4.5 CONFIDENCE INTERVALS AND HYPOTHESIS TESTS FOR THE MEAN

Let $X_1, X_2, \ldots, X_n$ be IID random variables with finite mean μ and finite variance σ^2. (Also assume that $\sigma^2 > 0$, so that the X_i's are not degenerate random variables.) In this section we discuss how to construct a confidence interval for μ and also the complementary problem of testing the hypothesis that $\mu = \mu_0$.

We begin with a statement of the most important result in probability theory, the classical central limit theorem. Let Z_n be the random variable $[\bar{X}(n) - \mu]/\sqrt{\sigma^2/n}$, and let $F_n(z)$ be the distribution function of Z_n for a sample size of n; that is, $F_n(z) = P(Z_n \le z)$. [Note that μ and σ^2/n are the mean and variance of $\bar{X}(n)$, respectively.] Then the *central limit theorem* is as follows [see Chung (1974, p. 169) for a proof].

THEOREM 4.1. $F_n(z) \to \Phi(z)$ as $n \to \infty$, where $\Phi(z)$, the distribution function of a normal random variable with $\mu = 0$ and $\sigma^2 = 1$ (henceforth called a *standard normal random variable*; see Sec. 6.2.2), is given by

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-y^2/2}\, dy \qquad \text{for } -\infty < z < \infty$$

The theorem says, in effect, that if n is "sufficiently large," the random variable Z_n will be approximately distributed as a standard normal random variable, regardless of the underlying distribution of the X_i's. It can also be shown for large n that the sample mean $\bar{X}(n)$ is approximately distributed as a normal random variable with mean μ and variance σ^2/n.

The difficulty with using the above results in practice is that the variance σ^2 is generally unknown. However, since the sample variance $S^2(n)$ converges to σ^2 as n gets large, it can be shown that Theorem 4.1 remains true if we replace σ^2 by $S^2(n)$ in the expression for Z_n. With this change the theorem says that if n is sufficiently large, the random variable $t_n = [\bar{X}(n) - \mu]/\sqrt{S^2(n)/n}$ is approximately distributed as a standard normal random variable. It follows for large n that

$$\begin{aligned} P\left(-z_{1-\alpha/2} \le \frac{\bar{X}(n) - \mu}{\sqrt{S^2(n)/n}} \le z_{1-\alpha/2}\right) &\\ = P\left[\bar{X}(n) - z_{1-\alpha/2}\sqrt{\frac{S^2(n)}{n}} \le \mu \le \bar{X}(n) + z_{1-\alpha/2}\sqrt{\frac{S^2(n)}{n}}\right] &\\ \approx 1 - \alpha & \qquad (4.10) \end{aligned}$$

where the symbol $\approx$ means "approximately equal" and $z_{1-\alpha/2}$ (for $0 < \alpha < 1$) is the upper $1 - \alpha/2$ critical point for a standard normal random variable (see Fig. 4.15 and the last line of Table T.1 of the Appendix at the back of the book). Therefore, if n is sufficiently large, an approximate $100(1 - \alpha)$ percent confidence interval for μ is given by

$$\bar{X}(n) \pm z_{1-\alpha/2}\sqrt{\frac{S^2(n)}{n}} \qquad (4.11)$$

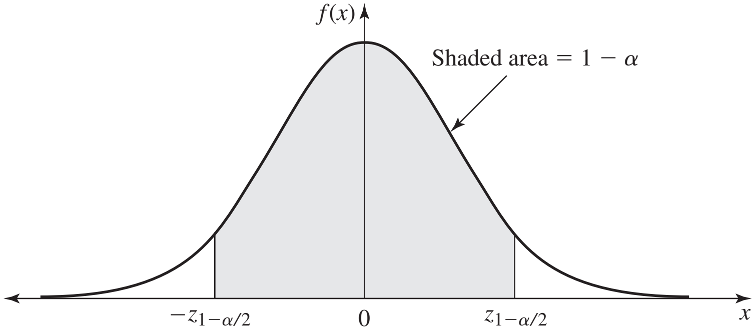

FIGURE 4.15
Density function for the standard normal distribution.

