Interval Estimation

7

Chapter Outline

In Chapter 5 we saw that variation occurs when we use a sample instead of the entire population. For example, in the presentation of the binomial distribution, we saw that the sample estimates of the population proportion varied considerably from sample to sample. In this chapter, we present *prediction, confidence*, and *tolerance intervals*, quantities that allow us to take the variation in sample results into account in describing the data. These intervals represent specific types of *interval estimation* — the provision of limits that are likely to contain either (1) the population parameter of interest or (2) future observations of the variable. Interval estimation thus provides more information about the population parameter than the *point estimation* approach that we met in Chapter 3. In that chapter, we provided a single value as the estimate of the population parameter without giving any information about the sampling variability of the estimator. For example, knowledge of the value of the sample mean, a point estimate of the population mean, does not tell us anything about the variability of the sample mean. Interval estimation addresses this variability.

7.1 Prediction, Confidence, and Tolerance Intervals

The material in this and the following section is based on material presented by Vardeman (1992) and Walsh (1962). To understand the difference between these three intervals (prediction, confidence, and tolerance), consider the following. Dairies add vitamin D to milk for the purpose of fortification. The recommended amount of vitamin D to be added to a quart of milk is 400 IUs (10 μg). If a dairy adds too much vitamin D, perhaps over 5000 IUs, the possibility exists that a consumer will develop hypervitaminosis D — that is, vitamin D toxicity.

A *prediction interval* focuses on a single observation of the variable — for example, the amount of vitamin D in the next bottle of milk. A *confidence interval* focuses on a population parameter — for example, the mean or median amount of vitamin D per bottle in a population of bottles of milk. Thus, the prediction interval is of more interest

to the consumer of the next bottle of milk, whereas the confidence interval is of more interest to the dairy. A *tolerance interval* provides limits such that there is a high level of confidence that a large proportion of values of the variable will fall within them. For example, besides being interested in the mean, the dairy owner or a regulatory agency also wants to be confident that a large proportion of the bottles' vitamin D contents are within a specified tolerance of the value of 400 IUs. We begin our treatment of these intervals with distribution-free intervals.

7.2 Distribution-Free Intervals

When the method for forming the different intervals is independent of how the data are distributed, the resultant intervals are said to be *distribution free*. Distribution-free intervals are based on the rank order of the sample values, with the following notations for rank order. The smallest of the x values is indicated by $x_{(1)}$, the second smallest by $x_{(2)}$, and so on, to the largest value that is denoted by $x_{(n)}$. The $x_{(i)}$ are called *order statistics*, since the subscripts show the order of the values.

We shall use hypothetical data showing the amount of vitamin D in 30 bottles of milk selected at random from one dairy. The values are shown in rank order in Table 7.1.

Based on this sample, $x_{(1)}$ equals 289 IUs, $x_{(2)}$ is 326 IUs and so on to $x_{(30)}$, which equals 485 IUs.

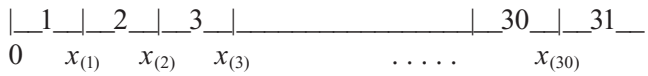
Table 7.1 Values of vitamin D (IUs) in a hypothetical sample of 30 bottles.

289	355	376	392	406	433
326	363	379	395	410	434
339	364	384	396	413	456
346	370	386	398	422	471
353	373	389	403	427	485

7.2.1 Prediction Interval

As a consumer of milk, our major concern about vitamin D is that the milk does not contain an amount of vitamin D that is toxic to us. We are not too concerned about there being too little vitamin D in the bottle. Based on the hypothetical sample of vitamin D contents in 30 bottles of milk, we can form a one-sided prediction interval — our concern focuses on the upper limit — for the amount of vitamin D in the bottle of milk that we are going to purchase.

A natural one-sided prediction interval in this case is from 0 to the maximum observed value of vitamin D (485 IUs) in the sample. The level of confidence associated with this interval, from 0 to 485 IUs, is 96.8 percent (= 30/31). This value can be found from the consideration of the order statistics and the real number line. For example, we have the line



and there are 31 intervals along this line. The vertical marks (|) indicate the location of the order statistics along the line, and the numbers above the line between the |'s indicate the interval number. There are 31 intervals, and the next observation can fall into any one of the intervals. Of these 31 intervals, 30 have values less than the maximum value. Hence, we are 96.8 percent confident that the vitamin D content in the next bottle will be between zero and the observed maximum value.

Note that we used the word *confidence* instead of *probability* here. We use confidence because we are using the sample data as the basis of estimating the probability distribution of the vitamin D content. If we used the probability distribution of the vitamin D content instead of using its sample estimate, the empirical distribution function, we would use the word *probability*. In repeated sampling, we expect that 96.8 percent of the prediction intervals, ranging from zero to the observed maximum in each sample of size 30, would contain the next observed vitamin D content.

The use of the second largest value, $x_{(29)}$, as the upper limit of the interval results in a prediction confidence level of 93.5 percent (= 29/31). An attraction of this interval is that it provides a slightly shorter interval with a maximum of 471 IUs, but we are slightly less confident about it. Based on either of these intervals, the consumer should not be worried about purchasing a bottle that has a value of vitamin D that would cause vitamin D poisoning.

For a two-sided interval, a natural interval would be from the minimum observed value, $x_{(1)}$, to the maximum observed value, $x_{(30)}$. In this case, the two-sided interval is from 289 to 485 IUs. The confidence level associated with this prediction interval is 93.5 percent (= 29/31). Of the 31 intervals just shown, there is one below the minimum value and one above the maximum value. Hence, there are 29 chances out of 31 that the next observed value will fall between the minimum and maximum values.

With a sample size of 30, it is not possible to have a distribution-free, two-sided, 95 percent prediction interval. The smallest sample size that attains the 95 percent level is 39. When n is 39, there are 40 intervals, and 2/40 equals 0.05. This calculation shows that it is easy to determine how large a sample is required to satisfy prediction interval requirements.

7.2.2 Confidence Interval

The dairy wants to know, on average, how much vitamin D is being added to the milk. If the interval estimate for the central tendency differs much from 400 IUs, the dairy may have to change its process for adding vitamin D. One way of obtaining the interval estimate is to use a distribution-free confidence interval.

Distribution-free confidence intervals are used to provide information about population parameters — for example, the median and other percentiles. There are two approaches to finding confidence intervals for percentiles: (1) the use of order statistics and (2) the use of the normal approximation to the binomial distribution. The first approach is generally used for smaller samples, whereas the second approach is used for larger samples.

***Use of Order Statistics and the Binomial Distribution*:** The lower and upper limits of the $(1 - \alpha)100$ percent confidence interval for the pth percentile of X are the order

statistics $x_{(j)}$ and $x_{(k)}$, where the values of j and k, j less than k, are to be determined. The limits of the confidence interval for the pth percentile of X are the values $x_{(j)}$ and $x_{(k)}$ that satisfy the following inequality:

$$\Pr\{x_{(j)} < p^{th}\ percentile < x_{(k)}\} \geq 1 - \alpha$$

and this is equivalently

$$\Pr\{x_{(j)} \geq p^{th}\ percentile\} + \Pr\{x_{(k)} \leq p^{th}\ percentile\} \leq \alpha$$

If we require that both terms in the sum be less than or equal to $\alpha/2$, from the first term, we have

$$\Pr\{at\ most\ j-1\ observations < p^{th}\ percentile\} \leq \alpha/2.$$

This is a situation with two outcomes: an observation is less than the pth percentile, or it is greater than or equal to the pth percentile. The probability that an observation is less than the pth percentile is p. The variable of interest is the number of observations, out of the n, that are less than the pth percentile. Thus, this variable follows a binomial distribution with parameters n and p. Knowing the values of n and p enables us to find the value of j because j must satisfy the following inequality:

$$\sum_{i=0}^{j-1} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \leq \alpha/2.$$

The inequality used to find the value of k is

$$\sum_{i=k}^{n} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \leq \alpha/2.$$

Putting these two inequalities together means that the binomial sum from j to $k-1$ must be greater than or equal to $1-\alpha$. Here we have dropped the requirement that the sums of the probabilities from 0 to $j-1$ and from k to n both must be less than $\alpha/2$. The values of j and k are found from the binomial table, Table B2, or by using a computer package such as SAS or Stata.

For example, suppose we want to find a 95 percent confidence interval for the median, the 50th percentile, for the vitamin D values from the dairy used in Table 7.1. The sample estimate of the median is the average of the 15th and 16th smallest values — that is, 390.5 IUs (= [389 + 392]/2).

To find the 95 percent confidence interval for the median in the population of bottles of milk from the selected dairy, we use the binomial distribution. For this problem we need a binomial distribution with $n = 30$ and $\pi = 0.5$, shown in Table 7.2. Since Table B2 does not have values for n larger than 20, we used SAS to obtain the distribution. The order and observations from Table 7.1 are also shown in the last two columns in Table 7.2. There may be more than one pair of values of j and k that satisfy the requirement that the sum of the binomial probabilities from j to $k-1$ is greater than or equal to $1-\alpha$. To choose from among these pairs, we shall select the pair whose difference $(k-j)$ is the smallest. In the special case of the median, we shall require that k equals $n-j+1$; this requirement gives the same number of observations in both tails of the distribution.

Table 7.2 Cumulative binomial distribution with $n = 30$ and $\pi = 0.5$ and sorted observations in Table 7.1.

x	Pr ($X \leq x$)	No.	Observation
0	0.0000	1	289
1	0.0000	2	326
2	0.0000	3	339
3	0.0000	4	346
4	0.0000	5	353
5	0.0002	6	355
6	0.0007	7	363
7	0.0026	8	364
8	0.0081	9	370
9	0.0214	10	373
10	0.0494	11	376
11	0.1002	12	379
12	0.1808	13	384
13	0.2923	14	386
14	0.4278	15	389
15	0.5722	16	392
16	0.7077	17	395
17	0.8192	18	396
18	0.8998	19	398
19	0.9506	20	403
20	0.9786	21	406
21	0.9919	22	410
22	0.9974	23	413
23	0.9993	24	422
24	0.9998	25	427
25	1.0000	26	433
26	1.0000	27	434
27	1.0000	28	456
28	1.0000	29	471
29	1.0000	30	485
30	1.0000		

The sum of the probabilities from j to $k - 1$ must be greater than or equal to 0.95. Examination of the cumulative probabilities tells us that j is 10 and k is 21. The sum of the probabilities between 10 and 20 is 0.9572 (= 0.9786 − 0.0214). If j were 11 and k were 20, the sum of the probabilities between 11 and 19 is 0.9012, less than the required value of 0.95. Thus, the approximate 95 percent (really closer to 96%) confidence interval for the median is from 373 IUs ($= x_{(10)}$) to 406 IUs ($= x_{(21)}$). The use of distribution-free intervals does not necessarily provide intervals that are symmetric about the sample estimator. For example, the sample median value, 390.5 IUs, is not in the exact middle of the confidence interval.

Note that the confidence interval for the median is much narrower than the approximate 95 percent prediction interval, from 289 to 485 IUs, for a single observation. As we saw in Chapter 3, there is much less variability associated with a mean or median than with a single observation, and this is additional confirmation of that.

As we can observe from the preceding, the use of distribution-free intervals does not provide exactly 95 percent levels. The level of confidence associated with these intervals is a function of the sample size as well as which order statistics are used in the creation of the interval.

It is also possible to create one-sided confidence intervals for parameters. For example, if the goal were to create an upper one-sided confidence interval for the median, we would find the value of k such that

$$\sum_{i=k}^{n} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \leq \alpha$$

for a p having the value of 0.50. The upper one-sided confidence interval for the median is from 0 to $x_{(k)}$ where k's value is found from the above inequality.

***Use of the Normal Approximation to the Binomial*:** For larger sample sizes, the normal approximation to the binomial distribution can be used to find the values of j and k. The sample size must be large enough to satisfy the requirements for the use of the normal approximation. Since p is 0.50, the sample size of 30 bottles from the dairy is large enough.

As before, we want to find the value of j such that the probability of the binomial variable, Y, being less than or equal to $j - 1$ is less than or equal to $\alpha/2$ — that is,

$$\Pr \{Y \leq j - 1\} \leq \alpha/2.$$

Use of the continuity correction converts this to

$$\Pr \{Y \leq j - 0.5\} \leq \alpha/2.$$

To convert Y to the standard normal variable, we must subtract np, the estimate of the mean, and divide by $\sqrt{np(1-p)}$, the estimate of the standard error. This yields

$$\Pr\left\{\frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{j - 0.5 - np}{\sqrt{np(1-p)}}\right\} \leq \frac{\alpha}{2}.$$

This can be rewritten as

$$\Pr\left\{Z \leq \frac{j - 0.5 - np}{\sqrt{np(1-p)}}\right\} \leq \frac{\alpha}{2}.$$

If we change this inequality to equality — that is, the probability is equal to $\alpha/2$ — we can find a unique value for j. The value of the term on the right side of the inequality inside the brackets is simply $z_{\alpha/2}$, and hence we can find the value of j from the equation

$$j - 0.5 - np = z_{\alpha/2}\sqrt{np(1-p)}$$

or

$$j = z_{\alpha/2}\sqrt{np(1-p)} + 0.5 + np.$$

In the preceding example, p was 0.50, n was 30, and α was 0.05. Since the value of $z_{0.025}$ is −1.96, we have

$$j = -1.96\sqrt{30(0.5)(0.5)} + 0.5 + 30(0.5)$$

or j is 10.13. To ensure that the level of the confidence interval is at least $(1 - \alpha) * 100$ percent, we must round down the value of j to the next smaller integer, 10, and we round up the value of k, found following, to the next larger integer.

