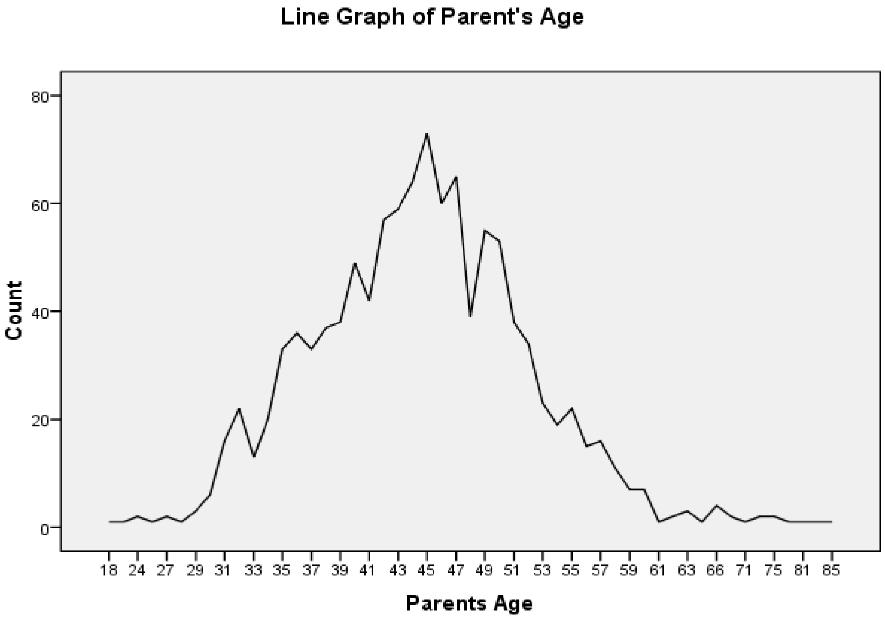
The large histogram is the distribution of individual values.

The small x-bars are distributions of sample averages for the individual values

If we collect enough sample averages, the averages will be normally distributed, and look like a normal curve

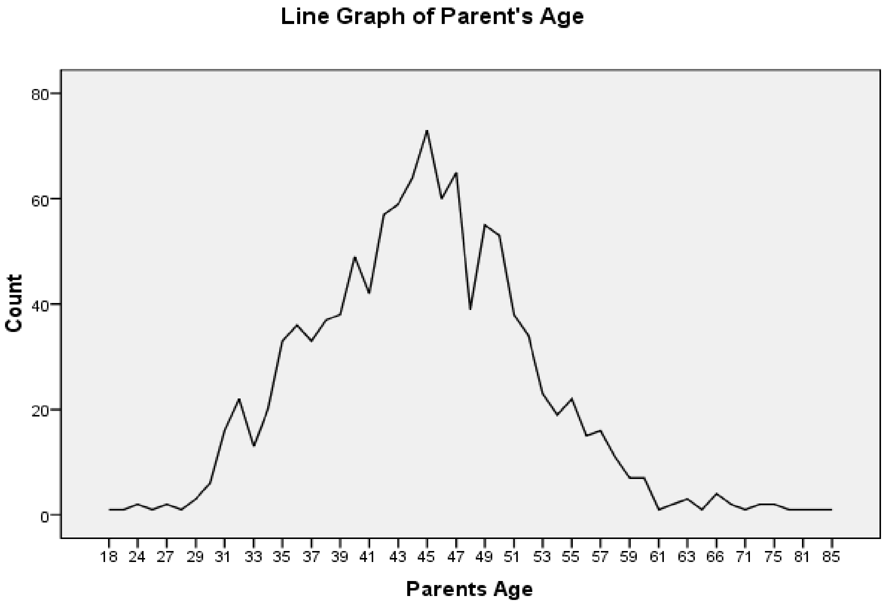
Because of this, distribution, we can make predictions

