Estimating Nonlinear Functions of Means

Peter J. Haas

CS 590M: Simulation
Spring Semester 2020

Nonlinear Functions of Means

Our focus up until now

- Estimate quantities of the form $\mu = E[X]$
- E.g., expected win/loss of gambling game
- We'll now focus on more complex quantities

Nonlinear functions of means:

$\alpha = g(\mu_1, \mu_2, \ldots, \mu_d)$, where

- g is a nonlinear function
- $\mu_i = E[X^{(i)}]$ for $1 \le i \le d$

- For simplicity, take $d = 2$ and focus on $\alpha = g(\mu_X, \mu_Y)$
- $\mu_X = E[X]$ and $\mu_Y = E[Y]$

Example: Retail Outlet

- Goal: Estimate $\alpha =$ long-run average revenue per customer
- $X_i = R_i =$ revenue generated on day i
- $Y_i =$ number of customers on day i
- Assume that pairs $(X_1, Y_1), (X_2, Y_2), \ldots$ are i.i.d.
- Set $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$ and $\bar{Y}_n = (1/n)\sum_{i=1}^{n} Y_i$

$$\alpha = \lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{Y_1 + \cdots + Y_n} = \lim_{n\to\infty} \frac{\bar{X}_n}{\bar{Y}_n} =$$

- So $\alpha = g(\mu_X, \mu_Y)$, where $g(x, y) =$

Example: Higher-Order Moments

- Let $R_1, R_2, \ldots$ be daily revenues as before
- Assume that the R_i's are i.i.d. (Critique?)
- $\alpha = \text{Var}[R] =$ variance of daily revenue
- Let $X = R^2$ and $Y = R$

$$\alpha = g(\mu_X, \mu_Y), \text{ where } g(x,y) =$$

Delta Method (Taylor Series)

Assume that function $g(x,y)$ is smooth

- Continuously differentiable in neighborhood of (μ_x, μ_y)
- I.e., g is continuous, as are $\partial g/\partial x$ and $\partial g/\partial y$

Point estimate

- Run simulation n times to get $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.
- Set $\alpha_n = g(\bar{X}_n, \bar{Y}_n)$
- This estimator is biased:
  - $E[\alpha_n] = E[g(\bar{X}_n, \bar{Y}_n)] \neq g(E[\bar{X}_n], E[\bar{Y}_n]) = g(\mu_x, \mu_y) = \alpha$
  - Jensen's inequality: $E[\alpha_n] = E[g(\bar{X}_n)] \geq g(\mu_X) = \alpha$ if g is convex
- By SLLN and continuity of g, we have bias $\to 0$ as $n \to \infty$ (Estimator α_n is asymptotically unbiased)

Delta Method, Continued

Confidence interval

- $(\bar{X}_n, \bar{Y}_n)$ should be "close" to (μ_X, μ_Y) for large n by SLLN
- $\alpha_n = g(\bar{X}_n, \bar{Y}_n)$ should be close to $g(\mu_X, \mu_Y) = \alpha$

$$\begin{aligned}\alpha_n - \alpha &= g(\bar{X}_n, \bar{Y}_n) - g(\mu_X, \mu_Y) \\ &= \frac{\partial g}{\partial x}(\mu_X, \mu_Y) \cdot (\bar{X}_n - \mu_X) + \frac{\partial g}{\partial y}(\mu_X, \mu_Y) \cdot (\bar{Y}_n - \mu_Y) \\ &= \bar{Z}_n\end{aligned}$$

- $Z_i = c(X_i - \mu_X) + d(Y_i - \mu_Y)$ and $\bar{Z}_n = (1/n)\sum_{i=1}^n Z_i$
- $c = \frac{\partial g}{\partial x}(\mu_X, \mu_Y)$ and $d = \frac{\partial g}{\partial y}(\mu_X, \mu_Y)$

Delta Method, Continued

Confidence interval, continued

- $\{Z_n : n \geq 1\}$ are i.i.d. as $Z = c(X - \mu_X) + d(Y - \mu_Y)$
- $E[Z] =$
- By CLT, $\sqrt{n}\bar{Z}_n/\sigma \overset{D}{\sim} N(0,1)$ approximately for large n
- Thus $\sqrt{n}(\alpha_n - \alpha)/\sigma \overset{D}{\sim} N(0,1)$ approximately for large n
- Here $\sigma^2 = \text{Var}[Z] = E[Z^2] = E\left[\left(c(X - \mu_X) + d(Y - \mu_Y)\right)^2\right]$
- So asymptotic $100(1-\delta)\%$ CI is $\alpha_n \pm z_\delta \sigma/\sqrt{n}$
  - z_δ is $1 - (\delta/2)$ quantile of standard normal distribution
  - Estimate c, d, and σ from data

