Small Samples

When sample size $N$ is small the estimator of a distributional property $a$ (mean, variance, 90 percentile, etc.) is generally not normal.

In this case, the CDF’s of the estimate $\hat{a}$ and standardized statistic $z$ (used to derive confidence intervals and hypothesis tests) can be approximated with stochastic simulation.

In order to generate random replicates in the stochastic simulation we need to specify the property $a$ (or parameters that are related to it):

For estimating confidence intervals we assume $a = \hat{a}$ (the estimate computed from the actual data).

For testing hypotheses we assume $a = a_0$ (the hypothesized parameter value).

The stochastic simulation uses many $N_{rep}$ random sample replicates, each of length $N$, to generate $N_{rep}$ estimates. The desired estimate and standardized statistic CDFs are derived from this ensemble of estimates.

**Example – Small-sample two-sided confidence Intervals for the mean of an exponential distribution**

Consider a small sample that is thought to be drawn from an exponential distribution with unknown parameter $a$:

$$[x_1, x_2, x_3, x_4, x_5] = [0.05, 1.46, 0.50, 0.72, 0.11]$$

The sample mean is an unbiased estimator of $a$:

$$\hat{a} = m_x = 0.57$$

As in the large sample case we derive the $a$ confidence interval from a standardized statistic $z$. Replicate $i$ of $z$ is:

$$z^i (\hat{a}^i, m_{\hat{a}}) = \frac{\hat{a}^i - m_{\hat{a}}}{s_{\hat{a}}}$$
where $m_\hat{a}$ and $s_\hat{a}$ are the sample mean and standard deviation computed over the ensemble of all estimate replicates $\hat{a}^i$ (e.g. $i = 1, \ldots, N_{\text{rep}}$). Each $\hat{a}^i$ is derived from $N = 5$ data values obtained from the MATLAB function `exprnd`, with $a = m_x = 0.57$.

The CDF $F_z(z)$ is obtained by plotting the $z^i$ replicates with `cdfplot` or `normplot`.

$F_z(z)$ for this example clearly deviates from the unit normal at both high and low values:

$F_z(z)$ is used as in the large sample case to identify the $z_L$ and $z_U$ values:

$$z_L = F_{z}^{-1}\left[\frac{\alpha}{2}\right] \quad z_U = F_{z}^{-1}\left[1 - \frac{\alpha}{2}\right]$$

The small-sample double-sided 95% confidence interval for $a$ is approximately:

$$a_L = \hat{a} - z_U s_\hat{a} = 0.56 - (+2.1)(0.255) = 0.02$$
$$a_U = \hat{a} - z_L s_\hat{a} = 0.56 - (-1.5)(0.255) = 0.94$$

0.02 ≤ $a$ ≤ 0.94

For comparison, the 95% doubled-sided large-sample (normal) interval:

0.06 ≤ $a$ ≤ +1.06
The difference is slight considering the small sample size. The difference in the small and large-sample 99% confidence intervals is greater.

**Example – Small-sample two-sided test of a hypothesis about the mean of an exponential distribution**

Consider in the above example the hypothesis:

\[ H_0: a=1.0 \]

We can derive the rejection region and \( p \) value for this hypothesis with a stochastic simulation similar to the performed above except that we use \( a = a_0 = 1.0 \) in `exprnd` and derive \( F_z(z) \) from replicates defined as follows:

\[
    z^i(a^i,a_0) = \frac{\hat{a}^i - a_0}{s_{\hat{a}}} = \frac{\hat{a}^i - 1.0}{s_{\hat{a}}}
\]

In this case the \( F_z(z) \) plot is the same as the one shown above.

The test statistic obtained from the observed sample mean is:

\[
    z(\hat{a},a_0) = \frac{\hat{a} - a_0}{s_{\hat{a}}} = \frac{0.56 -1.0}{0.255} = -1.73
\]

This gives a \( p \) value of approximately 0.004 (see figure), leading us reject the hypothesis.

**Special Case: Normally Distributed Samples**

If random sample(s) are normally distributed it is possible to derive the exact small sample CDFs of certain standardized statistics

**Two-sided Confidence Intervals for Small Normally Distributed Samples**

**Confidence Intervals for the Mean \( E(x) \):**

Standardized statistic:

\[
    t(m_x,a) = \frac{m_x - a}{s_x / \sqrt{N}}
\]

This has a \( t \) distribution with \( n=N-1 \) degrees of freedom.

Confidence interval:
\[ m_x - \frac{1}{2} \left(1 - \frac{\alpha}{2}\right) \frac{s_x}{\sqrt{N}} \leq \alpha \leq m_x - \frac{1}{2} \left(1 + \frac{\alpha}{2}\right) \frac{s_x}{\sqrt{N}} \]

Evaluate \(F_{t,v}^{-1}\) with MATLAB function `tinv`.

**Confidence Intervals for the Variance \(Var(x)\):**

Standardized statistic:
\[
\chi^2(s_x^2, \sigma_x^2) = \frac{(N - 1)s_x^2}{\sigma_x^2}
\]

This has a **Chi-squared distribution** with \(\nu = N - 1\) degrees of freedom.

Confidence interval:
\[
\frac{(N - 1)s_x^2}{F_{\chi^2, \nu}^{-1}(1 - \frac{\alpha}{2})} \leq Var[x] \leq \frac{(N - 1)s_x^2}{F_{\chi^2, \nu}^{-1}(\frac{\alpha}{2})}
\]

Evaluate \(F_{\chi^2, \nu}^{-1}\) with MATLAB function `chi2inv`.

**Two-sided Hypothesis Tests for Small Normally Distributed Samples**

**Hypothesis Tests about the Mean \(E(x)\):**

\(H_0 : E(x) = a_0\)

Use t test statistic \((\nu = N - 1)\):
\[
t(m_x, a_0) = \frac{m_x - a_0}{s_x \sqrt{N}}
\]

p value:
\[
\frac{P}{2} = F_{t,v}[t(m_x, a_0)] \quad \text{for} \quad F_{t,v} \leq 0.5
\]
\[
1 - \frac{P}{2} = F_{t,v}[t(m_x, a_0)] \quad \text{for} \quad F_{t,v} > 0.5
\]

Evaluate \(F_{t,v}\) with MATLAB function `tcdf`.

**Hypothesis Tests about the Variance \(Var(x)\)**
\( H_0: \text{Var}(x) = \sigma_x^2 = a_0 \)

Use Chi-squared test statistic \((v = N - 1): \chi^2(s_x^2, a_0) = \frac{(N - 1)s_x^2}{a_0} \)

\[
\frac{p}{2} = F_{\chi^2,v}[\chi^2(s_x^2, a_0)] \text{ for } F_{\chi^2,v} \leq 0.5
\]

\[
1 - \frac{p}{2} = F_{\chi^2,v}[\chi^2(s_x^2, a_0)] \text{ for } F_{\chi^2,v} > 0.5
\]

Evaluate \( F_{t,v} \) with MATLAB function \texttt{chi2cdf}. 

\[ \begin{align*}
\text{t CDF} \\
\text{Chi-squared} \\
\end{align*} \]

\[ \begin{align*}
\nu = 1 \\
\nu = 4 \\
\nu = 10 \\
\end{align*} \]

\[ \begin{align*}
\text{Probability} \\
\text{Probability} \\
\text{Probability} \\
\end{align*} \]

\[ \begin{align*}
\text{Data} \\
\text{Data} \\
\text{Data} \\
\end{align*} \]