The value of k is found from the equation

$$k = z_{1-\alpha/2}\sqrt{np(1-p)} + 0.5 + np$$

which yields a k equal to 20.87, which is rounded to 21. Thus, the 95 percent confidence interval is from 373 IUs ($= x_{(10)}$) to 406 IUs ($= x_{(21)}$). In this case, the binomial and the normal approximation approaches resulted in the same confidence limits.

7.2.3 Tolerance Interval

As we said before, tolerance intervals are of most interest to the dairy or to a regulatory agency. The tolerance limits are values such that we have a high level of confidence that a large proportion of the bottles have vitamin D contents located between the lower and upper tolerance limits. These upper and lower limits of the tolerance interval can be used in determining whether or not the process for adding vitamin D is under control. If the limits are too wide, the dairy may have to modify its process for adding vitamin D to the milk.

The dairy does not want to add too much vitamin D to the milk because of the possible problems for the consumer and the extra cost associated with using more vitamin D than required. At the same time, the dairy must add enough vitamin D to be in compliance with truth in advertising legislation.

As with the prediction interval, it is reasonable to use the smallest and largest observed values for the lower and upper limits of the tolerance interval, although other values could be used. We also have to specify the proportion of the population, p, that we want to include within the tolerance interval. Given the tolerance interval limits and the proportion of values to be included within it, we can calculate the confidence level, γ, associated with the interval.

In symbols, the tolerance interval limits are the order statistics $x_{(j)}$ and $x_{(k)}$ such that

$$\Pr\left[\Pr\{X \le x_{(k)}\} - \Pr\{X \le x_{(j)}\} \ge p\right] = \gamma.$$

The quantity, $\Pr\{X \le x_{(k)}\} - \Pr\{X \le x_{(j)}\}$, is the proportion of the population values contained in the tolerance interval for this sample. Let us call the above quantity W_{kj}. In symbols we then have $\Pr\{W_{kj} \ge p\} = \gamma$. The variable W_{kj} is either less than p or greater than or equal to p. This is a binomial situation, and, therefore, we can use the same approach as in the confidence interval section to find the value of γ. The value of γ can be expressed in terms of the binomial summation as

$$\gamma = \sum_{i=0}^{k-j-1} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}.$$

If we use the minimum, $x_{(1)}$, and the maximum, $x_{(n)}$, for the limits, $k - j - 1$ becomes $n - 1 - 1$, which equals $n - 2$. It is therefore easy to find the value of this summation for i ranging from 0 to $n - 2$ because that sum is equal to 1 minus the binomial sum from $n - 1$ to n. In symbols, the value of γ is

$$1 - [p^n] - [np^{n-1}(1-p)].$$

Suppose we want our tolerance interval to contain 95 percent of the observations. Let's calculate the confidence level associated with the tolerance interval of 289 to 485 IUs. In this case, n is 30 and p is 0.95. The value of γ is found by taking $1 - 0.95^{30}$

$- 30(0.95)^{29}(1 - 0.95)$, which equals 0.4465. There is not a high level of confidence associated with this tolerance interval. This confidence level is contrasted with the 0.935 level associated with the prediction interval. It is not surprising that the confidence level of the prediction interval is much higher than that of the tolerance interval because the prediction interval is based on the location of a single future value whereas the tolerance interval is based on the location of a large proportion of the population values.

The interval from 289 to 485 IUs is the widest interval we can have using the sample data since these are the minimum and maximum observed values. We can increase our confidence by either (1) decreasing p, the proportion of the population to be included in the tolerance interval or (2) by taking a larger sample.

Let us reduce p to 90 percent. The confidence level for this interval is increased to 0.8162, a much more reasonable value. Instead of reducing p, let us increase the sample size from 30 to 60. The confidence level associated with the increased sample size is 0.8084, also a much more reasonable value. Table 7.3 shows the sample size required to have 90, 95, and 99 percent confidence associated with tolerance intervals that have 80, 90, 95, and 99 percent coverage of the distribution, based on the use of $x_{(1)}$ and $x_{(n)}$.

Table 7.3 Sample size required for the tolerance interval to have the indicated confidence level for the specified coverage proportions based on the use of $x_{(1)}$ and $x_{(n)}$.

Coverage Proportion	Confidence Level		
	90%	95%	99%
0.80	18	22	31
0.90	38	46	64
0.95	77	93	130
0.99	388	473	662

From these calculations and the general formula for calculating, we can see the relationships between p, the values of k and j, n and γ. We can investigate the values of these quantities before we have performed the study and can modify the proposed study design if we are not satisfied with the values of p and γ.

A one-sided tolerance interval is sometimes of interest. Suppose that there was interest in the upper one-sided tolerance interval. In this case, the tolerance interval ranges from 0 to $x_{(n)}$ and the confidence associated with this interval is found by taking $1 - p^n$ — that is, one minus the binomial term calculated for i equal to n.

7.3 Confidence Intervals Based on the Normal Distribution

If the data are from a known probability distribution, knowledge of this distribution allows more informative (smaller) intervals to be constructed for the parameters of interest or for future values. We begin this presentation by showing how to create confidence intervals for a variety of population parameters, assuming that the data come from a normal distribution. The central limit theorem and the sampling distribution of statistics (e.g., sample mean) presented in Chapter 5 provide the rationale for interval estimation based on the normal distribution. Following the material on confidence

intervals, we show how to use the normal distribution in the creation of prediction and tolerance intervals. We begin the confidence interval presentation with the population mean and follow it with the confidence interval for the population proportion that can also be viewed as a mean.

7.3.1 Confidence Interval for the Mean

In the preceding material, we saw how to construct a confidence interval for the population median. That confidence interval gave information to the dairy about the amount of vitamin D being added to the milk. As an alternative to the median, a confidence interval for the mean could have been used. To find a confidence interval for the mean, assuming that the data follow a specific distribution, we must know the sampling distribution of its estimator. We must also specify how confident we wish to be that the interval contains the population parameter. The sample mean is the estimator of the population mean, and the sampling distribution of the sample mean is easily found.

Since we are assuming the data follow a normal distribution, the sample mean — the average of the sample values — also follows a normal distribution. However, this assumption is not crucial. Even if the data are not normally distributed, the central limit theorem states that the sample mean, under appropriate conditions, will approximately follow a normal distribution.

To specify the normal distribution completely, we also have to provide the mean and variance of the sample mean. First we develop the confidence interval for the mean assuming population variance is known and extend it to the situation where population variance is unknown and it is estimated from the sample.

***Known Variance*:** In Chapter 5, we saw that the mean of the sample mean was μ, the population mean, and its variance was σ^2/n. The standard deviation of the sample mean is thus $\sigma/\sqrt{n}$, and it is called the *standard error* of the sample mean ($\bar{x}$). The use of the word *error* is confusing, since no mistake has been made. However, it is the traditional term used in this context. The term *standard error* is used instead of standard deviation when we are discussing the variation in a sample statistic. The term *standard deviation* is usually reserved for discussion of the variation in the sample data themselves. Thus, the standard deviation measures the unit-to-unit variation, while the standard error measures the sample-to-sample variation.

We now address the issue of how confident we wish to be that the interval contains the population mean (μ). From the material on the normal distribution in Chapter 5, we know that

$$\Pr\{-1.96 < Z < 1.96\} = 0.95$$

where Z is the standard normal variable. In terms of the sample mean, this is

$$\Pr\left\{-1.96 < \frac{\bar{x} - \mu}{\left(\sigma/\sqrt{n}\right)} < 1.96\right\} = 0.95.$$

But we want an interval for μ, not for Z. Therefore, we must perform some algebraic manipulations to convert this to an interval for μ. First we multiply all three terms inside the braces by $\sigma/\sqrt{n}$. This yields

$$\Pr\left\{-1.96\left(\frac{\sigma}{\sqrt{n}}\right)<\bar{x}-\mu<1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right\}=0.95.$$

We next subtract $\bar{x}$ from all the expressions inside the braces, and this gives

$$\Pr\left\{-1.96\left(\frac{\sigma}{\sqrt{n}}\right)-\bar{x}<-\mu<1.96\left(\frac{\sigma}{\sqrt{n}}\right)-\bar{x}\right\}=0.95.$$

This interval is about $-\mu$; to convert it to an interval about μ, we must multiply each term in the brackets by −1. Before doing this, we must be aware of the effect of multiplying an inequality by a minus number. For example, we know that 3 is less than 4. However, −3 is greater than −4, so the result of multiplying both sides of an inequality by −1 changes the direction of the inequality. Therefore, we have

$$\Pr\left\{1.96\left(\frac{\sigma}{\sqrt{n}}\right)+\bar{x}>\mu>-1.96\left(\frac{\sigma}{\sqrt{n}}\right)+\bar{x}\right\}=0.95.$$

We reorder the terms to have the smallest of the three quantities to the left — that is,

$$\Pr\left\{\bar{x}-1.96\left(\frac{\sigma}{\sqrt{n}}\right)<\mu<\bar{x}+1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right\}=0.95$$

or, more generally,

$$\Pr\left\{\bar{x}-z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)<\mu<\bar{x}+z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right\}=1-\alpha.$$

The $(1-\alpha)*100$ percent confidence interval limits for the population mean can be expressed as

$$\bar{x}\pm z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$

The result of these manipulations is an interval for μ in terms of σ, n, 1.96 (or some other z value), and $\bar{x}$. The sample mean, $\bar{x}$, is the only one of these quantities that varies from sample to sample. However, once we draw a sample, the interval is fixed as the sample mean's value, $\bar{x}$, is known. Since the interval will either contain or not contain μ, we no longer talk about the probability of the interval containing μ.

Although we do not talk about the probability of an interval containing μ, we do know that in repeated sampling, intervals of the preceding form will contain the parameter, μ, 95 percent of the time. Thus, instead of discussing the probability of an interval containing μ, we say that we are 95 percent confident that the interval from $\left[\bar{x}-1.96\left(\sigma/\sqrt{n}\right)\right]$ to $\left[\bar{x}+1.96\left(\sigma/\sqrt{n}\right)\right]$ will contain μ. Intervals of this type are therefore called *confidence intervals*. This reason for the use of the word *confidence* is the same as that discussed in the preceding distribution-free material. The limits of the confidence interval usually have the form of the sample estimate plus or minus some distribution percentile — in this case, the normal distribution — times the standard error of the sample estimate.

Example 7.1

The 95 percent confidence interval for the mean systolic blood pressure for 200 patients can be found based on the dig200 data set introduced in Chapter 3. We assume that the standard deviation for this patient population is 20 mmHg. As the sample mean, $\bar{x}$, based on a sample size of 199 (one missing value) observations, was found to be 125.8 mmHg, the 95 percent confidence interval for the population mean ranges from $\left[125.8 - 1.96\left(20/\sqrt{199}\right)\right]$ to $\left[125.8 - 1.96\left(20/\sqrt{199}\right)\right]$ — that is, from 123.0 to 128.6 mmHg.

Table 7.4 illustrates the concept of confidence intervals. It shows the results of drawing 50 samples of size 60 from a normal distribution with a mean of 94 and a standard deviation of 11. These values are close to the mean and standard deviation of the systolic blood pressure variable for 5-year-old boys in the United States as reported by the NHLBI Task Force on Blood Pressure Control in Children (1987).

In this demonstration, 4 percent (2 out of 50 marked in the table) of the intervals did not contain the population mean, and 96 percent did. If we draw many more samples, the proportion of the intervals containing the mean will be 95 percent. This is the basis for the statement that we are 95 percent confident that the confidence interval, based on our single sample, will contain the population mean.

If we use a different value for the standard normal variable, the level of confidence changes accordingly. For example, if we had started with a value of 1.645, $z_{0.95}$, instead

Table 7.4 Simulation of 95% confidence intervals for 50 samples of $n = 60$ from the normal distribution with $\mu = 94$ and $\sigma = 11$ (standard error = 1.42).

Sample	Mean	Std	95% CI	Sample	Mean	Std	95% CI
1	94.75	10.25	(91.96, 97.54)	26	94.61	11.49	(91.82, 97.39)
2	94.85	10.86	(92.06, 97.63)	27	92.79	9.36	(90.00, 95.58)
3	94.71	10.09	(91.92, 97.50)	28	96.00	12.19	(93.22, 98.79)
4	94.03	12.27	(91.24, 96.82)	29	95.99	11.36	(93.20, 98.78)
5	93.77	10.05	(90.98, 96.56)	30	93.98	11.74	(91.19, 96.76)
6	92.54	9.32	(89.76, 95.33)	31	95.36	13.08	(92.57, 98.15)
7	93.40	12.07	(90.62, 96.19)	32	91.10	8.69	(88.31, 93.89)*
8	93.97	11.02	(91.18, 96.75)	33	93.85	12.94	(91.06, 96.63)
9	96.33	9.26	(93.54, 99.12)	34	96.01	9.63	(93.22, 98.79)
10	93.56	12.01	(90.78, 96.35)	35	95.20	8.94	(92.41, 97.99)
11	94.94	10.81	(92.15, 97.73)	36	95.64	9.41	(92.85, 98.43)
12	94.66	12.08	(91.88, 97.45)	37	94.74	10.31	(91.95, 97.53)
13	94.21	11.02	(91.42, 97.00)	38	93.52	10.30	(90.73, 96.31)
14	94.55	9.98	(91.76, 97.34)	39	92.92	10.27	(90.13, 95.71)
15	93.57	11.50	(90.79, 96.36)	40	95.08	10.07	(92.30, 97.87)
16	95.99	12.01	(93.20, 98.78)	41	93.88	10.53	(91.09, 96.66)
17	93.86	12.53	(91.08, 96.65)	42	95.38	9.98	(92.59, 98.17)
18	92.02	13.58	(89.23, 94.81)	43	94.38	11.65	(91.59, 97.17)
19	95.16	12.03	(92.38, 97.95)	44	91.55	10.63	(88.76, 94.33)
20	94.99	12.00	(92.20, 97.78)	45	95.41	12.79	(92.62, 98.20)
21	94.65	11.18	(91.86, 97.43)	46	92.40	10.57	(89.62, 95.19)
22	92.86	12.52	(90.07, 95.64)	47	96.00	11.45	(93.21, 98.78)
23	93.99	11.76	(91.20, 96.78)	48	95.39	10.56	(92.60, 98.18)
24	91.44	10.75	(88.65, 94.22)	49	97.69	10.89	(94.90, 100.47)*
25	96.07	11.89	(93.28, 98.86)	50	95.01	10.61	(92.22, 97.79)

*Does not contain 94

of 1.96, $z_{0.975}$, the confidence level would be 90 percent instead of 95 percent. The $z_{0.95}$ value is used with the 90 percent level because we want 5 percent of the values to be in each tail. The lower and upper limits for the 90 percent confidence interval for the population mean for the data in the first sample of 60 observations are 92.41 [= 94.75 − 1.645(1.42)] and 97.09 [= 94.75 + 1.645(1.42)], respectively. This interval is narrower than the corresponding 95 percent confidence interval of 91.96 to 97.54. This makes sense, since, if we wish to be more confident that the interval contains the population mean, the interval will have to be wider. The 99 percent confidence interval uses $z_{0.995}$, which is 2.576, and the corresponding interval is 91.09 [= 94.75 − 2.576(1.42)] to 98.41 [= 94.75 + 2.576(1.42)].