Delta Method, Continued

Delta Method CI Algorithm

1. Simulate to get $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.
2. $\alpha_n \leftarrow g(\bar{X}_n, \bar{Y}_n)$
3. $c_n \leftarrow \frac{\partial g}{\partial x}(\bar{X}_n, \bar{Y}_n)$ and $d_n \leftarrow \frac{\partial g}{\partial y}(\bar{X}_n, \bar{Y}_n)$
4. $s_n^2 = (n-1)^{-1}\sum_{i=1}^n \left(c_n(X_i - \bar{X}_n) + d_n(Y_i - \bar{Y}_n)\right)^2$
5. Return asymptotic $100(1-\delta)\%$ CI:

$$\left[\alpha_n - \frac{z_\delta s_n}{\sqrt{n}}, \alpha_n + \frac{z_\delta s_n}{\sqrt{n}}\right]$$

- SLLN and continuity assumptions imply that, with prob. 1, $c_n \to c$, $d_n \to d$, and $s_n^2 \to \sigma^2$

Example: Ratio Estimation: $g(x,y) = x/y$

Multi-pass method (apply previous algorithm directly)

$\alpha =$ $\quad c =$ $\quad d =$ $\quad \alpha_n =$ $\quad c_n =$ $\quad d_n =$

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

Example: Ratio Estimation: $g(x,y) = x/y$

Single-pass method

$$\sigma^2 = \text{Var}[Z] = \text{Var}[c(X - \mu_X) + d(Y - \mu_Y)]$$
$$= \frac{\text{Var}[X] - 2\alpha\,\text{Cov}[X,Y] + \alpha^2\,\text{Var}[Y]}{\mu_Y^2}$$

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

- $s_n(1,1) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$
- $s_n(2,2) = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$
- $s_n(1,2) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$
- Set $S_n^X = \sum_{i=1}^n X_i$ and $S_n^Y = \sum_{i=1}^n Y_i$

Use single-pass formulas

$$(k-1)v_k = (k-1)v_{k-1} + \left(\frac{S_{k-1}^X - (k-1)X_k}{k}\right)\left(\frac{S_{k-1}^Y - (k-1)Y_k}{k-1}\right)$$

Delta Method for Stochastic Root-Finding

Problem:

Find $\bar{\theta}$ such that $E[g(X,\bar{\theta})] = 0$ (can replace 0 with any fixed constant)

Applications:

- Process control, risk management, finance, quantiles, ...
- Stochastic optimization: $\min_\theta E[h(X,\theta)]$
  - Optimality condition: $\frac{\partial}{\partial\theta}E[h(X,\theta)] = 0$
  - Can often show that $\frac{\partial}{\partial\theta}E[h(X,\theta)] = E\left[\frac{\partial}{\partial\theta}h(X,\theta)\right]$
  - So take $g(X,\theta) = \frac{\partial}{\partial\theta}h(X,\theta)$

Point Estimate (Stochastic Average Approximation)

- Generate $X_1,\ldots,X_n$ i.i.d. as X
- Find θ_n s.t. $\frac{1}{n}\sum_{i=1}^n g(X_i,\theta_n) = 0$ (deterministic problem)

Delta Method for Stochastic Root-Finding

Problem:

Find $\bar{\theta}$ such that $E[g(X,\bar{\theta})] = 0$

Point Estimate (Stochastic Average Approximation)

- Generate $X_1,\ldots,X_n$ i.i.d. as X
- Find θ_n s.t. $\frac{1}{n}\sum_{i=1}^n g(X_i,\theta_n) = 0$

How to find a confidence interval for $\bar{\theta}$?

- Taylor series: $g(X_i,\theta_n) \approx g(X_i,\bar{\theta}) + \frac{\partial g}{\partial\theta}(X_i,\bar{\theta})(\theta_n - \bar{\theta})$
- Implies: $\frac{1}{n}\sum_{i=1}^n g(X_i,\theta_n) \approx \frac{1}{n}\sum_{i=1}^n g(X_i,\bar{\theta}) - c_n(\bar{\theta} - \theta_n)$
  - where $c_n = \frac{1}{n}\sum_{i=1}^n \frac{\partial g}{\partial\theta}(X_i,\bar{\theta}) \approx E\left[\frac{\partial g}{\partial\theta}(X,\bar{\theta})\right]$
- Implies: $\bar{\theta} - \theta_n \approx \frac{1}{c_n}\frac{1}{n}\sum_{i=1}^n g(X_i,\bar{\theta})$
- Implies: $\theta_n - \bar{\theta} \approx N(0,\sigma^2/n)$, where $\sigma^2 = \mathrm{Var}[g(X,\bar{\theta})]/c_n^2 = E[g(X,\bar{\theta})^2]/c_n^2$

