The Bootstrap[1]

STA442/2101 Fall 2017

[1] See last slide for copyright information.

Overview

Sampling distributions

- Let $\mathbf{x} = (X_1, \ldots, X_n)$ be a random sample from some distribution F.
- $T = T(\mathbf{x})$ is a statistic (could be a vector of statistics).
- Need to know about the distribution of T.
- Sometimes it's not easy, even asymptotically.

Sampling distribution of T: The elementary version

For example $T = \overline{X}$

- Sample repeatedly from this population (pretend).
- For each sample, calculate T.
- Make a relative frequency histogram of the T values you observe.
- As the number of samples becomes very large, the histogram approximates the distribution of T.

What is a bootstrap?

Pull yourself up by your bootstraps

The (statistical) Bootstrap

Bradley Efron, 1979

- Select a random sample from the population.
- If the sample size is large, the sample is similar to the population.
- Sample repeatedly from the sample. This is called *resampling*.
- Sample from the sample? Think of putting the sample data values in a jar . . .
- Calculate the statistic for every bootstrap sample.
- A histogram of the resulting values approximates the shape of the sampling distribution of the statistic.

Notation

- Let $\mathbf{x} = (X_1, \ldots, X_n)$ be a random sample from some distribution F.
- $T = T(\mathbf{x})$ is a statistic (could be a vector of statistics).
- Form a "bootstrap sample" $\mathbf{x}^*$ by sampling n values from $\mathbf{x}$ *with replacement*.
- Repeat this process B times, obtaining $\mathbf{x}_1^*, \ldots, \mathbf{x}_B^*$.
- Calculate the statistic for each bootstrap sample, obtaining $T_1^*, \ldots, T_B^*$.
- Relative frequencies of $T_1^*, \ldots, T_B^*$ approximate the sampling distribution of T.

Why does it work?

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\} \stackrel{a.s.}{\rightarrow} E(I\{X_i \leq x\}) = F(x)$$

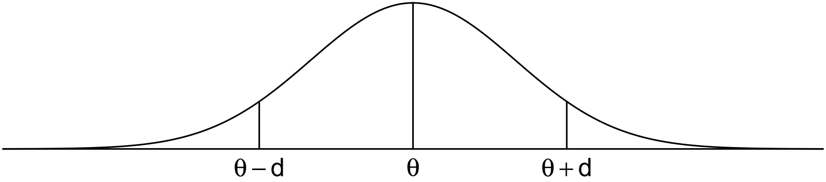
- Resampling from $\mathbf{x}$ with replacement is the same as simulating a random variable whose distribution is the empirical distribution function $\widehat{F}(x)$.
- Suppose the distribution function of T is a nice smooth function of F.
- Then as $n \rightarrow \infty$ and $B \rightarrow \infty$, bootstrap sample moments and quantiles of $T_1^*, \ldots, T_B^*$ converge to the corresponding moments and quantiles of the distribution of T.
- If the distribution of $\mathbf{x}$ is discrete and supported on a finite number of points, the technical issues are minor.

Quantile Bootstrap Confidence Intervals

- Suppose T_n is a consistent estimator of $g(\theta)$.
- And the distribution of T_n is approximately symmetric around $g(\theta)$.
- Then the lower $(1-\alpha)100\%$ confidence limit for $g(\theta)$ is the $\alpha/2$ sample quantile of $T_1^*, \ldots, T_B^*$, and the upper limit is the $1-\alpha/2$ sample quantile.
- For example, the 95% confidence interval ranges from the 2.5th to the 97.5th percentile of $T_1^*, \ldots, T_B^*$.

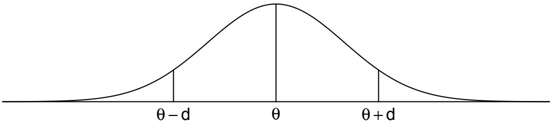
Symmetry

A requirement that is often ignored



The distribution of T symmetric about θ means for all $d > 0$,
$P\{T > \theta + d\} = P\{T < \theta - d\}$.

Why Symmetry?



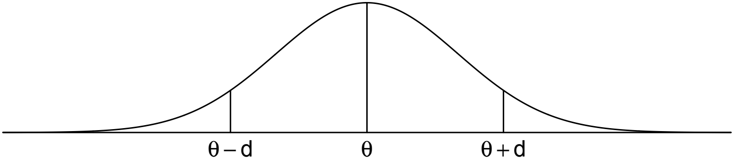
- The distribution of T symmetric about θ means for all $d > 0$, $P\{T > \theta + d\} = P\{T < \theta - d\}$.
- Select d so that the probability equals $\alpha/2$.

$$
\begin{aligned}
1 - \alpha &= P\{\theta - d < T < \theta + d\} \\
&= P\{T - d < \theta < T + d\}
\end{aligned}
$$

Need to estimate d.

Estimating d

There are two natural estimates



$$1 - \alpha = P\{\theta - d < T < \theta + d\} = P\{Q_{1-\alpha/2} < T < Q_{\alpha/2}\}$$

$$\widehat{\theta} - \widehat{d}_1 = \widehat{Q}_{\alpha/2} \quad \Rightarrow \quad \widehat{d}_1 = T - \widehat{Q}_{\alpha/2}$$

$$\widehat{\theta} + \widehat{d}_2 = \widehat{Q}_{1-\alpha/2} \quad \Rightarrow \quad \widehat{d}_2 = \widehat{Q}_{1-\alpha/2} - T$$

I would average them:

$$\widehat{d} = \frac{1}{2}(\widehat{d}_1 + \widehat{d}_2) = \frac{1}{2}(\widehat{Q}_{1-\alpha/2} - \widehat{Q}_{\alpha/2})$$

