8.3.2 Finding Interval Estimators

Here we would like to discuss how we find interval estimators. Before doing so, let's review a simple fact from random variables and their distributions. Let X be a continuous random variable with CDF $F_X(x) = P(X \leq x)$. Suppose that we are interested in finding two values x_h and x_l such that

$$P\bigg(x_l \leq X \leq x_h\bigg) = 1 - \alpha.$$

One way to do this, is to chose x_l and x_h such that

$$P(X \leq x_l) = \frac{\alpha}{2}, \quad \text{and} \quad P(X \geq x_h) = \frac{\alpha}{2}.$$

Equivalently,

$$F_X(x_l) = \frac{\alpha}{2}, \quad \text{and} \quad F_X(x_h) = 1 - \frac{\alpha}{2}.$$

We can rewrite these equations by using the inverse function F_X^{-1} as

$$x_l = F_X^{-1}\left(\frac{\alpha}{2}\right), \quad \text{and} \quad x_h = F_X^{-1}\left(1 - \frac{\alpha}{2}\right).$$

We call the interval $[x_l, x_h]$ a $(1-\alpha)$ interval for X. Figure 8.2 shows the values of x_l and x_h using the CDF of X, and also using the PDF of X.

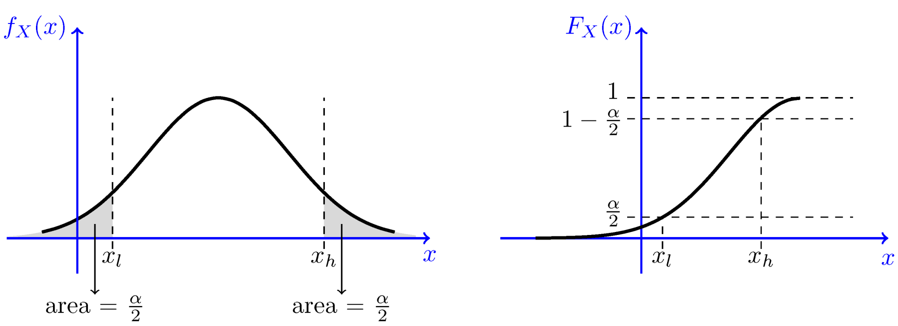

Figure 8.2 - $[x_l, x_h]$ is a $(1-\alpha)$ interval for X, that is,

$$P\bigg(x_l \leq X \leq x_h\bigg) = 1 - \alpha.$$

Example 8.12

Let $Z \sim N(0,1)$, find x_l and x_h such that

$$P\Big(x_l \leq Z \leq x_h\Big) = 0.95$$

Solution

Here, $\alpha = 0.05$ and the CDF of Z is given by the Φ function. Thus, we can choose

$$x_l = \Phi^{-1}(0.025) = -1.96, \quad \text{and} \quad x_h = \Phi^{-1}(1-0.025) = 1.96$$

Thus, for a standard normal random variable Z, we have

$$P\Big(-1.96 \leq Z \leq 1.96\Big) = 0.95$$

More generally, we can find a $(1-\alpha)$ interval for the standard normal random variable. Assume $Z \sim N(0,1)$. Let us define a notation that is commonly used. For any $p \in [0,1]$, we define z_p as the real value for which

$$P(Z > z_p) = p.$$

Therefore,

$$\Phi(z_p) = 1 - p, \quad z_p = \Phi^{-1}(1-p).$$

By symmetry of the normal distribution, we also conclude

$$z_{1-p} = -z_p.$$

Figure 8.3 shows z_p and $z_{1-p} = -z_p$ on the real line. In MATLAB, to compute z_p you can use the following command: `norminv(1 - p)`.

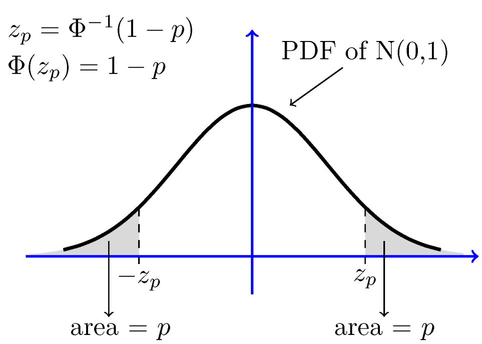

Figure 8.3 - By definition, z_p is the real number, for which we have $\Phi(z_p) = 1 - p$.

Now, using the z_p notation, we can state a $(1-\alpha)$ interval for the standard normal random variable Z as

$$P\left(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Figure 8.4 shows the $(1-\alpha)$ interval for the standard normal random variable Z.

