Computer Science

CSC380: Principles of Data Science

Statistics 5

Quantile method and bootstrapping

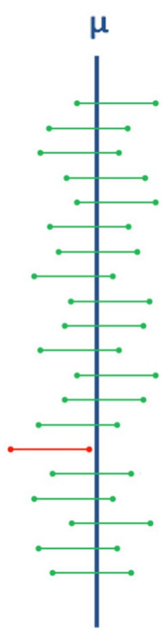
Credit:
- Jason Pacheco,
- Kwang-Sung Jun,
- Chicheng Zhang
- Xinchen yu

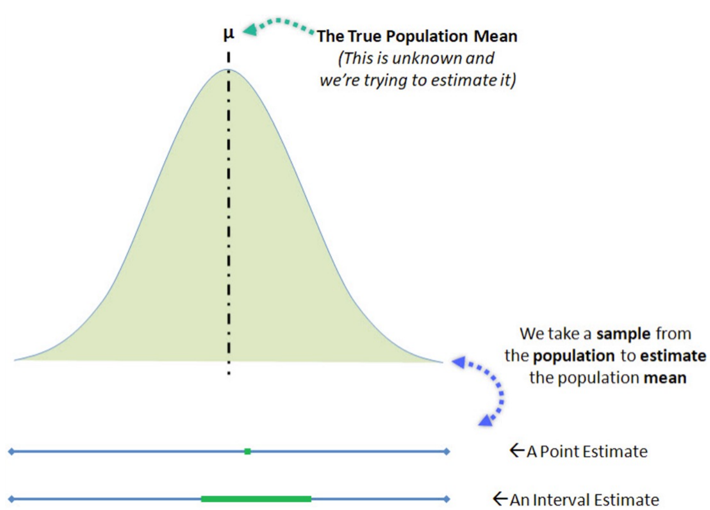
1

1

Review: Interval estimate

2





2

Method 2: Bootstrap

Suppose $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ & known σ^2.

(Fact 1) $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ $\sqrt{n}\frac{\boxed{\hat{\mu}-\mu}}{\sigma} \sim \mathcal{N}(0,1)$

Directly approximate distributions of $\hat{\mu} - \mu$

(Fact 2) If $Z \sim \mathcal{N}(0,1)$,

$$P(Z \in [-z, z]) = 1 - 2(1 - \Phi(z))$$

where $\Phi(z) \coloneqq \mathrm{P}(Z \le z)$ is the CDF of Z.

z = 1.96: RHS ≈ .95, 95% confident

z = 2.58: RHS ≈ .99,

Let: $Z \longrightarrow \sqrt{n}\,\frac{\hat{\mu}-\mu}{\sigma}$

$$P\left(\hat{\mu} \in \left[\mu - \frac{1.96\sigma}{\sqrt{n}}, \mu + \frac{1.96\sigma}{\sqrt{n}}\right]\right) \ge 0.95$$

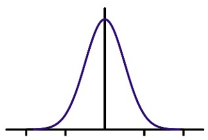
$$P\left(\hat{\mu} \in \left[\mu - \frac{2.58\sigma}{\sqrt{n}}, \mu + \frac{2.58\sigma}{\sqrt{n}}\right]\right) \ge 0.99$$

=> Compute $\left[\hat{\mu} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu} + \frac{1.96\sigma}{\sqrt{n}}\right]$. Done!