The fifty samples shown in Table 7.4 had sample means, based on 60 observations, ranging from a low of 91.1 to a high of 97.7. This is the amount of variation in sample means expected if the data came from the same normal population with a mean of 94 and a standard deviation of 11. The Second National Task Force on Blood Pressure Control in Children (1987) had study means ranging from 85.6 (based on 181 values) to 103.5 mmHg (based on 61 values), far outside the range just shown. These extreme values suggest that these data do not come from the same population, and this then calls into question the Task Force's combination of the data from these diverse studies.

The size of the confidence interval is also affected by the sample size that appears in the $\sigma/\sqrt{n}$ term. Since n is in the denominator, increasing n decreases the size of the confidence interval. For example, if we doubled the sample size from 60 to 120 in the preceding example, the standard error of the mean changes from $1.42\left(=11/\sqrt{60}\right)$ to $1.004\left(=11/\sqrt{120}\right)$. Doubling the sample size reduces the confidence interval to about 71 percent $\left(=1/\sqrt{2}\right)$ of its former width. Thus, we know more about the location of the population mean, since the confidence interval is shorter as the sample size increases.

The size of the confidence interval is also a function of the value of σ, but to change σ means that we are considering a different population. However, if we are willing to consider homogeneous subgroups of the population, the value of the standard deviation for a subgroup should be less than that for the entire population. For example, instead of considering the blood pressure of 5-year-old boys, we consider the blood pressure of 5-year-old boys grouped according to height intervals. The standard deviation of systolic blood pressure in the different height subgroups should be much less than the overall standard deviation.

Another factor affecting the size of the confidence interval is whether it is a one-sided or a two-sided interval. If we are only concerned about higher blood pressure values, we could use an upper one-sided confidence interval. The lower limit would be zero, or $-\infty$ for a variable that had positive and negative values, and the upper limit is

$$\bar{x} + z_{1-\alpha}\left(\frac{\sigma}{\sqrt{n}}\right).$$

This is similar to the two-sided upper limit except for the use of $z_{1-\alpha}$ instead of $z_{1-\alpha/2}$.

***Unknown Variance*:** When the population variance, σ^2, is unknown, it is reasonable to substitute its sample estimator, s^2, in the confidence interval calculation. There is a problem in doing this, though. Although $(\bar{x}-\mu)/\left(\sigma/\sqrt{n}\right)$ follows the standard normal distribution, $(\bar{x}-\mu)/\left(s/\sqrt{n}\right)$ does not. In the first expression, there is only one random

variable, $\bar{x}$, whereas the second expression involves the ratio of two random variables, $\bar{x}$ and s. We need to know the probability distribution for this ratio of random variables.

Fortunately, Gosset, who we encountered in Chapter 5, already discovered the distribution of $(\bar{x}-\mu)/(s/\sqrt{n})$. The distribution is called *Student's t* — crediting Student, the pseudonym used by Gosset — or, more simply, the *t distribution*. For large values of n, sample values of s are very close to σ, and, hence, the t distribution looks very much like the standard normal. However, for small values of n, the sample values of s vary considerably, and the t and standard normal distributions have different appearances. Thus, the t distribution has one parameter, the number of independent observations used in the calculation of s. In Chapter 3, we saw that this value was $n - 1$, and we called this value the degrees of freedom. Hence, the parameter of the t distribution is the degrees of freedom associated with the calculation of the standard error. The degrees of freedom are shown as a subscript — that is, as t_{df}. For example, a t with 5 degrees of freedom is written as t_5.

Figure 7.1 shows the distributions of t_1 and t_5 compared with the standard normal distribution over the range of –3.8 to 3.8. As we can see from these plots, the t distribution with one degree of freedom, the lowest curve, is considerably flatter — that is, there is more variability than for the standard normal distribution, the top curve in the figure. This is to be expected, since the sample mean divided by the sample standard deviation is more variable than the sample mean alone. As the degrees of freedom increase, the t distributions become closer and closer to the standard normal in appearance. The tendency for the t to approach the standard normal distribution as the number of degrees of freedom increases can also be seen in Table 7.5, which shows selected percentiles for several t distributions and the standard normal distribution. A more complete t table is found in Appendix Table B5.

Now that we know the distribution of $(\bar{x}-\mu)/(s/\sqrt{n})$, we can form confidence intervals for the mean even when the population variance is unknown. The form for the confidence interval is similar to that preceding for the mean with known variance except that s replaces σ and the t distribution is used instead of the standard normal

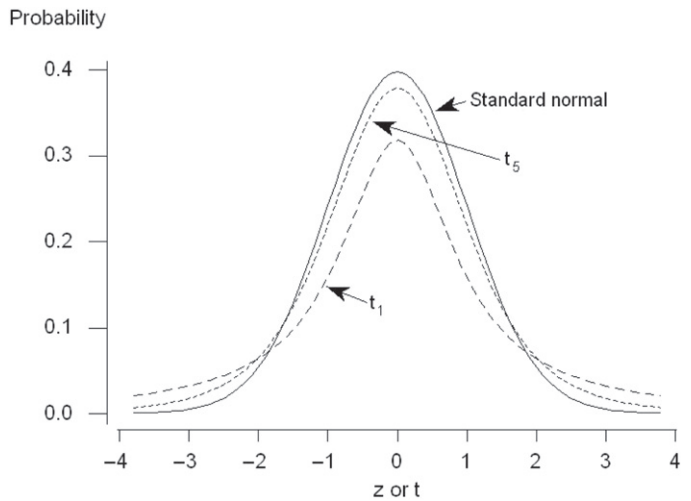


Figure 7.1
Distributions of t_1 and t_5 compared with z distribution.

Table 7.5 Selected percentiles for several *t* distributions and the standard normal distribution.

	Percentiles			
Distribution	**0.80**	**0.90**	**0.95**	**0.99**
t_1	1.376	3.078	6.314	31.821
t_5	0.920	1.476	2.015	3.365
t_{10}	0.879	1.372	1.813	2.764
t_{30}	0.854	1.310	1.697	2.457
t_{60}	0.848	1.296	1.671	2.390
t_{120}	0.845	1.289	1.658	2.358
Standard normal	0.842	1.282	1.645	2.326

distribution. Therefore, the lower and upper limits for the $(1 - \alpha) * 100$ percent confidence interval for the mean when the variance is unknown are $\{\bar{x} - t_{n-1,1-\alpha/2}(s/\sqrt{n})\}$ and $\{\bar{x} + t_{n-1,1-\alpha/2}(s/\sqrt{n})\}$, respectively.

Let us calculate the 90 percent confidence interval for the population mean of the systolic blood pressure for 5-year-old boys based on the first sample data in Table 7.4 (row 1). A 90 percent [= $(1 - \alpha) * 100$ percent] confidence interval means that α is 0.10. Based on a sample of 60 observations, the sample mean was 94.75 and the sample standard deviation was 10.25 mmHg. Thus, we need the 95th (= $1 - \alpha/2$) percentile of a *t* distribution with 59 degrees of freedom. However, neither Table 7.5 nor Table B5 shows the percentiles for a *t* distribution with 59 degrees of freedom. Based on the small changes in the *t* distribution for larger degrees of freedom, there should be little error if we use the 95th percentile for a t_{60} distribution. Therefore, the lower and upper limits are approximately

$$94.75 - 1.671\left(\frac{10.25}{\sqrt{60}}\right) \quad \text{and} \quad 94.75 + 1.671\left(\frac{10.25}{\sqrt{60}}\right)$$

or 92.54 and 96.96 mmHg, respectively.

If we use a computer package (see **Program Note 7.1** on the website) to find the 95th percentile value for a t_{59} distribution, we find its value is 1.6711. Hence, there is little error introduced in this example by using the percentiles from a t_{60} instead of a t_{59} distribution.

7.3.2 Confidence Interval for a Proportion

We are frequently exposed to the confidence interval for a proportion. Most surveys about opinions or voting intentions today report the margin of error. This quantity is simply one half the width of the 95 percent confidence interval for the proportion. Finding the confidence interval for a proportion, π, can be based on either the binomial or normal distribution. The binomial distribution is generally used for smaller samples and it provides an exact interval whereas the normal distribution is used with larger samples and provides an approximate interval. Let us examine the exact interval first.

***Use of the Binomial Distribution*:** Suppose we wish to find a confidence interval for the proportion of restaurants that are in violation of local health ordinances. A simple random sample of 20 restaurants is selected, and, of those, four are found to have violations. The sample proportion, *p*, which is equal to 0.20 (= 4/20), is the point estimate

of π, the population proportion. How can we use this sample information to create the $(1 - \alpha) * 100$ percent confidence interval for the population proportion?

This is a binomial situation, since there are only two outcomes for a restaurant — that is, a restaurant either does or does not have a violation. The binomial variable is the number of restaurants with a violation and we have observed its value to be 4 in this sample.

The limits of the confidence interval for the proportion are those values that make this outcome appear to be unusual. Another way of stating this is that the lower limit is the proportion for which the probability of 4 or more restaurants is equal to $\alpha/2$. Correspondingly, the upper limit is the proportion for which the probability of 4 or fewer restaurants is equal to $\alpha/2$. The two charts in Appendix Table B6 can be used to find the 95 and 99 percent confidence intervals.

Example 7.2

Suppose that we want the 95 percent confidence interval for $p = 0.20$ and $n = 20$. We use the first chart (Confidence Level 95 Percent) of Table B6, and, since the sample proportion is less than 0.50, we read across the bottom until we find the sample proportion value of 0.20. We then move up along the line corresponding to 0.20 until it intersects the first curve for a sample size of 20. Since p is less than 0.50, we read the value of the lower limit from the left vertical axis; it is slightly less than 0.06. To find the upper limit, we continue up the vertical line corresponding to 0.20 until we reach the second curve for a sample size of 20. We read the upper limit from the left vertical axis, and its value is slightly less than 0.44. The approximate 95 percent confidence limits are 0.06 and 0.44. Note that this interval is not symmetric about the point estimation. If p is greater than 0.5, we locate p across the top and read the limits from the right vertical axis.

Another method of finding the upper and lower limits of a confidence interval based on a binomial distribution is to find these values by trial and error.

Example 7.3

Suppose that we wish to find the 90 percent confidence interval for $p = 0.20$ ($x = 4$) and $n = 20$. This means that α is 0.10 and $\alpha/2$ is 0.05. We wish to find the probability of being less than or equal to 4 and being greater than or equal to 4 for different binomial proportions. For the upper limit, we can try some value above 0.20, say, 0.35 and calculate Pr ($X \leq x$). If Pr ($X \leq x$) is larger than $\alpha/2$, then we will try a larger value of p — say, 0.4. We can try this process until Pr ($X \leq$ x) is close enough to $\alpha/2$. For the lower limit, we try some value of p smaller than 0.20, say 0.1 and calculate Pr ($X \geq x$), which is 1 − Pr ($X \leq x - 1$). If 1 − Pr ($X \leq x - 1$) is smaller than $\alpha/2$, then we try a smaller value of p — say, 0.07. Continue this process until 1 − Pr ($X \leq x - 1$) is close enough to $\alpha/2$. Computers can perform this iterative process quickly. An SAS program produced the 90 percent confidence interval (0.0714, 0.4010). An option of getting a binomial confidence interval is available in most programs (see **Program Note 7.2** on the website).

***Use of the Normal Approximation to the Binomial*:** Let us now consider the use of the normal approximation to the binomial distribution. The sample proportion, p, is the binomial variable, x, divided by a constant, the sample size. Since the normal distribution was shown in Chapter 5 to be a good approximation for the distribution of x when the sample size was large enough, it also serves as a good approximation to the distribution of p. The variance of p is expressed in terms of the population proportion, π, and it is $\pi(1 - \pi)/n$. Because π is unknown, we estimate the variance by substituting p for π in the formula.

The sample proportion can also be viewed as a mean as was discussed in Chapter 5. Therefore, the confidence interval for a proportion has the same form as that of the mean, and the limits of the interval are

$$\left(p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} + \frac{1}{2n}, \quad p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} + \frac{1}{2n} \right).$$

The $1/(2n)$ is the continuity correction term required because a continuous distribution is used to approximate a discrete distribution. For large values of n, the term has little effect and many authors drop it from the presentation of the confidence interval.

