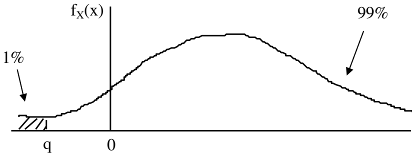
$g(E[X])$

Quantile Estimation

Peter J. Haas

CS 590M: Simulation
Spring Semester 2020

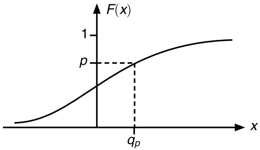
Quantiles



Example: Value-at-Risk

- $X =$ return on investment, want to measure downside risk
- $q =$ return s.t. $P(\text{worse return than } q) \leq 0.01$
  - q is called the 0.01-quantile of X
  - "Probabilistic worst case scenario"

Quantile Definition



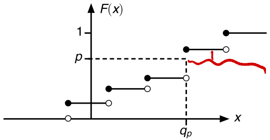
Definition of p-quantile q_p

$q_p = F_X^{-1}(p)$ (for $0 < p < 1$)

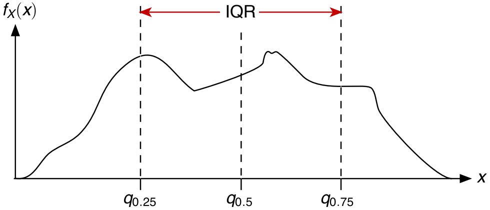
- When F_X is continuous and increasing: solve $F(q) = p$
- In general: Use our generalized definition of F^{-1} (as in inversion method)

Alternative Definition of p-quantile q_p

$q_p = \min\{q : F_X(q) \geq p\}$



Example: Robust Statistics



Median

- Median $= q_{0.5}$
- Alternative to means as measure of central tendency
- Robust to outliers

Inter-quartile range (IQR)

- Robust measure of dispersion
- IQR $= q_{0.75} - q_{0.25}$

Point Estimate of Quantile

- Given i.i.d. observations $X_1, \ldots, X_n \overset{D}{\sim} F$
- Natural choice is pth sample quantile:

$$Q_n = \hat{F}_n^{-1}(p)$$

- I.e., generalized inverse of empirical cdf $\hat{F}_n$
- Q: Can you ever use the simple (non-generalized) inverse here?
- Equivalently, sort data as $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ and set

$$Q_n = X_{(j)}, \quad \text{where} \quad j = \lceil np \rceil$$

- Ex: $q_{0.5}$ for $\{6, 8, 4, 2\} =$
- Other definitions are possible (e.g., interpolating between values), but we will stick with the above defs

Quantile Estimation

Confidence Interval Attempt #1: Direct Use of CLT

CLT for Quantiles (Bahadur Representation)

Suppose that $X_1, \ldots, X_n$ are i.i.d. with pdf f_X. Then for large n

$$Q_n \overset{D}{\sim} N\left(q_p, \frac{\sigma^2}{n}\right) \quad \text{with} \quad \sigma = \frac{\sqrt{p(1-p)}}{f_X(q_p)}$$

Can derive via Delta Method for stochastic root-finding

- Recall: to find $\bar{\theta}$ such that $E[g(X, \bar{\theta})] = 0$
  - Point estimate θ_n solves $\frac{1}{n}\sum_{i=1}^{n} g(X_i, \theta_n) = 0$
  - For large n, we have $\theta_n \approx N(\bar{\theta}, \sigma^2/n)$, where $\sigma^2 = \text{Var}[g(X, \bar{\theta})]/c^2$ with $c = E[\partial g(X, \bar{\theta})/\partial\theta]$
- For quantile estimation take $g(X, \theta) = I(X \le \theta) - p$
  - $\bar{\theta} = q_p$ and $\theta_n = Q_n$, since $E[g(X, \bar{\theta})] = P(X \le \bar{\theta}) - p = 0$
  - $E[\partial g(X, \bar{\theta})/\partial\theta] = \partial E[g(X, \bar{\theta})]/\partial\theta = \partial\big(F_X(\bar{\theta}) - p\big)/\partial\theta = f_X(\bar{\theta})$
  - $\text{Var}[g(X, \bar{\theta})] = E[g(X, \bar{\theta})^2] = E[I^2 - 2pI + p^2]$
    $= E[I - 2pI + p^2] = p - 2p^2 + p^2 = p(1-p)$