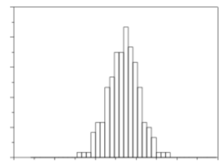
Method 2: Bootstrap

- Key idea: approximate ν, the distribution of $\hat{\theta}_n - \theta$
- Insight:

θ

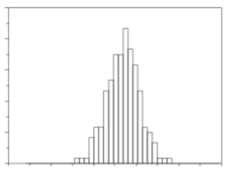


$D_\theta \xrightarrow{n \text{ iid samples}} S$

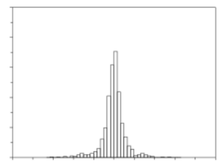


$\hat{\theta}_n$

$\hat{\theta}_n$



$\text{Uniform(S)} \xrightarrow{n \text{ iid samples}} S_b$



$\hat{\theta}_{n,b}$

- Use empirical distribution of $\hat{\theta}_{n,b} - \hat{\theta}_n$'s to approximate ν, obtaining approximations of $\nu_{\alpha/2}$ and $\nu_{1-\alpha/2}$
- This empirical distribution can be obtained by drawing multiple S_b's (bootstrap subsample)

10

Bootstrapping

Another method for estimating confidence intervals

Actually, this method is useful to estimate robustly all types of statistics (medians, quantiles, moments..)

Remember – if we know σ and that the distribution is Gaussian, we can do with a small sample (≤ 30)

If we don't know σ but sample is larger, we can use the central limit thm – in particular, obtain σ from the samaple.

What if not normal distribution and small n.

11

Bootstrapping – convert a small sample into a many sample

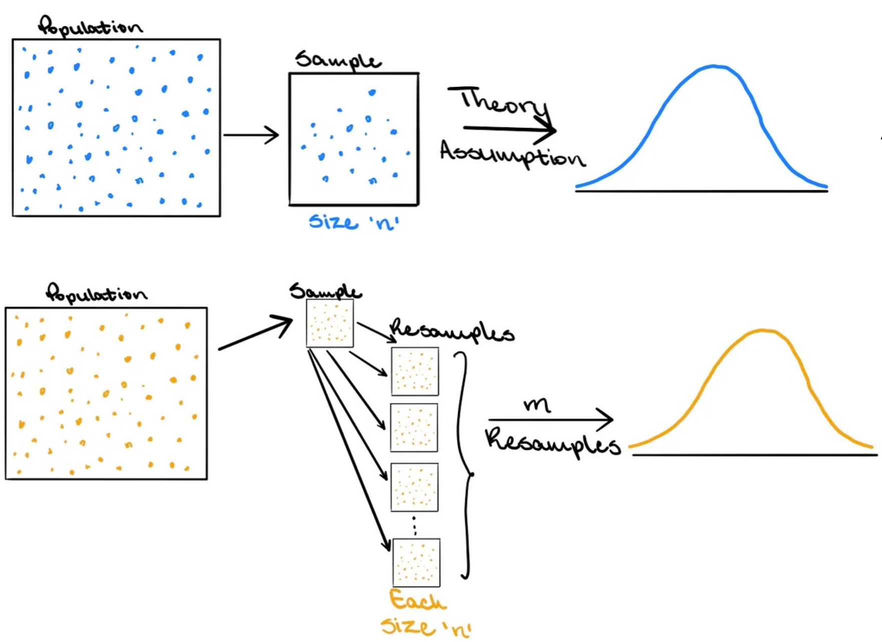
S- random sample of data – say S= {9,17,25}
Pick a random sample from S, - but with repetitions.

Sort each such sample

$$S_1 = \{9, 17, 17\}$$
$$S_2 = \{9, 9, 9\}$$
$$S_3 = \{9, 9, 9\}$$
$$S_4 = \{9, 25, 25\}$$
$$S_5 = \{9, 17, 25\}$$

12

Method 2: Bootstrap



Approximate distribution

$\hat{\theta} - \theta$

$\left\{\hat{\theta}_{n,b} - \frac{1}{n}\sum_{i=1}^{n} X_i\right\}_{b=1}^{B}$

13

Method 2: Bootstrap example

Sample data: `30, 37, 36, 43, 42, 43, 43, 46, 41, 42`

Sample mean: $\overline{x} = 40.3$

We want to know the distribution of: $\delta = \overline{x} - \mu$

Can approximate the distribution: $\delta^* = \overline{x}^* - \overline{x}$

Let's resample data with same size and generate 20 bootstrap samples:

