

15 Statistical Inference: Estimation

Advisor: Congratulations, Mr. Burns, the latest polls show you are up six points.

Burns: Ah, giving me a total of...

Advisor: Six.

From: *The Simpsons*

15.1 Empirical Sampling Distribution

Statistical inference brings together the threads of data analysis and probability theory. The inference process is concerned not simply with describing a particular sample (the data), but with using this sample to make a prediction about some underlying population. For example, in polling we sample the opinions of a random selection of the population and use characteristics of the sample to infer statements about a far larger group. Judging by the enormous monetary resources spent on polling each year, the process must yield useful information to the people paying for the surveys. In this chapter we describe how probability theory allows one to extract the desired information along with some measure of confidence in the predictions.

It is often helpful in thinking about statistical problems to bear in mind a simple mental model of what may be in reality a complicated physical process. To this end we propose a “bowl model” of the sort that we have used at times in the previous chapters.

Definition 15.1 (Bowl Model): Suppose we have one or more bowls each containing numbered balls. A *bowl model* consists of making successive random selections from the bowls, recording the number on the ball and then selecting another ball. ■

At any stage, the randomness condition means that each ball remaining must have an equal chance of being selected. In this definition we do not specify whether the selected ball is replaced in the bowl (sampling with replacement) or discarded. The mathematical treatment in the latter case is a bit more complicated and will not be discussed in detail. However, for now we leave open this aspect of the scheme and illustrate how the bowl model can be used to describe some typical inference scenarios.

Example 15.1: Describe bowl models for the following inference problems:

- a) Patients undergo a new treatment for cancer. On average how long does a patient live after receiving the treatment?
- b) What fraction of the voting population supports candidate A?
- c) A pharmaceutical company wants to evaluate the effectiveness of a new antidepressant drug by testing it on a random group of patients.

Solution:

- a) Imagine a bowl in which each ball represents a person receiving the new cancer treatment. On every ball a number is written with the number of years the patient survives after receiving the treatment. This imaginary bowl would hold very many balls, many of which would only be “visible” in the future. Nonetheless, we would like to know the average of the numbers on each ball, as this represents the average survival time for a random patient. To estimate this population average, we draw a certain number of balls (patients) from the bowl at random and find the average of the numbers on the balls from this sample.
- b) We imagine a large bowl containing a ball for each voter. A “1” is written on a ball if the person favors candidate *A* and a “0” otherwise. Of course the number on a ball may change over time, but at any particular moment we imagine that the distribution of zeroes and ones is fixed. We want to know the proportion of the balls that have a one. To estimate this proportion we randomly select balls and record the number on each. We then find the average number appearing on the selected balls. For example, if we selected 10 balls numbered 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0 then the average would be $\frac{0+1+1+1+0+1+0+1+1+0}{10} = \frac{6}{10} = 0.6$. In other words, the average is just the relative frequency of the outcome “1” (supporter of *A*) in the sample.
- c) The bowl model here is similar to b). The balls in the bowl represent depressed patients. After a number of weeks, a “1” (improvement) or “0” (no improvement) is written on each ball depending on the treatment outcome. To evaluate the medication a certain number of patients are selected at random and after treatment we record the “0” or “1” evaluation. As in b) averaging the recorded numbers gives the frequency of improved patients in the sample. ■

In this and the next chapter we will be concerned with the mathematical analysis of the results of sampling from a bowl model, assuming the sampling has been done correctly. In the bowl model this means that the balls that are examined in a sample are chosen in such a way that each ball has an equal chance of being selected from the entire population of balls remaining in the bowl. The goal of this process is to produce a representative sample of the population. Statisticians call this *simple random sampling*, but in practice it is not a simple matter to ensure that we have produced a representative sample of the intended population. In chapter 7 we have already cited the *Literary Digest* poll of 1936 which, in spite of its large size, produced inaccurate predictions because of its non-random selection criteria. The solution to Example 15.1c) proposed above is also defective in this regard. Perhaps unwittingly the group of patients from whom we are selecting our subjects contains people with unknown characteristics that might affect the outcome of the treatment. To guard against this possibility the study should be performed with a control group who are randomly selected from the same patients, but are treated with a known therapeutic agent. When possible, this is done in a so-called *double-blind* study in which neither the patient or examining physician knows which treatment a given subject is receiving. In such a scheme any difference in outcomes can usually be attributed to the difference in treatment between the groups. We refer the interested reader to the bibliography for further references on this topic.

In the remainder of this section we describe the outcome of some computer experiments with bowl models. In these experiments we work with bowls whose composition is known and empirically examine the characteristics of the samples. In section 15.3 we describe the more precise

theoretical properties of samples that are needed to make exact predictions about bowls whose composition is unknown.

Example 15.2: A bowl contains 800 balls. 100 balls have a “1” written on them, 200 have a “2” and 500 have a “10”.

- What is the average number written on the balls?
- Draw 20 samples (with replacement) each consisting of 10 balls. Compute the average of the numbers on the 10 balls in the each sample. Describe the distribution of the 20 averages. In particular, how close are these averages to the population average found in a)?
- Repeat the instructions in b) but using 20 replicates each consisting of a selection of 100 balls from the bowl.

Solution:

- The average is the sum of all the numbers on the 800 balls divided by the total number of balls, 800. Taking into account the balls with the same number we have

$$\mu = \frac{1(100) + 2(200) + 10(500)}{800} = \frac{5500}{800} = 6.875$$

We use μ here, rather than, \bar{x} because the average computed here refers to a value associated with the entire population, not simply a sample. Usually, this number will not be known and our objective is to estimate its value from the average of a sample.

- We can use *Excel* to simulate sampling with replacement by selecting random numbers according to the probabilities given in a discrete probability distribution table. In section 13.5.2 we described how this is done. Here the random variable X is the number on the selected ball. If the individual balls have equal chance of being selected then, because we are sampling with replacement, we are using the probability distribution for X given in the following table:

X	1	2	10
$P(X = x)$	$\frac{100}{800} = .125$	$\frac{200}{800} = .25$	$\frac{500}{800} = .625$

Table 15.1

We draw 20 columns, each with 10 random numbers (1, 2 or 10) selected according to this distribution, and then find the averages of the 10 numbers in each selection. For example, the first set of 10 random values that were selected by *Excel* were the numbers

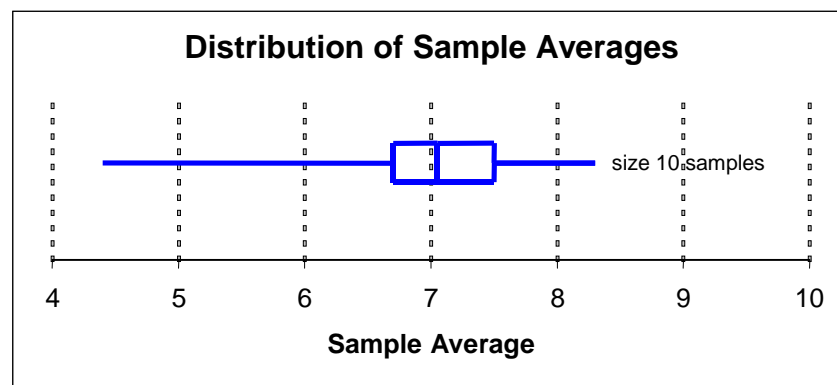
10	2	1	10	10	2	10	10	10	10
----	---	---	----	----	---	----	----	----	----

whose average is 7.5. The averages for the remaining 19 samples are shown in the following table.

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
7.5	7.5	8.3	6.7	8.3	8.3	5.8	7.5	7.4	6.8
Sample 11	Sample 12	Sample 13	Sample 14	Sample 15	Sample 16	Sample 17	Sample 18	Sample 19	Sample 20
6.7	5.7	8.2	4.4	6.7	7.3	6.7	7.5	6.6	6.8

Table 15.2

Some of the sample averages are quite far from the population average. In fact in samples 3, 5, 6 and 13 the sample average is more than 20% higher than the population average of 6.875. We can get a better sense of the dispersion of these sample averages using a graphical representation. As we have only 20 data points to consider (the 20 sample averages) a box plot would be a better pictorial representation than a histogram.



The median for the sample averages is a little above 7 so it is fairly close to the population average. In fact, the average of the 20 sample averages in Table 15.2 is 7.035, also fairly close to the population average of 6.875.

- c) We repeat the analysis above but using samples of size 100. The averages for the 20 samples are listed in Table 15.3.

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
5.91	7.02	7.24	6.96	5.68	7.21	6.74	7.17	7.03	6.73
Sample 11	Sample 12	Sample 13	Sample 14	Sample 15	Sample 16	Sample 17	Sample 18	Sample 19	Sample 20
6.25	6.83	7.08	7.23	6.91	6.54	6.78	7.27	7.18	7.75

Table 15.3

The variation from sample to sample appears much less pronounced than for samples of size 10. Comparing box plots for the two sets of numbers in Table 15.2 and Table 15.3 makes this apparent.