For a given set of data $X_1, X_2, \ldots, X_n$, the lower confidence-interval endpoint $l(n, \alpha) = \bar{X}(n) - z_{1-\alpha/2}\sqrt{S^2(n)/n}$ and the upper confidence-interval endpoint $u(n, \alpha) = \bar{X}(n) + z_{1-\alpha/2}\sqrt{S^2(n)/n}$ are just numbers (actually, specific realizations of random variables) and the confidence interval $[l(n, \alpha), u(n, \alpha)]$ either contains μ or does not contain μ. Thus, there is nothing probabilistic about the single confidence interval $[l(n, \alpha), u(n, \alpha)]$ *after* the data have been obtained and the interval's endpoints have been given numerical values. The correct interpretation to give to the confidence interval (4.11) is as follows [see (4.10)]: If one constructs a very large number of independent $100(1 - \alpha)$ percent confidence intervals, each based on n observations, where n is sufficiently large, the proportion of these confidence intervals that contain (cover) μ should be $1 - \alpha$. We call this proportion the *coverage* for the confidence interval.

EXAMPLE 4.26. To further amplify the correct interpretation to be given to a confidence interval, we generated 15 independent samples of size $n = 10$ from a normal distribution with mean 5 and variance 1. For each data set we constructed a 90 percent confidence interval for μ, which we know has a true value of 5. In Fig. 4.16 we plot the 15 confidence intervals vertically (the dot at the center of the confidence interval is the sample mean), and we see that all intervals other than 7 and 13 cover the mean value at height 5. In general, if we were to construct a very large number of such 90 percent confidence intervals, we would find that 90 percent of them will, in fact, contain (cover) μ.

The difficulty in using (4.11) to construct a confidence interval for μ is in knowing what "n sufficiently large" means. It turns out that the more skewed (i.e., nonsymmetric) the underlying distribution of the X_i's, the larger the value of n needed for the distribution of t_n to be closely approximated by $\Phi(z)$. (See the discussion later in this section.) If n is chosen too small, the actual coverage of a desired $100(1 - \alpha)$ percent confidence interval will generally be less than $1 - \alpha$. This is why the confidence interval given by (4.11) is stated to be only approximate.

In light of the above discussion, we now develop an alternative confidence-interval expression. If the X_i's are *normal* random variables, the random variable $t_n = [\bar{X}(n) - \mu]/\sqrt{S^2(n)/n}$ has a t distribution with $n - 1$ degrees of freedom (df)

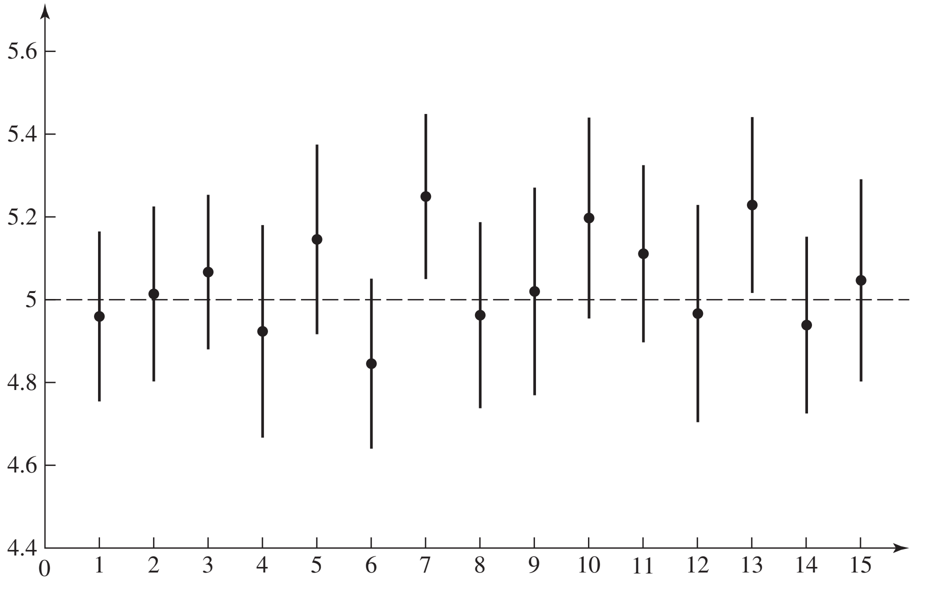

FIGURE 4.16
Confidence intervals each based on a sample of $n = 10$ observations from a normal distribution with mean 5 and variance 1.