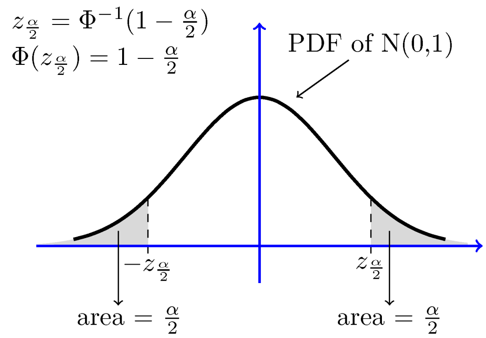

Figure 8.4 - A $(1-\alpha)$ interval for $N(0,1)$ distribution. In particular, in this figure, we have $P\left(Z \in \left[-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right]\right) = 1-\alpha$.

Now, let's talk about how we can find interval estimators. A general approach is to start with a point estimator $\hat{\Theta}$, such as the MLE, and create the interval $[\hat{\Theta}_l, \hat{\Theta}_h]$ around it such that $P\left(\theta \in [\hat{\Theta}_l, \hat{\Theta}_h]\right) \geq 1-\alpha$. How do we do this? Let's look at an example.

Example 8.13

Let X_1, X_2, X_3, $\ldots$, X_n be a random sample from a normal distribution $N(\theta, 1)$. Find a 95% confidence interval for θ.

Solution

Let's start with a point estimator $\hat{\Theta}$ for θ. Since θ is the mean of the distribution, we can use the sample mean

$$\hat{\Theta} = \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Since $X_i \sim N(\theta, 1)$ and the X_i's are independent, we conclude that

$$\overline{X} \sim N\left(\theta, \frac{1}{n}\right).$$

By normalizing $\overline{X}$, we conclude that the random variable

$$\frac{\overline{X}-\theta}{\frac{1}{\sqrt{n}}} = \sqrt{n}(\overline{X}-\theta)$$

has a $N(0,1)$ distribution. Therefore, by Example 8.12, we conclude

$$P\bigg(-1.96 \leq \sqrt{n}(\overline{X}-\theta) \leq 1.96\bigg) = 0.95$$

which is equivalent to (by rearranging the terms)

$$P\bigg(\overline{X} - \frac{1.96}{\sqrt{n}} \leq \theta \leq \overline{X} + \frac{1.96}{\sqrt{n}}\bigg) = 0.95$$

Therefore, we can report the interval

$$[\hat{\Theta}_l, \hat{\Theta}_h] = \left[\overline{X} - \frac{1.96}{\sqrt{n}}, \overline{X} + \frac{1.96}{\sqrt{n}}\right]$$

as our 95% confidence interval for θ.

At first, it might seem that our solution to Example 8.13 is not based on a systematic method. You might have asked: "How should I know that I need to work with the normalized $\overline{X}$?" However, by thinking more deeply about the way we solved this example, we can suggest a general method to solve confidence interval problems. The crucial fact about the random variable

$$\overline{X}-\theta$$

is that its distribution does not depend on the unknown parameter θ. Thus, we could easily find a 95% interval for the random variable $\sqrt{n}(\overline{X}-\theta)$ that did not depend on θ. Such a random variable is called a **pivot** or a **pivotal quantity**. Let us define this more precisely.

Pivotal Quantity

Let X_1, X_2, X_3, ..., X_n be a random sample from a distribution with a parameter θ that is to be estimated. The random variable Q is said to be a *pivot* or a *pivotal quantity*, if it has the following properties:

1. It is a function of the observed data X_1, X_2, X_3, ..., X_n and the unknown parameter θ, but it does not depend on any other unknown parameters:
$$Q = Q(X_1, X_2, \cdots, X_n, \theta).$$
2. The probability distribution of Q does not depend on θ or any other unknown parameters.

Example 8.14

Check that the random variables $Q_1 = \overline{X} - \theta$ and $Q_2 = \sqrt{n}(\overline{X} - \theta)$ are both valid pivots in Example 8.13.

Solution

We note that Q_1 and Q_2 by definitions are functions of $\overline{X}$ and θ. Since

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n},$$

we conclude Q_1 and Q_2 are both functions of the observed data X_1, X_2, X_3, ..., X_n and the unknown parameter θ, and they do not depend on any other unknown parameters. Also,

$$Q_1 \sim N(0, \frac{1}{n}), \quad Q_2 \sim N(0, 1).$$

Thus, their distributions do not depend on θ or any other unknown parameters. We conclude that Q_1 and Q_2 are both valid pivots.