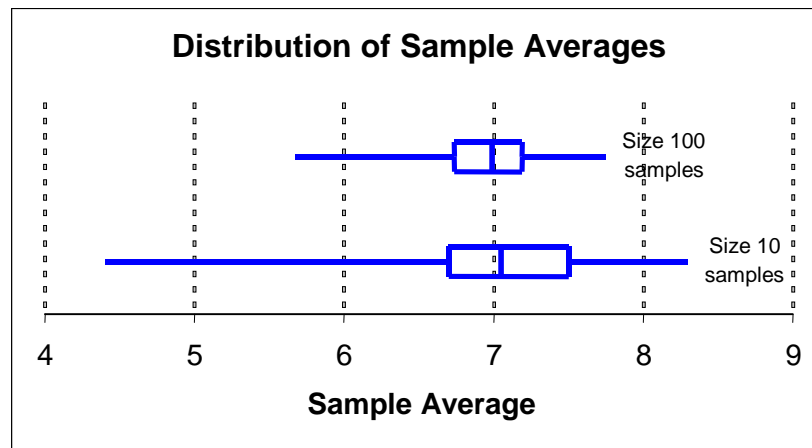


Figure 15.1

Notice that in either situation the center of the distribution lies close to the population average. For the larger sample size the distribution of sample averages appears more symmetric and more concentrated near the center. We will make these statements more precise in the next section. ■

15.2 Theoretical Sampling Distributions

The reader should have gleaned three ideas from the previous section.

- Although we are accustomed to thinking of the average for a data set as simply a number, we now want to consider how this number itself varies as we select different samples from a bowl model. *Thus, if we select n values, x_1, x_2, \dots, x_n from a bowl model then the average \bar{x} of these n values will vary from sample to sample*, as demonstrated in Table 15.1 and Table 15.2 above. Thus these *sample averages define a new random variable* which we denote by \bar{X}_n or simply \bar{X} if the reference to the number of sampled values is not needed.
- The center of the distribution of values of the random variable \bar{X} lies close to the mean μ of the population. (See Figure 15.1)
- As n increases the values of \bar{X} are more tightly compressed around the center. (See Figure 15.1)

In this section we will make the last two statements more precise. We will not attempt to prove any of the stated results, but we hope the reader will be convinced of their plausibility through the

evidence presented from simulations. In section 15.3 these precise characterizations will be used to construct confidence interval estimations for various population statistics.

Theorem 15.1 (Expected Value of \bar{X}): Suppose we have a bowl model for which the number X appearing on the balls has a (population) average of μ . If \bar{X} is the mean of a random sample of size n drawn from the bowl then the expected value of \bar{X} is also μ . Symbolically we write this as $\mu_{\bar{X}} = \mu_X$ or $E(\bar{X}) = E(X)$. ■

Note that this result is correct whether we draw from the bowl with replacement or without replacement. We illustrate the conclusion by referring to the data in Example 15.2.

Example 15.3: Compute the overall mean of the random variable X giving the value of the balls in the bowl model described in Example 15.2 and then compare this with the mean for the 20 values of \bar{X}_{10} and \bar{X}_{100} given in Table 15.2 and Table 15.3 respectively.

Solution:

In Example 15.2 we have already computed the mean μ_X for the balls in the bowl. We found that $\mu_X = 6.875$. For the samples of size 10 the mean for the 20 replicates listed in Table 15.2 (which we denote by \bar{x}_{10}) is $\bar{x}_{10} = 7.035$. In particular, the mean is quite close to the population mean. If we had taken more replicates we would probably have found that the sample means produced an average value even closer to the value of μ . For the samples of size 100 the 20 replicates listed in Table 15.3 have a mean $\bar{x}_{100} \approx 6.875$. Here we see vivid confirmation of Theorem 15.1. ■

Theorem 15.1 asserts that if we take repeated random samples of a fixed size from a bowl model, then the averages for these replications will themselves average out to the population average. From a practical viewpoint this result is not very useful since we rarely have the resources to investigate more than one sample from our bowl model. In order that this one sample provide a reliable estimate of the population average, we must be confident that the mean from an individual replicate is very close to the population average. Figure 15.1 certainly suggests that as the sample size increases the sample averages tend to cluster more tightly around the center of the distribution for \bar{X} . We can measure this effect precisely in terms of the standard deviation. The important theorem below is one of the main reasons why the rather complicated standard deviation (and the associated variance) plays such a prominent role in probability theory.

Theorem 15.2 (Standard Error of the Mean): Suppose we have a bowl model for which the number X appearing on the balls has a (population) standard deviation of σ . If we draw n independent random samples from the bowl, then the standard deviation of the sample average \bar{X} satisfies $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Remarks:

The expression “*Standard Error of the Mean*” or simply “*standard error (SE)*” is an old-fashioned (and sometimes confusing) bit of terminology. The word “error” in the expression signifies not a mistake, but rather variation. The term should be viewed as an abbreviation for the more precise expression “standard deviation of the sample mean” or $\sigma_{\bar{x}}$. **Note that while the standard deviation σ is a number measuring the variability of the underlying population, the standard error measures the variability of the sampling process.** ■

Theorem 15.2 requires that the sampling from the bowl be done in such a way that the chance of selecting a specific ball does not depend on which balls have already been selected. Sampling with replacement certainly meets this condition. Sampling without replacement does not satisfy it. For example, in the latter scheme selecting a ball precludes its ever being selected again. However, if the number of balls in the bowl is much greater than the total number of balls in a sample we can for practical purposes consider that samples drawn without replacement are approximately independent.

Qualitatively, the theorem shows that the spread of values of \bar{X}_n diminishes to zero as $n \rightarrow \infty$. Thus, when we take a sample average using a large n it is likely that any single sample average is close to the average for the population. This enables us to use a single such sample average as a usually reliable estimate for the population mean.

Theorem 15.3 below gives a more precise measure of how likely it is for a value of the sample average to fall a specified distance from μ .

Example 15.4: Compare the standard deviation for the data in Table 15.2 and Table 15.3 with the values predicted by Theorem 15.2.

Solution:

We first must find the standard deviation for the random variable X . This can be done using Table 15.1, the probability distribution table for the random variable X , and Definition 13.5 in Chapter 13. We get

$$\sigma_x = \sqrt{.125(1 - 6.875)^2 + .25(2 - 6.875)^2 + .5(10 - 6.875)^2} \approx 4.045.$$

According to Theorem 15.2 the standard deviation for the sample means based on 10 sample values should be $\sigma_{\bar{x}_{10}} = \frac{\sigma}{\sqrt{10}} \approx \frac{4.045}{3.16} \approx 1.28$. The 20 replications listed in Table 15.2 produced a standard deviation (which we denote by s_{10}) of $s_{10} \approx .98$. Similarly, for the samples of size 100 the theoretical value of the standard deviation for \bar{X}_{100} is $\sigma_{\bar{x}_{100}} = \frac{\sigma}{\sqrt{100}} \approx \frac{4.045}{10} = .4045$. The standard deviation computed from the data in Table 15.3 was $s_{100} \approx .49$. We observe here the

general effect predicted by Theorem 15.2, though a more impressive verification would require collecting a larger number of replicates. ■

Theorem 15.1 and Theorem 15.2 give us some information about the sample average random variable, \bar{X} . Our next and last theorem tells us much more about the actual probability distribution of \bar{X} . This theorem is a special case of a more general result in probability theory known as the *Central Limit Theorem*, and so we refer to it by that name.

Theorem 15.3 (Central Limit Theorem): Suppose we have a bowl model for which the number X appearing on the balls has a (population) mean of μ and a (population) standard deviation of σ . For independently drawn random samples of size $n > 30$, the sample average \bar{X}_n has an approximately normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. In other words,

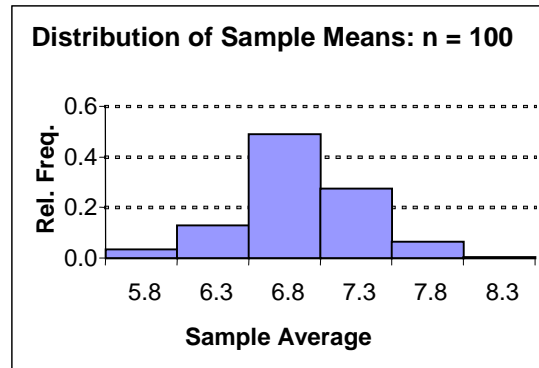
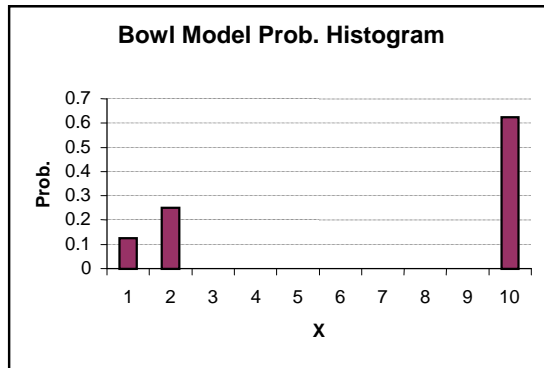
$$\bar{X}_n \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right). \blacksquare$$

A truly remarkable feature of this result is that the sample averages approach a normal distribution no matter what the distribution of values X on the balls in the bowl, as long as the samples are drawn independently and the sample size is larger than approximately 30. We will also use this result and its corollary Theorem 15.4 below, when our bowl contains very many balls and we draw a random sample without replacement that is comparatively small compared with the number of balls in the bowl. When the distribution of values in the bowl is not very mound-shaped it may require sample sizes considerably larger than 30 before the normal approximation described in Theorem 15.3 gives accurate estimates for the distribution of \bar{X} .

