The Central Limit Theorem and confidence interval

The Central Limit Theorem

Distribution of Sample Means

In addition to knowing how individual data values vary about the mean for a population, statisticians are interested in knowing about the distribution of the means of samples taken from a population.

Let's look at an example to illustrate the situation. Assume that the population under consideration consists of the heights of the starting five players on a basketball team:

Player	1	2	3	4	5
Height (in.)	67	73	75	76	84

Suppose we decide to take a sample of size n = 2. There are ten possible samples of size 2. They are listed in Table below

Sample	1	2	3	4	5	6	7	8	9	10
Players selected	1@2	1@3	1@4	1@5	2@3	2@4	2@5	3@4	3@5	4@5
Heights (in.)	67@73	67@75	67@76	67@84	73@75	73@76	73@84	75@76	75@76	76@84

Using our usual formula, we find $\mu=75$ and $\sigma=5.48$

If we consider example of size n = 4, then there are five possible outcomes:

Sample	1	2	3	4	5
Players selected	1, 2, 3, 4	1, 2, 3, 5	1, 2, 4, 5	1, 3, 4, 5	2, 3, 4, 5
Heights (in.)	67, 73, 75, 76	67, 73, 75, 84	67, 73, 76, 84	67, 75, 76, 84	73, 75, 76, 84

The mean and Standard Deviation of $\bar{x}$

The next step for describing the sampling distribution of the mean is to learn how to find the mean and standard deviation of the random variable $\bar{x}$**.** This is necessary in order to use normal curve methods to find probabilities for $\bar{x}$.

Let's consider random sample of size n = 2. There are ten possible samples of size 2. These ten samples are listed below, along with their $\bar{x}$ values

Sample	1	2	3	4	5	6	7	8	9	10
Players selected	1,2	1,3	1,4	1,5	2,3	2,4	2,5	3,4	3,5	4,5
Heights (in.)	67,73	67,75	67,76	67,84	73,75	73,76	73,84	75,76	75,76	76,84
$\bar{x}$	70.0	71.0	71.5	75.5	74.0	74.5	78.5	75.5	79.5	80.0

Since each sample has probability $\frac{1}{10}$ of being the one selected, the probability distribution of $\bar{x}$ is:

$\bar{x}$	70.0	71.0	71.5	74.0	74.5	75.5	78.5	79.5	80.0
P($\bar{x}$)	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1

To estimate $\mu_{\bar{x}}$ (mean of $\bar{x}$) and $\sigma_{\bar{x}}$ (standard deviation of $\bar{x}$), we use expected value method as shown below

$$\mu_{\bar{x}} = \sum \bar{x}P(\bar{x}) = (70.0).(0.1) + (71.0).(0.1) + \cdots + (80.0).(0.1) = 75.0$$

and $\sigma_{\bar{x}} = \sqrt{\sum(\bar{x} - \mu_{\bar{x}})P(\bar{x})} = 3.35$

Note that $\mu_{\bar{x}}$ = 75.0, which is the same as μ.

Is this a coincidence? Let's try samples size n = 4. There are five possible samples of size 4, as shown in Table below:

Sample	1	2	3	4	5
Players selected	1, 2, 3, 4	1, 2, 3, 5	1, 2, 4, 5	1, 3, 4, 5	2, 3, 4, 5
Heights (in.)	67, 73, 75, 76	67, 73, 75, 84	67, 73, 76, 84	67, 75, 76, 84	73, 75, 76, 84
$\bar{x}$	72.75	74.75	75.00	75.50	77.00

Since each of the five samples has probability 1/5 of being the one selected, the probability distribution of x is:

$\bar{x}$	72.75	74.75	75.00	75.50	77.00
$P(\bar{x})$	0.2	0.2	0.2	0.2	0.2

We computed the mean and standard deviation of $\boldsymbol{x}$ and found that $\mu=75$ *and* $\sigma=1.37$

CONCLUSIONS:

Suppose a random sample of size n is taken from a population with mean μ. Then the mean of $\bar{x}$ always equals to the mean of the population (regardless of sample size). The standard deviation of $\bar{x}$ is approximately equal to the standard deviation of the population divided by the square root of the sample size. That is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Central – limit Theorem

Suppose a random sample of size n (where n is at least 30) is taken from a population. Then $\bar{x}$ is (approximately) normally distributed with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. That is, probabilities for $\bar{x}$ can be found approximately by using areas under the normal curve with parameters μ and $\frac{\sigma}{\sqrt{n}}$.