[see, for example, Hogg and Craig (1995, pp. 181–182)], and an *exact* (for any $n \geq 2$) $100(1 - \alpha)$ percent confidence interval for μ is given by

$$\overline{X}(n) \pm t_{n-1,\,1-\alpha/2}\sqrt{\frac{S^2(n)}{n}} \tag{4.12}$$

where $t_{n-1,1-\alpha/2}$ is the upper $1 - \alpha/2$ critical point for the t distribution with $n - 1$ df. These critical points are given in Table T.1 of the Appendix at the back of the book. Plots of the density functions for the t distribution with 4 df and for the standard normal distribution are given in Fig. 4.17. Note that the t distribution is less peaked and

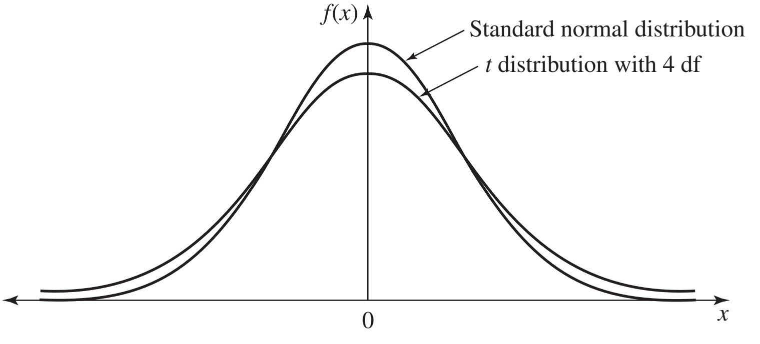

FIGURE 4.17
Density functions for the t distribution with 4 df and for the standard normal distribution.

has longer tails than the normal distribution, so, for any finite n, $t_{n-1,1-\alpha/2} > z_{1-\alpha/2}$. We call (4.12) the *t confidence interval*.

The quantity that we add to and subtract from $\overline{X}(n)$ in (4.12) to construct the confidence interval is called the *half-length* of the confidence interval. It is a measure of how precisely we know μ. It can be shown that if we increase the sample size from n to $4n$ in (4.12), then the half-length is decreased by a factor of approximately 2 (see Prob. 4.20).

In practice, the distribution of the X_i's will rarely be normal, and the confidence interval given by (4.12) will also be approximate in terms of coverage. Since $t_{n-1,1-\alpha/2} > z_{1-\alpha/2}$, the confidence interval given by (4.12) will be larger than the one given by (4.11) and will generally have coverage closer to the desired level $1 - \alpha$. For this reason, we recommend using (4.12) to construct a confidence interval for μ. Note that $t_{n-1,1-\alpha/2} \to z_{1-\alpha/2}$ as $n \to \infty$; in particular, $t_{40,0.95}$ differs from $z_{0.95}$ by less than 3 percent. However, in most of our applications of (4.12) in Chaps. 9, 10, and 12, n will be small enough for the difference between (4.11) and (4.12) to be appreciable.

EXAMPLE 4.27. Suppose that the 10 observations 1.20, 1.50, 1.68, 1.89, 0.95, 1.49, 1.58, 1.55, 0.50, and 1.09 are from a normal distribution with unknown mean μ and that our objective is to construct a 90 percent confidence interval for μ. From these data we get

$$\overline{X}(10) = 1.34 \qquad \text{and} \qquad S^2(10) = 0.17$$

which results in the following confidence interval for μ:

$$\overline{X}(10) \pm t_{9,0.95}\sqrt{\frac{S^2(10)}{10}} = 1.34 \pm 1.83\sqrt{\frac{0.17}{10}} = 1.34 \pm 0.24$$

Note that (4.12) was used to construct the confidence interval and that $t_{9,0.95}$ was taken from Table T.1. Therefore, subject to the interpretation stated above, we claim with 90 percent confidence that μ is in the interval [1.10, 1.58].