To summarize, here are the steps in the pivotal method for finding confidence intervals:

1. First, find a pivotal quantity $Q(X_1, X_2, \cdots, X_n, \theta)$.
2. Find an interval for Q such that

$$P\big(q_l \leq Q \leq q_h\big) = 1 - \alpha.$$

3. Using algebraic manipulations, convert the above equation to an equation of the form

$$P\big(\hat{\Theta}_l \leq \theta \leq \hat{\Theta}_h\big) = 1 - \alpha.$$

You are probably still not sure how exactly you can perform these steps. The most crucial one is the first step. How do we find a pivotal quantity? Luckily, for many important cases that appear frequently in practice, statisticians have already found the pivotal quantities, so we can use their results directly. In practice, many of the interval estimation problems you encounter are of the forms for which general confidence intervals have been found previously. Therefore, to solve many confidence interval problems, it suffices to write the problem in a format similar to a previously solved problem. As you see more examples, you will feel more confident about solving confidence interval problems.

Example 8.15

Let X_1, X_2, X_3, $\ldots$, X_n be a random sample from a distribution with known variance $\mathrm{Var}(X_i) = \sigma^2$, and unknown mean $EX_i = \theta$. Find a $(1 - \alpha)$ confidence interval for θ. Assume that n is large.

Solution

As usual, to find a confidence interval, we start with a point estimate. Since $\theta = EX_i$, a natural choice is the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Since n is large, by the Central Limit Theorem (CLT), we conclude that

$$Q = \frac{\overline{X} - \theta}{\frac{\sigma}{\sqrt{n}}}$$

has approximately $N(0,1)$ distribution. In particular, Q is a function of the X_i's and θ, and its distribution does not depend on θ, or any other unknown parameters. Thus, Q is a pivotal quantity. The next step is to find a $(1-\alpha)$ interval for Q. As we saw before, a $(1-\alpha)$ interval for the standard normal random variable Q can be stated as

$$P\left(-z_{\frac{\alpha}{2}} \leq Q \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha.$$

Therefore,

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\overline{X}-\theta}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha.$$

which is equivalent to

$$P\left(\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \theta \leq \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

We conclude that $\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$ is a $(1-\alpha)100\%$ confidence interval for θ.

The above example is our first important case of known interval estimators, so let's summarize what we have shown:

Assumptions: A random sample X_1, X_2, X_3, $\ldots$, X_n is given from a distribution with known variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$; n is large.

Parameter to be Estimated: $\theta = EX_i$.

Confidence Interval: $\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$ is approximately a $(1-\alpha)100\%$ confidence interval for θ.

Note that to obtain the above interval, we used the CLT. Thus, what we found is an approximate confidence interval. Nevertheless, for large n, the approximation is very good.

Example 8.16

An engineer is measuring a quantity θ. It is assumed that there is a random error in each measurement, so the engineer will take n measurements and report the average of the measurements as the estimated value of θ. Here, n is assumed to be large enough so that the central limit theorem applies. If X_i is the value that is obtained in the ith measurement, we assume that

$$X_i = \theta + W_i,$$

where W_i is the error in the ith measurement. We assume that the W_i's are i.i.d. with $EW_i = 0$ and $\mathrm{Var}(W_i) = 4$ units. The engineer reports the average of the measurements

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

How many measurements does the engineer need to make until he is 90% sure that the final error is less than 0.25 units? In other words, what should the value of n be such that

$$P\big(\theta - 0.25 \leq \overline{X} \leq \theta + 0.25\big) \geq .90\,?$$

Solution

Note that, here, the X_i's are i.i.d. with mean

$$\begin{aligned} EX_i &= \theta + EW_i \\ &= \theta, \end{aligned}$$

and variance

$$\begin{aligned} \mathrm{Var}(X_i) &= \mathrm{Var}(W_i) \\ &= 4. \end{aligned}$$

Thus, we can restate the problem using our confidence interval terminology: "Let X_1, X_2, X_3, $\ldots$, X_n be a random sample from a distribution with known variance $\mathrm{Var}(X_i) = \sigma^2 = 4$. How large n should be so that the interval

$$\big[\overline{X} - 0.25, \overline{X} + 0.25\big]$$

is a 90% confidence interval for $\theta = EX_i$?"