Example 15.5: Draw the theoretical probability histogram for the random variable X described in Table 15.1. Using a computer, generate 200 samples each of size 100 from the bowl model described by this probability distribution and assess (via the Bell-Curve rule) how well the normal distribution described in Theorem 15.3 fits the 200 sample averages.

Solution:

The histograms are shown below. The distribution of the sample means appears reasonably bell-shaped.



To test the Bell Curve Rule, we list in the following table the sample averages for each of the 200 replicates. For convenience in checking the Bell Curve Rule these have been arranged in ascending order from left to right.

5.7	5.77	5.79	5.82	5.86	5.89	5.97	6.11	6.15	6.18	6.19	6.2	6.26	6.26	6.28	6.29	6.29	6.31	6.32	6.34
6.36	6.37	6.38	6.39	6.39	6.4	6.42	6.42	6.44	6.45	6.48	6.5	6.5	6.51	6.51	6.52	6.53	6.53	6.54	6.54
6.55	6.55	6.56	6.57	6.57	6.57	6.58	6.58	6.58	6.59	6.59	6.59	6.6	6.6	6.6	6.61	6.61	6.64	6.64	6.65
6.65	6.66	6.66	6.66	6.67	6.67	6.67	6.68	6.68	6.72	6.72	6.72	6.73	6.74	6.74	6.74	6.75	6.76	6.77	6.77
6.79	6.79	6.79	6.79	6.8	6.8	6.81	6.82	6.82	6.82	6.83	6.83	6.83	6.83	6.83	6.84	6.84	6.85	6.85	6.86
6.87	6.88	6.88	6.89	6.89	6.89	6.89	6.89	6.9	6.91	6.91	6.92	6.92	6.92	6.93	6.93	6.93	6.94	6.94	6.95
6.95	6.95	6.96	6.97	6.98	6.98	7	7	7	7	7.01	7.03	7.03	7.05	7.05	7.06	7.07	7.09	7.09	
7.1	7.11	7.13	7.14	7.14	7.15	7.15	7.15	7.16	7.16	7.16	7.18	7.18	7.19	7.19	7.19	7.2	7.21	7.21	7.23
7.24	7.24	7.25	7.25	7.25	7.26	7.26	7.28	7.29	7.3	7.3	7.31	7.32	7.33	7.33	7.33	7.36	7.36	7.39	7.42
7.43	7.45	7.46	7.46	7.47	7.49	7.51	7.52	7.52	7.55	7.55	7.58	7.59	7.59	7.62	7.66	7.79	7.81	7.82	8.02

Table 15.4: Sample Means

The average value of the sample means in Table 15.4 is 6.87 (recall from Example 15.2 that the mean for the population of all balls in the bowl is $\mu = 6.875$) and the standard deviation of the 200 sample mean replicates in Table 15.4 is 0.422. This compares well with the theoretical value (Theorem 15.2) of $\frac{\sigma_x}{\sqrt{n}} = \frac{4.045}{\sqrt{100}} \approx .404$ (remember, the n in the latter formula refers to the size of each sample, not the number of replicates, 200). We now find the fraction of replicates for which the sample average falls in the intervals $[\bar{x} - s, \bar{x} + s]$ and $[\bar{x} - 2s, \bar{x} + 2s]$, where $\bar{x} = 6.87$ and $s = 0.422$. The reader can verify the results in the following table:

Interval	# of sample means	% of sample means
$[\bar{x} - s, \bar{x} + s] = [6.448, 7.292]$	140	$140/200 = 70\%$
$[\bar{x} - 2s, \bar{x} + 2s] = [6.026, 7.714]$	189	$189/200 = 94.5\%$

The percentages in column three are in close agreement with the Bell Curve Rule, as we might have anticipated from the histogram given earlier. ■

The result of Theorem 15.3 is particularly important for the bowl model in which each ball has either a “one” or a “zero”. Recall that this provides a model of the situation in which we wish to determine the fraction of a population that has a certain property, A . A ball numbered “one” denotes a member of the population having the property, while a ball numbered “zero” denotes a member without this property.

Theorem 15.4 (Central Limit Theorem for Proportions): Suppose the fraction p of a population has a property A . If we draw n independent random samples from this population, where $n > 30$, then the fraction of the sample that has property A , which we denote by \hat{p} (read “ p hat”) has an approximately normal distribution with $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$, where $q = 1 - p$.

Solution:

In our bowl model for this scenario the bowl contains balls numbered either “one” or “zero”. We let the random variable X denote the number on a selected ball. This random variable takes on only two values, 1 or 0, with probabilities p and $q = 1 - p$, respectively. In other words, the probability distribution table for X is given by

X	0	1
$P(X = k)$	q	p

We then have that $\mu_X = 0(q) + 1(p) = p$ and

$$\sigma_X^2 = (0 - \mu_X)^2 q + (1 - \mu_X)^2 p = p^2 q + (1 - p)^2 p = p^2 q + q^2 p = pq(p + q) = pq$$

so that $\sigma_X = \sqrt{pq}$. When we take a sample from the bowl with replacement of size n , the sample mean \bar{X}_n is just the number of balls in the sample with a “1”, divided by the sample size n . In other words, it is the relative frequency of “ones” in the sample, or \hat{p} . The conclusion of the theorem is then a direct consequence of Theorem 15.3, using $\mu_X = p$ and $\sigma_X = \sqrt{pq}$. ■

While the precise statement of Theorem 15.4 will be used in the next section, the reader should appreciate the qualitative significance of the result. The value of \hat{p} is always between 0 and 1; it is centered around the true frequency p with which property A is present in the population and the spread of \hat{p} around this central value p becomes increasingly narrow as the sample size n increases. These ideas are illustrated in the histograms below that show the exact distribution of \hat{p} when the true $p = .5$. The graphs are scaled so that areas under each graph represent the probability that \hat{p} will fall in a given range. The decrease in the standard deviation, $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$, as n increases is evident from the histograms. As n increases, it becomes less likely that the

frequency of success in a sample, \hat{p} , will deviate significantly from the true population frequency p .

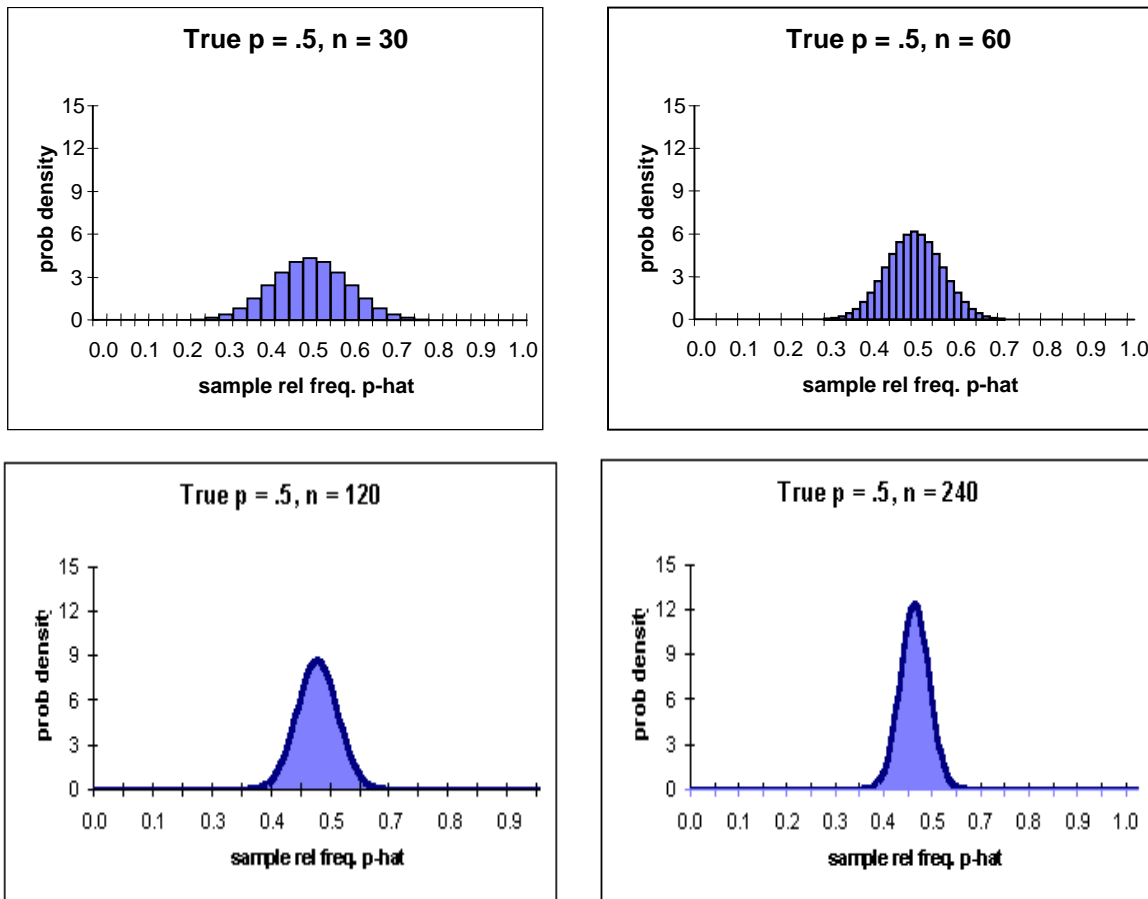


Figure 15.2(Distribution of \hat{p})

The random variable \hat{p} introduced in Theorem 15.4 is simply the relative frequency of picking a “one” in the sample of n balls. If we think of selection of a “one” as a success, and let Y be the number of successes in the sample of size n , then $\hat{p} = \frac{Y}{n}$. From Chapter 14 section 14.5 we know that Y has a binomial distribution with probability of success p and therefore for large n behaves like a normal distribution. Since \hat{p} is just a rescaled version of Y , the fact that \hat{p} also has a normal distribution should not be so surprising and in fact can be derived from the result concerning Y . Often, as in the following example, it is more direct to use Theorem 15.4 than to convert the analysis back to a question regarding Y , the number of successes.