Delta Method for Stochastic Root-Finding

Algorithm

1. Simulate to get $X_1,\ldots,X_n$ i.i.d.
2. Find θ_n s.t. $\frac{1}{n}\sum_{i=1}^n g(X_i,\theta_n) = 0$
3. $\hat{c}_n \leftarrow \frac{1}{n}\sum_{i=1}^n \frac{\partial g}{\partial\theta}(X_i,\theta_n)$
4. $s_n^2 \leftarrow \frac{1}{n}\sum_{i=1}^n g(X_i,\theta_n)^2/\hat{c}_n^2$
5. Return asymptotic $100(1-\delta)\%$ CI:

$$\left[\theta_n - \frac{z_\delta s_n}{\sqrt{n}}, \theta_n + \frac{z_\delta s_n}{\sqrt{n}}\right]$$

- Can use pilot runs, etc. in the usual way

Jackknife Method

Overview

- Goal: estimate $\alpha = g(\mu_x, \mu_y)$
- Naive point estimator $\alpha_n = g(\bar{X}_n, \bar{Y}_n)$ is biased
- Jackknife estimator has lower bias
- Avoids need to compute partial derivatives as in Delta method
- More computationally intensive

Starting point: Taylor series + expectation

$$E[\alpha_n] = \alpha + \frac{b}{n} + \frac{c}{n^2} + \cdots$$

- Thus bias is $O(n^{-1})$
- Estimate b and adjust? $\alpha_n^* = \alpha_n - \frac{b_n}{n}$
  - Messy partial derivative calculation, adds noise

Jackknife, Continued

- Observe that

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

- and so

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

- Bias reduced to $O(n^{-2})$!
- Q: What is special about deleting the nth data point?

Jackknife, Continued

- Delete each data point in turn to get a low-bias estimator
- Average the estimators to reduce variance

Jackknife CI Algorithm for $\alpha = g(\mu_X, \mu_Y)$

1. Choose n and δ, and set $z_\delta = 1 - (\delta/2)$ quantile of $N(0,1)$
2. Simulate to get $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.
3. $\alpha_n \leftarrow g(\bar{X}_n, \bar{Y}_n)$
4. For $i = 1$ to n
   - 4.1 $\alpha_n^i \leftarrow g\left(\frac{1}{n-1}\sum_{j=1, j\neq i}^{n} X_j, \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} Y_j\right)$ (leave out X_i)
   - 4.2 $\alpha_n(i) \leftarrow n\alpha_n - (n-1)\alpha_n^i$ (ith pseudovalue)
5. Point estimator: $\alpha_n^J \leftarrow (1/n)\sum_{i=1}^n \alpha_n(i)$
6. $v_n^J = \frac{1}{n-1}\sum_{i=1}^n \left(\alpha_n(i) - \alpha_n^J\right)^2$
7. $100(1-\delta)\%$ CI: $\left[\alpha_n^J - z_\delta\sqrt{v_n^J/n}, \alpha_n^J + z_\delta\sqrt{v_n^J/n}\right]$

Jackknife, Continued

Observations

- Not obvious that CI is correct (why?)
- Substitutes computational brute force for analytical complexity
- Not a one-pass algorithm
- Basic jackknife breaks down for "non-smooth" statistics like quantiles, maximum (but can fix—see next lecture)

Estimating Nonlinear Functions of Means

Overview

Delta Method

Jackknife Method

Bootstrap Confidence Intervals

Complete Bias Elimination

Bootstrap Confidence Intervals

Another brute force method

- Key idea: analyze variability of estimator using samples of original data
- More general than jackknife (estimates entire sampling distribution of estimator, not just mean and variance)
- Jackknife is somewhat better empirically at variance estimates
- "Non-repeatable", unlike jackknife
- OK for quantiles, still breaks down for maximum

Bootstrap Samples

- Given data $\mathbf{X} = (X_1, \ldots, X_n)$: i.i.d. samples from cdf F
- Bootstrap sample $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$: i.i.d. samples from $\hat{F}$
  - Recall: empirical distribution $\hat{F}_n(x) = (1/n)(\# \text{ obs} \le x)$
  - Same as n i.i.d. samples with replacement from $\{X_1, \ldots, X_n\}$

