Section 8.1: Interval Estimation

Discrete-Event Simulation: A First Course

©2006 Pearson Ed., Inc. 0-13-142917-5
Theorem (Central Limit Theorem)

If $X_1, X_2, \ldots, X_n$ is an iid sequence of RVs with

- common mean $\mu$
- common standard deviation $\sigma$

and if $\bar{X}$ is the (sample) mean of these RVs

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

then $\bar{X}$ approaches a Normal($\mu, \sigma/\sqrt{n}$) RV as $n \rightarrow \infty$
Choose one of the random variate generators in \texttt{rvgs} to generate a sequence of random variable samples with fixed sample size $n > 1$

With the $n$-point samples indexed $j = 1, 2, \ldots$, the corresponding sample mean $\bar{X}_j$ and sample standard deviation $s_j$ can be calculated using Algorithm 4.1.1

A continuous-data histogram can be created using program \texttt{cdh}
Properties of Sample Mean Histogram

- Independent of $n$,  
  - the histogram mean is approximately $\mu$  
  - the histogram standard deviation is approximately $\sigma/\sqrt{n}$
- If $n$ is sufficiently large,  
  - the histogram density approximates the $Normal(\mu, \sigma/\sqrt{n})$ pdf
Example 8.1.2: 10000 $n$-point Exponential($\mu$) samples

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ n = 9 \]

\[ n = 36 \]
The histogram mean and standard deviation are approximately $\mu$ and $\sigma/\sqrt{n}$.

The histogram density corresponding to the 36-point sample means is closely matched by the pdf of a $Normal(\mu, \sigma/\sqrt{n})$ RV.

For $Exponential(\mu)$ samples, $n = 36$ is large enough for the sample mean to be approximately $Normal(\mu, \sigma/\sqrt{n})$.

The histogram density corresponding to the 9-point sample means matches relatively well, but with a skew to the left.

$n = 9$ is not large enough.
Essentially all of the sample means are within an interval of width of $4\sigma/\sqrt{n}$ centered about $\mu$

Because $\sigma/\sqrt{n} \to 0$ as $n \to \infty$, if $n$ is large, all the sample means will be close to $\mu$

In general:

- The accuracy of the $\text{Normal}(\mu, \sigma/\sqrt{n})$ pdf approximation is dependent on the shape of a fixed population pdf.
- If the samples are drawn from a population with
  - a highly asymmetric pdf (like the $\text{Exponential}(\mu)$ pdf): $n$ may need to be as large as 30 or more for good fit.
  - a pdf symmetric about the mean (like the $\text{Uniform}(a, b)$ pdf): $n$ as small as 10 or less may produce a good fit.
We can standardize the sample means \( \bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots \) by subtracting \( \mu \) and dividing the result by \( \sigma / \sqrt{n} \) to form the standardized sample means \( z_1, z_2, z_3, \ldots \) defined by

\[
z_j = \frac{\bar{x}_j - \mu}{\sigma / \sqrt{n}} \quad j = 1, 2, 3, \ldots
\]

Generate a continuous-data histogram for the standardized sample means by program \text{cdh}

\[ z_1, z_2, z_3, \ldots \rightarrow \text{cdh} \rightarrow \text{histogram mean} \rightarrow \text{histogram standard deviation} \rightarrow \text{histogram density} \]
Properties of Standardized Sample Mean Histogram

- Independent of $n$,
  - the histogram mean is approximately 0
  - the histogram standard deviation is approximately 1
- If $n$ is sufficiently large,
  - the histogram density approximates the $\text{Normal}(0, 1)$ pdf
Example 8.1.4

The sample means from Example 8.1.2 were standardized.

\[
z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}
\]

For \( n = 9 \):

- Mean: \( \bar{x} \)
- Standard error: \( \sigma / \sqrt{n} \)
- Standardized value: \( z \)

For \( n = 36 \):

- Mean: \( \bar{x} \)
- Standard error: \( \sigma / \sqrt{n} \)
- Standardized value: \( z \)
Properties of the Histogram in Example 8.1.4