Example 15.6: Based on past experience an airline expects that 95% of those booking a seat on a flight will actually show up. If the airline sells 300 tickets for a flight what is the probability that fewer than 92% of the purchasers will show up at departure?

Solution:

We imagine a bowl with a large number of balls representing the population of all ticket purchasers. Balls with a “one” correspond to ticket holders who show up for the particular flight and those with a “zero” represent purchasers who do not show up. The assumption stated in the problem implies that 95% of the balls have a “one”. The 300 ticket holders constitute a sample of size 300 drawn from the population of all ticket holders. In order to apply Theorem 15.4 the individual sample members must be drawn independently and at random. In practice this is not always the case as people often travel together. We will ignore this complication. In addition, since people buy only one ticket the selection occurs without replacement, which also violates the assumptions of Theorem 15.4. Since the sample is small compared to the large number of possible purchasers, the latter effect is negligible. Thus we assume the ticket buyers are drawn randomly from typical customers and their decisions regarding whether to show up or not are taken independently.

The fraction of ticket buyers that actually show for the flight is a value \hat{p} for a sample of size 300. According to Theorem 15.4, \hat{p} has an approximately normal distribution with mean equal to the population proportion .95 and a standard deviation of $\sqrt{\frac{(.95)(.05)}{300}} \approx .013$, i.e. $\hat{p} = N(.95, .013)$. We want to find the probability that $\hat{p} \leq .92$. This we can now do using the usual procedures for finding probabilities for a normal distribution. Namely, using the tables in Appendix B we obtain

$$P(\hat{p} \leq .92) = P\left(Z \leq \frac{.92 - .95}{.013}\right) = P(Z \leq -2.31) \approx .01,$$

so that in approximately 1 such flight in 100 the rate of ticket holders that show up will fall below 92%. ■

15.3 Confidence Intervals

A typical poll of voters might draw a random sample of about 1200 from a much larger population of potential voters. Let us assume that the sampling has been carried out so that those selected have been chosen independently. In practice, this means that the sampling technique must allow the possibility that an individual can be polled more than once, although this will seldom actually happen. If 800 voters in the sample favor candidate A, what can we conclude regarding the true percentage p of the voting population that favors this candidate?

The quantity $\hat{p} = \frac{800}{1200} \approx .67$ can be taken as an estimate of p . But how good is this estimate? Since \hat{p} has a normal distribution with mean equal to p (unknown) and standard deviation

$\sigma_{\hat{p}} = \sqrt{\frac{pq}{1200}}$ (also unknown), we can for instance say that there is a 95% chance that \hat{p} lies within $1.96\sigma_{\hat{p}}$, or about two standard errors, from p . In the last section we viewed this statement

as saying the interval of length $1.96\sigma_{\hat{p}}$ around p , in other words $p - 1.96\sigma_{\hat{p}}$ to $p + 1.96\sigma_{\hat{p}}$, has a 95% chance of containing the sample frequency \hat{p} . Expressing it slightly differently there is about a 95% chance that \hat{p} and p differ by at most $1.96\sigma_{\hat{p}}$. This last statement, however, is symmetric in p and \hat{p} , i.e. if I say that you and I are within 10 feet of each other then an interval 10 feet wide drawn around either of us will contain the other. In this case, this means that if instead of drawing an interval of length $1.96\sigma_{\hat{p}}$ around p , we draw an interval of the same length around \hat{p} , then 95 percent of the time this will contain the true p . Thus, we can say with 95% confidence that the true value of p lies in the interval

$$\hat{p} - 1.96\sqrt{\frac{pq}{1200}} \text{ to } \hat{p} + 1.96\sqrt{\frac{pq}{1200}} .$$

We know \hat{p} from the sample data, but we do not know p , so the interval above would seem useless. Note though that p will be close to \hat{p} . In the formula for the standard error, if we replace the unknowns p and q by the sample values \hat{p} and $\hat{q} = 1 - \hat{p}$ the small error we make will be further diminished because we are dividing the term pq by the large denominator 1200. This leads us to the important 95% confidence interval construction for an unknown population frequency.

Rule 15.1 (95% Confidence Interval for p): If a random sample of size $n > 30$ is drawn from a population in which an unknown fraction p possess property A , then if \hat{p} is the frequency of the property A in the sample, there is a 95% chance that the interval $\hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$ to $\hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$ contains the true value of p . ■

Example 15.7: Compute the 95% confidence interval for a population frequency p based on a sample of size 1200, if the sample frequency $\hat{p} = .67$.

Solution:

Using Rule 15.1 the 95% confidence interval extends from $\hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = .643$ to $\hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = .696$. Expressing the answer in percentage terms, we are 95% confident that the true p lies between 64.3% and 69.6%. Another way to express this is to say that we are 95% confident that $p = 67\% \pm 2.7\%$. The 2.7% is sometimes called the *sampling error* of the poll. It provides an estimate of the likely uncertainty of our answer due to random variation in the selected

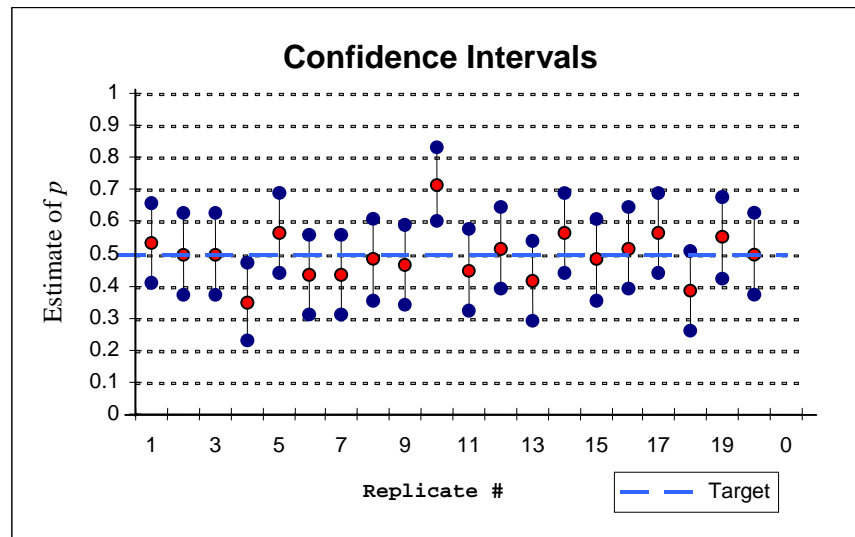
sample. The lower limit 64.3% placed on the true p is called the *lower confidence limit* (LCL) and the upper estimate, 69.6%, the *upper confidence limit* (UCL).■

⚡ **Note that the width of the confidence interval obtained in Example 15.7 (i.e. the sampling error) only depends on the size of \hat{p} and n .** It does not depend on the size of the underlying population. Thus a random sample of size 1200 from a population of 100,000 will have the same predictive value as the same size sample from a population of 10,000,000. This is the underlying reason why opinion sampling is so widely used. Accurate prediction is possible for huge populations with rather modest sized samples, if only the members in the sample are selected randomly and independently.

⚡ **However, we all know that predictions of polls are sometimes wrong. Often this may be due to the fact that there has been a change in frequency of the underlying population after the poll was conducted. Or, as with the infamous Literary Digest poll (see Chapter 7), the sampling may have been biased. *It is important to understand though that even correctly done polls must sometimes be wrong; in fact, the very meaning of 95% confidence implies that 5% of time the interval produced by Rule 15.1 will not contain the true value of p .***

What we mean when we say, as in Rule 15.1, that there is a 95% chance that a confidence interval contains the true value of p , is that when we repeatedly produce such intervals, 95% of them will contain the true value of the population parameter p . Whether a specific interval, such as the one computed in Example 15.7, really contains the true p is not a question that can be answered based on the sample. When we talk of a “confidence interval,” the confidence level with which we are working describes the reliability of the method, not the reliability of any specific interval the method produces.

For example, in the course of a long election campaign hundreds of polls will be produced in which a 95% confidence interval is reported. By the very nature of a 95% confidence interval, approximately 95% of these predictions will be correct and approximately 5% will be wrong. Unfortunately, we don’t usually have any way of deciding which predictions are correct and which are not. In the graphic below from the file *confidence intervals.xls* we have additional information that enables us to determine which intervals actually contain the true p . The simulation produces twenty 95% confidence intervals from samples of size 60 from a population with known frequency of $p = .5$ for some unspecified property. Notice that two of these intervals fail to contain the true value of p . This is only a 90% success rate, but the 95% success that we refer to for our construction is only meaningful when we produce a large number of such intervals.