$1 - \alpha = P\{T - d < \theta < T + d\}$

Plug in an estimate of d

- $\widehat{d}_1 = T - \widehat{Q}_{\alpha/2}$
- $\widehat{d}_2 = \widehat{Q}_{1-\alpha/2} - T$
- $\widehat{d} = \frac{1}{2}(\widehat{d}_1 + \widehat{d}_2)$

Using $\widehat{d}_1$ on the left yields

$$T - \widehat{d}_1 = T - (T - \widehat{Q}_{\alpha/2}) = \widehat{Q}_{\alpha/2}$$

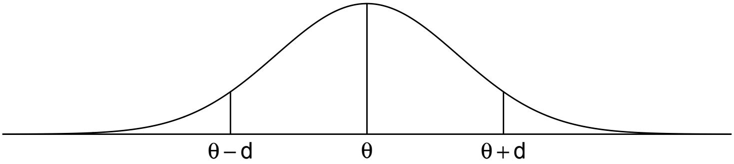
Using $\widehat{d}_2$ on the right yields

$$T + \widehat{d}_2 = T + (\widehat{Q}_{1-\alpha/2} - T) = \widehat{Q}_{1-\alpha/2},$$

which is the quantile confidence interval.

Maybe more reasonable: $T \pm \widehat{d}$

But this is just me



where

- $\widehat{d}_1 = T - \widehat{Q}_{\alpha/2}$
- $\widehat{d}_2 = \widehat{Q}_{1-\alpha/2} - T$
- $\widehat{d} = \frac{1}{2}(\widehat{d}_1 + \widehat{d}_2)$

Justifying the Assumption of Symmetry

- Smooth functions of asymptotic normals are asymptotically normal.
- This includes functions of sample moments and MLEs.
- Delta method:

  $\sqrt{n}\,(T_n - \theta) \stackrel{d}{\to} T \sim N(0, \sigma^2)$ means T_n is asymptotically normal.

  $\sqrt{n}\,(g(T_n) - g(\theta)) \stackrel{d}{\to} Y \sim N\left(0, g'(\theta)^2\,\sigma^2\right)$ means $g(T_n)$ is asymptotically normal too.
- Univariate and multivariate versions.

Can use asymptotic normality directly

Suppose T is asymptotically normal.

- Sample standard deviation of $T_1^*, \ldots, T_B^*$ is a good standard error.
- Confidence interval is $T \pm 1.96\, SE$.
- If T is a vector, the sample variance-covariance matrix of $T_1^*, \ldots, T_B^*$ is useful.

Example

Let $Y_1, \ldots, Y_n$ be a random sample from an unknown distribution with expected value μ and variance σ^2. Give a point estimate and a 95% confidence interval for the coefficient of variation $\frac{\sigma}{\mu}$.

- Point estimate is $T = S/\overline{Y}$.
- If $\mu \neq 0$ then T is asymptotically normal and therefore symmetric.
- Resample from the data urn n times with replacement, and calculate T_1^*.
- Repeat B times, yielding $T_1^*, \ldots, T_B^*$.
- Percentile confidence interval for $\frac{\sigma}{\mu}$ is $(\widehat{Q}_{\alpha/2}, \widehat{Q}_{1-\alpha/2})$.
- Alternatively, since T is approximately normal, calculate $\widehat{\sigma}_T = \frac{1}{B-1} \sum_{i=i}^{B} (T_i^* - \overline{T}^*)^2$
- And a 95% confidence interval is $T \pm 1.96\, \widehat{\sigma}_T$.

Example: Distribution-free regression

Independently for $i = 1, \ldots, n$, let

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where

- X_i and ϵ_i come from unknown distributions,
- $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$,
- X_i and ϵ_i are independent.
- Moments of X_i will be denoted $E(X)$, $E(X^2)$, etc.

Observable data consist of the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Estimation

Estimate β_0 and β_1 as usual by

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\ &= \frac{\sum_{i=1}^n X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2} \text{ and}\end{aligned}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

- Consistency follows from the Law of Large Numbers and continuous mapping.
- Looks like $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are asymptotically normal.
- Use this to get tests and confidence intervals.

Bootstrap approach: All by computer

- Earlier discussion implies $\widehat{\boldsymbol{\beta}}$ is asymptotically multivariate normal.
- Say $\widehat{\boldsymbol{\beta}} \stackrel{.}{\sim} N_p(\boldsymbol{\beta}, \mathbf{V})$.
- All we need is a good $\widehat{\mathbf{V}}$.
- Put data vectors $\mathbf{d}_i = (\mathbf{x}_i, Y_i)$ in a jar.
- Sample n vectors with replacement, yielding $\mathbf{D}_1^*$. Fit the regression model, obtaining $\widehat{\boldsymbol{\beta}}_1^*$.
- Repeat B times. This yields $\widehat{\boldsymbol{\beta}}_1^* \ldots \widehat{\boldsymbol{\beta}}_B^*$.
- The sample covariance matrix of $\widehat{\boldsymbol{\beta}}_1^* \ldots \widehat{\boldsymbol{\beta}}_B^*$ is $\widehat{\mathbf{V}}$.
- Under $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$,

$$(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})^\top (\mathbf{L}\widehat{\mathbf{V}}^{-1}\mathbf{L}^\top)^{-1} (\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h}) \stackrel{.}{\sim} \chi^2(r)$$

Remark

This is not a typical bootstrap regression.

- Usually people fit a model and then bootstrap the residuals, not the whole data vector.
- Bootstrapping the residuals applies to conditional regression (conditional on $\mathbf{X} = \mathbf{x}$).
- Our regression model is unconditional.
- The large-sample arguments are simpler in the unconditional case.

Copyright Information