If the sample size is sufficiently large, the central limit theorem can be used to answer questions about sample means in the same manner that the normal distribution can be used to answer questions about individual values. The only difference is that a new formula must be used for the z values. It is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Hours That Children Watch Television

A. C. Neilsen reported that children between the ages of 2 and 5 watch an average of 25 hours of television per week. Assume the variable is normally distributed and the standard deviation is 3 hours. If 20 children between the ages of 2 and 5 are randomly selected, find the probability that the mean of the number of hours they watch television will be greater than 26.3 hours.

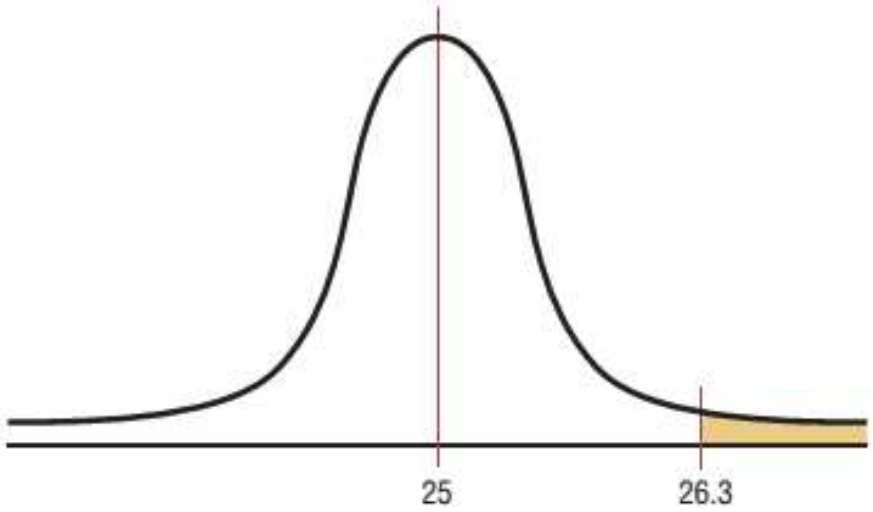
Source: Michael D. Shook and Robert L. Shook, *The Book of Odds*.

Solution

Since the variable is approximately normally distributed, the distribution of sample means will be approximately normal, with a mean of 25. The standard deviation of the sample means is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{20}} = 0.671$$

The distribution of the means is shown in Figure 6–32, with the appropriate area shaded.



The z value is

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{26.3 - 25}{3/\sqrt{20}} = \frac{1.3}{0.671} = 1.94$$

The area to the right of 1.94 is $1.000 - 0.9738 = 0.0262$, or 2.62%.

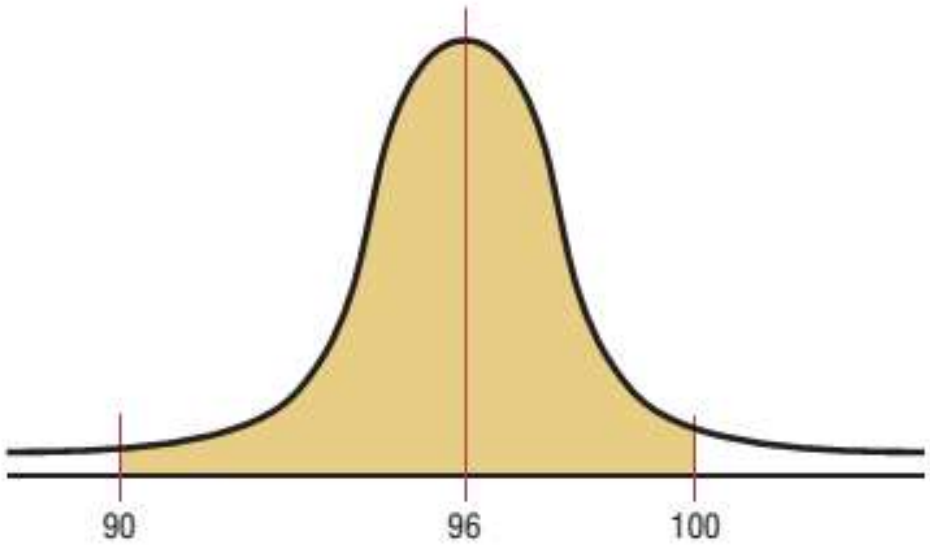
One can conclude that the probability of obtaining a sample mean larger than 26.3 hours is 2.62% [i.e., $P(\overline{X} > 26.3) = 2.62\%$].