The 95% confidence interval is certainly the most commonly used interval estimate for an unknown population frequency p . However, investigators may want less uncertainty or sometimes are willing to accept more uncertainty in their results. For example, suppose we would be satisfied with an interval that has a 90% chance of containing the true frequency p . We have only to refer to the standard normal distribution and ask what interval from $-z_0$ to z_0 has a total area of about .90. Referring to the figure below

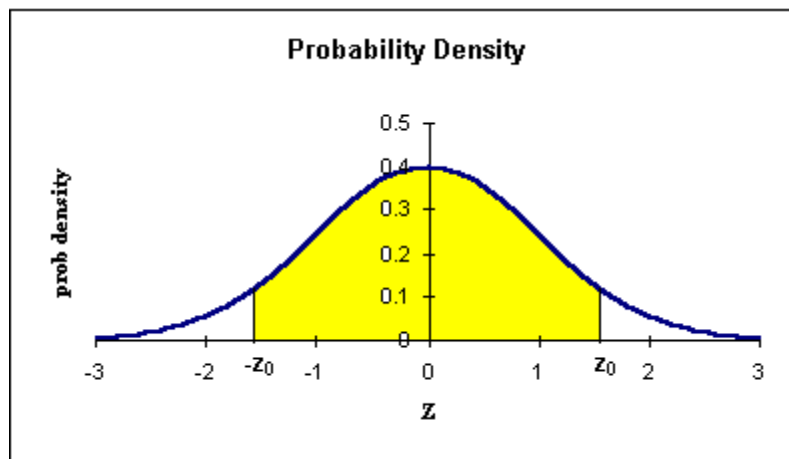


Figure 15.3

the shaded area is to have area 0.9 and therefore the area below z_0 must be .95. From Appendix B we find that $z_0 = 1.65$. In a similar way we can find the coefficients required to construct the confidence intervals listed in the following table. The tech notes at the end of the chapter show how these and other confidence coefficients may be found directly using *Excel*.

Rule 15.2 (Confidence Coefficients): If z^* is the confidence coefficient in Table 15.5, the interval $\hat{p} \pm z^* \sigma_{\hat{p}}$ gives a confidence interval with confidence level as specified in the table.

Confidence Level (%)	Confidence coefficient (# of standard errors)
90	1.65
95	1.96
98	2.33
99	2.57

Table 15.5 ■

Example 15.8: Referring to the Example 15.7 construct a 99% confidence interval for the true value of p .

Solution:

According to Rule 15.2 the 99% confidence interval extends from $\hat{p} - 2.57\sigma_{\hat{p}}$ to $\hat{p} + 2.57\sigma_{\hat{p}}$.

Using the estimate for the standard error $\sigma_{\hat{p}} \approx \sqrt{\frac{(0.33)(0.67)}{1200}} \approx .014$ given in Example 15.7 we obtain the confidence interval $67\% \pm 3.5\%$. Notice that while we have higher confidence than in Example 15.7 that the interval contains the true value of p , the price for that increased confidence is a less precise determination of the location. In general, if the sample size is fixed there is a tradeoff between the narrowness of the confidence interval and the certainty (confidence level) that it contains the quantity you wish to estimate. This is clear from the fact that the confidence coefficient (and therefore the size of the confidence interval) increases as the confidence percentage rises. ■

So far we have examined the process of estimating a population frequency. Similar considerations apply using Central Limit Theorem (Theorem 15.3) to estimation of a population mean.

Rule 15.3 (Confidence Intervals for Means): Suppose we independently draw a random sample of size $n > 30$ from a bowl model with mean μ and standard deviation σ . If the sample mean is \bar{x} and the sample standard deviation is s then confidence intervals for μ have the form $\bar{x} - z^* \frac{s}{\sqrt{n}}$ to $\bar{x} + z^* \frac{s}{\sqrt{n}}$, where z^* is the confidence coefficient in Table 15.5 for the particular confidence level desired. ■

The reasoning behind this result parallels that of Rule 15.1 for proportions. The random variable \bar{X}_n has an approximately normal distribution with mean μ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Although σ is not known, it only appears in a formula in which it is divided by a large number

\sqrt{n} , so that approximating it with s does not cause an appreciable error. Hence the given estimates.

Example 15.9: 100 patients are treated with a new cancer therapy. The average length of time the patients are in remission is 3.7 years, with a standard deviation of 1.3 years. Assuming the patients represent a random selection of patients with this type of cancer, find a 95% confidence interval for the expected time in remission for all patients receiving the treatment.

Solution:

According to Rule 15.3 the 95% confidence interval extends from $3.7 - 1.96 \frac{1.3}{\sqrt{100}} \approx 3.4$ to

$3.7 + 1.96 \frac{1.3}{\sqrt{100}} \approx 4.0$. We have 95% confidence that the true average time in remission is

between 3.4 and 4.0 years. Explicitly stating the confidence interval is good practice, but it is not universal. For instance, many investigators express “error bounds” as

$\bar{x} \pm SE = 3.7 \pm \frac{1.3}{\sqrt{100}} = 3.7 \pm 0.13$. This of course gives a smaller interval than the 95% confidence

interval, but only has about a 68% chance of containing the true value of μ . Sometimes the estimate is stated as $\bar{x} \pm 2s = 3.7 \pm 2.6$. In this case the investigator is trying to convey information about the variation in the individual times to recurrence, rather than the precision in the determination of μ , for which the standard error is the appropriate yardstick. The use of the $\pm 2s$ implicitly assumes these times are approximately normally distributed, which is probably incorrect in this case. ■

We have seen above that the width of the confidence interval depends on the sample size n as well as the variability of the population. As long as the sample size is larger than 30 the methods discussed above can be used to produce confidence intervals. However, the sampling error in these intervals (i.e. the width of the confidence interval) may be too large to be useful. For example, if we are trying to estimate the incidence of some environmentally caused illness and we obtain a sample value of $\hat{p} = 0.04$ with a sampling error of ± 0.05 , we really have not learned anything. In fact, the confidence interval contains the value zero and therefore the incidence of the disease could actually be extremely low or it could be as high as 9 or 10%. In this case we would have to redo the study with a more appropriate sample size, if that is feasible. How large a sample size is needed?

Example 15.10: Suppose in a target population a disease has an incidence of about 10%. If you want to do a study to estimate the incidence and to produce a 95% confidence interval with a sampling error of at most 1%, how large a sample must you include in your study?

Solution:

The 95% confidence interval has a sampling error of $1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$. Based on the prior information we take $\hat{p}=.1$ and $\hat{q}=.9$. We then need to find the smallest value for n such that $1.96\sqrt{\frac{(.1)(.9)}{n}}=.01$. Squaring both sides of this equation yields $3.8\left(\frac{.09}{n}\right)=.0001$, and solving for n we find $n \approx 3420$. ■

15.4 Small Samples

The discussion to this point has required that the sample size be relatively large. However, many investigations by experimenters with limited resources of time and/or money are necessarily done with small samples. Is it possible to draw conclusions from such studies? This is usually the case, but additional assumptions may be necessary and the conclusions are usually not as accurate as those based on large samples. We describe some of the ideas here but refer the reader to more specialized texts in statistics for a comprehensive treatment.

Estimating a population proportion from a small sample does not produce very reliable results. For example, from section B.4 we see that if we find 10 successes in 20 trials then the value of $\hat{p}=.50$, but the 95% confidence interval for the unknown p extends from .30 to .70, so we have very little certainty as to the true location of p . The results in section B.4 were obtained using the exact binomial distribution, rather than the normal approximation that is appropriate for larger sample sizes. Nonetheless, Rule 15.1 yields approximately the same results. For example, in the case considered above the 95% confidence interval derived from Rule 15.1 is the interval $[.28, .72]$, fairly close to the exact 95% confidence interval. When using small samples to estimate proportions, one often is willing to accept a smaller level of confidence to obtain somewhat sharper bounds for the unknown probability. In this case, 10 successes in 20 trials, an 80% confidence interval for the true p extends from .38 to .66.

When we wish to find a confidence interval for an unknown population mean based on a small random sample, we can no longer rely on the applicability of Central Limit Theorem (Theorem 15.3) to the distribution of the sample mean \bar{X} . Rather, we must make the additional assumption that the quantity X , whose expected value μ we are interested in, has itself a normal distribution with mean μ and standard deviation σ . In this case, it turns out that regardless of the sample size n the sample mean \bar{X}_n will also have a normal distribution with mean μ and standard deviation σ/\sqrt{n} .

Example 15.11: Classification of a person as mildly hypertensive (high blood pressure) is often based on a diastolic pressure reading in excess of 90 mm Hg. Assume a person's blood pressure readings show a normal distribution when taken on different occasions.

- If a person is classified as hypertensive based on a single reading and the true average diastolic pressure for that person is 80 mm Hg with a standard deviation of 7 mm Hg, what is the likelihood of a misclassification?
- If we use the average on four independent pressure readings as the basis of the classification, what is the likelihood of misclassifying the individual in a)?