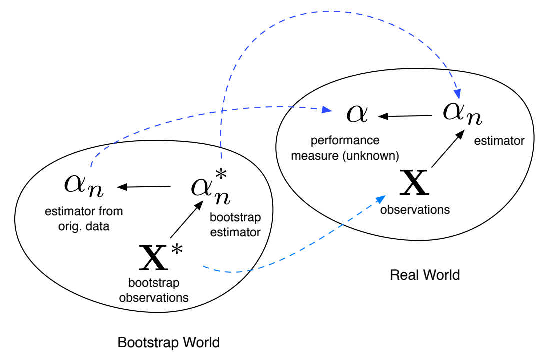
Creating a Bootstrap Sample $\mathbf{X}^*$ from $\mathbf{X} = (X_1, \ldots, X_n)$

For $i = 1$ to n:

1. Generate $U \overset{D}{\sim} \text{Uniform}(0, 1)$
2. Set $J = \lceil nU \rceil$ //Random integer between 1 and n
3. Add X_J to $\mathbf{X}^*$

Data	4	2	7	6	8	3
Sample 1	6	2	2	8	7	6
Sample 2	3	8	8	8	2	4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

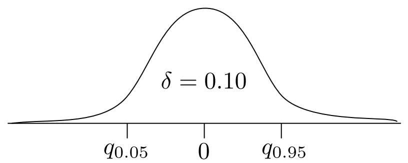
Bootstrap "Imitates" the Real World



- Boostrap world approaches real world as $n \to \infty$
  - Glivenko-Cantelli: $\sup_x |\hat{F}_n(x) - F(x)| \to 0$ wp1 as $n \to \infty$
- So distribution of $\alpha_n^* - \alpha_n$ approximates distribution of $\alpha_n - \alpha$
  - For small n, better than dist'n of α_n^* approximates dist'n of α_n
  - Hence pivot method instead of direct "percentile method"
  - Can estimate distribution of $\alpha_n^* - \alpha_n$ by sampling from it

Bootstrap Confidence Intervals: Pivot Method

Distribution of $\bar{X}_n - \mu$ is approx. $N(0, \sigma^2/n)$ by CLT



$z_{0.95} = 0.95$ quantile of $N(0,1)$

Revisit usual 90% **confidence interval for the mean**

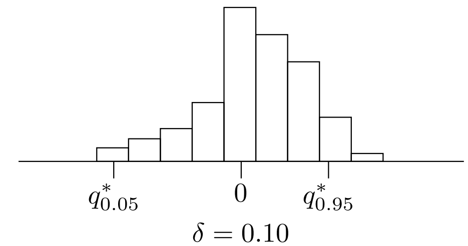
$$P(q_{0.05} \le \bar{X}_n - \mu \le q_{0.95}) \approx 0.9$$
$$\Rightarrow P(\bar{X}_n - q_{0.95} \le \mu \le \bar{X}_n - q_{0.05}) \approx 0.9$$
$$\Rightarrow 90\% \text{ CI} = [\bar{X}_n - q_{0.95}, \bar{X}_n - q_{0.05}]$$

To recover usual formulas, observe that $q_{0.05} = -q_{0.95}$ and $q_{0.95} = (\sigma/\sqrt{n})z_{0.95}$ because $P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le z_{0.95}\right) = P\left(\bar{X}_n - \mu \le q_{0.95}\right)$

$$90\% \text{ CI} = [\bar{X}_n - z_{0.95}\sigma/\sqrt{n}, \bar{X}_n + z_{0.95}\sigma/\sqrt{n}]$$

Bootstrap Confidence Intervals: Pivot Method

Distribution of $\bar{X}_n^* - \bar{X}_n$



Bootstrap approach for mean (no normality assumption)

- 90% CI $= [\bar{X}_n - q^*_{0.95}, \bar{X}_n - q^*_{0.05}]$
- Approximate quantiles of $\bar{X}_n - \mu$ by quantiles of $\pi^* = \bar{X}_n^* - \bar{X}_n$
- Generate many replicates of π^* to estimate the latter quantiles
- Technique applies to other statistics such as $\alpha = g(\mu_X, \mu_Y)$

Pivot Method for Nonlinear Functions of Means

Bootstrap Confidence Intervals (Pivot Method)