The average age of a vehicle registered in the United States is 8 years, or 96 months. Assume the standard deviation is 16 months. If a random sample of 36 vehicles is selected, find the probability that the mean of their age is between 90 and 100 months.

Source: *Harper's Index.*

Solution

Since the sample is 30 or larger, the normality assumption is not necessary. The desired area is shown in Figure 6–33.



The two z values are

$$z_1 = \frac{90 - 96}{16/\sqrt{36}} = -2.25$$

$$z_2 = \frac{100 - 96}{16/\sqrt{36}} = 1.50$$

To find the area between the two z values of -2.25 and 1.50, look up the corresponding area in Table E and subtract one from the other. The area for $z = -2.25$ is 0.0122, and the area for $z = 1.50$ is 0.9332. Hence the area between the two values is $0.9332 - 0.0122 = 0.9210$, or 92.1%.

Hence, the probability of obtaining a sample mean between 90 and 100 months is 92.1%; that is, $P(90 < \overline{X} < 100) = 92.1\%$.

Meat Consumption

The average number of pounds of meat that a person consumes per year is 218.4 pounds. Assume that the standard deviation is 25 pounds and the distribution is approximately normal.

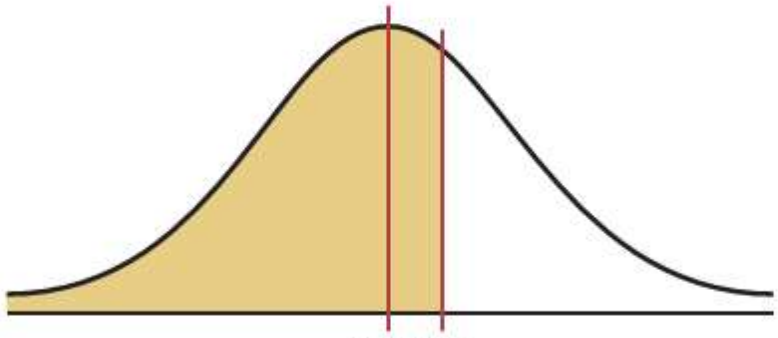
Source: Michael D. Shook and Robert L. Shook, *The Book of Odds*.

a. Find the probability that a person selected at random consumes less than 224 pounds per year.

b. If a sample of 40 individuals is selected, find the probability that the mean of the sample will be less than 224 pounds per year.

Solution

a. Since the question asks about an individual person, the formula $z = (X - \mu)/\sigma$ is used. The distribution is shown in Figure 6–34.



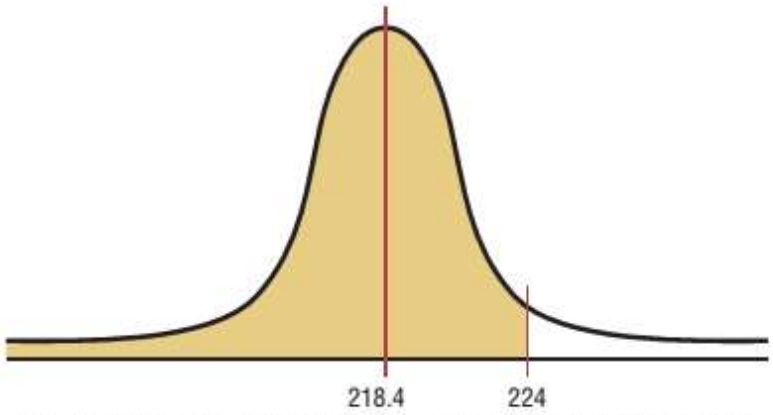
Distribution of individual data values for the population