Example 7.4

The local health department is concerned about the protection of children against diphtheria, pertussis, and tetanus (DPT). To determine if there is a problem in the level of DPT immunization, the health department decides to estimate the proportion immunized by drawing a simple random sample of 150 children who are 5 years old. If the proportion of children in the community who are immunized against DPT is clearly less than 75 percent, the health department will mount a campaign to increase the immunization level. If the proportion is clearly greater than 75 percent, the health department will shift some resources from immunization to prenatal care. The department decides to use a 99 percent confidence interval for the proportion to help it reach its decision.

Based on the sample, 86 families claimed that their child was immunized, and 54 said their child was not immunized. There were 10 children for whom the immunization status could not be determined. As was mentioned in Chapter 6, there are several approaches to dealing with the unknowns. Since there are only 10 unknowns, we shall ignore them in the calculations. Thus, the value of p is 0.614 (= 86/140), much lower than the target value of 0.75. If all 10 of the children with unknown status had been immunized, then p would have been 0.640, not much different from the value of 0.614, and still much less than the target value of 0.75.

Applying the preceding formula, the 99 percent confidence interval ranges from

$$0.614 - 2.576\sqrt{\frac{0.614(0.386)}{140}} + \frac{1}{2(140)} \quad \text{to} \quad 0.614 + 2.576\sqrt{\frac{0.614(0.386)}{140}} + \frac{1}{2(140)}$$

or from 0.504 to 0.724. Since the upper limit of the 99 percent confidence interval is less than 0.75, the health department decides that it is highly unlikely that the proportion of 5-year-old children who are immunized is as large as 0.75. Therefore, the health department will mount a campaign to increase the level of DPT immunization in the community.

If the issue facing the health department was whether or not to add resources to the immunization program, not to shift any resources away from the program, a one-sided interval could have been used. The 99 percent upper one-sided interval uses $z_{0.99}$ instead of $z_{0.995}$ in its calculation and it ranges from 0 to 0.713. This interval also does not contain 0.75. Therefore, resources should be added to the immunization program.

Let us use the normal approximation to find the confidence for the data in Example 7.3. The confidence interval for π based on $p = 0.2$ and $n = 20$ using the normal distribution is (0.0779, 0.3721) compared to (0.0714, 0.4010) based on the binomial distribution (Example 7.3). The former interval is symmetric, while the latter interval is not symmetric. The use of the normal distribution can give a negative lower limit when used with a small p and a small n. For this extreme case the binomial distribution is recommended. The charts in Table B6 suggest that the normal approximation is satisfactory for a large n and can be used even for a relatively small n when p is close to 0.5.

7.3.3 Confidence Intervals for Crude and Adjusted Rates

In Chapter 3, we presented crude, specific, and direct and indirect adjusted rates. However, we did not present any estimate for the variance or standard deviation of a rate, quantities that are necessary for the calculation of the confidence interval. Therefore, we begin this material with a section on how to estimate the variance of a rate.

Rates are usually based on the entire population. If this is the case, there is really no need to calculate their variances or confidence intervals for them. However, we often view a population rate in some year as a sample in location or time. From this perspective, there is justification for calculating variances and confidence intervals. If the value of the rate is estimated from a sample, as is often done in epidemiology, then it is important to estimate the variance and the corresponding confidence interval for the rate. If the rate is based on the occurrence of a very small number of events — for example, deaths — the rate may be unstable and it should not be used in this case. We shall say more about this later.

***Variances of Crude and Adjusted Rates*:** The crude rate is calculated as the number of events in the population during the year divided by the midyear population. This rate is not really a proportion, but it is very similar to a proportion, and we shall treat it as if it were a proportion. The variance of a sample proportion, p, is $\pi(1 - \pi)/n$. Thus, the variance of a crude rate is approximated by the product of the rate (converted to a decimal value) and one minus the rate divided by the population total.

From the data on rates in Chapter 3, we saw that the crude death rate for American Indian/Alaskan Native males in 2002 was 439.6 per 100,000. The corresponding estimated 2002 American Indian/Alaskan Native male population was 1,535,000. Thus the estimated standard error, the square root of the variance estimate, for this crude death rate is

$$\sqrt{\frac{0.004396(1-0.004396)}{1535000}} = 0.0000534$$

or 5.3 deaths per 100,000 population.

The direct age-adjusted rate is a sum of the age-specific rates, $(sr)_i$'s, in the population under study weighted by the age distribution, w_i's, in the standard population. In symbols, this is $\Sigma[w_i(sr)_i]$, where w_i is the proportion of the standard population in the ith age group and $(sr)_i$ is the age-specific rate in the ith age category. The age-specific rate is calculated as the number of events in the age category divided by the midyear population in that age category. Again, this rate is not a proportion, but it is very similar to a proportion. We shall approximate the variance of the age-specific rates by treating them as if they were proportions. Since the w_i's are from the standard population that is usually very large and stable, we shall treat the w_i's as constants as far as the variance calculation is concerned. Since the age-specific rates are independent of one another, the variance of the direct adjusted rate, that is, the variance of this sum, is simply the sum of the individual variances

$$Var\left(\sum w_i (sr)_i\right) = \sum Var\left(w_i (sr)_i\right) = \sum w_i^2 \left(\frac{(sr)_i \left(1-(sr)_i\right)}{n_i}\right)$$

where n_i is the number of persons in the ith age subgroup in the population under study.

Considering the U.S. mortality data as a sample in time, we can calculate the approximate variance of the direct age-adjusted death rate. The data to be used are the 2002 U.S. male age-specific death rates along with the U.S. male population totals and the 2000 U.S. population proportions by age from Table 3.14. Table 7.6 repeats the relevant data and shows the calculations. The entries in the last column are all quite small, less than 0.00000001, and therefore, only their sum is shown. The standard error of the direct age-adjusted mortality rate is 0.0000117 (= square root of variance). The direct age-adjusted rate was 1013.7 deaths per 100,000 population, and the standard error of the rate is 1.2 deaths per 100,000. The magnitude of the standard error here is not unusual, and it shows why the sampling variation of the adjusted rate is often ignored in studies involving large population bases.

For the indirect method, the adjusted rate can be viewed as the observed crude rate in the population under study multiplied by a ratio. The ratio is the standard population's

Table 7.6 Calculation of the approximate variance for the age-adjusted death rate by the direct method for U.S. males in 2002.

Age i	U.S. Male Age-Specific Rates $(sr)_i$	U.S. Male Population n_i	U.S. Population Proportion[a] w_i	$\Sigma[w_i^2(sr)_i(1-(sr)_i)/n_i]$
Under 1	0.007615	2,064,000	0.013818	
1–4	0.000352	7,962,000	0.055317	
5–14	0.000200	21,013,000	0.145565	
15–24	0.001173	20,821,000	0.138645	
25–34	0.001422	20,203,000	0.135573	
35–44	0.002575	22,367,000	0.162613	
45–54	0.005475	19,676,000	0.134834	
55–64	0.011840	12,784,000	0.087242	
65–74	0.028553	8,301,000	0.066037	
75–84	0.067605	5,081,000	0.044842	
85 & over	0.162545	1,390,000	0.015508	
Total		141,661,000	1.000000	1.37×10^{10}

[a]U.S. total population proportion in 2000 (the standard)

crude rate divided by the rate obtained by weighting the standard population's age-specific rates by the age distribution from the study population. This ratio is viewed as a constant in terms of approximating the variance. Hence, the approximation of the variance of the indirect adjusted rate is simply the square of the ratio multiplied by the variance of the study population's crude rate.

Using the data from Chapter 3, the standard population's (the 2000 U.S. population) crude rate was 854.0 deaths per 100,000 population. The combination of the standard population's age-specific rates with the study population's (the 2002 American Indian/Alaskan Native male) age distribution yielded 413.6 deaths per 100,000 population. The crude rate for American Indian/Alaskan Native male was 439.6 deaths per 100,000 population. Thus, the approximate standard error, the square root of the variance, of the indirect age-adjusted death rate is

$$\sqrt{\left(\frac{0.008540}{0.00413.6}\right)^2\left(\frac{0.004396(1-0.004396)}{1535000}\right)} = 0.00011025$$

or 11 per 100,000.

***Formation of the Confidence Interval*:** To form the confidence interval for a rate, we require knowledge of its sampling distribution. Since we are treating crude and specific rates as if they are proportions, the confidence intervals for these rates will be based on the normal approximation as just shown for the proportion. Therefore, the confidence interval for the population crude rate (θ) is

$$cr - z_{1-\alpha/2}\sqrt{\frac{cr(1-cr)}{n}} < \theta < cr + z_{1-\alpha/2}\sqrt{\frac{cr(1-cr)}{n}}$$

where *cr* is the value of the crude rate based on the observed sample.

For example, the 95 percent confidence interval for the 2002 American Indian/Alaskan Native male crude death rate is

$$0.00439.6 - 1.96(0.0000534) < \theta < 0.00439.6 + 1.96(0.0000534)$$

or from 0.004291 to 0.004501. Thus, the confidence interval for the crude death rate is from 429.1 to 450.1 deaths per 100,000 population.

The confidence intervals for the rates from the direct and indirect methods of adjustment have the same form as that of the crude rate. For example, the 95 percent confidence interval for the indirect age-adjusted death rate for 2002 American Indian/Alaskan Native male is found by taking

$$907.8 \pm 1.96(11.0) = 907.8 \pm 21.6$$

and thus the limits are from 886.2 to 929.4 deaths per 100,000 population.

***Minimum Number of Events Required for a Stable Rate*:** As we just mentioned, rates based on a small number of occurrences of the event of interest may be unstable. To deal with this instability, a health agency for a small area often will combine its mortality data over several years. By using the estimated coefficient of variation, the estimated standard error of the estimate divided by the estimate and multiplied by 100 percent, we can determine when there are too few events for the crude rate to be stable.

Recall that in Chapter 3 we said that if the coefficient of variation was large, the data had too much variability for the measure of central tendency to be very informative. Values of the coefficient of variation greater than 30 percent — others might use slightly larger or smaller values — are often considered to be large. We shall use this idea with the crude rate to determine how many events are required so that the rate is stable.

For example, the coefficient of variation for the 1986 crude mortality rate of Harris County is 0.904 percent (= [0.0000479/0.005296] ∗ 100). This rate, less than 1 percent, is very reliable from the coefficient of variation perspective. It turns out that the coefficient of variation of the crude rate can be approximated by $(1/\sqrt{d}) * 100$ percent, where d is the number of events. For example, the total number of deaths for Harris County in 1986 was 12,152 and (1/12152) ∗ 100 is 0.907 percent, essentially the same result as above.

Thus, we can use the approximation $(1/\sqrt{d}) * 100$ percent for the coefficient of variation. Setting the coefficient of variation to 20, 30, and 40 percent, yields 25, 12, and 7 events, respectively. If the crude rate is based on fewer than seven events, it certainly should not be reported. If we require that the coefficient of variation be less than 20 percent, there must be at least 25 occurrences of the event for the crude rate to be reported.

7.4 Confidence Interval for the Difference of Two Means and Proportions

We often wish to compare the mean or proportion from one population to that of another population. The confidence interval for the difference of two means or proportions facilitates the comparison. As will be seen the following sections, the method of constructing the confidence interval is different, depending on whether the two means or proportions are independent or not and depending on what assumptions are made.

7.4.1 Difference of Two Independent Means

Examples of comparing two independent means include the following. Is the mean change in blood pressure for men with mild to moderate hypertension the same for men taking different doses of an angiotensin-converting enzyme inhibitor? Is the mean length of stay in a psychiatric hospital equal for patients with the same diagnosis but under the care of two different psychiatrists? Given the following, there is an interest in the mean change in air pollution — specifically, in carbon monoxide — from 1991 to 1992 for neighboring states A and B. There was no change in gasoline formulation in State A, whereas State B required on January 1, 1992, that gasoline must consist of 10 percent ethanol during the November to March period.

One reason for interest in the confidence interval for the difference of two means is that it can be used to address the question of the equality of the two means. If there is no difference in the two population means, the confidence interval for their difference is likely to include zero.

***Known Variances*:** The confidence interval for the difference of two means has the same form as that for a single mean; that is, it is the difference of the sample means

plus or minus some distribution percentile multiplied by the standard error of the difference of the sample means. Let's convert these words to symbols. Suppose that we draw samples of sizes n_1 and n_2 from two independent populations. All the observations are assumed to be independent of one another — that is, the value of one observation does not affect the value of any other observation. The unknown population means are μ_1 and μ_2, the sample means are $\bar{x}_1$ and $\bar{x}_2$, and the known population variances are σ_1^2 and σ_2^2, respectively. The variances of the sample means are σ_1^2/n_1 and σ_2^2/n_2, respectively. Since the means are from two independent populations, the standard error of the difference of the sample means is the square root of the sum of the variances of the two sample means — that is,

$$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}.$$

The central limit theorem implies that the difference of the sample means will approximately follow the normal distribution for reasonable sample sizes. Thus, we have

$$Z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}.$$

Therefore, the $(1-\alpha)*100$ percent confidence interval for the difference of population means, $\mu_1-\mu_2$, is

$$\left((\bar{x}_1-\bar{x}_2)-z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},(\bar{x}_1-\bar{x}_2)+z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right).$$

Example 7.5

Suppose we wish to construct a 95 percent confidence interval for the effect of different doses of Ramipril, an angiotensin-converting enzyme converting inhibitor, used in treating high blood pressure. A study reported changes in diastolic blood pressure using the values at the end of a four-week run-in period as the baseline and measured blood pressure after two, four, and six weeks of treatment (Walter, Forthofer, and Witte 1987). We shall form a confidence interval for the difference in mean decreases from baseline to two weeks after treatment was begun between doses of 1.25 mg and 5 mg of Ramipril. The sample mean decreases are 10.6 ($\bar{x}_1$) and 14.9 mmHg ($\bar{x}_2$) for the 1.25 and 5 mg doses, respectively, and n_1 and n_2 are both equal to 53. Both σ_1 and σ_2 are assumed to be 9 mmHg. The 95 percent confidence interval for $\mu_1-\mu_2$ is calculated as follows:

$$\left((10.6-14.9)-1.96\sqrt{\frac{81}{53}+\frac{81}{53}},(10.6-14.9)+1.96+\sqrt{\frac{81}{53}+\frac{81}{53}}\right)$$

or ranging from −7.98 to −0.62. The value of 0 is not contained in this interval. Since the difference in mean decreases is negative, it appears that the 5 mg dose of Ramipril is associated with a greater decrease in diastolic blood pressure during the first two weeks of treatment when considering only these two doses.