- The histogram mean and standard deviation are approximately 0.0 and 1.0 respectively.
- The histogram density corresponding to the 36-point sample means matches the pdf of a Normal$(0, 1)$ random variable almost exactly.
- The histogram density corresponding to the 9-point sample means matches the pdf of a Normal$(0, 1)$ random variable, but not as well.
Want to replace *population* standard deviation $\sigma$ with *sample* standard deviation $s_j$ in standardization equation

$$z_j = \frac{\bar{x}_j - \mu}{\sigma/\sqrt{n}} \quad j = 1, 2, 3, \ldots$$

**Definition 8.1.1**

- Each sample mean $\bar{x}_j$ is a *point estimate* of $\mu$
- Each sample variance $s_j^2$ is a *point estimate* of $\sigma^2$
- Each sample standard deviation $s_j$ is a *point estimate* of $\sigma$
The sample mean is an *unbiased* point estimate of $\mu$

- The mean of $\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots$ will converge to $\mu$

The sample variance is a *biased* point estimate of $\sigma^2$

- The mean of $s_1^2, s_2^2, s_3^2, \ldots$ will converge to $(n - 1)\sigma^2 / n$, not $\sigma^2$

To remove this $(n - 1)/n$ bias, it is conventional to multiply the sample variance by a *bias correction* $n/(n - 1)$

The point estimate of $\sigma/\sqrt{n}$ is

$$\frac{\left(\sqrt{\frac{n}{n-1}}\right) s_j}{\sqrt{n}} = \frac{s_j}{\sqrt{n-1}}$$
Example 8.1.5

Calculate the $t$-statistic

$$t_j = \frac{\bar{x}_j - \mu}{s_j / \sqrt{n - 1}} \quad j = 1, 2, 3, \ldots$$

Generate a continuous-data histogram using cdh

$t_1, t_2, t_3, \ldots \xrightarrow{\text{cdh}} \text{histogram mean}$

$\text{cdh} \xrightarrow{\text{histogram standard deviation}} \text{histogram density}$
If $n > 2$, the histogram mean is approximately 0.

If $n > 3$, the histogram standard deviation is approximately $\sqrt{\frac{n-1}{n-3}}$.

If $n$ is sufficiently large, the histogram density approximates the pdf of a $\text{Student}(n-1)$ random variable.
Example 8.1.6

Generate $t$-statistics from Example 8.1.2

\[
t = \frac{\bar{x} - \mu}{s / \sqrt{n - 1}}
\]

**n = 9**

**n = 36**
Properties of the Histogram in Example 8.1.6

- The histogram mean and standard deviation are approximately 0.0 and \( \sqrt{\frac{n - 1}{n - 3}} \approx 1.0 \) respectively.
- The histogram density corresponding to the 36-point sample means matches the pdf of a \( \text{Student}(35) \) RV relatively well.
- The histogram density corresponding to the 9-point sample means matches the pdf of a \( \text{Student}(8) \) RV, but not as well.
Theorem (8.1.2)

If $x_1, x_2, \ldots, x_n$ is an (independent) random sample from a “source” of data with unknown mean $\mu$, if $\bar{x}$ and $s$ are the mean and standard deviation of this sample, and if $n$ is large, it is approximately true that

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

is a Student$(n-1)$ random variate

- Theorem 8.1.2 provides the justification for estimating an interval that is likely to contain the mean $\mu$
- As $n \to \infty$, the Student$(n-1)$ distribution becomes indistinguishable from Normal$(0, 1)$
Suppose

- $T$ is a $\text{Student}(n - 1)$ random variable
- $\alpha$ is a “confidence parameter” with $0.0 < \alpha < 1.0$

Then there exists a corresponding positive real number $t^*$

$$\Pr(-t^* \leq T \leq t^*) = 1 - \alpha$$
Suppose \( \mu \) is unknown. Since \( t \approx \text{Student}(n - 1) \),

\[
-t^* \leq \frac{\bar{x} - \mu}{s/\sqrt{n - 1}} \leq t^*
\]

will be approximately true with probability \( 1 - \alpha \)

- **Right inequality:** \( \frac{\bar{x} - \mu}{s/\sqrt{n - 1}} \leq t^* \iff \bar{x} - \frac{t^*s}{\sqrt{n - 1}} \leq \mu \)
- **Left inequality:** \( -t^* \leq \frac{\bar{x} - \mu}{s/\sqrt{n - 1}} \iff \mu \leq \bar{x} + \frac{t^*s}{\sqrt{n - 1}} \)

So, with probability \( 1 - \alpha \) (approximately),

\[
\bar{x} - \frac{t^*s}{\sqrt{n - 1}} \leq \mu \leq \bar{x} + \frac{t^*s}{\sqrt{n - 1}}
\]
Theorem 8.1.3