The z value is

$$z = \frac{X - \mu}{\sigma} = \frac{224 - 218.4}{25} = 0.22$$

The area to the left of $z = 0.22$ is 0.5871. Hence, the probability of selecting an individual who consumes less than 224 pounds of meat per year is 0.5871, or 58.71% [i.e., $P(X < 224) = 0.5871$].

b. Since the question concerns the mean of a sample with a size of 40, the formula $z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ is used. The area is shown in Figure 6–35.



Distribution of means for all samples of size 40 taken from the population

The z value is

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{224 - 218.4}{25/\sqrt{40}} = 1.42$$

The area to the left of $z = 1.42$ is 0.9222.

Confidence Intervals and Sample Size

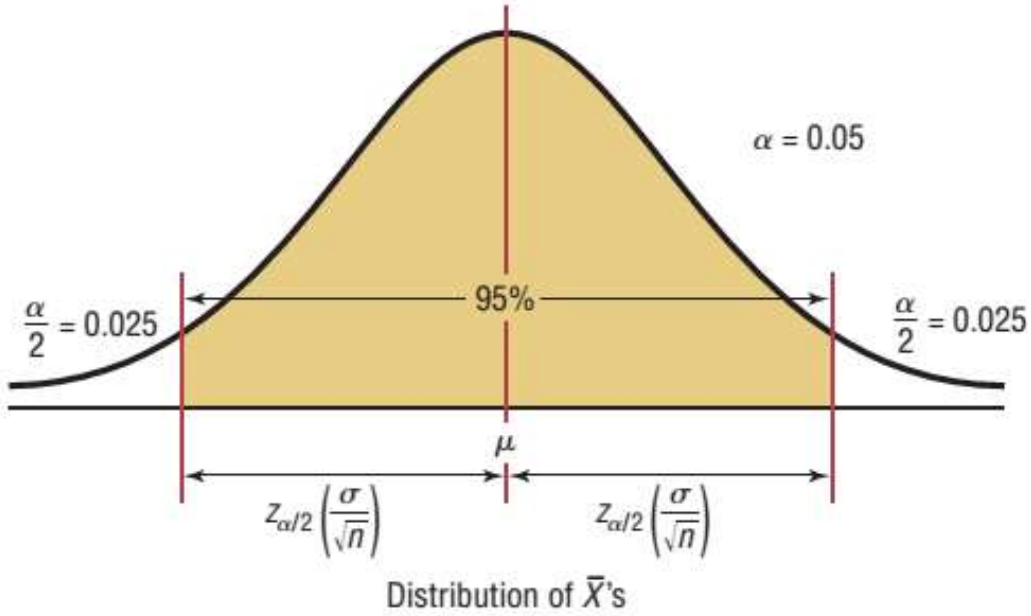
Confidence Intervals for the Mean When σ Is Known

Formula for the Confidence Interval of the Mean for a Specific α When σ is Known

$$\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

For a 90% confidence interval, $z_{\alpha/2} = 1.65$; for a 95% confidence interval, $z_{\alpha/2} = 1.96$; and for a 99% confidence interval, $z_{\alpha/2} = 2.58$.

The term $z_{\alpha/2}(\sigma/\sqrt{n})$ is called the *margin of error* (also called the *maximum error of the estimate*). For a specific value, say, $\alpha = 0.05$, 95% of the sample means will fall within this error value on either side of the population mean, as previously explained. See Figure 7–1.



When $n \geq 30$, s can be substituted for σ, but a different distribution is used.