Solution:

- If X denotes the pressure reading at an arbitrary occasion we are assuming that $X = N(80, 7)$. The probability that $X > 90$ is then given by

$$P(X > 90) = P\left(Z > \frac{90 - 80}{7}\right) = P(Z > 1.43) \approx 0.08$$

- Under the stated method the classification is done with the sample average $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4$. The assumption that each reading is normal and that the readings are made independently implies that \bar{X} also has a normal distribution with mean 80 and standard deviation $\sigma/\sqrt{4} = \sigma/2 = 3.5$. Repeating the calculation in a) we have

$$P(\bar{X} > 90) = P\left(Z > \frac{90 - 80}{3.5}\right) = P(Z > 2.86) \approx 0.002$$

Thus only about 0.2% (1 out of 500) of such patients would be misclassified, as opposed to about 8% using the first method. ■

If the standard deviation happens to be known, then confidence intervals for the unknown mean μ can be constructed from values of \bar{X} according to Rule 15.3, with s replaced by the known value of the standard deviation σ . However, in most situations of this type σ is also not known, and then using the recipe of Rule 15.3 does not produce correct results. The trouble arises because we are using the quantity s in the formula $\bar{x} \pm z^* \frac{s}{\sqrt{n}}$, rather than the unknown σ . When n is small we cannot neglect the error incurred through this approximation. To continue using the sample standard deviation s the formula must be modified by replacing the coefficient z^* that is taken from Table 15.5, with a coefficient based on areas beneath a family of bell-shaped (non-normal) distributions known as *Student's t*, or simply *t distributions*. (Student was the pen name used by the discoverer of these distributions, W. S. Gossett.) Figure 15.4 below shows three examples of these density curves (labeled $f = 1, f = 4, f = 16$) together with the standard normal density. Following the traditional practice, we denote a random variable with a Student distribution using lower case t , rather than capital T as our conventional notation would require.

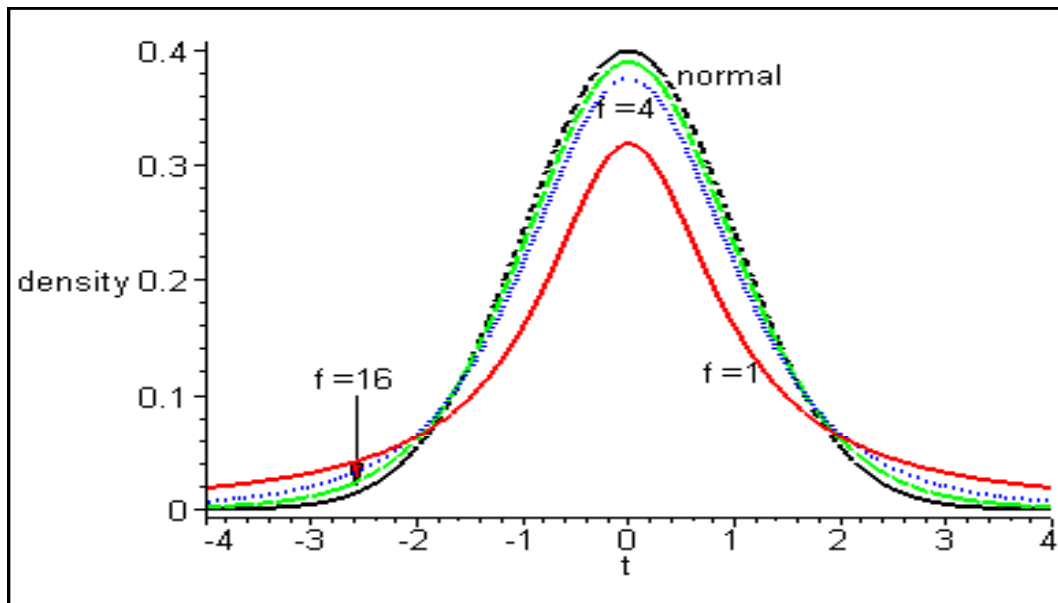


Figure 15.4 (Student's- t & Normal Density)

If our sample has size n then instead of z^* we must use a coefficient t_{n-1}^* based on the confidence level and the sample size. The quantity $f = n - 1$ is referred to as the number of *degrees of freedom* for the particular Student distribution. The precise construction of the confidence interval is given in Rule 15.4.

Rule 15.4 (Confidence Intervals for Means from a Normal Distribution): Suppose we are sampling from a bowl model which has a normal or approximately normal distribution, but for which both the mean μ and standard deviation σ are unknown. If a sample of size n has sample mean \bar{x} and sample standard deviation s then confidence intervals for μ have the form $\bar{x} - t_{n-1}^* \frac{s}{\sqrt{n}}$ to $\bar{x} + t_{n-1}^* \frac{s}{\sqrt{n}}$, where t_{n-1}^* is the confidence coefficient for the particular confidence level desired, based on a t distribution with $f = n - 1$ degrees of freedom. ■

The table below gives representative values for the coefficients t_f^* for various degrees of freedom and confidence levels of 95 and 98 %. These numbers are defined for the Student distributions as they were for the standard normal distribution (see Figure 15.3). The tech notes describe the *Excel* command needed to find these quantities. Section B.5 contains a more extensive table of these values. The column headed ∞ refers to the confidence coefficients for the normal distribution. As is evident from the table, the corresponding coefficients for the t distribution exceed these, but approach the latter values as the number of sample values increases. When n exceeds 30 there is usually little practical difference in using the coefficient from the t distribution or the coefficient based on the normal distribution, in which case Rule 15.4 reduces to the previously stated Rule 15.3.

Confidence %	Degrees of Freedom f							
	1	5	10	15	20	25	30	∞
95	12.71	2.57	2.23	2.13	2.09	2.06	2.04	1.96
98	31.82	3.36	2.76	2.60	2.53	2.49	2.46	2.33

Table 15.6 (Confidence Coefficients for Student's- t)

Example 15.12: A random sample of 21 grades (below) was selected from a collection of 285 grades on a final exam. See data file *grades.xls*. Find a 95% confidence interval for the mean grade of the entire set.

Solution:

The selected grades are given below:

27	30	80	86	54	80	52	60	88	56	97
70	57	60	69	39	48	37	5	60	78	

The mean and standard deviation for these grades are 58.7 and 22.8 respectively. Assuming the totality of grades is normally distributed or close to normally distributed, the upper and lower confidence limits can be computed from Rule 15.4, using Table 15.6. Since the sample size $n = 21$ we use the entry in Table 15.6 for the number of degrees of freedom $f = n - 1 = 20$ and the confidence level 95%. We then obtain

$$\text{LCL} = 58.7 - 2.09 \frac{22.8}{\sqrt{20}} \approx 48.0$$

$$\text{UCL} = 58.7 + 2.09 \frac{22.8}{\sqrt{20}} \approx 69.4$$

In this case, the actual mean for the entire population can be computed. The result is 63.9, which lies inside the confidence interval from the LCL to the UCL, as we have strong reason to believe it will. ■

As a final remark, we recall that in Chapter 7 when we introduced the standard deviation we noted in its definition the somewhat unnatural choice of denominator $n - 1$, rather than n . The choice of $n - 1$ gives a quantity that better approximates the standard deviation of the population and the accuracy of this approximation is important in the small-sample estimation we have discussed in this section.

15.5 Tech Notes

The confidence coefficients that we have exhibited in Table 15.5 and Table 15.6 can be easily computed using the *Excel* functions *normsinv* and *tinvs*.

Example 15.13:

- Find the confidence coefficient for the normal distribution corresponding to a confidence percent of 98%.
- Find the confidence coefficient for a Student- t with 8 degrees of freedom and a confidence percent of 98%.

Solution:

- Using a diagram similar to Figure 15.3, we need a value z_0 such that $P(-z_0 \leq Z \leq z_0) = .98$, where Z has a standard normal distribution. Using the symmetry of the distribution this is equivalent to the condition $P(Z \leq z_0) = .99$. The function *normsinv* accepts as an argument an area (probability) and produces the value of z_0 for which the area to the left of z_0 has the desired probability. In this case enter *=normsinv(.99)* and you get the answer 2.326.
- The reasoning is similar as in a) except we use the function *tinvs*. This function requires two inputs. The first is the area in the two tails, i.e. the value of $1 - P(-t_0 \leq t \leq t_0) = .02$. The second is the number of degrees of freedom. Thus we enter *=tinvs(.02,8)* and obtain the value 2.896 for the coefficient. ■

Excel can also be used to compute exact confidence intervals for the binomial parameter p (see section B.4). However, the computational methods are beyond the level of these notes.

In the next chapter we will want to compute probabilities associated with Student's distribution, rather than simply the coefficients associated with tail probabilities. Since there are a large number of Student distributions it is impractical to provide tables similar to those used for the normal distribution (section B.3). However, *Excel* has a command that can be used for this purpose. Specifically,

Example 15.14: If t has a Student distribution with 5 d.f. (degrees of freedom), find

- $P(t \geq 2.5)$
- $P(|t| \geq 2.5)$
- $P(t < 2.5)$

Solution:

- Enter the command *=tdist(2.5, 5, 1)*. This gives the area in the right tail of the distribution, i.e. for $t \geq 2.5$. You obtain .027. The last argument, 1, indicates that only the area in one tail is computed.