Taking Samples from Populations



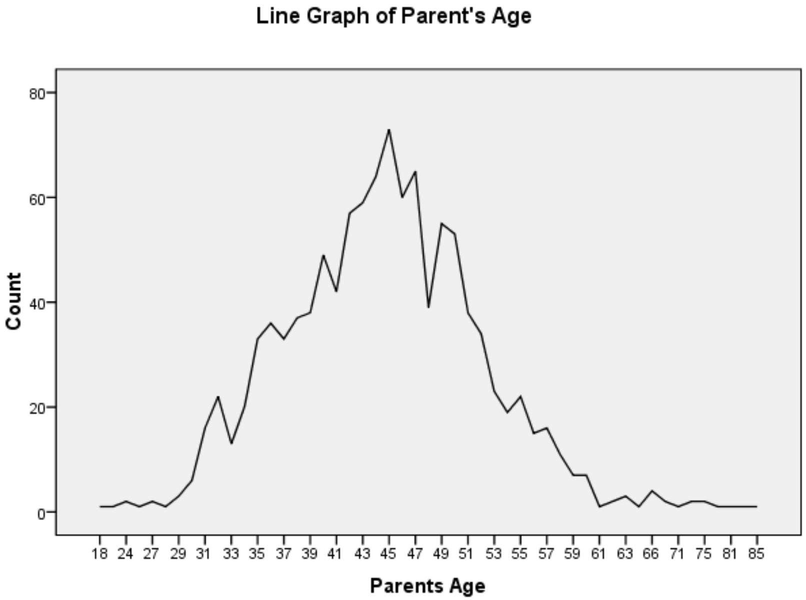
This is Parent’s Age from the data we used for our project. μ = 44.74. Let’s imagine that this is the true population value. (The N is large, so we can assume this.)

Taking Samples from Populations



Because the true population is normal, any sample we draw will have approx. the same mean as the true population, and a stand. dev. that decreases as the N of our sample increases.

Taking Samples from Populations



Pop. Mean = 44.74

Sample N = 92

$\mu_{\bar{X}}$ = 44.86

$\sigma_{\bar{X}} =$ .715

Sample N = 266

$\mu_{\bar{X}}$ = 44.32

$\sigma_{\bar{X}} =$ .453

Sample N = 502

$\mu_{\bar{X}}$ = 44.46

$\sigma_{\bar{X}} =$ .326

Standard deviation changes as N changes

Means of sample close to means of population

Keys to the Central Limit Theorem

- Here are two main things you should remember and use when proving agreement with the Central Limit Theorem:

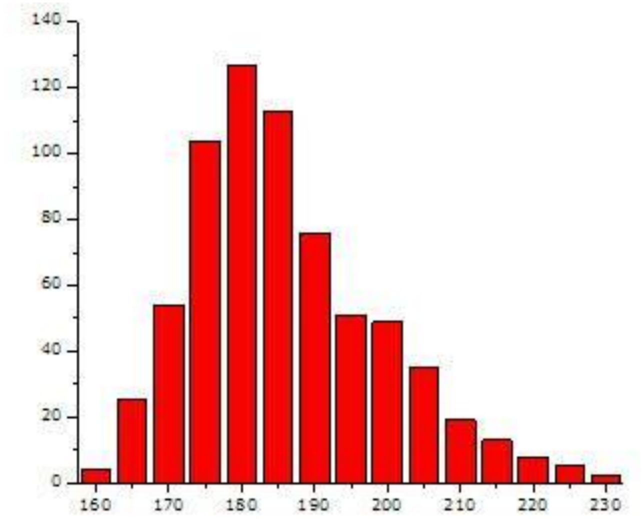
1) If the sample size is greater than 30, the Distribution of Sample Means will approximately follow a normal distribution REGARDLESS of the underlying population distribution.
   - If the underlying population distribution is Normally Distributed, the Distribution of Sample Means will be normally distributed REGARDLESS of the sample sizes used.
2) The mean of the Distribution of Sample Means will be the same as the underlying population’s mean. The standard deviation of the Distribution of Sample Means, however, will be the underlying population’s standard deviation divided by the square root of the sample size.

