Estimating Nonlinear Functions of Means

Back at the beginning of the course, we discussed how to obtain point estimates and confidence intervals for quantities of the form $\mu = E[X]$. (E.g., $X$ is the random reward from a play of the gambling game.) In many applications, we need to estimate quantities of the form $\alpha = g(\mu_1, \mu_2, \ldots, \mu_d)$, where $g$ is a nonlinear function and $\mu_i = E[X^{(i)}]$ for $1 \leq i \leq d$.

To keep notation simple, we will focus on the case $k = 2$, and look at quantities of the form $\alpha = \frac{X}{Y}$ with $\mu_X = E[X]$ and $\mu_Y = E[Y]$.

1. Examples

Example 1: Suppose that one is running a retail outlet in which one wishes to compute the long-run average revenue generated per customer. Let $R_i$ denote the revenue generated on day $i$, and set

$$X_i = R_i, \quad Y_i = \text{total number of customers on day } i.$$  

It is often reasonable to assume that the $(X_i, Y_i)$ pairs are i.i.d. as a random pair $(X, Y)$. Then, the average revenue over $n$ days is equal to

$$\sum_{i=1}^{n} \frac{X_i}{Y_i} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i} = \frac{\bar{X}_n}{\bar{Y}_n},$$  

where $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

Observe that

$$\frac{\bar{X}_n}{\bar{Y}_n} \to \frac{\mu_X}{\mu_Y}$$  

with probability 1 as $n \to \infty$ by the SLLN (applied to both the numerator and denominator).

Thus, the long-run average revenue per customer, $\alpha$, takes the form $\alpha = g(\mu_X, \mu_Y)$, where $g(x, y) = x/y$. Here, the $(X_i, Y_i)$ replicates can be simulated by repeatedly and independently simulating one day’s operation of the retail outlet.

Example 2: Suppose that one wishes to estimate the variance of the revenue generated per day. As above, suppose that $\{R_i : i \geq 1\}$ are i.i.d. as a random variable $R$. Set

$$X = R^2 \quad \text{and} \quad Y = R$$

Then we wish to compute $\alpha = E[R^2] - (E[R])^2 = g(\mu_X, \mu_Y)$, where $g(x, y) = x - y^2$. 

Page 1 of 7
2. Taylor-Series Approach (Delta-Method)

We need to assume here that $g$ is continuously differentiable, i.e., $g$ is continuous and its partial derivatives exist and are continuous.

Point Estimate:

To estimate $\alpha$, simulate $n$ i.i.d. replicates $(X_1, Y_1), \ldots, (X_n, Y_n)$ of $(X, Y)$. Use the natural estimator

$$\alpha_n = g\left(\bar{X}_n, \bar{Y}_n\right).$$

Observe that, in general, $E[\alpha_n] = E[g(\bar{X}_n, \bar{Y}_n)] \neq g(E[\bar{X}_n], E[\bar{Y}_n]) = g(\mu_X, \mu_Y) = \alpha$; i.e., the estimator is biased.

Confidence Intervals:

Intuitively, for large $n$, $\bar{X}_n$ should be close to $\mu_X$ and $\bar{Y}_n$ should be close to $\mu_Y$, so that

$$\alpha_n - \alpha = g(\bar{X}_n, \bar{Y}_n) - g(\mu_X, \mu_Y)$$

$$\approx \frac{\partial g}{\partial X}(\mu_X, \mu_Y) \cdot (\bar{X}_n - \mu_X) + \frac{\partial g}{\partial Y}(\mu_X, \mu_Y) \cdot (\bar{Y}_n - \mu_Y)$$

$$= \bar{Z}_n,$$

where $Z_i = c(X_i - \mu_X) + d(Y_i - \mu_Y)$ for $i \geq 1$ and $\bar{Z}_n = (1/n) \sum_{i=1}^n Z_i$. Here

$$c = \frac{\partial g}{\partial X}(\mu_X, \mu_Y) \quad \text{and} \quad d = \frac{\partial g}{\partial Y}(\mu_X, \mu_Y).$$

Observe that the sequence $\{Z_n : n \geq 0\}$ is i.i.d. as the random variable $Z = c(X - \mu_X) + d(Y - \mu_Y)$, and that $E[Z] = 0$. By the CLT for i.i.d. random variables,

$$\frac{\sqrt{n} \bar{Z}_n}{\sigma} \xrightarrow{d} N(0,1), \quad \text{so that} \quad \frac{\sqrt{n}(\alpha_n - \alpha)}{\sigma} \xrightarrow{d} N(0,1)$$

where $\sigma^2 = \text{Var}[Z] = E[Z^2]$. This suggests using the approximate $100(1-\delta)\%$ confidence interval for $\alpha$ given by

$$\left[\alpha_n - \frac{Z_{\delta/2}}{\sqrt{n}}, \alpha_n + \frac{Z_{\delta/2}}{\sqrt{n}}\right]$$
where $z_\delta$ is the $1 - (\delta/2)$ quantile of the standard normal distribution, so that $P(-z_\delta \leq N(0,1) \leq z_\delta) = 1 - \delta$.