***Unknown but Equal Population Variances*:** If the variances are unknown but assumed to be equal, data from both samples can be combined to form an estimate of the common population variance. Use of the sample estimator of the variance calls for the use of the t instead of the normal distribution in the formation of the confidence interval. The pooled estimator of the common variance, s_p^2, is defined as

$$s_p^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

and this can be rewritten as

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

The pooled estimator is a weighted average of the two sample variances, weighted by the respective degrees of freedom associated with the individual sample variances and divided by sum of the degrees of freedom associated with each of the two sample variances.

Now that we have an estimator of σ^2, we can use it in estimating the standard error of the difference of the sample means, $\bar{x}_1$ and $\bar{x}_2$. Since we are assuming that the population variances for the two groups are the same, the standard error of the difference of the sample means is

$$\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and its estimator is

$$s_p = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

The corresponding t statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{1/n_1 + 1/n_2}}$$

and the $(1 - \alpha) * 100$ percent confidence interval for $(\mu_1 - \mu_2)$ is

$$\left((\bar{x}_1 - \bar{x}_2) - t_{n-2,1-\alpha/2}s_p\sqrt{1/n_1 + 1/n_2}\,,\,(\bar{x}_1 - \bar{x}_2) + t_{n-2,1-\alpha/2}s_p\sqrt{1/n_1 + 1/n_2}\right)$$

where n is the sum of n_1 and n_2.

Example 7.6

Suppose that we wish to calculate the 95 percent confidence interval for the difference in the proportion of caloric intake that comes from fat for fifth- and sixth-grade boys compared to seventh- and eighth-grade boys in suburban Houston. The sample data are shown in Table 7.7. The proportion of caloric intake that comes from fat is found by converting the grams of fat to calories by multiplying by nine (9 calories result from 1 gram of fat) and then dividing by the number of calories consumed.

Table 7.7 Total fat,[a] calories, and the proportion of calories from total fat for the 33 boys.

Grades 7 and 8			Grades 5 and 6		
Total Fat	Calories	Prop. from Fat	Total Fat	Calories	Prop. from Fat
567	1,823	0.311	1,197	3,277	0.365
558	2,007	0.278	891	2,039	0.437
297	1,053	0.282	495	2,000	0.248
1,818	4,322	0.421	756	1,781	0.424
747	1,753	0.426	1,107	2,748	0.403
927	2,685	0.345	792	2,348	0.337
657	2,340	0.281	819	2,773	0.295
2,043	3,532	0.578	738	2,310	0.319
1,089	2,842	0.383	738	2,594	0.285
621	2,074	0.299	882	1,898	0.465
225	1,505	0.150	612	2,400	0.255
783	2,330	0.336	252	2,011	0.125
1,035	2,436	0.425	702	1,645	0.427
1,089	3,076	0.354	387	1,723	0.225
621	1,843	0.337			
666	2,301	0.289			
1,116	2,546	0.438			
531	1,292	0.411			
1,089	3,049	0.357			

[a]Total fat has been converted to calories by multiplying the number of grams by 9.

The sample mean for the 14 fifth- and sixth-grade boys is 0.329 compared to 0.353 for the 19 seventh- and eighth-grade boys. These values of percent of intake from fat are slightly above the recommended value of 30 percent (Life Sciences Research Office 1989). The corresponding standard deviations are 0.0895 and 0.0974, which support the assumption of equal variances.

The estimate of the pooled standard deviation is therefore

$$s_p = \sqrt{\frac{13(0.0895^2) + 18(0.0974^2)}{14 + 19 - 2}} = 0.094.$$

The estimate of the standard error of the difference of the sample means is

$$0.094\sqrt{1/14 + 1/19} = 0.033.$$

To find the confidence interval, we require $t_{31,\,0.975}$. This value is not shown in Table B5, but, based on the values for 29 and 30 degrees of freedom, an approximate value for it is 2.04. Therefore, the lower and upper limits are

$$[(0.329 - 0.353) - 2.04\ (0.033)] \text{ and } [(0.329 - 0.353) + 2.04\ (0.033)]$$

or −0.092 and 0.044. Since zero is contained in the 95 percent confidence interval, there does not appear to be a difference in the mean proportions of calories that come from fat for fifth- and sixth-grade boys compared to seventh- and eighth-grade boys in suburban Houston.

***Unknown and Unequal Population Variances*:** If the population variances are different, this poses a problem. There is a procedure for obtaining an exact confidence interval for the difference in the means when the population variances are unequal, but it is much more complex than the other methods in this book (Kendall and Stuart 1967).

Because of this complexity, most researchers use an approximate approach to the problem. The following shows one of the approximate approaches.

Since the population variances are unknown, we again use a *t*-like statistic. This statistic is

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

The *t* distribution with the degrees of freedom shown next can be used to obtain the percentiles of the t' statistic. The degrees of freedom value, *df*, is

$$df = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/(n_1 - 1) + \left(s_2^2/n_2\right)^2/(n_2 - 1)}.$$

This value for the degrees of freedom was suggested by Satterthwaite (1946). It is unlikely to be an integer, and it should be rounded to the nearest integer.

The approximate $(1 - \alpha) * 100$ percent confidence interval for the difference of two independent means when the population variances are unknown and unequal is

$$(\bar{x} - \bar{x}_2) - t_{df,1-\alpha/2} s_{\bar{x}_1 - \bar{x}_2} < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + t_{df,1-\alpha/2} s_{\bar{x}_1 - \bar{x}_2}$$

where the estimate of the standard error of the difference of the two sample means is

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Example 7.7

In Exercise 3.8, we presented survival times from Exercise Table 3.3 in Lee (1980) on 71 patients who had a diagnosis of either acute myeloblastic leukemia (AML) or acute lymphoblastic leukemia (ALL). In one part of the exercise, we asked for additional variables that should be considered before comparing the survival times of these two diagnostic groups of patients. One such variable is age. Let us examine these two groups to determine if there appears to be an age difference. If there is a difference, it must be taken into account in the interpretation of the data. To examine if there is a difference, we find the 99 percent confidence interval for the difference of the mean ages of the AML and ALL patients. Since we have no knowledge about the variation in the ages, we shall assume that the variances will be different. Table 7.8 shows the ages and survival times for these 71 patients.

The sample mean age for the AML patients, $\bar{x}_1$, is 49.86 and s_1 is 16.51 based on the sample size, n_1, of 51 patients. The sample mean, $\bar{x}_2$, for the 20 ALL patients is 36.65 years and s_2 is 17.85. This is the information needed to calculate the confidence interval. Let's first calculate the sample estimate of the standard error of the difference of the means:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{16.51^2}{51} + \frac{17.85^2}{20}} = 4.61.$$

We next calculate the degrees of freedom, *df*, to be used and we find it from

Table 7.8 Ages and survival times of the AML and ALL patients (age and survival times are in the same order).

AML Patients													
Age	20	25	26	26	27	27	28	28	31	33	33	33	34
	36	37	40	40	43	45	45	45	45	47	48	50	50
	51	52	53	53	56	57	59	59	60	60	61	61	61
	62	63	65	71	71	73	73	74	74	75	77	80	
Survival Time	18	31	31	31	36	01	09	39	20	04	45	36	12
in Months	08	01	15	24	02	33	29	07	00	01	02	12	09
	01	01	09	05	27	01	13	01	05	01	03	04	01
	18	01	02	01	08	03	04	14	03	13	13	01	
ALL Patients													
Age	18	19	21	22	26	27	28	28	28	28	34	36	37
	47	55	56	59	62	83	19						
Survival Time	16	25	01	22	12	12	74	01	16	09	21	09	64
in Months	35	01	07	03	01	01	22						

$$df = \frac{\left(16.51^2/51 + 17.85^2/20\right)^2}{\left(\frac{\left(16.51^2/51\right)^2}{51-1} + \frac{\left(17.85^2/20\right)^2}{20-1}\right)} = 32.501.$$

This value is rounded to 33. The 99.5 percentile of the t distribution with 33 degrees of freedom is about midway between the value of 2.750 (30 degrees of freedom) and 2.724 (35 degrees of freedom) in Appendix Table B5. We shall interpolate and use a value of 2.7344 for the 99.5 percentile of the t distribution with 33 degrees of freedom. Therefore, the approximate 99 percent confidence interval for the difference of the mean ages is

$$(49.86 - 36.65) - 2.7344\ (4.61) < \mu_1 - \mu_2 < (49.86 - 36.65) + 2.7344\ (4.61)$$

or

$$0.60 < \mu_1 - \mu_2 < 25.82.$$

Since zero is not contained in this confidence interval, there is an indication of a difference in the mean ages. If the survival patterns differ between patients with these two diagnoses, it may be due to a difference in the age of the patients.

How large would the confidence interval have been if we had assumed that the unknown population variances were equal? Using the approach in the previous section, the pooled estimate of the standard deviation, s_p, is

$$\sqrt{\frac{(51-1)61.51^2 + (20-1)17.85^2}{51+20-2}} = 16.89.$$

This leads to an estimate of the standard error of the difference of the two means of

$$16.89\sqrt{\frac{1}{51} + \frac{1}{20}} = 4.456.$$

Thus the confidence interval, using an approximation of 2.65 to the 99.5 percentile of the t distribution with 69 degrees of freedom, is

$$(49.86 - 36.65) - 2.65\ (4.456) < \mu_1 - \mu_2 < (49.86 - 36.65) + 2.65\ (4.456)$$

or

$$1.20 < \mu_1 - \mu_2 < 25.02.$$

This interval is slightly narrower than the preceding confidence interval found. However, both intervals lead to the same conclusion about the ages in the two diagnosis groups. For the use of a computer for this calculation, see **Program Note 7.3** on the website.

In practice, we usually know little about the magnitude of the population variances. This makes it difficult to decide which approach, equal or unequal variances, should be used. We recommend that the unequal variances approach be used in those situations when we have no knowledge about the variances and no reason to believe that they are equal. Fortunately, as we just saw, often there is little difference in the results of the two approaches. Some textbooks and computer packages recommend that we first test to see if the two population variances are equal and then decide which procedure to use. Several studies have been conducted recently and conclude that this should not be done (Gans 1991; Markowski and Markowski 1990; Moser and Stevens 1992).

7.4.2 Difference of Two Dependent Means

Dependent means occur in a variety of situations. One situation of interest occurs when there is a preintervention measurement of some intervention and a postintervention measurement. Another dependent mean situation occurs when there is a matching or pairing of subjects with similar characteristics. One subject in the pair receives one type of treatment and the other member in the pair receives another type of treatment. Measurements on the variable of interest are made on both members of the pair. In both of these situations, there is some relation between the values of the observations in a pair. For example, the preintervention measurement for a subject is likely to be correlated with the postintervention measurement on the same subject. If there is a nonzero correlation, this violates the assumption of independence of the observations. To deal with this relation (dependency), we form a new variable that is the difference of the observations in the pair. We then analyze the new variable, the difference of the paired observations.

Consider the blood pressure example just presented. Suppose that we focus on the 1.25 mg dose of Ramipril. We have a value of the subject's blood pressure at the end of a four-week run-in period and the corresponding value after two weeks of treatment for 53 subjects. There are 106 measurements, but only 53 pairs of observations and only 53 differences for analysis. The mean decrease in diastolic blood pressure after two weeks of treatment for the 53 subjects is 10.6 mmHg, and the sample standard deviation of the difference is 8.5 mmHg. The confidence interval for this difference has the form of the confidence interval for the mean from a single population. If the population variance is known, we use the normal distribution; otherwise we use the *t* distribution. We assumed that the population standard deviation was 9 mmHg previously, and we shall use that value here. Thus, the confidence interval will use the normal distribution — that is,

$$\bar{x}_d - z_{1-\alpha/2}\left(\frac{\sigma_d}{\sqrt{n}}\right) < \mu_d < \bar{x}_d + z_{1-\alpha/2}\left(\frac{\sigma_d}{\sqrt{n}}\right)$$

where the subscript d denotes difference.

Let us calculate the 90 percent confidence interval for the mean decrease in diastolic blood pressure. Table B4 shows that the 95th percentile of the standard normal is 1.645. Thus, the confidence interval is

$$10.6 - 1.645\left(\frac{9}{\sqrt{53}}\right) < \mu_d < 10.6 + 1.645\left(\frac{9}{\sqrt{53}}\right)$$

which gives an interval ranging from 8.57 to 12.63 mmHg. Since zero is not contained in the interval, it appears that there is a decrease from the end of the run-in period to the end of the first two weeks of treatment.

If we had ignored the relation between the pre- and postintervention values and used the approach for independent means, how would that have changed things? The mean difference between the pre- and postvalues does not change, but the standard error of the mean difference does change. We shall assume that the population variances are known and that σ_1, for the preintervention value, is 7 mmHg and σ_2 is 8 mmHg. The standard error of the differences, wrongly ignoring the correlation between the pre- and postmeasures, is then

$$\sqrt{\frac{7^2}{53} + \frac{8^2}{53}} = 1.46.$$

This is larger than the value of $9/\sqrt{53}$ (= 1.236) just found when taking the correlation into account. This larger value for the standard error of the difference (1.46 versus 1.236) makes the confidence interval larger than it would be had the correct method been used.