Keys to the Central Limit Theorem

- Proving agreement with the Central Limit Theorem

1) Show that the distribution of Sample Means is approximately normal (you could do this with a histogram)
    1) Remember this is true for any type of underlying population distribution if the sample size is greater than 30
    2) If the underlying population distribution is known to be Normally distributed ANY SAMPLE size will suffice
2) Show that the standard deviation of the Distribution of Sample Means is approximately the same as the underlying population's standard deviation divided by the square root of the sample size.

Keys to the Central Limit Theorem



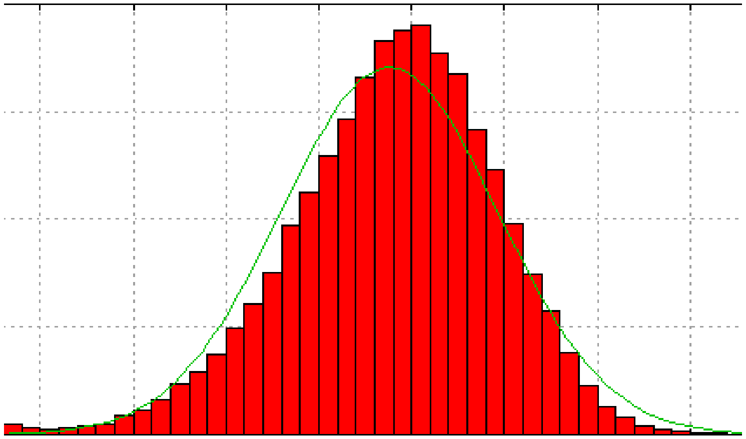
Let’s say we have a right skewed population distribution as shown on the left

Let’s say that we know the Population mean is 185 and the Population standard deviation is 15.

We take a group of 36 samples and find the mean is 184 and the standard deviation of the samples is 2.43.

On the following pages we will see if this follows the Central Limit Theorem.

Keys to the Central Limit Theorem



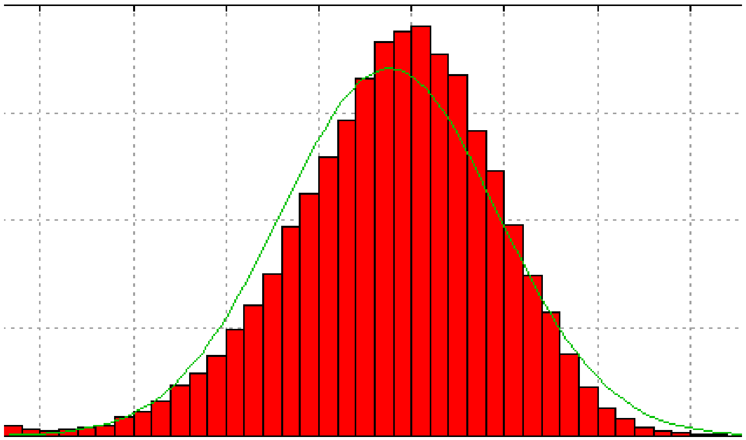
First let’s check the distribution of the sample means by looking at the histogram at the right. It is approximately normal, so YES this agrees with the Central Limit Theorem for the Population distribution even though it was “skewed”. The reason is that we had a sample of greater than 30 (we had 36)

Comparing the means, the Population’s was 185 versus the samples of 184 (about the same, yes it AGREES)

One more step on the following page.

Keys to the Central Limit Theorem

- Example Continued



The last step is to check the Standard deviations. Remember the POPULATION mean was 185 and the its standard deviation was 15.

We had a group of 36 samples and the mean was 184 and the standard deviation of the samples was 2.43.

We COMPARE the following:

Is the Population standard deviation divided by the square root of the number of samples approximately equal to the standard deviation of the samples?
15/sqrt(36) = 15/6 = 2.5
Is 2.5 close to 2.43?
YES, so again it agrees with the Central Limit Theorem!