Days It Takes to Sell an Aveo

A researcher wishes to estimate the number of days it takes an automobile dealer to sell a Chevrolet Aveo. A sample of 50 cars had a mean time on the dealer's lot of 54 days. Assume the population standard deviation to be 6.0 days. Find the best point estimate of the population mean and the 95% confidence interval of the population mean.

Source: Based on information obtained from Power Information Network.

Solution

The best point estimate of the mean is 54 days. For the 95% confidence interval use $z = 1.96$.

$$\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$54 - 1.96\left(\frac{6.0}{\sqrt{50}}\right) < \mu < 54 + 1.96\left(\frac{6.0}{\sqrt{50}}\right)$$

$$54 - 1.7 < \mu < 54 + 1.7$$

$$52.3 < \mu < 55.7 \text{ or } 54 \pm 1.7$$

Hence one can say with 95% confidence that the interval between 52.3 and 55.7 days does contain the population mean, based on a sample of 50 automobiles.

Waiting Times in Emergency Rooms

A survey of 30 emergency room patients found that the average waiting time for treatment was 174.3 minutes. Assuming that the population standard deviation is 46.5 minutes, find the best point estimate of the population mean and the 99% confidence of the population mean.

Source: Based on information from Press Ganey Associates Inc.

Solution

The best point estimate is 174.3 minutes. The 99% confidence is interval is

$$\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$174.3 - 2.58\left(\frac{46.5}{\sqrt{30}}\right) < \mu < \bar{X} + 2.58\left(\frac{46.5}{\sqrt{30}}\right)$$

$$174.3 - 21.9 < \mu < 174.3 + 21.9$$

$$152.4 < \mu < 196.2$$

Hence, one can be 99% confident that the mean waiting time for emergency room treatment is between 152.4 and 196.2 minutes.

Credit Union Assets

The following data represent a sample of the assets (in millions of dollars) of 30 credit unions in southwestern Pennsylvania. Find the 90% confidence interval of the mean.

12.23	16.56	4.39
2.89	1.24	2.17
13.19	9.16	1.42
73.25	1.91	14.64
11.59	6.69	1.06
8.74	3.17	18.13
7.92	4.78	16.85
40.22	2.42	21.58
5.01	1.47	12.24
2.27	12.77	2.76

Source: *Pittsburgh Post Gazette.*

Solution

Step 1 Find the mean and standard deviation for the data. Use the formulas shown in Chapter 3 or your calculator. The mean $\overline{X} = 11.091$. Assume the standard deviation of the population is 14.405.

Step 2 Find $\alpha/2$. Since the 90% confidence interval is to be used, $\alpha = 1 - 0.90 = 0.10$, and

$$\frac{\alpha}{2} = \frac{0.10}{2} = 0.05$$

Step 3 Find $z_{\alpha/2}$. Subtract 0.05 from 1.000 to get 0.9500. The corresponding z value obtained from Table E is 1.65. (*Note:* This value is found by using the z value for an area between 0.9495 and 0.9505. A more precise z value obtained mathematically is 1.645 and is sometimes used; however, 1.65 will be used in this textbook.)

Step 4 Substitute in the formula

$$\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$11.091 - 1.65\left(\frac{14.405}{\sqrt{30}}\right) < \mu < 11.091 + 1.65\left(\frac{14.405}{\sqrt{30}}\right)$$

$$11.091 - 4.339 < \mu < 11.091 + 4.339$$

$$6.752 < \mu < 15.430$$

Hence, one can be 90% confident that the population mean of the assets of all credit unions is between \$6.752 million and \$15.430 million, based on a sample of 30 credit unions.

Sample Size

Sample size determination is closely related to statistical estimation. Quite often you ask, How large a sample is necessary to make an accurate estimate? The answer is not simple, since it depends on three things: the margin of error, the population standard deviation, and the degree of confidence. For example, how close to the true mean do you want to be (2 units, 5 units, etc.), and how confident do you wish to be (90, 95, 99%, etc.)? For the purpose of this chapter, it will be assumed that the population standard deviation of the variable is known or has been estimated from a previous study.