43	36	46	30	43	43	43	37	42	42	43	37	36	42	43	43	42	43	42	43
43	41	37	37	43	43	46	36	41	43	43	42	41	43	46	36	43	43	43	42
42	43	37	43	46	37	36	41	36	43	41	36	37	30	46	46	42	36	36	43
37	42	43	41	41	42	36	42	42	43	42	43	41	43	36	43	43	41	42	46
42	36	43	43	42	37	42	42	42	46	30	43	36	43	43	42	37	36	42	30
36	36	42	42	36	36	43	41	30	42	37	43	41	41	43	43	42	46	43	37
43	37	41	43	41	42	43	46	46	36	43	42	43	30	41	46	43	46	30	43
41	42	30	42	37	43	43	42	43	43	46	43	30	42	30	42	30	43	43	42
46	42	42	43	41	42	30	37	30	42	43	42	43	37	37	37	42	43	43	46
42	43	43	41	42	36	43	30	37	43	42	43	41	36	37	41	43	42	43	43

14

Method 2: Bootstrap example

43	36	46	30	43	43	43	37	42	42	43	37	36	42	43	43	42	43	42	43
43	41	37	37	43	43	46	36	41	43	43	42	41	43	46	36	43	43	43	42
42	43	37	43	46	37	36	41	36	43	41	36	37	30	46	46	42	36	36	43
37	42	43	41	41	42	36	42	42	43	42	43	41	43	36	43	43	41	42	46
42	36	43	43	42	37	42	42	42	46	30	43	36	43	43	42	37	36	42	30
36	36	42	42	36	36	43	41	30	42	37	43	41	41	43	43	42	46	43	37
43	37	41	43	41	42	43	46	46	36	43	42	43	30	41	46	43	46	30	43
41	42	30	42	37	43	43	42	43	43	46	43	30	42	30	42	30	43	43	42
46	42	42	43	41	42	30	37	30	42	43	42	43	37	37	37	42	43	43	46
42	43	43	41	42	36	43	30	37	43	42	43	41	36	37	41	43	42	43	43

Calculate sample mean for each column (bootstrap sample), compute: $\delta^* = \overline{x}^* - \overline{x}$

Sort the 20 differences:

-1.6, -1.4, -1.4, -0.9, -0.5, -0.2, -0.1, 0.1, 0.2, 0.2, 0.4, 0.4, 0.7, 0.9, 1.1, 1.2, 1.2, 1.6, 1.6, 2.0

If confidence level is 80%, find out top 10% and bottom 10%:

-1.6, $\boxed{-1.4}$ -1.4, -0.9, -0.5, -0.2, -0.1, 0.1, 0.2, 0.2, 0.4, 0.4, 0.7, 0.9, 1.1, 1.2, 1.2, 1.6, $\boxed{1.6,}$ 2.0

The bootstrap confidence interval is:

$$[\overline{x} - \delta^*_{.1},\ \overline{x} - \delta^*_{.9}] \ = \ [40.3 - 1.6,\ 40.3 + 1.4] \ = \ [38.7,\ 41.7]$$

15

Method 2: Bootstrap

16

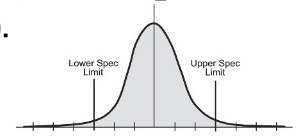
Suppose we observe data $X_1, X_2, \dots, X_n \sim P(X;\theta)$:

1. Sample new "dataset" $X_1^*, \dots, X_n^*$ uniformly from $X_1, \dots, X_n$ **with replacement**

2. Compute estimate $\hat{\theta}_n(X_1^*, \dots, X_n^*)$

3. Repeat B times to get the estimators $\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,B}$

4. Consider the **empirical distribution** of $\left\{\hat{\theta}_{n,b} - \frac{1}{n}\sum_{i=1}^{n} X_i\right\}_{b=1}^{B}$ and find its top $\frac{\alpha}{2}$ quantile and bottom $\frac{\alpha}{2}$ quantile (denoted by Q_U and Q_L respectively).