Confidence Interval Attempt #1: Direct Use of CLT

CLT for Quantiles (Bahadur Representation)

Suppose that $X_1, \ldots, X_n$ are i.i.d. with pdf f_X. Then for large n

$$Q_n \overset{D}{\sim} N\left(q_p, \frac{\sigma^2}{n}\right) \quad \text{with} \quad \sigma = \frac{\sqrt{p(1-p)}}{f_X(q_p)}$$

- So if we can find an estimator s_n of σ, then $100(1-\delta)\%$ CI is

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

- Problem: Estimating a pdf f_X is hard (e.g., need to choose "bandwidth" for "kernel density estimator")
- So we want to avoid estimation of σ

Confidence Interval Attempt #2: Sectioning

- Assume that $n = mk$ and divide $X_1, \ldots, X_n$ into m sections of k observations each
- m is small (around 10–20) and k is large
- Let $Q_n(i)$ be estimator of q_p based on data in ith section
- Observe that $Q_n(1), \ldots, Q_n(m)$ are i.i.d.
- By prior CLT, each $Q_n(i)$ is approx. distributed as $N\left(q_p, \frac{\sigma^2}{k}\right)$
- For i.i.d. normals, standard $100(1-\delta)\%$ CI for mean is

$$\left[\bar{Q}_n - t_{m-1,\delta}\sqrt{\frac{v_n}{m}}, \bar{Q}_n + t_{m-1,\delta}\sqrt{\frac{v_n}{m}}\right]$$

  - $\bar{Q}_n = (1/m)\sum_{i=1}^m Q_n(i)$
  - $v_n = \frac{1}{m-1}\sum_{i=1}^m \left(Q_n(i) - \bar{Q}_n\right)^2$
  - $t_{m-1,\delta}$ is $1-(\delta/2)$ quantile of Student-t distribution with $m-1$ degrees of freedom

Sectioning: So What's the Problem?

$n = mk$

- Can show, as with nonlinear functions of means, that

$$E[Q_n] \approx q_p + \frac{b}{n} + \frac{c}{n^2}$$

- It follows that

$$E[Q_n(i)] \approx q_p + \frac{b}{k} + \frac{c}{k^2} = q_p + \frac{mb}{n} + \frac{m^2 c}{n^2}$$

- So

$$E[\bar{Q}_n] \approx q_p + \frac{mb}{n} + \frac{m^2 c}{n^2}$$

- Bias of $\bar{Q}_n$ is roughly m times larger than bias of Q_n!

Attempt #3: Sectioning + Jackknifing

Sectioning + Jackknifing: General Algorithm for a Statistic α

1. Generate $n = mk$ i.i.d. observations $X_1, \ldots, X_n$
2. Divide observations into m sections, each of size k
3. Compute point estimator α_n based on all observations
4. For $i = 1, 2, \ldots, m$:
   4.1 Compute estimator $\tilde{\alpha}_n(i)$ using all observations except those in section i
   4.2 Form pseudovalue $\alpha_n(i) = m\alpha_n - (m-1)\tilde{\alpha}_n(i)$
5. Compute point estimator: $\alpha_n^J = \frac{1}{m}\sum_{i=1}^{m} \alpha_n(i)$
6. Set $v_n^J = \frac{1}{m-1}\sum_{i=1}^{m} \left(\alpha_n(i) - \alpha_n^J\right)^2$
7. Compute $100(1-\delta)\%$ CI: $\left[\alpha_n^J - t_{m-1,\delta}\sqrt{\frac{v_n^J}{m}}, \alpha_n^J + t_{m-1,\delta}\sqrt{\frac{v_n^J}{m}}\right]$