This experiment was to examine the dose-response relation of Ramipril. It consisted of a comparison of the changes in the pre- and postintervention blood pressure values for three different doses of Ramipril. If the purpose had been different — for example, to determine whether or not the 1.25 mg dose of Ramipril had an effect — this type of design may not have been the most appropriate. One problem with this type of design — measurement, treatment, measurement — when used to establish the existence of an effect is that we have to assume that nothing else relevant to the subjects' blood pressure values occurred during the treatment period. If this assumption is reasonable, then we can attribute the decrease to the treatment. However, if this assumption is questionable, then it is problematic to attribute the change to the treatment. In this case, the patients received a placebo — here, a capsule that looked and tasted liked the medication to be taken later — during the four-week run-in period. There was little evidence of a placebo effect, a change that occurs because the subject believes that something has been done. A placebo effect, when it occurs, is real and may reflect the power of the mind to affect disease conditions. This lack of a placebo effect here lends credibility to attributing the decrease to the medication, but it is no guarantee.

7.4.3 Difference of Two Independent Proportions

In this section, we want to find the $(1 - \alpha) * 100$ percent confidence interval for the difference of two independent proportions — that is, π_1 minus π_2. We shall assume that the sample sizes are large enough so that it is appropriate to use the normal distribution as an approximation to the distribution of p_1 minus p_2. In this case, the confidence interval for the difference of the two proportions is approximate. Its form is very similar to that for the difference of two independent means when the variances are not equal.

The variance of the difference of the two independent proportions is

$$\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}.$$

Since the population proportions are unknown, we shall substitute the sample proportions, p_1 and p_2, for them in the variance formula. The $(1 - \alpha) * 100$ percent confidence interval for $\pi_1 - \pi_2$ then is

$$(p_1 - p_2) - z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} < \pi_1 - \pi_2$$
$$< (p_1 - p_2) + z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Because we are considering the difference of two proportions, the continuity correction terms cancel out in taking the difference.

Example 7.8

Holick et al. (1992) conducted a study of 13 milk processors in five eastern states. They found that only 12 of 42 randomly selected samples of milk that they collected contained 80 to 120 percent of the amount of vitamin D stated on the label. Suppose that 10 milk processors in the Southwest are also studied and that 21 of 50 randomly selected samples of milk contained 80 to 120 percent of the amount of vitamin D stated on the label. Construct a 99 percent confidence interval for the difference of proportions of milk that contain 80 to 120 percent of the amount of vitamin D stated on the label between these eastern and southwestern producers.

Since the sample sizes and the proportions are relatively large, the normal approximation can be used. The estimate of the standard error of the sample difference is

$$\sqrt{\frac{(12/42)(1-12/42)}{42} + \frac{(21/50)(1-21/50)}{50}} = 0.0987.$$

The value of $z_{0.995}$ is found from Table B4 to be 2.576. Therefore, the 99 percent confidence interval is

$$(0.286 - 0.420) - 2.576\ (0.0987) < \pi_1 - \pi_2 < (0.286 - 0.420) + 2.576\ (0.0987)$$

which is (−0.388, 0.120).

Since zero is contained in the confidence interval, there is little indication of a difference in the proportion of milk samples with vitamin D content within the 80 to 120 percent range of the amount stated on the label between these eastern and southwestern milk producers.

7.4.4 Difference of Two Dependent Proportions

Suppose that a sample of n subjects has been selected to examine the relationship between the presences of a particular attribute at two time points for the same individuals (paired observations). The situation could also be used to examine the relationship between two different attributes for the same individuals. The sample data for these situations can be arranged as follows:

Attribute at Time		
1	2	Number of Subjects
Present	Present	a
Present	Absent	b
Absent	Present	c
Absent	Absent	d
Total		n

Then the estimated proportion of subjects with the attribute at time 1 is $p_1 = (a + b)/n$, and the estimated proportion with the attribute at time 2 is $p_2 = (a + c)/n$. The difference between the two estimated proportions is

$$p_d = p_1 - p_2 = \frac{a+b}{n} - \frac{a+c}{n} = \frac{b-c}{n}.$$

Since the two population probabilities are dependent, we cannot use the same approach for estimating the standard error of the difference that we used in the previous section. Instead of showing the steps in the derivation of the formula, we simply present the formula for the estimated standard error (Fleiss 1981).

$$Estimated\ SE(p_d) = \frac{1}{n}\sqrt{(b+c) - \frac{(b-c)^2}{n}}.$$

The confidence interval for the difference of two dependent proportions, $\pi_d\ (= \pi_1 - \pi_2)$, is then given by

$$p_d - z_{1-\alpha/2}\mathrm{SE}(p_d) < \pi_d < p_d + z_{1-\alpha/2}\mathrm{SE}(p_d).$$

Example 7.9

Suppose that 100 students took both biostatistics and epidemiology tests, and 18 failed in biostatistics ($p_1 = 0.18$) and 10 failed in epidemiology ($p_2 = 0.10$). There is an 8 percentage point difference ($p_d = 0.08$). The confidence interval for the difference of these two failure rates cannot be constructed using the method in the previous

subsection because the two rates are dependent. We need additional information to assess the dependency. Nine students failed both tests ($p_{12} = 0.09$), and this reflects the dependency. The dependency between p_1 and p_2 can be seen more clearly when the data are presented in a 2 by 2 table.

	Epidemiology		
Biostatistics	**Failed**	**Passed**	**Total**
Failed	**9** (*a*)	9 (*b*)	**18**
Passed	1 (*c*)	81 (*d*)	82
Total	**10**	90	**100** (*n*)

The marginal totals reflect the two failure rates. The numbers in the diagonal cells (*a, d*) are concordant pairs of test scores (those who passed or failed both tests), and those in the off-diagonal cells (*b, c*) are discordant pairs (those who passed one test but failed the other). Important information for comparing the two dependent failure rates is contained in discordant pairs, as the estimated difference of the two proportions and its estimated standard error are dependent on *b* and *c*.

Using the standard error equation, we have

$$\frac{1}{100}\sqrt{(9+1)-\frac{(9-1)^2}{100}} = 0.0306.$$

Then the 95 percent confidence interval for the difference of these two dependent proportions is

$$0.08 - 1.96\ (0.0306) < \pi_d < 0.08 + 1.96\ (0.0306)$$

or (0.0200, 0.1400). This interval does not include 0, suggesting that the failure rates of these two tests are significantly different. However, this method is not recommended for small frequencies and further discussion will follow in conjunction with hypothesis testing in the next chapter.

7.5 Confidence Interval and Sample Size

One important point about the confidence interval for the population mean is that its width can be calculated before the sample is selected. The half-width of the confidence interval is

$$z_{1-\alpha/2}\sqrt{\frac{\sigma}{\sqrt{n}}}.$$

When σ and n are known, the width can be calculated. If the interval is viewed as being too wide to be informative, we can change one of the values used (*z, n,* or σ) in calculating the width to see if we can reduce it to an acceptable value. The two most common ways of reducing its width are by decreasing our level of confidence (reducing the *z* value) or by increasing the sample size (*n*); however, there are limits for both of these choices. Most researchers prefer to use at least the 95 percent level for the confidence interval although the use of the 90 percent level is not uncommon. To drop below the

90 percent level is usually unacceptable. Researchers may be able to increase the sample size somewhat, but the increase requires additional resources that are often limited.

Example 7.10

Suppose that we wish to estimate the mean systolic blood pressure of girls who are 120 to 130 cm (approximately 4 feet to 4 feet 3 inches) tall. We assume that the standard deviation of the systolic blood pressure variable for girls in this height group is 7 mmHg. Given this information, how large a sample is required so that the half-width of the 95 percent confidence interval is no more than 3 mmHg wide?

The half-width of the confidence interval can be equated to the specified half-width — that is

$$1.96\left(\frac{7}{\sqrt{n}}\right)=3.$$

This equation can be solved for n, multiplying both sides by $\sqrt{n}$ and squaring both sides, which gives

$$n=\left(\frac{1.96(7)}{3}\right)^2=20.9.$$

Since n must be an integer, the next highest integer value, 21, is taken to be the value of n.

The formula for n, given a specified half-width, d, for the $(1-\alpha)*100$ percent confidence interval is

$$n=\left(\frac{z_{1-\alpha/2}\sigma}{d}\right)^2.$$

So far, we have been assuming that σ is known; however, in practice, we seldom know the population standard deviation. Sometimes the literature or a pilot study provides an estimate of its value that we may use for σ.

For the case of proportion, the sample size can be calculated by the following formula:

$$n=\left(\frac{z_{1-\alpha/2}\sqrt{\pi(1-\pi)}}{d}\right)^2.$$

In this formula π is the population proportion and $\pi(1-\pi)/n$ is the variance of binomial distribution as shown in Chapter 4. The population proportion is seldom known when calculating the sample size. Again, the literature or a pilot study may provide an estimate. In cases when we have no information for π, we can use $\pi=0.5$. This practice is based on the fact that $\pi(1-\pi)$ is the maximum when $\pi=0.5$ and the calculated sample size will be sufficient for any value of π.

The confidence interval for the difference between two independent means, μ_1 and μ_2, can be used to determine the sample size required when there are two equal-size

experimental groups. We assume that the same known population variance is σ^2 and two equal random samples of size n are to be taken. Then the half-width of the confidence interval for the difference of two means simplifies to

$$z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{n}}.$$

As before, let d be the half-width of the desired confidence interval. Equating the preceding quantity to d and solving for n we have

$$n=2\left(\frac{z_{1-\alpha/2}\sigma}{d}\right)^2.$$

For the case of the difference of two independent proportions, the required sample size can be calculated by

$$n=2\left(\frac{z_{1-\alpha/2}\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}}{d}\right)^2.$$

Example 7.11

A researcher wants to be 99 percent confident ($z = 2.567$) that the difference in the mean systolic blood pressure of boys and girls be estimated within plus and minus 2 mmHg ($d = 2$). How large a sample size does the researcher need in each group? We will assume that the sample size is large enough that the normal distribution approximation can be used. We also assumed that the standard deviation of the systolic blood pressure for boys and girls are the same, and it is 8 mmHg. The required sample size is

$$n=2\left(\frac{2.567(8)}{2}\right)^2=210.9.$$

The required sample size is 211 in each group.

In the planning of a statistical study, the determination of sample size is not as simple as the preceding example may suggest. If you want a high level of confidence and a small interval, a very large sample size is required. The difficulty lies in deciding what level of confidence to aim for within the limit imposed by available resources. The balancing of the level of confidence against availability of resources may require an iterative process until a solution is found that satisfies both requirements.

7.6 Confidence Intervals for Other Measures

We next consider confidence intervals for the variance and the Pearson correlation coefficient. Interval estimation for other measures such as the odds ratio and regression coefficient will be discussed in subsequent chapters.

7.6.1 Confidence Interval for the Variance

Besides being useful in describing the data, the variance is also frequently used in quality control situations. It is one way of stating how reliable the process under study is. For example, in Chapter 2 we presented data on the measurement of blood lead levels by different laboratories. We saw from that example that great variability in the measurements made by laboratories exists, and the variance is one way to characterize that variability. Variability within laboratories can be due to different technicians, failure to calibrate the equipment, and so forth. It is critically important that measurements of the same sample within a laboratory have variability less than or equal to a prespecified small amount. Thus, based on the sample variance for a laboratory for measuring blood lead, we wish to determine whether or not the laboratory's variance is in compliance with the standards. The confidence interval for the population variance provides one method of doing this.

To construct the confidence interval for the population variance, we need to know the sampling distribution of its estimator, the sample variance, s^2. The sampling distribution of s^2 can be examined by (1) taking a repeated random sample from a normal distribution, (2) calculating a sample variance from each sample, and (3) plotting a histogram of sample variances. When we take a repeated random sample of size 3, the distribution of sample variances looks like the black line in Figure 7.2. The distribution for df = 2 is very asymmetric with a long tail to the right, suggesting that there is tremendous variability in the sample variances. This large variation is expected as each sample variance was based on only three observations. When we increase the sample size to 6 (df = 5), the distribution of sample variances is not so asymmetric and the tail to the right is much shorter than in the first distribution. When we increase the sample size to 11 (df = 10), the distribution of sample variances is almost symmetric. We can see that the sampling distributions for the three samples sizes are very different; that is, they depend on the sample size.

It appears that the distribution of the sample variance does not match any of the probability distributions we have encountered so far. Fortunately, when the data come from a normal distribution, the distribution of the sample variance is known. The sample variance (s^2), multiplied by $(n - 1)/\sigma^2$, follows a *chi-square* (χ^2) *distribution*. Two

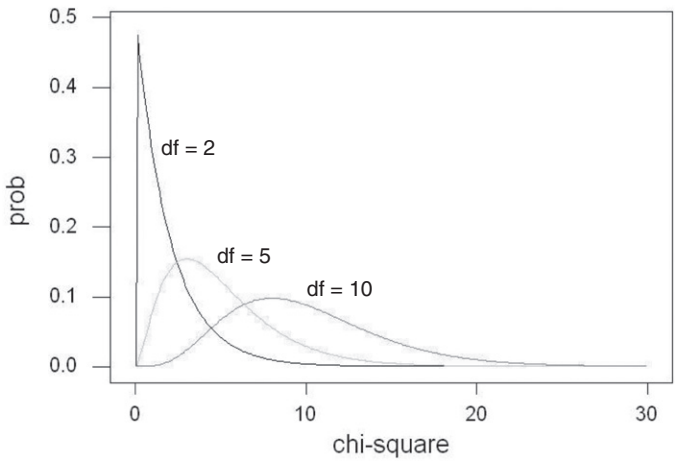


Figure 7.2 Chi-square distributions with df = 2, df = 5, and df = 10.

eminent 19th-century French mathematicians, Laplace and Bienaymé, played important roles in the development of the chi-square distribution. Karl Pearson, an important British statistician previously encountered in connection with the correlation coefficient, popularized the use of the chi-square distribution in the early 20th century. As we just saw, the distribution of the sample variance depends of the sample size, actually on the number of independent observations (degrees of freedom) used to calculate s^2. Therefore, Appendix Table B7 shows percentiles of the chi-square distribution for different values of the degrees of freedom parameter.