The formula for sample size is derived from the margin of error formula

$$E = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

and this formula is solved for n as follows:

$$E\sqrt{n} = z_{\alpha/2}(\sigma)$$

$$\sqrt{n} = \frac{z_{\alpha/2} \cdot \sigma}{E}$$

Hence, $$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

Depth of a River

A scientist wishes to estimate the average depth of a river. He wants to be 99% confident that the estimate is accurate within 2 feet. From a previous study, the standard deviation of the depths measured was 4.33 feet.

Solution

Since $\alpha = 0.01$ (or $1 - 0.99$), $z_{\alpha/2} = 2.58$ and $E = 2$. Substituting in the formula,

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left[\frac{(2.58)(4.33)}{2}\right]^2 = 31.2$$

Round the value 31.2 up to 32. Therefore, to be 99% confident that the estimate is within 2 feet of the true mean depth, the scientist needs at least a sample of 32 measurements.

In most cases in statistics, we round off. However, when determining sample size, we always round up to the next whole number.

Confidence Intervals for the Mean When σ Is Unknown

When σ is known and the sample size is 30 or more, or the population is normally distributed if the sample size is less than 30, the confidence interval for the mean can be found by using the z distribution as shown in Section 7–1. However, most of the time, the value of σ is not known, so it must be estimated by using s, namely, the standard deviation of the sample. When s is used, especially when the sample size is small, critical values greater than the values for $z_{\alpha/2}$ are used in confidence intervals in order to keep the interval at a given level, such as the 95%. These values are taken from the *Student t distribution,* most often called the ***t* distribution.**

To use this method, the samples must be simple random samples, and the population from which the samples were taken must be normally or approximately normally distributed, or the sample size must be 30 or more.

Some important characteristics of the t distribution are described now.

Confidence Intervals for the Mean When σ Is Unknown

Characteristics of the *t* Distribution

The *t* distribution shares some characteristics of the normal distribution and differs from it in others. The *t* distribution is similar to the standard normal distribution in these ways:

1. It is bell-shaped.
2. It is symmetric about the mean.
3. The mean, median, and mode are equal to 0 and are located at the center of the distribution.
4. The curve never touches the *x* axis.

The *t* distribution differs from the standard normal distribution in the following ways:

1. The variance is greater than 1.
2. The *t* distribution is actually a family of curves based on the concept of *degrees of freedom,* which is related to sample size.
3. As the sample size increases, the *t* distribution approaches the standard normal distribution. See Figure 7–6.

Confidence Intervals for the Mean When σ Is Unknown

Formula for a Specific Confidence Interval for the Mean When σ Is Unknown

$$\bar{X} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{X} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

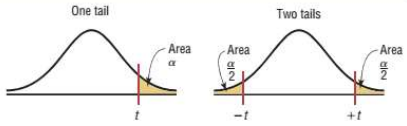
The degrees of freedom are $n - 1$.

Table F The t Distribution

	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169
11		1.363	1.796	2.201	2.718	3.106
12		1.356	1.782	2.179	2.681	3.055
13		1.350	1.771	2.160	2.650	3.012
14		1.345	1.761	2.145	2.624	2.977
15		1.341	1.753	2.131	2.602	2.947
16		1.337	1.746	2.120	2.583	2.921
17		1.333	1.740	2.110	2.567	2.898
18		1.330	1.734	2.101	2.552	2.878
19		1.328	1.729	2.093	2.539	2.861
20		1.325	1.725	2.086	2.528	2.845
21		1.323	1.721	2.080	2.518	2.831
22		1.321	1.717	2.074	2.508	2.819
23		1.319	1.714	2.069	2.500	2.807
24		1.318	1.711	2.064	2.492	2.797
25		1.316	1.708	2.060	2.485	2.787
26		1.315	1.706	2.056	2.479	2.779
27		1.314	1.703	2.052	2.473	2.771
28		1.313	1.701	2.048	2.467	2.763
29		1.311	1.699	2.045	2.462	2.756
30		1.310	1.697	2.042	2.457	2.750
32		1.309	1.694	2.037	2.449	2.738
34		1.307	1.691	2.032	2.441	2.728
36		1.306	1.688	2.028	2.434	2.719
38		1.304	1.686	2.024	2.429	2.712
40		1.303	1.684	2.021	2.423	2.704
45		1.301	1.679	2.014	2.412	2.690
50		1.299	1.676	2.009	2.403	2.678
55		1.297	1.673	2.004	2.396	2.668
60		1.296	1.671	2.000	2.390	2.660
65		1.295	1.669	1.997	2.385	2.654
70		1.294	1.667	1.994	2.381	2.648
75		1.293	1.665	1.992	2.377	2.643
80		1.292	1.664	1.990	2.374	2.639
90		1.291	1.662	1.987	2.368	2.632
100		1.290	1.660	1.984	2.364	2.626
500		1.283	1.648	1.965	2.334	2.586
1000		1.282	1.646	1.962	2.330	2.581
$(z)\ \infty$		1.282[a]	1.645[b]	1.960	2.326[c]	2.576[d]