We now discuss how the coverage of the confidence interval given by (4.12) is affected by the distribution of the X_i's. In Table 4.1 we give estimated coverages for 90 percent confidence intervals based on 500 independent experiments for each of the sample sizes $n = 5$, 10, 20, and 40 and each of the distributions normal, exponential, chi square with 1 df (a standard normal random variable squared; see the discussion of the gamma distribution in Sec. 6.2.2), lognormal (e^Y, where Y is a

TABLE 4.1
Estimated coverages based on 500 experiments

Distribution	Skewness ν	$n = 5$	$n = 10$	$n = 20$	$n = 40$
Normal	0.00	0.910	0.902	0.898	0.900
Exponential	2.00	0.854	0.878	0.870	0.890
Chi square	2.83	0.810	0.830	0.848	0.890
Lognormal	6.18	0.758	0.768	0.842	0.852
Hyperexponential	6.43	0.584	0.586	0.682	0.774

standard normal random variable; see Sec. 6.2.2), and hyperexponential whose distribution function is given by

$$F(x) = 0.9F_1(x) + 0.1\,F_2(x)$$

where $F_1(x)$ and $F_2(x)$ are the distribution functions of exponential random variables with means 0.5 and 5.5, respectively. For example, the table entry for the exponential distribution and $n = 10$ was obtained as follows. Ten observations were generated from an exponential distribution with a *known* mean μ, a 90 percent confidence interval was constructed using (4.12), and it was determined whether the interval contained μ. (This constituted one experiment.) Then the whole procedure was repeated 500 times, and 0.878 is the proportion of the 500 confidence intervals that contained μ. Note that the coverage for the normal distribution and $n = 10$ is 0.902 rather than the expected 0.900, since the table is based on 500 rather than an infinite number of experiments.

Observe from the table that for a particular distribution, coverage generally gets closer to 0.90 as n gets larger, which follows from the central limit theorem (see Prob. 4.22). (The results for the exponential distribution would also probably follow this behavior if the number of experiments were larger.) Notice also that for a particular n, coverage decreases as the skewness of the distribution gets larger, where skewness is defined by

$$\nu = \frac{E[(X-\mu)^3]}{(\sigma^2)^{3/2}} \qquad -\infty < \nu < \infty$$

The skewness, which is a measure of symmetry, is equal to 0 for a symmetric distribution such as the normal. We conclude from the table that the larger the skewness of the distribution in question, the larger the sample size needed to obtain satisfactory (close to 0.90) coverage.

*We saw in Table 4.1 that there is still significant degradation in coverage probability for sample sizes as large as 40 if the data come from a highly skewed distribution such as the lognormal, which is not at all uncommon in practice. As a result we now discuss an improved confidence developed by Willink (2005), which computes an estimate of the skewness ν and uses this to obtain a confidence interval with coverage closer to the nominal value $1 - \alpha$ than that for the standard t confidence given by (4.12). Let

$$\hat{\mu}_3 = \frac{n\sum_{i=1}^{n}[X_i - \overline{X}(n)]^3}{(n-1)(n-2)}, \qquad a = \frac{\hat{\mu}_3}{6\sqrt{n}[S^2(n)]^{3/2}},$$

and

$$G(r) = \frac{[1 + 6a(r-a)]^{1/3} - 1}{2a}$$

where $\hat{\mu}_3/[S^2(n)]^{3/2}$ is an estimator for the skewness ν. Then an approximate $100(1 - \alpha)$ percent confidence interval for μ is given by

$$\left[\overline{X}(n) - G(t_{n-1,1-\alpha/2})\sqrt{S^2(n)/n},\ \overline{X}(n) - G(-t_{n-1,1-\alpha/2})\sqrt{S^2(n)/n}\right] \qquad (4.13)$$

*The discussion of the Willink confidence interval may be skipped on a first reading.

EXAMPLE 4.28. For the data of Example 4.27, we now construct a 90 percent confidence interval for μ using the Willink confidence interval given by (4.13). We get

$$\hat{\mu}_3 = -0.062, \quad a = -0.048, \quad G(r) = \frac{[1 - 0.288(r + 0.048)]^{1/3} - 1}{-0.096}$$

and the following 90 percent confidence interval for μ:

$$[1.34 - 0.31, 1.34 + 0.20] \qquad \text{or} \qquad [1.04, 1.54]$$

In order to get an idea how much improvement in coverage probability might be obtained by using the Willink confidence interval given by (4.13) instead of the t confidence interval given by (4.12), we regenerated using different random numbers the observations for the entry in Table 4.1 corresponding to the lognormal distribution and $n = 10$. Based again on 500 experiments, the estimated coverages for the Willink and t confidence intervals were 0.872 and 0.796, respectively. Thus, the Willink confidence interval produces a coverage probability "close" to the nominal level 0.90 even for the highly skewed lognormal distribution and a sample size of only 10. On the other hand, the average half-length for the Willink confidence interval was 76 percent larger than the average half-length for the t confidence interval in this case. The decision whether to use the t or Willink confidence interval should depend on the relative importance one places on coverage close to the nominal level $1 - \alpha$ and a small half-length.