To create a confidence interval for the population variance, we begin with the probability statement

$$\Pr\left\{\chi^2_{n-1,\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-1,1-\alpha/2}\right\} = 1-\alpha.$$

This statement indicates that the confidence interval will be symmetric in the sense that the probability of being less than the lower limit is the same as that of being greater than the upper limit. However, the confidence limit will not be symmetric about s^2. This probability statement is in terms of s^2, however, and we want a statement about σ^2. To convert it to a statement about σ^2, we first divide all three terms in the braces by $(n-1)s^2$. This yields

$$\Pr\left\{\frac{\chi^2_{n-1,1-\alpha/2}}{(n-1)s^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{n-1,\alpha/2}}{(n-1)s^2}\right\} = 1-\alpha.$$

The interval is now about $1/\sigma^2$, not σ^2. Therefore, we next take the reciprocal of all three terms, which changes the direction of the inequalities. For example, we know that 3 is greater than 2, but the reciprocal of 3, which is 1/3 or 0.333, is less than the reciprocal of 2, which is 1/2 or 0.500. Thus, we have

$$\Pr\left\{\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} > \sigma^2 > \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right\} = 1-\alpha$$

and reversing the directions of the inequalities to have the smallest term on the left, yields

$$\Pr\left\{\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right\} = 1-\alpha.$$

It is also possible to create one-sided confidence intervals for the population variance. For example, the lower one-sided confidence interval for the population variance is

$$\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}} < \sigma^2 < \infty.$$

Example 7.12

Let's apply this formula to an example. From 1988 to 1991, eight persons in Massachusetts were identified as having vitamin D intoxication due to receiving large doses of vitamin D_3 in fortified milk (Jacobus, Holick, and Shao 1992). The problem was

traced to a local dairy that had tremendous variability in the amount of vitamin D added to individual bottles of milk. Homogenized whole milk showed the greatest variability based on measurements made in April and June 1991, with a low value of less than 40 IUs and a high of 232,565 IUs of vitamin D_3 per quart. These values are contrasted with the requirement for at least 400 IUs (10 μg) to no more than 500 IUs of vitamin D per quart of milk in Massachusetts.

The Food and Drug Administration (FDA) found poor compliance with the requirement for 400 IUs of vitamin D per quart of vitamin D fortified milk in a 1988 survey (Holick et al. 1992). Based on this poor compliance, the FDA urged that the problem be corrected; otherwise it would institute a regulatory program. Suppose that compliance is defined in terms of the mean and standard error of the mean vitamin D concentration in milk. The mean concentration should be 400 IUs with a variance of less than 1600 IUs. To determine if a milk producer is in compliance, a simple random sample of milk cartons from the producer is selected and the amount of vitamin D in the milk is ascertained. It is decided that if the 90 percent lower one-sided confidence interval for the variance contains 1600 IUs, the process used by the producer to add vitamin D is said to be within the acceptable limits for variability. This is an approach for determining compliance that greatly favors the producer.

A random sample of 30 cartons is selected and the sample variance for the vitamin D in the milk is found to be 1700 IUs. The 90 percent confidence interval uses $\chi^2_{29,0.90}$, where the first subscript is the degrees of freedom parameter and the second subscript is the percentile value. The value from Table B7 is 39.09. The lower limit is found from [29(1700)]/39.09, which gives the value of 1261.3. Since the 90 percent confidence interval does contain 1600 IUs, the producer is said to be in compliance with the variability requirement. To find that a producer is not in compliance requires a sample variance to be at least 2156.5.

A key assumption in calculating the confidence interval for the population variance is that the data come from a normal distribution. If the data are from a very nonnormal distribution, the use of the preceding formula for calculating the confidence interval can be very misleading.

To find the confidence interval for the population standard deviation, we take the square root of the variance's confidence interval limits. Thus, the lower limit of the confidence interval for σ in the above example is 35.5 IUs.

7.6.2 Confidence Interval for the Pearson Correlation Coefficient

In Chapter 3, we presented ρ, the Pearson correlation coefficient, which is used in assessing the strength of the linear relation between two jointly normally distributed variables. We presented a formula for finding r, the sample Pearson correlation coefficient. We also found the correlation between systolic and diastolic blood pressures, based on the 12 adults in Example 3.18, to be 0.894, suggestive of a strong positive relation. Although this point estimate of ρ is informative, more information is provided by the interval estimate. For example, if the sampling variation of r were so large that

the 95 percent confidence interval for ρ contains zero, we would not be impressed by the strength of the relation between total fat and protein.

It turns out that the sampling distribution of r is not easily characterized. However, the father of modern statistics, Ronald Fisher, showed that a transformation of r approximately follows a normal distribution. This transformation is

$$z' = 0.5[\log_e(1 + r) - \log_e(1 - r)]$$

and it provides the basis for the confidence interval for ρ. The mean of z′ is $[\log_e (1 + \rho) - \log_e (1 - \rho)]$ and its standard deviation, $\sigma_{z'}$, is $1/\sqrt{(n-3)}$. Note that for convenience, $\log_e$ is often written as ln, and we shall do that following. Thus, we can employ the procedures we have just used for finding the confidence interval for the transformed value of ρ — that is,

$$z' - z_{1-\alpha/2}\sigma_{z'} < 0.5[\ln(1 + \rho) - \ln(1 - \rho)] < z' + z_{1-\alpha/2}\sigma_z.$$

There is one simplification we can make that allows us to have to take only one natural logarithm in the calculation instead of finding two natural logarithms. In the presentation of the geometric mean in Chapter 3, we saw that the sum of logarithms of two terms is the logarithm of the product of the terms — that is,

$$\ln x_1 + \ln x_2 = \ln(x_1 x_2).$$

In the same way, the difference of logarithms of two terms is the logarithm of the quotient of the terms — that is,

$$\ln x_1 - \ln x_2 = \ln\left(\frac{x_1}{x_2}\right).$$

Thus, we have the relation

$$z' = 0.5[\ln(1+r) - \ln(1-r)] = 0.5\ln\left(\frac{1+r}{1-r}\right).$$

Let us apply these formulas for finding the 95 percent confidence interval for the correlation between systolic and diastolic blood pressure for 12 adults just mentioned. Since r is 0.894, z' is

$$(0.5)\ln\left(\frac{1+0.894}{1-0.894}\right) = (0.5)\ln 17.8679 = (0.5)2.8830 = 1.4415.$$

The standard deviation of z' is $1/\sqrt{12-3}$, which is 0.3333. Thus the interval for $(0.5)\ln[(1 + \rho)/(1 - \rho)]$ is from 01.4415 − 1.96(0.3333) to 1.4415 + 1.96(0.3333) or from 0.7882 to 2.0948.

To find the confidence interval for ρ, we first perform the inverse transformation on twice the lower and upper limits of the interval just calculated. The inverse transformation of the natural logarithm, ln, is the exponential transformation. This means that

$$\exp(\ln x) = x.$$

After obtaining the exponential of twice a limit, call it a, further manipulation leads to the following equation:

$$\text{limit for } \rho = \frac{a-1}{a+1}.$$

The exponential of twice the lower limit — that is, two times 0.7882 — is the exponential of 1.5764, which is 4.83785, and this is the value used for a for the lower limit. The lower limit for ρ is

$$\frac{\exp[2(0.7882)]-1}{\exp[2(0.7882)]+1} = \frac{4.8375-1}{4.8375+1} = 0.657.$$

Similarly, the upper limit for ρ is

$$\frac{\exp[2(2.0948)]-1}{\exp[2(2.0948)]+1} = \frac{65.9926-1}{65.9926+1} = 0.970.$$

Therefore, the 95 percent confidence interval for the Pearson correlation coefficient between systolic and diastolic blood pressure in the population is from 0.657 to 0.970. The interval does not include 0. Thus, it is reasonable to conclude that there is a strong positive association between systolic and diastolic blood pressures among patients in the DIG clinical trial. These calculations are easily performed by a program (see **Program Note 7.4** on the website). The preceding material also applies to the Spearman correlation coefficient for sample sizes greater than or equal to 10.

7.7 Prediction and Tolerance Intervals Based on the Normal Distribution

As we have seen, knowledge that the data follow a specific distribution can be used effectively in the creation of confidence intervals. This knowledge can also be used in the formation of prediction and tolerance intervals, and this use is shown next.

7.7.1 Prediction Interval

The distribution-free method for forming intervals used specific observed values of the variable under study. In contrast, the formation of intervals based on the normal distribution uses the sample estimates of its parameters: the mean and standard deviation. Assuming that the data follow the normal distribution, the prediction interval is formed by taking the sample mean plus or minus some value. This form is the same as that used in the construction of the confidence interval for the population mean. However, we know that the prediction interval will be much wider than the confidence interval, since the prediction interval focuses on a single future observation.

The confidence interval for the mean, when the population variance is unknown, is

$$\bar{x} \pm t_{n-1,1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right).$$

The estimated standard error of the sample mean, $s/\sqrt{n}$, can also be expressed as $\sqrt{[s^2(1/n)]}$. The variance of a future observation is the sum of the variance of an observation about the sample mean and the variance of the sample mean itself, that is, $\sigma^2 + \sigma^2/n$. Thus, the estimated standard error of a future observation is $\sqrt{[s^2(1+1/n)]}$ and the corresponding prediction interval is

$$\bar{x} \pm t_{n-1,1-\alpha/2} s \sqrt{1+\frac{1}{n}}.$$

Let us calculate the prediction interval for the systolic blood pressure data just used in the calculation of the 90 percent confidence interval for the mean. The sample mean was 94.75 mmHg, and the sample standard deviation was 10.25 mmHg, based on a sample size of 60. The value of $t_{59,0.95}$ used in the 90 percent confidence interval was 1.671. The value of $s\sqrt{(1+1/n)}$ is $10.335\left(=10.25\sqrt{[1+1/60]}\right)$. Therefore, the prediction interval is

$$94.75 \pm 1.671\ (10.335)$$

and the lower and upper limits are 77.48 and 112.02 mmHg, respectively. These values are contrasted with 92.54 and 96.96 mmHg, limits of the confidence interval for the mean. Thus, as expected, the 90 percent prediction interval for a single future observation is much wider than the corresponding 90 percent confidence interval for the mean.

7.7.2 Tolerance Interval

The tolerance interval is also formed by taking the sample mean plus or minus some quantity, k, multiplied by the estimate of the standard deviation. Since the derivation of k is beyond the level of this book, we shall simply use its value found in Table B8. In symbols, the $(1 - \alpha) * 100$ percent tolerance interval containing p percent of the population based on a sample of size n is

$$\bar{x} \pm k_{n,p,1-\alpha}\, s.$$

Let us use Table B8 to find the 90 percent tolerance interval containing 95 percent of the systolic blood pressure values in the population based on the first sample of 60 observations from above. From Table B8 we find that $k_{60,0.95,0.90}$'s value is 2.248. Therefore, the tolerance interval is

$$94.75 \pm 2.248\ (10.25)$$

which gives limits of 71.71 and 117.79. One-sided prediction and tolerance intervals based on the normal distribution are also easy to construct.

Conclusion

In this chapter, the concept of interval estimation was introduced. We presented prediction, confidence, and tolerance intervals and explained their applications. We showed how distribution-free intervals and intervals based on the normal distribution were calculated. The idea and use of confidence intervals discussed in this chapter will be explored further to introduce methods of testing statistical hypotheses in the next two chapters. Parenthetically, it is worth pointing out that the idea of confidence interval is often expressed as a margin of error in journalistic reporting, which refers to one-half of the width of a two-sided confidence interval.

We also pointed out that characteristics — for example, size — of the intervals could be examined before actually conducting the experiment. If the characteristics of the

interval are satisfactory, the investigator uses the proposed sample size. If the characteristics are unsatisfactory, the design of the experiment, the topic of the next chapter, needs to be modified.

EXERCISES

7.1 Assume that the AML patients shown in Exercise 3.7 can be considered a simple random sample of all AML patients.
   a. Calculate the 95 percent confidence interval for the population mean survival time after diagnosis for AML patients.
   b. Interpret this confidence interval so that someone who knows no statistics can understand it.
   c. Calculate the approximate 95 percent confidence interval for the median survival time. Compare the intervals for the population mean and median.
   d. There are two methods for forming the tolerance interval. Use both methods to form the approximate 95 percent tolerance interval containing 90 percent of the survival times for the population of AML patients. Which method do you think is the more appropriate one to use here? Provide your rationale.

7.2 Calculate a 90 percent confidence interval for the population median length of stay based on the data from the patient sample shown in Exercise 3.10. Is it appropriate to calculate a confidence interval for the population mean based on these data? Support your answer.

7.3 Find a study from the health literature that uses confidence intervals for one of the statistics covered in this chapter. Provide a reference for the study and briefly explain how confidence intervals were used.

7.4 The following table shows the average annual fatality rate per 100,000 workers based on the 1980–1988 period by state along with the state's composite score on a scale created by the National Safe Workplace Institute (NSWI). The scale takes into account prevention and enforcement activities and compensation paid to the victim. The data are taken from the Public Citizen Health Research Group (1992).