CENTRAL LIMIT THEOREM

- specifies a theoretical distribution
- formulated by the selection of all possible random samples of a fixed size n
- a sample mean is calculated for each sample and the distribution of sample means is considered

SAMPLING DISTRIBUTION OF THE MEAN

- The mean of the sample means is equal to the mean of the population from which the samples were drawn.
- The variance of the distribution is σ divided by the square root of n. (the standard error.)

STANDARD ERROR

Standard Deviation of the Sampling Distribution of Means

$\sigma_x = \sigma / \sqrt{n}$

How Large is Large?

- If the sample 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.

Population Parameters and Sample Statistics

Population parameter	*Value*	*Sample statistic used to estimate*
p *proportion of population with a certain characteristic*	*Unknown*	$\hat{p}$
μ *mean value of a population variable*	*Unknown*	$\overline{x}$

- The value of a population parameter is a **fixed** number, it is NOT random; its value is **not known.**
- The value of a sample statistic is calculated from sample data
- The value of a sample statistic will vary from sample to sample (sampling distributions)

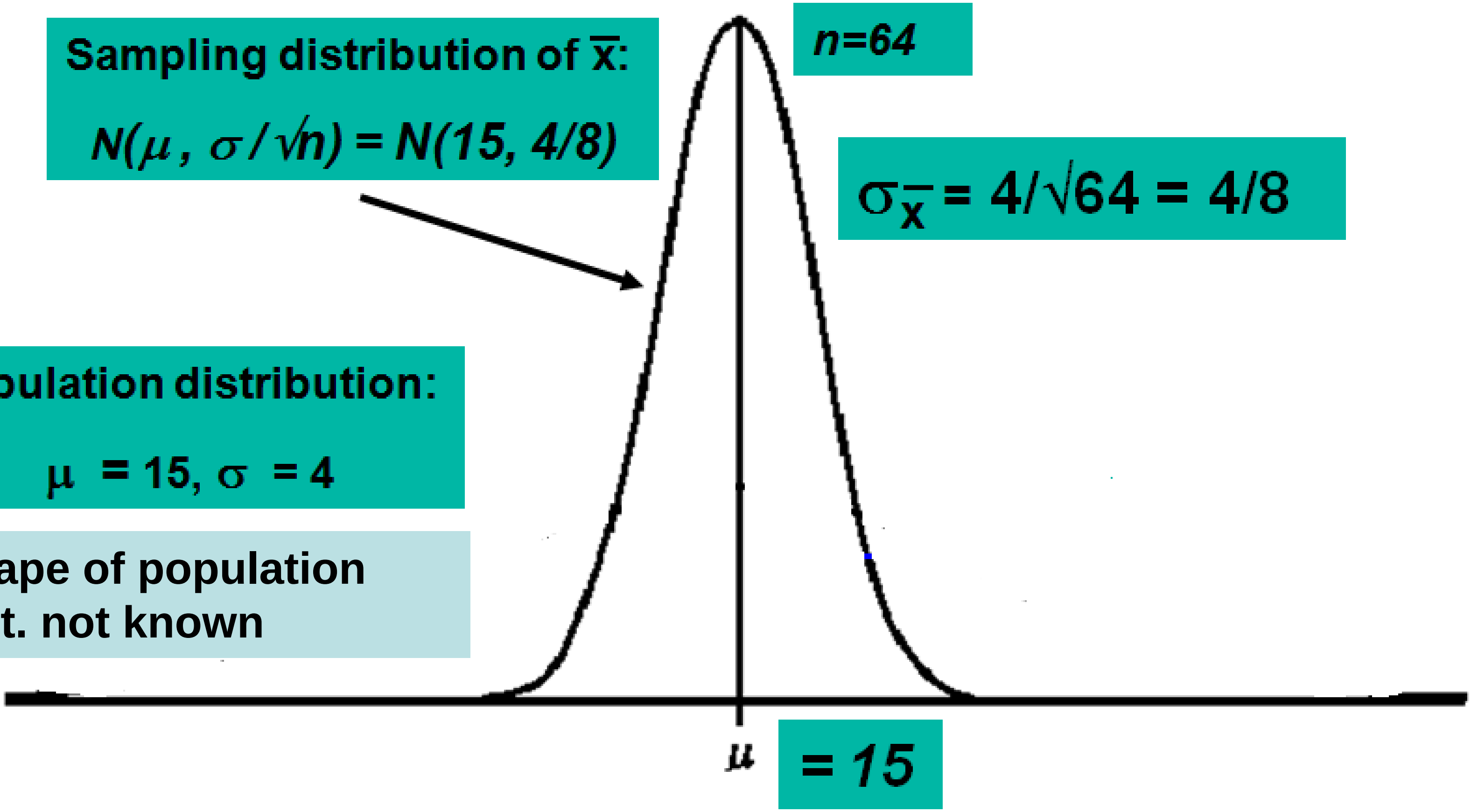
Example

A random sample of *n*=64 observations is drawn from a population with mean μ=15 and standard deviation σ=4.

a. $E(\bar{X}) = \mu = 15; SD(\bar{X}) = \frac{SD(X)}{\sqrt{n}} = \frac{4}{8} = .5$

b. The shape of the sampling distribution model for $\bar{x}$ is approx. normal (by the CLT) with mean $\mathrm{E}(\bar{\mathrm{X}}) = 15$ and $SD(\bar{X}) = .5$. The answer depends on the sample size since $SD(\bar{X}) = \frac{SD(X)}{\sqrt{n}}$.

Graphically



Sampling distribution of $\bar{x}$:

$N(\mu , \sigma / \sqrt{n}) = N(15, 4/8)$

n=64