If

- $x_1, x_2, \ldots, x_n$ is an independent random sample from a “source” of data with unknown mean $\mu$
- $\bar{x}$ and $s$ are the sample mean and sample standard deviation
- $n$ is large

Then, given a confidence parameter $\alpha$ with $0.0 < \alpha < 1.0$, there exists an associated positive real number $t^*$ such that

$$\Pr\left( \bar{x} - \frac{t^*s}{\sqrt{n - 1}} \leq \mu \leq \bar{x} + \frac{t^*s}{\sqrt{n - 1}} \right) \approx 1 - \alpha$$
Example 8.1.7

If $\alpha = 0.05$, we are 95% confident that $\mu$ lies somewhere between

$$\bar{x} - \frac{t^* s}{\sqrt{n-1}} \quad \text{and} \quad \bar{x} + \frac{t^* s}{\sqrt{n-1}}$$

For a fixed sample size $n$ and level of confidence $1 - \alpha$, use `rvms` to determine $t^* = \text{idfStudent}(n - 1, 1 - \alpha/2)$

For example, if $n = 30$ and $\alpha = 0.05$, then $t^* = \text{idfStudent}(29, 0.975) \approx 2.045$
Definition 8.1.2

- The interval defined by the two endpoints

\[ \bar{x} \pm \frac{t^* s}{\sqrt{n - 1}} \]

is a \((1 - \alpha) \times 100\%\) confidence interval estimate for \(\mu\)

- \((1 - \alpha)\) is the level of confidence associated with this interval estimate and \(t^*\) is the critical value of \(t\)
Algorithm 8.1.1

To calculate an interval estimate for the unknown mean $\mu$ of the population from which a random sample $x_1, x_2, x_3, \ldots, x_n$ was drawn:

- Pick a level of confidence $1 - \alpha$ (typically $\alpha = 0.05$)
- Calculate the sample mean $\bar{x}$ and standard deviation $s$ (use Algorithm 4.1.1)
- Calculate the critical value $t^* = \text{idfStudent}(n - 1, 1 - \alpha/2)$
- Calculate the interval endpoints

$$
\bar{x} \pm \frac{t^* s}{\sqrt{n - 1}}
$$

If $n$ is sufficiently large, then you are $(1 - \alpha) \times 100\%$ confident that the mean $\mu$ lies within the interval. The midpoint of the interval is $\bar{x}$. 
Example 8.1.8

The random sample of size $n = 10$:

$$
\begin{array}{cccccc}
1.051 & 6.438 & 2.646 & 0.805 & 1.505 \\
0.546 & 2.281 & 2.822 & 0.414 & 1.307 \\
\end{array}
$$

is drawn from a population with unknown mean $\mu$

- $\bar{x} = 1.982$ and $s = 1.690$
- To calculate a 90% confidence interval estimate:
  - Determine $t^* = t_{df \text{Student}}(9, 0.95) \approx 1.833$
  - Interval: $1.982 \pm (1.833)(1.690/\sqrt{9}) = 1.982 \pm 1.032$
- We are approximately 90% confident that $\mu$ is between 0.950 and 3.014
To calculate a 95% confidence interval estimate:

- Determine \( t^* = \text{idfStudent}(9, 0.975) \approx 2.262 \)
- Interval: \( 1.982 \pm (2.262)(1.690 / \sqrt{9}) = 1.982 \pm 1.274 \)

We are approximately 95% confident that \( \mu \) is between 0.708 and 3.256

To calculate a 99% confidence interval estimate:

- Determine \( t^* = \text{idfStudent}(9, 0.995) \approx 3.250 \)
- Interval: \( 1.982 \pm (3.250)(1.690 / \sqrt{9}) = 1.982 \pm 1.832 \)

We are approximately 99% confident that \( \mu \) is between 0.150 and 3.814

Note: \( n = 10 \) is not large
For a fixed sample size
- More confidence can be achieved *only* at the expense of a larger interval
- A smaller interval can be achieved *only* at the expense of less confidence

The only way to make the interval smaller without lessening the level of confidence is to increase the sample size

Good news: with simulation, we can collect more data

Bad news: interval size decreases with $\sqrt{n}$, not $n$
How Much More Data Is Enough?