Assume that $X_1, X_2, \ldots, X_n$ are normally distributed (or are approximately so) and that we would like to test the *null hypothesis* H_0: $\mu = \mu_0$ against the *alternative hypothesis* H_1: $\mu \neq \mu_0$, where μ_0 is a fixed, hypothesized value for μ. Intuitively, we would expect that if $|\overline{X}(n) - \mu_0|$ is large [recall that $\overline{X}(n)$ is the point estimator for μ], H_0 is not likely to be true. However, to develop a test with known statistical properties, we need a statistic (a function of the X_i's) whose distribution is known when H_0 is true. It follows from the above discussion that if H_0 is true, the statistic $t_n = [\overline{X}(n) - \mu_0]/\sqrt{S^2(n)/n}$ will have a t distribution with $n - 1$ df. Therefore, consistent with our intuitive discussion above, the form of our (two-tailed) hypothesis test for H_0 is

$$\left[\begin{array}{l}\text{If } |t_n| > t_{n-1,1-\alpha/2}, \text{ reject } H_0 \\ \text{If } |t_n| \leq t_{n-1,1-\alpha/2}, \text{ fail to reject } H_0\end{array}\right. \tag{4.14}$$

The portion of the real line that corresponds to rejection of H_0, namely, the set of all x such that $|x| > t_{n-1,1-\alpha/2}$, is called the *rejection* (or critical) *region* for the test, and the probability that the statistic falls in the rejection region given that H_0 is true, which is clearly equal to α, is called the *level* (or size) of the test. Typically, an experimenter will choose the level equal to 0.05 or 0.10. We call the hypothesis test given by (4.14) the *t test*.

When one performs a hypothesis test, two types of errors can be made. If one rejects H_0 when in fact it is true, this is called a *Type I error*. The probability of Type I error is equal to the level α and is thus under the experimenter's control. If one fails to reject H_0 when it is false, this is called a *Type II error*. For a fixed level α and sample size n, the probability of a Type II error, which we denote by β, depends on

TABLE 4.2
Hypothesis-testing situations and their corresponding probabilities of occurrence

Outcome \ H_0	True	False
Reject	α	$\delta = 1 - \beta$
Fail to reject	$1 - \alpha$	β

what is actually true (other than H_0: $\mu = \mu_0$), and is usually unknown. We call $\delta = 1 - \beta$ the *power* of the test, and it is equal to the probability of rejecting H_0 when it is false. There are four different situations that can occur when one tests the null hypothesis H_0 against the alternative hypothesis H_1, and these are delineated in Table 4.2 along with their probabilities of occurrence.

Clearly, a test with high power is desirable. If α is fixed, the power can only be increased by increasing n. Since the power of a test may be low and unknown to us, this is why we say "fail to reject H_0" (instead of "accept H_0") when the statistic t_n does not lie in the rejection region. (When H_0 is not rejected, we generally do not know with any certainty whether H_0 is true or whether H_0 is false, since our test may not be powerful enough to detect any difference between H_0 and what is actually true.)

EXAMPLE 4.29. For the data of Example 4.27, suppose that we would like to test the null hypothesis H_0: $\mu = 1$ against the alternative hypothesis H_1: $\mu \neq 1$ at level $\alpha = 0.1$. Since

$$t_{10} = \frac{\overline{X}(10) - 1}{\sqrt{S^2(10)/10}} = \frac{0.34}{\sqrt{0.17/10}} = 2.65 > 1.83 = t_{9,0.95}$$

we reject H_0.