State	Fatality[a] Rate	NSWI[b] Score	State	Fatality Rate	NSWI Score	State	Fatality Rate	NSWI Score
CT	1.9	65	SC	6.7	26	LA	11.2	31
MA	2.4	73	VT	6.8	38	NE	11.3	27
NY	2.5	76	IL	6.9	76	NV	11.5	30
RI	3.3	59	NC	7.2	47	TX	11.7	72
NJ	3.4	80	WA	7.7	55	KY	11.9	32
AZ	4.1	40	IN	7.8	47	NM	12.0	14
MN	4.3	64	ME	7.8	67	AR	12.5	11
NH	4.5	56	TN	8.1	24	UT	13.5	26
OH	4.8	55	OK	8.7	53	ND	13.8	21
MI	5.3	63	AL	9.0	25	MS	14.6	25
MO	5.3	42	KS	9.1	15	SD	14.7	25
MD	5.7	46	IA	9.2	54	WV	16.2	47
DE	5.8	40	CO	9.3	52	ID	17.2	22
HI	6.0	25	FL	9.3	48	MT	21.6	28
PA	6.1	55	VA	9.9	60	WY	29.5	12
WI	6.3	58	GA	10.3	36	AK	33.1	59
CA	6.5	81	OR	11.0	63			

[a]Average annual fatality rate per 100,000 workers based on 1980–1988 data
[b]National Safe Workplace Institute Score (116 is the maximum and a higher score is better)

During the 1980–1988 period, the National Institute of Occupational Safety and Health reported that there were 56,768 deaths in the workplace. The preceding rates are based on that number. The National Safety Council reported 105,500 deaths for the same period. Do you think that there should be any relationship between the fatality rates and the NSWI scores? If you think that there is a nonzero correlation, will it be positive or negative? Explain your reasoning. Calculate the Pearson correlation coefficient for these data. Is there any reason to calculate a confidence interval based on the correlation value you calculated? Why or why not?

7.5 There is some concern today about excessive intakes of vitamins and minerals, possibly leading to nutrient toxicity. For example, many persons take vitamin and mineral supplements. It is estimated that 35 percent of the adult U.S. population consumes vitamin C in the form of supplements (LSRO 1989). Based on survey results, among users of vitamin C supplements, the median intake was 333 percent of the recommended daily allowance. Suppose that you take a tablet that claims to contain 500 mg vitamin C. Which type of interval — prediction, confidence, or tolerance — about the vitamin C content in the tablets is of most interest to you? Explain your reasoning.

7.6 In a test of a laboratory's measurement of serum cholesterol, 15 samples containing the same known amount (190 mg/dL) of serum cholesterol are submitted for measurement as part of a larger batch of samples, one sample each day over a three-week period. Suppose that the following daily values in mg/dL for serum cholesterol for these 15 samples were reported from the laboratory:

180	190	197	199	210	187	192	199	214	237	188	197	208	220	239

Assume that the variance for the measurement of serum cholesterol is supposed to be no larger than 100 mg/dL. Construct the 95 percent confidence interval for this laboratory's variance. Does 100 mg/dL fall within the confidence interval? What might be an explanation for the pattern shown in the reported values?

7.7 The percentage of persons in the United States without health insurance in 1991 was 14.1 percent, or approximately 35.5 million persons. The following data show the percent of persons without health insurance in 1991 by state (PCHRG 1993) along with the 1990 population of the state (U.S. Bureau of the Census 1991). The District of Columbia is treated as a state in this presentation. Calculate the sample Pearson correlation coefficient between the state population total and its percent without health insurance. How can these counts be viewed as a sample? Calculate a 95 percent confidence interval for the Pearson correlation coefficient in the population. Does there appear to be a strong linear relation between these two variables? Provide at least one additional variable that may be related to the proportion without health insurance in each state and provide a rationale for your choice.

State	Population[a]	Percent without Health Insurance	State	Population	Percent without Health Insurance
New England			East South Central		
ME	1.23	11.1	KY	3.69	13.1
NH	1.11	10.1	TN	4.88	13.4
VT	0.56	12.7	AL	4.04	17.9
MA	6.02	10.9	MS	2.57	18.9
RI	1.00	10.2			
CT	3.29	7.5	West South Central		
			AR	2.35	15.7
Mid Atlantic		LA	4.22	20.7	
NY	17.99	12.3	OK	3.15	18.2
NJ	7.73	10.8	TX	16.99	22.1
PA	11.88	7.8			
			Mountain		
East North Central			MT	0.80	12.7
OH	10.85	10.3	ID	1.01	17.8
IN	5.54	13.0	WY	0.45	11.3
IL	11.43	11.5	CO	3.29	10.1
MI	9.30	9.0	NM	1.52	21.5
WI	4.89	8.0	AZ	3.67	16.9
			UT	1.72	13.8
West North Central			NV	1.20	18.7
ND	0.64	7.6			
SD	0.70	9.9	Pacific		
NE	1.58	8.3	WA	4.87	10.4
KS	2.48	11.4	OR	2.84	14.2
MN	4.38	9.3	CA	29.76	18.7
IA	2.78	8.8	AK	0.55	13.2
MO	5.12	12.2	HI	1.11	7.0
South Atlantic					
DE	0.67	13.2			
MD	4.78	13.1			
VA	6.19	16.3			
WV	1.79	15.7			
FL	12.94	18.6			
NC	6.63	14.9			
SC	3.49	13.2			
GA	6.48	14.1			
DC	0.61	25.7			

[a]Population is expressed in millions

7.8 Calculate the mean state proportion of those without health insurance from data in Exercise 7.7. Is this number the same as the overall U.S. percentage? Explain how the state information can be used to obtain the overall U.S. percentage of 14.1.

7.9 Suppose you are planning a simple random sample survey to estimate the mean family out-of-pocket expenditures for health care in your community during the last year. In 1990, the approximate per capita (not per family) out-of-pocket expenditure was $525 (NCHS 1992). From previous studies in the literature, you think that the population standard deviation for family out-of-pocket expenditures is $500. You want the 90 percent confidence interval for the community mean family out-of-pocket expenditures to be no wider than $100.

a. How many families do you require in the sample to satisfy your requirement for the width of the confidence interval for the mean?

b. Do you believe that family out-of-pocket expenditures follow the normal distribution? Support your answer.

c. Regardless of your answer, assume that you said that the family out-of-pocket expenditures do not follow a normal distribution. Discuss why it is still appropriate to use the material based on the normal distribution in finding the confidence interval for the population mean.

d. In the conduct of the survey, how would you overcome reliance on a person's memory for out-of-pocket expenditures for health care for the past year?

7.10 In 1979, the Surgeon General's *Report on Health Promotion and Disease Prevention* and its follow-up in 1980 established health objectives for 1990. One of the objectives was that the proportion of 12- to 18-year-old adolescents who smoked should be reduced to below 6 percent (NCHS 1992). Suppose that you have monitored progress in your community toward this objective. In a survey conducted in 1983, you found that 17 of 90 12- to 18-year-old adolescents admitted that they were smokers. In your 1990 simple random sample survey, you found 11 of 85 12- to 18-year-old adolescents who admitted that they smoked.

a. Construct a 95 percent confidence interval for the proportion of smokers among 12- to 18-year-old adolescents in your community. Is 6 percent contained in the confidence interval?

b. Construct a 99 percent confidence interval for the difference in the proportion of smokers among 12- to 18-year-old adolescents from 1983 to 1990. Do you believe that there is a difference in the proportion of smokers among the 12- to 18-year-old adolescents between 1983 and 1990? Explain your answer.

c. Briefly describe how you would conduct a simple random sample of 12- to 18-year-old adolescents in your community. Do you have confidence in the response to the question about smoking? Provide the rationale for your answer. What is a method that might improve the accuracy of the response to the smoking question?

7.11 Construct the 95 percent confidence interval for the difference in the population mean survival times between the AML and ALL patients shown in Table 7.6. Since there appears to be a difference in mean ages between the AML and ALL patients, perhaps we should adjust for age. One way to do this is to calculate age-specific confidence intervals. For example, calculate the confidence interval for the difference in population mean survival times for AML and ALL patients who are less than or equal to 40 years old. Is the confidence interval for those less than or equal to 40 years of age consistent with the confidence interval which has ignored the ages? How else might we adjust for the age variable in the comparison of the AML and ALL patients?

7.12 Suppose we wish to investigate the claims of a weight loss clinic. We randomly select 20 individuals who have just entered the program, and we follow them for six weeks. The clinic claims that its members will lose on the average 10 pounds during the first six weeks of membership. The beginning weights and the weights after six weeks are shown following. Based on this sample of 20 individuals, is the clinic's claim plausible?

Person	Beginning Weight	Weight at 6 Weeks	Person	Beginning Weight	Weight at 6 Weeks
1	147	143	11	246	239
2	163	151	12	218	222
3	198	184	13	143	135
4	261	245	14	129	124
5	233	229	15	154	136
6	227	220	16	166	159
7	158	161	17	278	263
8	154	147	18	228	205
9	162	155	19	173	164
10	249	254	20	135	122

7.13 In a study of aplastic anemia patients, 16 of 41 patients on one treatment achieved complete or partial remission after three months of treatment compared to 28 of 43 patients on another treatment (Frickhofen et al. 1991). Construct a 99 percent confidence interval on the difference in proportions that achieved complete or partial remission. Does there appear to be a difference in the population proportions of the patients who would achieve complete or partial remission on these two treatments?

7.14 In 1970, Japanese American women had a fertility rate (number of live births per 1000 women ages 15–44) of 51.2, considerably lower than the rate of 87.9 for all U.S. women in this age group. Use the following data to calculate an age-adjusted fertility rate for Japanese American women and approximate the standard deviation of the age-adjusted rate.

Age	U.S. Age-Specific Fertility Rate	Number of Japanese American Women
15–19	69.6	24,964
20–24	167.8	23,435
25–29	145.1	22,093
30–34	73.3	23,055
35–39	31.7	32,935
40–44	8.6	34,044

Source: U.S. Population Census, 1970, P(2)-1G and U.S. Vital Statistics, 1970

REFERENCES

Fleiss, J. L. *Statistical Methods for Rates and Proportions*, 2nd ed. New York: Wiley, 1981.

Frickhofen, N., J. P. Kaltwasser, H. Schrezenmeier, "Treatment of Aplastic Anemia with Antilymphocyte Globulin and Methylprednisolone with or without Cyclosporine." *The New England Journal of Medicine* 324:1297–1304, 1991.

Gans, D. J. "Letter to the Editor — Preliminary Test on Variances." *The American Statistician* 45:258, 1991.

Holick, M. F., Q. Shao, W. W. Liu, "The Vitamin D Content of Fortified Milk and Infant Formula." *The New England Journal of Medicine* 326:1178–1181, 1992.

Jacobus, C. H., M. F. Holick, Q. Shao, et al. "Hypervitaminosis D Associated with Drinking Milk." *The New England Journal of Medicine* 326:1173–1177, 1992.

Kendall, M. G., and A. Stuart. *The Advanced Theory of Statistics, Volume 2, Inference and Relationship*, 2nd edition. New York: Hafner Publishing Company, 1967.

Lee, E. T. *Statistical Methods for Survival Data Analysis.* Belmont, CA: Wadsworth, 1980.

Life Sciences Research Office (LSRO), Federation of American Societies for Experimental Biology. *Nutrition Monitoring in the United States — An Update Report on Nutrition Monitoring*. Prepared for the U.S. Department of Agriculture and the U.S. Department of Health and Human Services. DHHS Pub. No. (PHS) 89-1255, 1989.

Markowski, C. A., and E. P. Markowski. "Conditions for the Effectiveness of a Preliminary Test of Variance." *The American Statistician* 44:322–326, 1990.

Moser B. K., and G. R. Stevens. "Homogeneity of Variance in the Two-Sample Means Test." *The American Statistician* 46:19–21, 1992.

National Center for Health Statistics. *Health, United States, 1991 and Prevention Profile*. Hyattsville, Maryland: Public Health Service. DHHS Pub. No. 92-1232, 1992.

The NHLBI Task Force on Blood Pressure Control in Children. "The Report of the Second Task Force on Blood Pressure Control in Children, 1987." *Pediatrics* 79:1–25, 1987.

Public Citizen Health Research Group. "Work-Related Injuries Reached Record Level Last Year." *Public Citizen Health Research Group Health Letter* 8(12):1–3, 9, 1992.

Public Citizen Health Research Group (PCHRG). "The Growing Epidemic of Uninsurance." *Public Citizen Health Research Group Health Letter* 9(1):1–2, 1993.

Satterthwaite, F. E. "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin* 2:110–114, 1946.

U.S. Bureau of the Census. 1970 Census of Population, Subject Reports: Japanese, Chinese, and Filipinos in the United States, P(2)-1G, Government Printing Office, Washington, D.C.; National Center for Health Statistics (1975). *U.S. Vital Statistics*, *1970*, Volume I: Natality. Government Printing Office, Washington, D.C., 1973.

U.S. Bureau of the Census. 1990 Census of Population and Housing, Summary Tape File 1A on CD-ROM Technical Documentation prepared by Bureau of the Census. Washington: The Bureau, 1991.

Vardeman, S. B. "What About the Other Intervals?" *The American Statistician* 46:193–197, 1992.

Walsh, J. E. "Nonparametric Confidence Intervals and Tolerance Regions." In *Contributions to Order Statistics*, edited by A. E. Sarhan and B. G. Greenberg. New York: John Wiley & Sons, 1962.

Walter U., R. Forthofer, and P. U. Witte. "Dose-Response Relation of Angiotensin Converting Enzyme Inhibitor Ramipril in Mild to Moderate Essential Hypertension." *American Journal of Cardiology* 59:125D–132D, 1987.