Clearly, $\sigma^2$ can be estimated by

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (c_n (X_i - \bar{X}_n) + d_n (Y_i - \bar{Y}_n))^2,$$

where $c_n = \frac{\partial g}{\partial x} (\bar{X}_n, \bar{Y}_n)$ and $d_n = \frac{\partial g}{\partial y} (\bar{X}_n, \bar{Y}_n)$.

**Algorithm** (to compute an approximate $100(1-\delta)$% confidence interval for $\alpha = g(\mu_X, \mu_Y)$).

1. Choose $n$ and $\delta$. Set $z_\delta = 1 - (\delta/2)$ quantile of $N(0,1)$ distribution.
2. Simulate $n$ i.i.d. replicates $(X_1, Y_1), \ldots, (X_n, Y_n)$ of $(X, Y)$.
3. Compute the confidence interval $\left[ \alpha_n - \frac{z_\delta s_n}{\sqrt{n}}, \alpha_n + \frac{z_\delta s_n}{\sqrt{n}} \right]$, where

   $$\alpha_n = g(\bar{X}_n, \bar{Y}_n)$$

   $$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (c_n (X_i - \bar{X}_n) + d_n (Y_i - \bar{Y}_n))^2$$

   with $c_n = \frac{\partial g}{\partial x} (\bar{X}_n, \bar{Y}_n)$ and $d_n = \frac{\partial g}{\partial y} (\bar{X}_n, \bar{Y}_n)$.

(Of course, $n$ can be determined using trial runs as discussed previously.)

**Example (ratio estimation):**

Suppose that $g(x, y) = x/y$. Then

$$\alpha = \frac{\mu_X}{\mu_Y}, \quad c = \frac{1}{\mu_Y}, \quad d = -\frac{\mu_X}{\mu_Y^2}, \quad \alpha_n = \frac{\bar{X}_n}{\bar{Y}_n}, \quad c_n = \frac{1}{\bar{Y}_n}, \quad d_n = -\frac{\bar{X}_n}{(\bar{Y}_n)^2}$$

and the above formulas can be used. Note that multiple passes over the data are required. For computational reasons, an alternative formula for $s_n^2$ can be useful. The idea is to use a little algebra and write

$$\sigma^2 = \frac{\text{Var}(X) - 2\alpha \text{Cov}(X, Y) + \alpha^2 \text{Var}(Y)}{\mu_Y^2}$$

so that

$$s_n^2 = s_n (1,1) - 2\alpha_n s_n (1,2) + \alpha_n^2 s_n (2,2) \left( \frac{1}{\bar{Y}_n} \right)^2,$$
where

\[
s_n(1,1) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2
\]

\[
s_n(2,2) = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2
\]

\[
s_n(1,2) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)
\]

Quantities such as \( s_n(1,1) \) can be computed using the numerically stable single-pass method described at the beginning of the course. E.g., set \( w_1(1,1) = 0 \), \( t_1 = X_1 \), and then use the recursion

\[
w_k(1,1) = w_{k-1}(1,1) + \frac{(t_{k-1} - (k-1)X_k)}{k} \cdot \frac{(t_{k-1} - (k-1)X_k)}{k-1}
\]

\[t_k = t_{k-1} + X_k.
\]

for \( k = 2, \ldots, n \). Finally, compute \( s_n(1,1) = \frac{w_n(1,1)}{n-1} \).

A similar recursion works for computing \( s_n(2,2) \), and the appropriate recursion for \( s_n(1,2) \) is

\[
w_1(1,2) = 0, \quad t_1^{(1)} = X_1, \quad t_1^{(2)} = Y_1.
\]

\[
w_k(1,2) = w_{k-1}(1,2) + \frac{(t_{k-1}^{(1)} - (k-1)X_k)}{k} \cdot \frac{(t_{k-1}^{(2)} - (k-1)Y_k)}{k-1}
\]

\[t_k^{(1)} = t_{k-1}^{(1)} + X_k \quad \text{and} \quad t_k^{(2)} = t_{k-1}^{(2)} + Y_k.
\]

for \( k = 2, \ldots, n \), and then \( s_n(1,2) = \frac{w_n(1,2)}{n-1} \).

3. The Jackknife Approach

From above, the naïve point estimator \( \alpha_n = g(\bar{X}_n, \bar{Y}_n) \) is biased. The jackknife approach provides an alternate point estimator with lower bias, along with an accompanying confidence interval. The approach avoids the need to analytically derive partial derivatives of \( g \), at the cost of increased computations.