By our discussion above, the 95% confidence interval for $\theta = EX_i$ is given by

$$\left[\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

Thus, we need

$$z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}=0.25,$$

where $\sigma=2$, $\alpha=1-0.90=0.1$. In particular,

$$z_{\frac{\alpha}{2}}=z_{0.05}=\Phi^{-1}(1-0.05)=1.645$$

Thus, we need to have

$$1.645\frac{2}{\sqrt{n}}=0.25$$

We conclude that $n \geq 174$ is sufficient.

Now suppose that X_1, X_2, X_3, $\ldots$, X_n is a random sample from a distribution with *unknown* variance $\mathrm{Var}(X_i)=\sigma^2$. Our goal is to find a $1-\alpha$ confidence interval for $\theta=EX_i$. We also assume that n is large. By the above discussion, we can say

$$P\left(\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \theta \leq \overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha.$$

However, there is a problem here. We do not know the value of σ. How do we deal with this issue? There are two general approaches: we can either find an upper bound for σ, or we can estimate σ.

1. An upper bound for σ^2: Suppose that we can somehow show that

$$\sigma \leq \sigma_{max},$$

where $\sigma_{max} < \infty$ is a real number. Then, if we replace σ in $\left[\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$ by σ_{max}, the interval gets bigger. In other words, the interval

$$\left[\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma_{max}}{\sqrt{n}}, \overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma_{max}}{\sqrt{n}}\right]$$

is still a valid $(1-\alpha)100\%$ confidence interval for θ.

2. Estimate σ^2: Note that here, since n is large, we should be able to find a relatively good estimate for σ^2. After estimating σ^2, we can use that estimate and $\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$ to find an approximate $(1-\alpha)100\%$ confidence interval for θ.

We now provide examples of each approach.

Example 8.17

(Public Opinion Polling) We would like to estimate the portion of people who plan to vote for Candidate A in an upcoming election. It is assumed that the number of voters is large, and θ is the portion of voters who plan to vote for Candidate A. We define the random variable X as follows. A voter is chosen uniformly at random among all voters and we ask her/him: "Do you plan to vote for Candidate A?" If she/he says "yes," then $X=1$, otherwise $X=0$. Then,

$$X \sim Bernoulli(\theta).$$

Let X_1, X_2, X_3, $\ldots$, X_n be a random sample from this distribution, which means that the X_i's are i.i.d. and $X_i \sim Bernoulli(\theta)$. In other words, we randomly select n voters (with replacement) and we ask each of them if they plan to vote for Candidate A. Find a $(1-\alpha)100\%$ confidence interval for θ based on X_1, X_2, X_3, $\ldots$, X_n.

Solution

Note that, here,

$$EX_i = \theta.$$

Thus, we want to estimate the mean of the distribution. Note also that

$$\mathrm{Var}(X_i) = \sigma^2 = \theta(1-\theta).$$

Thus, to find σ, we need to know θ. But θ is the parameter that we would like to estimate in the first place. By the above discussion, we know that if we can find an upper bound for σ, we can use it to build a confidence interval for θ. Luckily, it is easy to find an upper bound for σ in this problem. More specifically, if you define

$$f(\theta) = \theta(1-\theta), \quad \text{for } \theta \in [0,1].$$

By taking derivatives, you can show that the maximum value for $f(\theta)$ is obtained at $\theta = \frac{1}{2}$ and that

$$f(\theta) \leq f\left(\frac{1}{2}\right) = \frac{1}{4}, \quad \text{for } \theta \in [0,1].$$

We conclude that

$$\sigma_{max} = \frac{1}{2}$$

is an upper bound for σ. We conclude that the interval

$$\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma_{max}}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma_{max}}{\sqrt{n}}\right]$$

is a $(1-\alpha)100\%$ confidence interval for θ, where $\sigma_{max} = \frac{1}{2}$. Thus,

$$\left[\overline{X} - \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}\right]$$

is a $(1-\alpha)100\%$ confidence interval for θ. Note that we obtained the interval by using the CLT, so it is an approximate interval. Nevertheless, for large n, the approximation is very good. Also, since we have used an upper bound for σ, this confidence interval might be too conservative, specifically if θ is far from $\frac{1}{2}$.

The above setting is another important case of known interval estimators, so let's summarize it:

Assumptions: A random sample X_1, X_2, X_3, $\ldots$, X_n is given from a $Bernoulli(\theta)$; n is large.

Parameter to be Estimated: θ

Confidence Interval: $\left[\overline{X} - \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}\right]$ is approximately a $(1-\alpha)100\%$ confidence interval for θ. This is a conservative confidence interval as it is obtained using an upper bound for σ.