- b) Enter the command `=tdist(2.5,5,2)`. This gives the area in both the right and left tails (hence the 2 as the last argument), i.e. when $t \geq 2.5$ or $t \leq -2.5$. By the symmetry of the t distribution the result is twice the probability associated with a single tail (part a).
- c) Enter `=1 -tdist(2.5,5,1)`, since the event $t < 2.5$ is the complement of $t \geq 2.5$. ■

15.6 Summary

In statistical estimation we try to obtain reliable estimates of population parameters, such as the mean or a probability, from a numerical quantity computed using a sample. In this chapter we discussed how the sample mean can be used to estimate the population mean, or, as a special case, how the sample proportion can be used to estimate a true proportion in a population.

A simple, but important idea, is to regard a sample mean as not just a number, but as a random variable whose value varies from sample to sample. Provided the samples are chosen in a suitably random fashion, the distribution of the sample mean has the same expected value as that of the underlying population, i.e. $\mu_{\bar{X}} = \mu_X$, but a smaller standard deviation. In fact, the standard

deviation of \bar{X} , also called the **standard error (SE)**, is $\frac{\sigma_X}{\sqrt{n}}$, where n is the sample size. When n

is large this implies that individual values of \bar{X} will tend to fall not far from their expected value μ_X . A more precise statement can be made based on the Central Limit Theorem: When $n > 30$ the random variable \bar{X} has an approximately normal distribution with mean μ_X and standard deviation $\frac{\sigma_X}{\sqrt{n}}$.

The latter result leads to the construction of **confidence intervals** for unknown population parameters such as μ and p . A confidence interval is an interval constructed from a sample that has a specified probability of containing the unknown population parameter. For any given construction of a confidence interval, the true value of the population parameter may or may not lie in the constructed interval. But, for example, approximately 95% of all 95% confidence intervals will actually contain the parameter that supposedly is being estimated by the construction.

When the sample size is small, we may not be able to use the normal distribution to construct confidence intervals. Under suitable assumptions, a similar construction of confidence intervals may be carried out based on a family of distributions known as **Student's-t**. The following table summarizes the schema we have developed in this chapter to estimate means.

Type of Random Variable → Sample Size ↓	X arbitrary non-normal	X normal
Small Sample ($n < 30$)	Use non-parametric methods (not covered). The data can sometimes be transformed (via logs for example) to get a normal distribution.	1. If σ is known, use z coefficients and normal theory 2. If σ is unknown, estimate with s and use Student- t_{n-1}
Large Sample	Estimate σ with sample standard deviation s and use normal theory (z coefficients)	

15.7 Exercises

- Each of the following describes a certain statistical investigation. In each case formulate the investigation in terms of a bowl model; i.e. describe the composition of the bowl, the information recorded on the balls and the computation(s) made by the investigator.
 - A certain species of plant produces flowers that are either red or white. Two red flowering plants are crossed and 500 seeds from the cross are raised to mature plants. To test an inheritance model we need to know the frequency of the two flower colors among these mature offspring.
 - A college with approximately 3000 first year students wishes to do a survey to estimate these students' average weekly beer consumption.
 - A school system wants to order some classroom seats with the writing surface on the left, which is more convenient for person's who are left-handed. The administrators want a survey performed that will provide an estimate of the number of left-handed students.
 - A person's blood pressure reading (either systolic or diastolic) is often different in the right and left arms, even when taken at the same time. You wish to determine an average value for this difference in adults with normal pressure.
 - Polyps are benign tumors of the large intestine that are implicated in the development of colon cancer. You wish to compare the incidence of polyp recurrence over a period of four years in 3000 men who had previously had such a growth removed and are placed on a strict low-fat, high fiber diet, with a similar (control) group who are simply told to follow a prudent diet.
- In Chapter 10 exercise 11 we described a tetrahedral die with four faces numbered 1,2, 3, 4 each of which has equal chance of facing down (counted) when the die is tossed.
 - If X is the value of a single toss, find the average and standard deviation for X .
 - Find the probability distribution for the average $\bar{X} = (X_1 + X_2)/2$ of two independent tosses of the die.

c) Using the probability distribution in b) compute $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ and verify that you get the same answers as those predicted by Theorem 15.1 and Theorem 15.2.

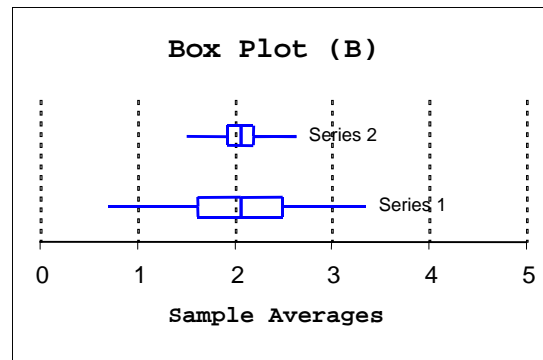
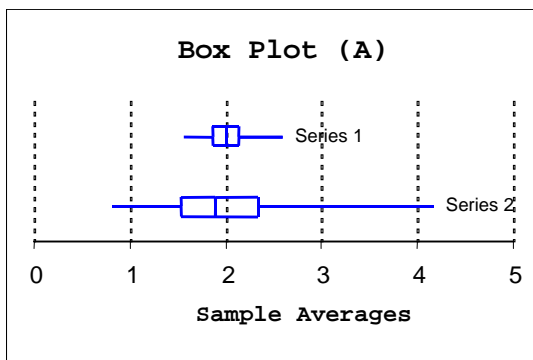
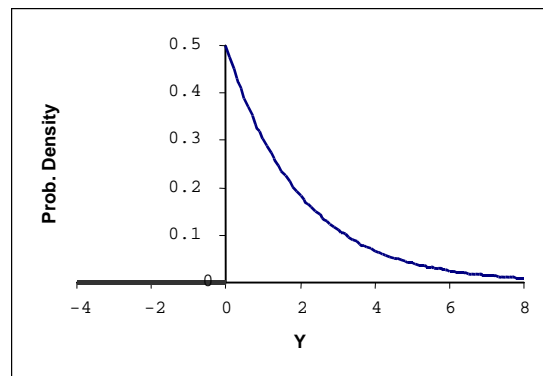
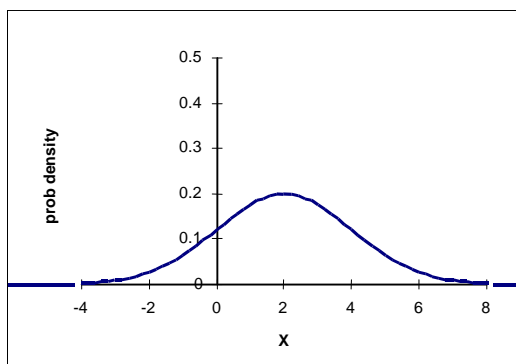


d) Use *Excel's* random number generator (see section 13.5.2) to simulate 100 pairs (two columns) of independent tosses of a single tetrahedral die. In a third column, compute the average tossed for each pair to obtain 100 samples of the random variable \bar{X} . Compare the sample average and standard deviation for these 100 values of \bar{X} with the theoretical values you found in c).

3. The top row below displays density functions for two continuous random variables labeled X and Y , each of which has mean two and standard deviation two. The bottom panel displays box plots for the average of 100 independent samples taken from each random variable. In each case, one box plot used samples of size 10 and the other used samples of size 100.

a) In each box plot displayed in the bottom row identify which plot (series 1 or series 2) came from the replications using samples of size 10 and which from samples of size 100. Justify your answer.

b) Identify which box plots came from sampling random variable X and which from Y . Justify your answer.



4. A certain game is played according to rules that produce the following probability distribution for the winnings X on a single play. (A negative value for X denotes a loss to the player.)

X (\$)	5	-5	-10
$P(X = x)$.65	.25	.10

- a) Find $E(X)$ and μ_X .
- b) You play 10 independent games. Let \bar{X} denote your average winnings in 10 games. What are the expected value and the standard deviation of \bar{X} .



- c) Use *Excel* to simulate 200 replications of playing this game 10 times, i.e. generate 200 columns of data in which each column uses the probability distribution in a) to give the results of 10 plays. Use these 200 replications to estimate the theoretical expected value and standard deviation found in b).


5. a) You toss a fair coin 100 times. Your friend tosses a similar coin 25 times. Which of you is more likely to have at least 60% of the tosses turn up heads?
- b) A small city has two hospitals. One delivers on average 100 babies a day, the other about 25. Which one has a greater chance that more than 60% of the babies delivered on a day will be boys?



6. Open the file *sampling.xls*. This file allows you to draw samples from a hidden sheet containing a large number of 0s and 1s, which represents a bowl model consisting of white beans (0s) and red beans (1s). Each sample of size n provides an estimate \hat{p}_n for the frequency p of 1s in the bowl.