- How large should $n$ be to achieve an interval estimate $\bar{x} \pm w$ where $w$ is user-specified?
- Answer: Use Algorithm 4.1.1 with Algorithm 8.1.1 to iteratively collect data until a specified interval width is achieved.
- Note: if $n$ is large then $t^*$ is essentially independent of $n$.

\[ t^* = \text{idfStudent}(n - 1, 1 - \alpha/2) \]

<table>
<thead>
<tr>
<th>$n$</th>
<th>$t^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2.576 ((\alpha = 0.01))</td>
</tr>
<tr>
<td>5</td>
<td>1.960 ((\alpha = 0.05))</td>
</tr>
<tr>
<td>10</td>
<td>1.645 ((\alpha = 0.10))</td>
</tr>
</tbody>
</table>
The asymptotic (large $n$) value of $t^*$ is

$$t^*_\infty = \lim_{n \to \infty} \text{idfStudent}(n-1, 1-\alpha/2) = \text{idfNormal}(0.0, 1.0, 1-\alpha/2)$$

Unless $\alpha$ is very close to 0.0, if $n > 40$, the asymptotic value $t^*_\infty$ can be used.

If $n > 40$ and wish to construct a 95% confidence interval estimate, $t^*_\infty = 1.960$ can be used in Algorithm 8.1.1.
Example 8.1.9

Given a reasonable guess for $s$ and a user-specified \textit{half-width} parameter $w$, if $t_\infty^*$ is used in place of $t^*$, $n$ can be determined by solving $w = \frac{t^*s}{\sqrt{n - 1}}$ for $n$:

$$n = \left\lfloor \left( \frac{t_\infty^*s}{w} \right)^2 \right\rfloor + 1$$

provided $n > 40$

For example, if $s = 3.0$ and want to estimate $\mu$ with 95% confidence to within $\pm 0.5$, a value of $n = 139$ should be used.
If a reasonable guess for $s$ is not available, $w$ can be specified as a proportion of $s$ thereby eliminating $s$ from the previous equation.

For example, if $w$ is 10% of $s$ and 95% confidence is desired, $n = 385$ should be used to estimate $\mu$ to within $\pm w$. 
Program estimate automates the interval estimation process

A typical application: estimate the value of an unknown population mean $\mu$ by using $n$ replications to generate an independent random variate sample $x_1, x_2, \ldots, x_n$

Function `Generate()` represents a discrete-event or Monte Carlo simulation program that returns a random variate output $x$

Using the Generate Method

```plaintext
for (i = 1; i <= n; i++)
    x_i = Generate();
return x_1, x_2, ..., x_n;
```

Given a level of confidence $1 - \alpha$, program estimate can be used with $x_1, x_2, \ldots, x_n$ to compute an interval estimate for $\mu$
Algorithm 8.1.2

Given an interval half-width $w$ and level of confidence $1 - \alpha$, the algorithm computes the interval estimate $\bar{x} \pm w$

\[
t = \text{idfNormal}(0.0, 1.0, 1-\alpha/2); /* t_\infty */
\]
\[
x = \text{Generate}();
\]
\[
n = 1; \; \nu = 0.0; \; \bar{x} = x;
\]
\[
\text{while } ((n<40) \text{ or } (t*\sqrt{\nu/n} > w * \sqrt{n-1}))\{
\]
\[
\quad x = \text{Generate}();
\]
\[
\quad n++;
\]
\[
\quad d = x - \bar{x};
\]
\[
\quad \nu = \nu + d * d * (n - 1) / n;
\]
\[
\quad \bar{x} = \bar{x} + d / n;
\]
\[
\}
\]
\[
\text{return } n, \bar{x};
\]

- It is important to appreciate the need for sample independence in Algorithms 8.1.1 and 8.1.2
The meaning of confidence

Incorrect:

- “For this 95% confidence interval, the probability that $\mu$ is within this interval is 0.95”
- Why incorrect?
  - $\mu$ is not a random variable; it is constant (but unknown)
  - The *interval endpoints* are random

Correct:

- “If I create many 95% confidence intervals, approximately 95% of them should contain $\mu$”
Example 8.1.11

- 100 samples of size $n = 9$ drawn from $\text{Normal}(6, 3)$ population
- For each sample, construct a 95% confidence interval
- 95 intervals contain $\mu = 6$
- Three intervals “too low”, two intervals “too high”

Discrete-Event Simulation: A First Course