Example 8.18

There are two candidates in a presidential election: Candidate A and Candidate B. Let θ be the portion of people who plan to vote for Candidate A. Our goal is to find a confidence interval for θ. Specifically, we choose a random sample (with replacement) of n voters and ask them if they plan to vote for Candidate A. Our goal is to estimate the θ such that the margin of error is 3 percentage points. Assume a 95% confidence level. That is, we would like to choose n such that

$$P\left(\overline{X} - 0.03 \leq \theta \leq \overline{X} + 0.03\right) \geq 0.95,$$

where $\overline{X}$ is the portion of people in our random sample that say they plan to vote for Candidate A. How large does n need to be?

Solution

Based on the above discussion,

$$\left[\overline{X} - \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}\right]$$

is a valid $(1-\alpha)100\%$ confidence interval for θ. Therefore, we need to have

$$\frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}} = 0.03$$

Here, $\alpha = 0.05$, so $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$. Therefore, we obtain

$$n = \left(\frac{1.96}{2 \times 0.03}\right)^2.$$

We conclude $n \geq 1068$ is enough. The above calculation provides a reason why most polls before elections are conducted with a sample size of around one thousand.

As we mentioned, the above calculation might be a little conservative. Another approach would be to estimate σ^2 instead of using an upper bound. In this example, the structure of the problem suggests a way to estimate σ^2. Specifically, since

$$\sigma^2 = \theta(1-\theta),$$

we may use

$$\begin{aligned}\hat{\sigma}^2 &= \hat{\theta}(1-\hat{\theta}) \\ &= \overline{X}(1-\overline{X})\end{aligned}$$

as an estimate for θ, where $\hat{\theta} = \overline{X}$. The rationale behind this approximation is that since n is large, $\overline{X}$ is likely a good estimate of θ, thus $\hat{\sigma}^2 = \hat{\theta}(1-\hat{\theta})$ is a good estimate of σ^2. After estimating σ^2, we can use $\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}\right]$ as an approximate $(1-\alpha)100\%$ confidence interval for θ. To summarize, we have the following confidence interval rule:

Assumptions: A random sample X_1, X_2, X_3, ..., X_n is given from a $Bernoulli(\theta)$; n is large.

Parameter to be Estimated: θ

Confidence Interval: $\left[\overline{X} - z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{X}(1-\overline{X})}{n}}, \overline{X} + z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right]$ is approximately a $(1-\alpha)100\%$ confidence interval for θ.

Again, the above confidence interval is an approximate confidence interval because we used two approximations: the CLT and an approximation for σ^2.

The above scenario is a special case ($Bernoulli(\theta)$) for which we could come up with a point estimator for σ^2. Can we have a more general estimator for σ^2 that we can use for any distribution? We have already discussed such a point estimator and we called it the sample variance:

$$S^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X})^2 = \frac{1}{n-1} \left(\sum_{k=1}^{n} X_k^2 - n\overline{X}^2 \right).$$

Thus, using the sample variance, S^2, we can have an estimate for σ^2. If n is large, this estimate is likely to be close to the real value of σ^2. So let us summarize this discussion as follows:

Assumptions: A random sample X_1, X_2, X_3, $\ldots$, X_n is given from a distribution with unknown variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$; n is large.

Parameter to be Estimated: $\theta = EX_i$.

Confidence Interval: If S is the sample standard deviation

$$S = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{k=1}^{n} X_k^2 - n\overline{X}^2\right)},$$

then the interval

$$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$

is approximately a $(1-\alpha)100\%$ confidence interval for θ.

Example 8.19

We have collected a random sample X_1, X_2, X_3, $\ldots$, X_{100} from an unknown distribution. The sample mean and the sample variance for this random sample are given by

$$\overline{X} = 15.6, S^2 = 8.4$$

Construct an approximate 99% confidence interval for $\theta = EX_i$.

Solution

Here, the interval

$$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$

is approximately a $(1-\alpha)100\%$ confidence interval for θ. Since $\alpha = 0.01$, we have

$$z_{\frac{\alpha}{2}} = z_{0.005} = 2.576$$

Using $n = 100$, $\overline{X} = 15.6$, $S^2 = 8.4$, we obtain the following interval

$$\left[15.6 - 2.576\frac{\sqrt{8.4}}{\sqrt{100}}, 15.6 + 2.576\frac{\sqrt{8.4}}{\sqrt{100}}\right] = [14.85, 16.34].$$