- a) On the top sheet (sheet 1) use the button provided to generate 20 samples of size 10. Compute the value of \hat{p}_{10} for each sample. Use the values of \hat{p}_{10} to estimate $E(\hat{p}_{10})$ and $\sigma_{\hat{p}_{10}}$.
- b) Repeat the steps in a) but using 20 samples of size 100 on sheet 2. Use the values of \hat{p}_{100} to estimate $E(\hat{p}_{100})$ and $\sigma_{\hat{p}_{100}}$. You should find that $E(\hat{p}_{10}) \approx E(\hat{p}_{100})$, but in theory $\sigma_{\hat{p}_{10}}$ should be about 3 times larger than $\sigma_{\hat{p}_{100}}$. Explain why and comment on any discrepancies between theory and data.
- c) Use the 95% confidence limits given in section B.4 to find confidence intervals for the true p for each of the samples \hat{p}_{10} in part a). Using Rule 15.1 find 95% confidence intervals associated with each value \hat{p}_{100} in b).
- d) In part c) you constructed forty, 95% confidence intervals. About how many of these will fail to contain the true value of p ? Using the best estimate for p based on the information you have gathered in the problem, determine which of the confidence intervals actually do not have the true p .

7. a) Explain the difference between the terms *standard deviation* and *standard error*.

- b) What is meant by the expression “*sampling error*”?
- c) What is meant by a *confidence interval*?
8. A sample of 50 girls in a junior high school had an average height of 60 inches, with a standard deviation $s = 2.5$ inches.
- a) What is the standard error (SE) for this sample?
- b) Explain the difference in meaning between the reported intervals $60 \pm 1.96s$ inches and $60 \pm 1.96 \text{ SE}$ inches.
9. On the basis of past performance a professor expects that the average score on the mid-term exam will be 75 out of 100. In previous years the standard deviation of the individual scores has been around 10 points. What is the probability that the average score for his current class of 40 students will be below 72? What assumptions about this group of students are you making when performing this calculation?
10. An airline claims that 92% of its flights meet the criterion for being classified as on-time. A consumer group wishing to monitor this claim checks the flight records for 75 randomly selected flights. They find that 60 of the sampled flights satisfy on-time status. If the airline’s claim is true, how likely is it that the sample would have included as few as 60 flights fulfilling the on-time status?
11. Using the table of the standard normal distribution (section B.3) derive the confidence coefficients in Table 15.5.
12. Using the normal approximation, estimate the 80% confidence interval for p when $n = 20$ and the observed frequency $\hat{p} = 0.3$. Compare your answer with the Table in section B.4.
13. Suppose that the true $p = 0.4$ and we use a sample of size 20. Using the construction of the 80% confidence intervals given in the table of section B.4 find the exact probability that such an interval contains the true value $p = 0.4$. Is your answer close to 0.80?
14. According to a *New York Times* report (3/21/00), in the fall of 1999 the New York City Department of Health tested blood samples of 677 anonymous donors in northern Queens. Nineteen tested positive for the antibodies to West Nile virus.
- a) Find a 95% confidence interval for the fraction of the entire population of that community that had been infected with the virus.
- b) If the area in which the donors lived had a total population of 46,000, what estimates could you give for the number of people in that area who had been infected with the virus?
-  15. The file *confidence intervals.xls* allows you to construct 20 confidence intervals for a selected population frequency by taking random samples from a population with a known value of p .


- a) Select two different values of p and run the simulation (with a 95 % confidence level) three times for each value of p . In total you will have constructed 120 confidence intervals. How many should you expect will not contain the true value of p ? Is your expectation confirmed?
 - b) If the true p is around 0.4, determine how large the sample size must be so that a 98% confidence interval will produce an error of at most three percentage points in the estimate of p . Run the simulation with the computed sample size, $p \approx .4$ and confidence level 98%. Verify that the results confirm your expectations, i.e. size of confidence intervals and frequency with which they contain the true value of p .
16. (Capture-recapture) How do ecologists estimate the number of fish in a lake? Suppose you are working in fishery management. You go out on the lake and capture 100 fish of the type you are interested in counting. These fish are marked and released. After a period of time to allow the fish to disperse, you catch say 75 of this type of fish and observe 7 of the ones you marked. (Assume you toss fish back into the water after capture and you move around the lake after each fish is caught.)
- a) Find a 95% confidence interval for the percent of fish in the lake that have been marked.
 - b) Using the answer to a) give upper and lower bounds for the fish population in the lake.
17. In estimating an unknown population frequency p using confidence intervals, comment on the truth or falsity of the following assertions.
- a) A 90% confidence interval using $n = 150$ has a higher chance of containing the true value of p than the same confidence interval based on a sample of size 50.
 - b) A 90% confidence interval using $n = 150$ gives a narrower estimate of p than the same confidence interval based on a sample of size 50.
 - c) If the sample size is the same, then a 95% confidence interval for p is more “accurate” than a 90% confidence interval.
18. A survey of 1500 voters across the country showed that 885 would be willing to pay \$200 more a year in taxes to support improved schools.
- a) Find a 95% confidence interval for the true fraction of all voters who support the proposal.
 - b) How many voters would have to be surveyed to obtain an estimate which with 95% confidence deviated from the true fraction by at most one percentage point.
19. The *NY Times* usually publishes a brief explanation of the statistical meaning of the results of surveys that it commissions. Typical wording is quoted below:

How the Poll Was Conducted

The latest NY Times/CBS News Poll is based on telephone interviews conducted Dec. 6th to 9th with (*a certain number*) of adults throughout the United States.

(There follows an explanation of how a random sample was obtained.)

In theory, in 19 cases out of 20 the results based on such a sample will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all American adults.

- a) Using the information in the quoted paragraph, what is the minimum number of adults that should have been included in the sample? (Hint: If we know nothing about p , then the maximum value of $pq = p(1 - q)$ is 0.25.)
 - b) Assuming the sample used was the size you computed in a), if the fraction that responded favorably to a question was 0.4, find a 98% confidence interval for the true value of p .
20. A study of blood cholesterol levels of 189 men in the age group 36 to 45 yielded a sample average of 2.42 mg/cc with a standard deviation of 0.43 mg/cc. Determine a 95% confidence interval for the mean cholesterol blood level for men in this age group.
 21. Construct a 98% confidence interval for the mean height of a population of males if a random sample of 125 yielded a mean of 70 inches with a standard deviation of 2 inches.
 22. A clairvoyant claims to be able to read people's minds. As a test, you agree to sit in one room and to concentrate for several minutes on a photo of the head or tail face of a coin. Sitting in an adjacent room and exercising similar concentration, the clairvoyant states his opinion of which face is being shown to you. The photo is then randomly "flipped" and the process repeated. Suppose that a test of this sort produced 15 correct responses from the clairvoyant out of 20 attempts. What statistical or probabilistic arguments would you use to interpret the results?
 23. a) Using the table in section B.5 find the confidence coefficients for the following Student distributions and confidence levels:
 - i) $f = 10$, confid. level = 90%
 - ii) $f = 15$, confid. level = 99%
 - iii) $f = 40$, confid. level = 95%
 - b)  Using *Excel* find the confidence coefficients for the following Student distributions and confidence levels:
 - i) $f = 5$, confid. level = 80%
 - ii) $f = 10$, confid. level = 75%
 - iii) $f = 25$, confid. level = 80%

24. An anti-arrhythmic drug is used to correct abnormal heart rhythms. For example, a patient in atrial fibrillation may have a resting heart rate of 150 - 200 beats/minute and this needs to be brought under control. A new anti-arrhythmic drug is tested on 8 patients experiencing elevated resting heart rates. Each patient is given the same dose of the medication and his or her heart rate is recorded before and one hour after drug administration:

	Rate Before	Rate After	Difference		Rate Before	Rate After	Difference
Patient 1	175	110	65	Patient 5	145	85	60
Patient 2	135	92	43	Patient 6	150	100	50
Patient 3	180	100	80	Patient 7	180	95	85
Patient 4	200	95	105	Patient 8	160	100	60

Use the data to find a 95% confidence interval for the mean difference in heart rate attributable to the medication. What probabilistic assumptions are we making when we perform this statistical analysis.

25. The need for prophylaxis to control blood pressure is based on measurements of pressure taken during office visits to a physician. To limit unnecessary use of medication or dietary restrictions, in borderline cases the diagnosis should not be made based on a single reading. A doctor records the following four diastolic readings for a patient over a four-week period { 95, 82, 79, 84 } (in standard pressure units of mm Hg). According to current standards, an adult with diastolic readings between 90 and 99 should be classified as Stage 1 (mildly) hypertensive.

- Using a 95% confidence interval, decide how to classify this patient.
- What statistical assumptions are we making about the blood pressure readings when carrying out the analysis in a)?



26. Open the file *ny_weather.xls*. Using the command `=randbetween(6,136)` select a random number between 6 and 136 (the row indices) and record the average temperature for the year in the row with that random number as its row index. Repeat this 10 times to obtain a random sample of 10 data values for the average annual temperature.

- Provide suitable evidence that the complete population of yearly temperatures has a normal distribution.
- In view of a), use your sample data values and a suitable t coefficient from table in section B.5 to find 90 and 95% confidence intervals for the average annual temperature in Central Park over the 130 years.
- Find the true average and see if the confidence intervals you constructed in b) contain the true value of the mean.
- Suppose we had conducted the sampling by selecting at random a data value from each decade (excluding the year 1869 from the sampling). Would this sampling process constitute what we have called a random sample from the entire population? Explain.