5. (1-α) Confidence Interval: $\left[\frac{1}{n}\sum_{i=1}^{n} X_i - |Q_U|, \frac{1}{n}\sum_{i=1}^{n} X_i + |Q_L|\right]$

counterintuitively, upper quantile for lower width, lower quantile for upper width. Why?

$$P\left(v_{\frac{\alpha}{2}} \le \hat{\theta}_n - \theta \le v_{1-\frac{\alpha}{2}}\right) \ge 1 - \alpha$$

16

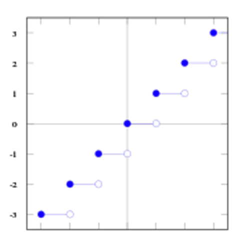
Method 2: Bootstrap

17

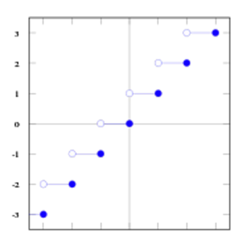
Pseudocode

Input: $X_1, \dots, X_n, B, \alpha$

- Compute $\bar{X}_n$
- Bootstrapping B times to obtain $\{\hat{\theta}_{n,b} - \bar{X}_n\}_{b=1}^{B}$; call this array S
- Sorted S in increasing order.
- $Q_U :=$ the top $\frac{\alpha}{2}$ quantile; i.e., `S[int(np.ceil( (1-alpha/2)*(B-1) ))]`
- $Q_L :=$ the bottom $\frac{\alpha}{2}$ quantile; i.e., `S[int(np.floor((  alpha/2)*(B-1) ))]`
- Return $[\bar{X}_n - |Q_U|, \bar{X}_n + |Q_L|]$



floor



ceil

17

Confidence Intervals Comparison

18

good = correct
bad = incorrect

	Gaussian (corrected)	Bootstrap
small n	Bad	Bad
moderate n	Okay / Bad	Okay
large n	Good	Very Good
Computational complexity	Low	High, depends on B

bad if the estimator takes a long time to compute

Q: When could it be bad?

When the distribution is far from Gaussian

18

Classical Statistics Review

19

- **Statistical Estimation** infers unknown parameters θ of a distribution $p(X;\theta)$ from observed data $X_1, \ldots, X_n$
- An estimator is a function of the data $\hat{\theta}(X_1, \ldots, X_n)$, it is a **random variable**, so it has a distribution
- **Confidence Intervals** measure uncertainty of an estimator, e.g.

$$P(\theta \in (a(X), b(X))) \geq 0.95$$

- **Bootstrap** A simple method for estimating confidence intervals

↑ Q: when is this good?

Caution

- Confidence intervals are often misinterpreted!
- Confidence intervals in practice may not be true for small n

19

Classical Statistics Review

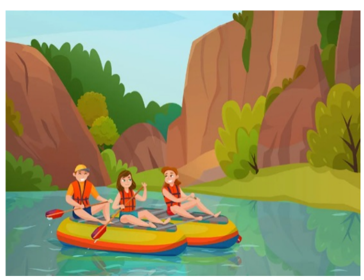
20

- **Estimator bias** describes systematic error of an estimator
- **Mean squared error (MSE)** measures estimator quality / efficiency,

$$\mathrm{MSE}(\hat{\theta}) = \mathbf{E}\left[(\hat{\theta} - \theta)^2\right] = \mathrm{bias}^2(\hat{\theta}) + \mathbf{Var}(\hat{\theta})$$

- **Law of Large Numbers (LLN)** guarantees that sample mean approaches (piles up near) true mean in the limit of infinite data
- **Central Limit Theorem (CLT)** says sample mean approaches a Normal distribution with enough data. Also means $\frac{1}{\sqrt{n}}$ convergence.
- **LLN** and **CLT** are *asymptotic statements* and do not hold for small/medium data in general

20



- Probability
- Statistics

- Data Visualization

- Predictive modeling
- Linear models
- Nonlinear models
- Clustering

21