$\sigma_{\bar{x}} = 4/\sqrt{64} = 4/8$

Population distribution:

$\mu = 15, \sigma = 4$

Shape of population dist. not known

= 15

Example (cont.)

c. $\bar{x} = 15.5;$

$$z = \frac{\bar{x} - \mu}{SD(\bar{X})} = \frac{15.5 – 15}{.5} = \frac{.5}{.5} = 1$$

This means that $\bar{x}$=15.5 is one standard deviation above the mean $E(\bar{X}) = 15$

EXAMPLE

A certain brand of tires has a mean life of 25,000 miles with a standard deviation of 1600 miles.

What is the probability that the mean life of 64 tires is less than 24,600 miles?

Example continued

The sampling distribution of the means has a mean of 25,000 miles (the population mean)

μ = 25000 mi.

and a standard deviation (i.e.. standard error) of:

1600/8 = 200

Example continued

Convert 24,600 mi. to a z-score and use the normal table to determine the required probability.

$$z = (24600-25000)/200 = -2$$

$$P(z < -2) = 0.0228$$

or 2.28% of the sample means will be less than 24,600 mi.

ESTIMATION OF POPULATION VALUES

- Point Estimates
- Interval Estimates

CONFIDENCE INTERVAL ESTIMATES for LARGE SAMPLES

- The sample has been randomly selected
- The population standard deviation is known or the sample size is at least 25.

Confidence Interval Estimate of the Population Mean

$$\bar{X} - z\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + z\frac{s}{\sqrt{n}}$$

$\bar{X}$: sample mean

s: sample standard deviation

n: sample size

EXAMPLE

Estimate, with 95% confidence, the lifetime of nine volt batteries using a randomly selected sample where:

--
X = 49 hours

s = 4 hours

n = 36

EXAMPLE continued

Lower Limit: 49 - (1.96)(4/6)
49 - (1.3) = 47.7 hrs

Upper Limit: 49 + (1.96)(4/6)
49 + (1.3) = 50.3 hrs

We are 95% confident that the mean lifetime of the population of batteries is between 47.7 and 50.3 hours.