Application to Quantile Estimation

- $\tilde{Q}_n(i) =$ quantile estimate ignoring section i
- Clearly, $\tilde{Q}_n(i)$ has same distribution as $Q_{(m-1)k}$, so

$$E[\tilde{Q}_n(i)] \approx q_p + \frac{b}{(m-1)k} + \frac{c}{(m-1)^2 k^2}$$

- It follows that, for pseudovalue $\alpha_n(i)$,

$$E[\alpha_n(i)] = E\left[mQ_n - (m-1)\tilde{Q}_n(i)\right] \approx q_p - \frac{c}{(m-1)mk^2}$$

- Averaging does not affect bias, so, since $n = mk$,

$$E[\bar{Q}_n] = q_p + O(1/n^2)$$

- General procedure is also called the "delete-k jackknife"

Further Comments

A confession

- There exist special-purpose methods for quantile estimation [Sections 2.6.1 and 2.6.3 in Serfling book]
- We focus on sectioning + jackknife because method is general
- Can also use bias elimination method from prior lecture

Conditioning the data for q_p when $p \approx 1$

- Fix $r > 1$ and get $n = rmk$ i.i.d. observations $X_1, \ldots, X_n$
- Divide data into blocks of size r
- Set $Y_j =$ maximum value in jth block for $1 \le j \le mk$
- Run quantile estimation procedure on $Y_1, \ldots, Y_{mk}$
- Key observation: $F_Y(q_p) = [F_X(q_p)]^r = p^r$
  - So p-quantile for X equals p^r-quantile for Y
  - Ex: if $r = 50$, then $q_{0.99}$ for X equals $q_{0.61}$ for Y
- Often, reduction in sample size outweighs cost of extra runs

$F_Y(q_p) = P(\max_i X_i \le q_p)$
$= P(X_1, X_2, \ldots, X_r \le q_p)$
$= P(X_1 \le q_p)^r$

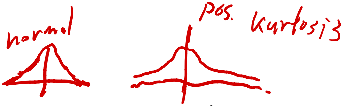
Checking Normality

Undercoverage

- E.g., when a "95% confidence interval" for the mean only brackets the mean 70% of the time
- Due to failure of CLT at finite sample sizes
- Note: If data is truly normal, then no error in CI for the mean

Simple diagnostics

- Skewness (measures symmetry, equals 0 for normal)
  - Definition: $\text{skewness}(X) = \frac{E[(X - E(X))^3]}{(\text{var } X)^{3/2}}$
  - Estimator: $\frac{n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^3}{\left(n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right)^{3/2}}$



- Kurtosis (measures fatness of tails, equals 0 for normal)
  - Definition: $\text{kurtosis}(X) = \frac{E[(X - E(X))^4]}{(\text{var } X)^2} - 3$
  - Estimator: $\frac{n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^4}{\left(n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right)^2} - 3$

Bootstrap Confidence Intervals

General method works for quantiles (no normality assumptions needed)

Bootstrap Confidence Intervals (Pivot Method)

1. Run simulation n times to get $\mathcal{D} = \{X_1, \ldots, X_n\}$
2. Compute Q_n = sample quantile based on $\mathcal{D}$
3. Compute bootstrap sample $\mathcal{D}^* = \{X_1^*, \ldots, X_n^*\}$
4. Set Q_n^* = sample quantile based on $\mathcal{D}^*$
5. Set pivot $\pi^* = Q_n^* - Q_n$ ("bootstrap world" estimate of "real world" quantity $Q_n - q_p$)
6. Repeat Steps 3–5 B times to create $\pi_1^*, \ldots, \pi_B^*$
7. Sort pivots to obtain $\pi_{(1)}^* \leq \pi_{(2)}^* \leq \cdots \leq \pi_{(B)}^*$
8. Set $l = \lceil (1 - \delta/2)B \rceil$ and $u = \lceil (\delta/2)B \rceil$
9. Return $100(1-\delta)\%$ CI $[Q_n - \pi_{(l)}^*, Q_n - \pi_{(u)}^*]$