1 Interval Estimation (Confidence Intervals)

Let $X_1, X_2, X_3, ..., X_n$ be a random sample from a distribution with a parameter $\theta$ that is to be estimated. An interval estimator with confidence level $1 - \alpha$ consists of two estimators $\hat{\Theta}_l(X_1, X_2, \cdots, X_n)$ and $\hat{\Theta}_h(X_1, X_2, \cdots, X_n)$ such that

$$P\left(\hat{\Theta}_l \leq \theta \text{ and } \hat{\Theta}_h \geq \theta\right) \geq 1 - \alpha,$$

for every possible value of $\theta$. Equivalently, we say that $[\hat{\Theta}_l, \hat{\Theta}_h]$ is a $(1 - \alpha)100\%$ confidence interval for $\theta$.

1.1 Definition of $z_p$

Assume $Z \sim N(0, 1)$. 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 1 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: \texttt{norminv}(1-p).
Figure 1: 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 2 shows the \((1 - \alpha)\) interval for the standard normal random variable \( Z \).

\[
z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2}) \quad \Phi(z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}
\]

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

\[
z_{0.05} = 1.645, \quad z_{0.025} = 1.96, \quad z_{0.005} = 2.576
\]
1.2 Different Cases:

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

Parameter to be Estimated: \( \theta = E X_i \).

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

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: \( \theta \)

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

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

Parameter to be Estimated: \( \theta = E X_i \).

Confidence Interval: If \( S \) is the sample standard deviation

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

then the interval

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

is approximately a \( (1 - \alpha)100\% \) confidence interval for \( \theta \).
2 Problems

1. A random sample $X_1, X_2, X_3, \ldots, X_{100}$ is given from a distribution with known variance $\text{Var}(X_i) = 81$. For the observed sample, the sample mean is $\bar{X} = 50.1$. Find an approximate 95% confidence interval for $\theta = E(X_i)$.

   **Solution:** Since $n$ is large, a 95% CI can be expressed as given by

   \[
   \left[ \bar{X} - z_{0.025} \sqrt{\frac{\text{Var}(X_i)}{n}}, \bar{X} + z_{0.025} \sqrt{\frac{\text{Var}(X_i)}{n}} \right].
   \]

   \[
   = \left[ 50.1 - 1.96 \sqrt{\frac{81}{100}}, 50.1 + 1.96 \sqrt{\frac{81}{100}} \right].
   \]

   If we plug in known values, the 95% CI is $(48.3, 51.9)$. 
2. To estimate the portion of voters who plan to vote for Candidate A in an election, a random sample of size $n$ from the voters is chosen. The sampling is done with replacement. Let $\theta$ be the portion of voters who plan to vote for Candidate A among all voters.

(a) How large does $n$ need to be so that we can obtain a 90% confidence interval with 3% margin of error?

(b) How large does $n$ need to be so that we can obtain a 99% confidence interval with 3% margin of error?

Solution: As discussed in the lectures, we can 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,

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

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

Solution: Note that, here,

$$EX_i = \theta.$$

Thus, we want to estimate the mean of the distribution.

(a) Here,

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

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

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

Here $\alpha = 0.10$, so $z_{0.05} = 1.645$. Therefore, we obtain

$$n = \left( \frac{1.645}{2 \times 0.03} \right)^2.$$

We conclude $n \geq 752$ is enough.
(b) Here,

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

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

\[
\frac{z_{\alpha/2}}{2\sqrt{n}} = 0.03
\]

Here \(\alpha = 0.01\), so \(z_{\alpha/2} = z_{0.005} = 2.576\). Therefore, we obtain

\[
n = \left( \frac{2.576}{2 \times 0.03} \right)^2.
\]

We conclude \(n \geq 1844\) is enough.
3. Let $X_1, X_2, X_3, \ldots, X_{100}$ be a random sample from a distribution with unknown variance $\text{Var}(X_i) = \sigma^2 < \infty$. For the observed sample, the sample mean is $\overline{X} = 110.5$, and the sample variance is $S^2 = 45.6$. Find a 95% confidence interval for $\theta = EX_i$.

Solution: Since $n$ is relatively large, 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 $\theta$. Here, $n = 100$, $\alpha = .05$, so we need

$$z_{\frac{\alpha}{2}} = z_{0.025} = \Phi^{-1}(1 - 0.025) = 1.96.$$

Thus, we can obtain a 95% confidence interval for $\mu$ as

$$\left[ \overline{X} - z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right] = \left[ 110.5 - 1.96 \cdot \frac{\sqrt{45.6}}{10}, 110.5 + 1.96 \cdot \frac{\sqrt{45.6}}{10} \right]$$

$$\approx [109.18, 111.82]$$

Therefore, $[109.18, 111.82]$ is an approximate 95% confidence interval for $\mu$. 