1. Run simulation n times to get $(X_1, Y_1), \dots, (X_n, Y_n)$
2. Compute $\alpha_n = g(\bar{X}_n, \bar{Y}_n)$
3. Compute bootstrap sample $(X_1^*, Y_1^*), \dots, (X_n^*, Y_n^*)$
4. Set $\alpha_n^* = g(\bar{X}_n^*, \bar{Y}_n^*)$
5. Set pivot $\pi^* = \alpha_n^* - \alpha_n$
6. Repeat Steps 3–5 B times to create $\pi_1^*, \dots, \pi_B^*$
7. Sort pivots to obtain $\pi^*_{(1)} \le \pi^*_{(2)} \le \cdots \le \pi^*_{(B)}$
8. Set $l = \lceil (1 - \delta/2)B \rceil$ and $u = \lceil (\delta/2)B \rceil$
9. Return $100(1-\delta)\%$ CI $[\alpha_n - \pi^*_{(l)}, \alpha_n - \pi^*_{(u)}]$

- Example: For $B = 100$, 90% CI is $[\alpha_n - \pi^*_{(95)}, \alpha_n - \pi^*_{(5)}]$
- Improvements include BCa bootstrap confidence interval [See Efron & Tibshirani book]

Complete Bias Elimination [Blanchet et al. 2015]

Idea: Construct X^* such that $E[X^*] = g(\mu_X)$

- Then can use usual estimation methods
- Assumes simulation cost not too expensive

Algorithm for Generating a Sample of X^*

1. Set $p = 1 - (1/2)^{3/2} \approx 0.65$ and $n_0 = 10$
2. Generate N s.t. $p(k) \stackrel{\text{def}}{=} P(N = k) = p(1-p)^{k-n_0}$ for $k \geq n_0$
3. Generate $X_1, X_2, \ldots, X_{2^{N+1}}$ i.i.d. copies of X and set

$$\bar{X}_{2^N}^{\text{odd}} = \frac{X_1 + X_3 + \cdots + X_{2^{N+1}-1}}{2^N} \quad \text{and} \quad \bar{X}_{2^N}^{\text{even}} = \frac{X_2 + X_4 + \cdots + X_{2^{N+1}}}{2^N}$$

4. Return

$$X^* = \frac{g(\bar{X}_{2^{N+1}}) - \left(g(\bar{X}_{2^N}^{\text{odd}}) + g(\bar{X}_{2^N}^{\text{even}})\right)/2}{p(N)} + g(\bar{X}_{2^{n_0}})$$

Unbiasedness of B-E Estimator

Since $E[g(\bar{X}_{2^n}^{\text{odd}})] = E[g(\bar{X}_{2^n}^{\text{even}})] = E[g(\bar{X}_{2^n})]$, for all $n \geq 1$, we have

$$E\left[g(\bar{X}_{2^{n+1}}) - \left(g(\bar{X}_{2^n}^{\text{odd}}) + g(\bar{X}_{2^n}^{\text{even}})\right)/2\right] = E[g(\bar{X}_{2^{n+1}})] - E[g(\bar{X}_{2^n})], \quad n \geq 1$$

and

$$E\left[\frac{g(\bar{X}_{2^{N+1}}) - \left(g(\bar{X}_{2^N}^{\text{odd}}) + g(\bar{X}_{2^N}^{\text{even}})\right)/2}{p(N)}\right]$$
$$= \sum_{n=n_0}^{\infty} E\left[\frac{g(\bar{X}_{2^{n+1}}) - \left(g(\bar{X}_{2^n}^{\text{odd}}) + g(\bar{X}_{2^n}^{\text{even}})\right)/2}{p(n)} \,\middle|\, N = n\right] \times p(n)$$
$$= \sum_{n=n_0}^{\infty} E\left[g(\bar{X}_{2^{n+1}}) - \left(g(\bar{X}_{2^n}^{\text{odd}}) + g(\bar{X}_{2^n}^{\text{even}})\right)/2\right] = \sum_{n=n_0}^{\infty} E[g(\bar{X}_{2^{n+1}})] - E[g(\bar{X}_{2^n})]$$
$$= E[g(\bar{X}_{2^{n_0+1}})] - E[g(\bar{X}_{2^{n_0}})] + E[g(\bar{X}_{2^{n_0+2}})] - E[g(\bar{X}_{2^{n_0+1}})] + E[g(\bar{X}_{2^{n_0+3}})] - \cdots$$
$$= E[g(\bar{X}_{2^\infty})] - E[g(\bar{X}_{2^{n_0}})] = g(\mu_x) - E[g(\bar{X}_{2^{n_0}})]$$

So

$$E\left[\frac{g(\bar{X}_{2^{N+1}}) - \left(g(\bar{X}_{2^N}^{\text{odd}}) + g(\bar{X}_{2^N}^{\text{even}})\right)/2}{p(N)} + g(\bar{X}_{2^{n_0}})\right] = g(\mu_X)$$

Can also show $\text{Var}[X^*] < \infty$ and $E[\text{simulation cost}] < \infty$