By expanding \( g \) in a Taylor series about \((\mu_X, \mu_Y)\) and taking expectations, we find that

\[E(\alpha_n) = \alpha + \frac{b}{n} + \frac{c}{n^2} + \ldots\]
We could try to reduce the bias of $\alpha_n$ by estimating the first-order bias constant $b$ (by, say, $b_n$) and then subtracting off the estimated term:

$$\hat{\alpha}_n = \alpha_n - \frac{b_n}{n}$$

in the hope that the bias of $\hat{\alpha}_n$ would be $O(n^{-2})$. Although this approach can work, computing the necessary partial derivatives of $g$ can be a bit messy and time-consuming. Estimating the bias constants also increases the variance of the estimator; the resulting loss of precision may outweigh the advantages of reducing the bias.

The *jackknife estimator* offers an alternative that avoids this messy computation on the analyst’s part. Note that

$$E(\alpha_n) = \alpha + \frac{b}{n} + \frac{c}{n^2} + \ldots$$

$$E(\alpha_{n-1}) = \alpha + \frac{b}{n-1} + \frac{c}{(n-1)^2} + \ldots$$

and so

$$E[n\alpha_n - (n-1)\alpha_{n-1}] = \alpha + c \left( \frac{1}{n} - \frac{1}{n-1} \right) + \ldots = \alpha - \frac{c}{n(n-1)} + \ldots$$

Hence, the linear combination $n\alpha_n - (n-1)\alpha_{n-1}$ has lower bias (for moderate to large values of $n$) than does $\alpha_n$. Now $\alpha_{n-1}$ is constructed on the basis of sample size $(n-1)$ obtained by deleting the $n$th observation. But the observations are i.i.d. so we’d get the same bias expansion if we deleted *any* single observation from the original sample. So (to reduce the variance of our final estimator) we will work with *all* such linear combinations.

Set

$$\bar{X}_n^i = \frac{1}{n-1} \sum_{j \neq i}^{n} X_j \quad \bar{Y}_n^i = \frac{1}{n-1} \sum_{j \neq i}^{n} Y_j \quad \alpha_n^i = g(\bar{X}_n^i, \bar{Y}_n^i) \quad \text{and}$$

$$\alpha_n(i) = n\alpha_n - (n-1)\alpha_n^i \quad \text{(the $n$th “pseudovalue”)}$$

The jackknife estimator is then given by

$$\alpha_n^J = \frac{1}{n} \sum_{i=1}^{n} \alpha_n(i).$$

Clearly,
so the jackknife estimator reduces bias. But the jackknife estimator has an additional nice property. It turns out that

\[ \frac{\sqrt{n}(\alpha_n^i - \alpha)}{\sigma} \approx \frac{\sqrt{n}(\alpha_n - \alpha)}{\sigma} \overset{\mathcal{D}}{\sim} N(0,1) \]

where \( \sigma^2 \) was defined earlier. Somewhat surprisingly, \( \sigma^2 \) can be estimated by the jackknife variance estimator

\[ v_n^j = \frac{1}{n-1} \sum_{i=1}^{n} (\alpha_n(i) - \alpha_n^i)^2. \]

This leads to the following confidence interval procedure:

**Algorithm** (to compute an approximate 100(1-\( \delta \)% confidence interval for \( \alpha = g(\mu_X, \mu_Y) \))

1. Choose \( n \) and \( \delta \). Set \( z_{\delta} = 1 - (\delta/2) \) quantile of \( N(0,1) \) distribution.
2. Simulate \( n \) i.i.d. replicates \( (X_1, Y_1), ..., (X_n, Y_n) \) of \( (X, Y) \).
3. Compute the following quantities:

\[ \alpha_n = g(\overline{X}_n, \overline{Y}_n) \]

\[ \alpha_n^i = g \left( \frac{1}{n-1} \sum_{j \neq i} X_j, \frac{1}{n-1} \sum_{j \neq i} Y_j \right) \]

\[ \alpha_n(i) = n\alpha_n - (n-1)\alpha_n^i \]

\[ \alpha_n^j = \frac{1}{n} \sum_{i=1}^{n} \alpha_n(i) \quad \text{and} \quad v_n^j = \frac{1}{n-1} \sum_{i=1}^{n} (\alpha_n(i) - \alpha_n^j)^2 \]

4. An approximate 100(1-\( \delta \)% confidence interval for \( \alpha \) is given by

\[ \left[ \alpha_n^j - z_{\delta} \sqrt{\frac{v_n^j}{n}}, \alpha_n^j + z_{\delta} \sqrt{\frac{v_n^j}{n}} \right] \]

This confidence interval procedure reduces bias, and does not require that the analyst compute any partial derivatives (however, extra computing is required to calculate the pseudovalues).
The jackknife procedure is an example of a modern approach to estimation that replaces complicated analytical calculations (e.g., of partial derivatives) with a large amount of “brute force” computation. Such computationally-intensive procedures are well-suited to modern computers and permit confidence-interval calculation for a large class of analytically intractable performance measures. See the book by Efron and Tibshirani for a general discussion of the jackknife and of another computationally intensive procedure called the bootstrap.