[a] This value has been rounded to 1.28 in the textbook.
[b] This value has been rounded to 1.65 in the textbook.
[c] This value has been rounded to 2.33 in the textbook.
[d] This value has been rounded to 2.58 in the textbook.

Source: Adapted from W. H. Beyer, *Handbook of Tables for Probability and Statistics*, 2nd ed., CRC Press, Boca Raton, Fla., 1986. Reprinted with permission.



Find the $t_{\alpha/2}$ value for a 95% confidence interval when the sample size is 22.

Solution

The d.f. = 22 − 1, or 21. Find 21 in the left column and 95% in the row labeled Confidence Intervals. The intersection where the two meet gives the value for $t_{\alpha/2}$, which is 2.080. See Figure 7–7.

Table F

The *t* Distribution

d.f.	Confidence Intervals	50%	80%	90%	95%	98%	99%
	One tail α	0.25	0.10	0.05	0.025	0.01	0.005
	Two tails α	0.50	0.20	0.10	0.05	0.02	0.01
1							
2							
3							
⋮							
21					2.080	2.518	2.831
⋮							
$(z)\infty$		0.674	1.282[a]	1.645[b]	1.960	2.326[c]	2.576[d]

Sleeping Time

Ten randomly selected people were asked how long they slept at night. The mean time was 7.1 hours, and the standard deviation was 0.78 hour. Find the 95% confidence interval of the mean time. Assume the variable is normally distributed.

Source: Based on information in *Number Freaking*.

Solution

Since σ is unknown and s must replace it, the t distribution (Table F) must be used for the confidence interval. Hence, with 9 degrees of freedom $t_{\alpha/2} = 2.262$. The 95% confidence interval can be found by substituting in the formula.

$$\bar{X} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{X} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$7.1 - 2.262\left(\frac{0.78}{\sqrt{10}}\right) < \mu < 7.1 + 2.262\left(\frac{0.78}{\sqrt{10}}\right)$$

$$7.1 - 0.56 < \mu < 7.1 + 0.56$$

$$6.54 < \mu < 7.66$$

Therefore, one can be 95% confident that the population mean is between 6.54 and 7.66 inches.

Home Fires Started by Candles

The data represent a sample of the number of home fires started by candles for the past several years. (Data are from the National Fire Protection Association.) Find the 99% confidence interval for the mean number of home fires started by candles each year.

5460 5900 6090 6310 7160 8440 9930

Solution

Step 1 Find the mean and standard deviation for the data. Use the formulas in Chapter 3 or your calculator. The mean $\overline{X} = 7041.4$. The standard deviation $s = 1610.3$.

Step 2 Find $t_{\alpha/2}$ in Table F. Use the 99% confidence interval with d.f. $= 6$. It is 3.707.

Step 3 Substitute in the formula and solve.

$$\overline{X} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \overline{X} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$7041.4 - 3.707\left(\frac{1610.3}{\sqrt{7}}\right) < \mu < 7041.4 + 3.707\left(\frac{1610.3}{\sqrt{7}}\right)$$

$$7041.4 - 2256.2 < \mu < 7041.4 + 2256.2$$

$$4785.2 < \mu < 9297.6$$