EXAMPLE 4.30. For the null hypothesis H_0: $\mu = 1$ in Example 4.29, we can estimate the power of the test when, in fact, the X_i's have a normal distribution with $\mu = 1.5$ and standard deviation $\sigma = 1$. (This is H_1.) We randomly generated 1000 independent observations of the statistic $t_{10} = [\overline{X}(10) - 1]/\sqrt{S^2(10)/10}$ under the assumption that H_1 is true. For 433 out of the 1000 observations, $|t_{10}| > 1.83$ and, therefore, the estimated power is $\hat{\delta} = 0.433$. Thus, if H_1 is true, we will only reject the null hypothesis H_0 approximately 43 percent of the time for a test at level $\alpha = 0.10$. To see what effect the sample size n has on the power of the test, we generated 1000 observations of t_{25} ($n = 25$) when H_1 is true and also 1000 observations of t_{100} ($n = 100$) when H_1 is true (all X_i's were normal). The estimated powers were $\hat{\delta} = 0.796$ and $\hat{\delta} = 0.999$, respectively. It is not surprising that the power is apparently an increasing function of n, since we would expect to have a better estimate of the true value of μ when n is large. [Note that in the case of normal sampling and a *known* standard deviation, as in this example, the power of the test can actually be computed numerically, obviating the need for simulation as done here; see, for example, Devore (2008, pp. 302–303).]

It should be mentioned that there is an intimate relationship between the confidence interval given by (4.12) and the hypothesis test given by (4.14). In particular, rejection of the null hypothesis H_0: $\mu = \mu_0$ is equivalent to μ_0 not being contained in the confidence interval for μ, assuming the same value of α for both the hypothesis test and the confidence interval (see Prob. 4.28). However, the confidence interval *also* gives you a range of possible values for μ, and in this sense it is the preferred methodology.

4.6 THE STRONG LAW OF LARGE NUMBERS

The second most important result in probability theory (after the central limit theorem) is arguably the strong law of large numbers. Let $X_1, X_2, \ldots, X_n$ be IID random variables with finite mean μ. Then the *strong law of large numbers* is as follows [see Chung (1974, p. 126) for a proof].

THEOREM 4.2. $\overline{X}(n) \to \mu$ w.p. 1 as $n \to \infty$.

The theorem says, in effect, that if one performs an infinite number of experiments, each resulting in an $\overline{X}(n)$, and n is sufficiently large, then $\overline{X}(n)$ will be arbitrarily close to μ for almost all the experiments.

EXAMPLE 4.31. Suppose that $X_1, X_2, \ldots$ are IID normal random variables with $\mu = 1$ and $\sigma^2 = 0.01$. Figure 4.18 plots the values of $\overline{X}(n)$ for various n that resulted from sampling from this distribution. Note that $\overline{X}(n)$ differed from μ by less than 1 percent for $n \geq 28$.

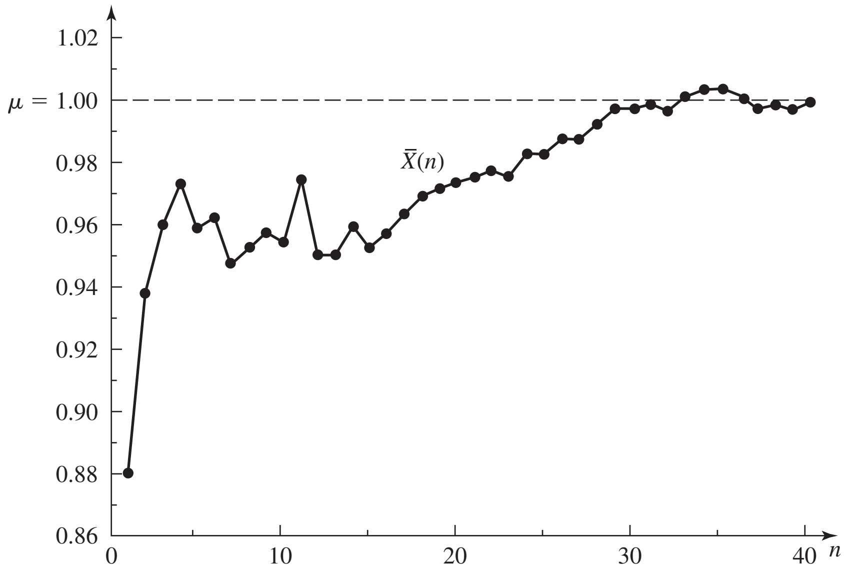

FIGURE 4.18
$\overline{X}(n)$ for various values of n when the X_i's are normal random variables with $\mu = 1$ and $\sigma^2 = 0.01$.