STA261: Week 11

Computational Methods: Jackknife and Bootstrap

Alex Stringer

March 26th - 30th, 2018

Disclaimer

The materials in these slides are intended to be a companion to the course textbook, *Mathematical Statistics and Data Analysis, Third Edition*, by John A Rice. Material in the slides may or may not be taken directly from this source. These slides were organized and typeset by Alex Stringer.

A big thanks to Jerry Brunner as well for providing inspiration for assignment questions.

License

Recap

In this course we have talked about

- How to estimate quantities of interest, given data
- How to estimate the variability in our estimates
- How to compare estimates to their variability to make decisions/inferences

The central idea behind the type of frequentist inference we have learned is the *standard deviation of the estimator*, sometimes called the **standard error**.

Standard Error

We have gotten standard errors for two big cases:

- *Exact*: when $X \sim N(\mu, \sigma^2)$ and $\hat{\mu} = \bar{X}$, then $SD(\hat{\mu}) = s/n$ exactly
- *Approximate*: the CLT for the MLE implies that in finite samples, $\hat{\theta}$ has standard deviation that is well approximated by $1/\sqrt{I(\theta_0)}$

What about standard errors for functions of $\mathbf{X}$ (note: vector $\mathbf{X}$, i.e. the whole sample) that aren't based on the above two situations?

Standard Error

In general, formulas don't exist for $SD(g(\mathbf{X}))$ for arbitrary g.

If g is smooth, can use more Taylor series arguments to linearize g and then compute approximate standard errors (the "delta" method).

Measures of variability are interpreted as the spread in values we would see in repeated sampling of a quantity- e.g. if we sampled $g(\mathbf{X})$ a bunch of times, we could estimate $SD(g(\mathbf{X}))$ using the sample standard deviation of this sample.

We can't do that... or can we?

In the 1960's, computers came along.

The Jackknife

One of the first computational algorithms for computing approximate standard errors of estimators was the Jackknife.

We have a parameter θ that we are estimating using an estimator $\hat{\theta}(\mathbf{X})$ based off our sample $\mathbf{x}$.

Let $\mathbf{x}_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ be the sample with the i^{th} value removed (the order is arbitrary).

Denote by $\hat{\theta}_{(i)}$ the value of $\hat{\theta}$ computed using $\mathbf{x}_{(i)}$

The Jackknife

The Jackknife estimator of the standard error of $\hat{\theta}$ is then

$$\hat{SE}_{Jack} = \left(\frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} \right)^2 \right)^{1/2}$$

where

$$\hat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)}$$

The Jackknife

The Jackknife creates fictional random sampling from the population that generated our sample $\mathbf{x}$.

Exercise: show that when $\hat{\theta} = \bar{X}$, the sample mean, the jackknife estimator is *exactly* equal to the actual sample estimate of the standard error of $\bar{X}$:

$$SD(\bar{X}) = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(n-1)}}$$

Hint: find closed-form expressions for $\hat{\theta}_{(i)}$ and $\hat{\theta}_{(\cdot)}$ in this case.

The Non-Paramteric Bootstrap

The jackknife uses samples of size $n-1$. Every sample used contains most of the data.

The **Non-parametric Bootstrap** extends this notion. Rather than using leave-one-out samples, why don't we take random samples from our dataset?

Suppose our original IID sample was $X_i \sim F_\theta, i = 1 \ldots n$.

We got our original sample and estimate according to the following:

$$F_\theta \to \mathbf{x} \to \hat{\theta}(\mathbf{x})$$

The Non-Parametric Bootstrap

To estimate the standard deviation of $\hat{\theta}$ using a non-parametric bootstrap,

- Choose $B \in \mathbb{N}$
- For b in $1 \ldots B$:
  - Obtain the bootstrap sample $\mathbf{x}_b$ by samping n points from $\mathbf{x}$, **with replacement**
  - Compute $\hat{\theta}_b = \hat{\theta}(\mathbf{x}_b)$

This gives sequences of bootstrap samples $\{\mathbf{x}_b\}_{b=1}^{B}$ and estimates $\left\{\hat{\theta}_b\right\}_{b=1}^{B}$.

The Non-Parametric Bootstrap

Compute

$$\hat{SE}_{boot}(\hat{\theta}) = \left(\frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}_b - \hat{\theta}_. \right)^2 \right)^{1/2}$$

where

$$\hat{\theta}_. = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b$$

is the bootstrapped mean of $\hat{\theta}$.

The Non-Parametric Bootstrap

What is it doing? Again, our original dataset was obtained according to the following scheme:

$$F_\theta \rightarrow \mathbf{x} \rightarrow \hat{\theta}(\mathbf{x})$$

The bootstrap uses the empirical distribution function, $\hat{F}$, in place of F:

$$\hat{F}_\theta \rightarrow \{\mathbf{x}_b\}_{b=1}^{B} \rightarrow \left\{\hat{\theta}\right\}_{b=1}^{B}$$

The resulting collection of $\hat{\theta}_b$'s is referred to as a bootstrap sample from the sampling distribution of $\hat{\theta}$.

The Non-Parametric Bootstrap

We can use this to

- Make a histogram of $\hat{\theta}$, i.e. estimate the empirical distribution of $\hat{\theta}$
- Estimate a standard deviation of $\hat{\theta}$
- Get a confidence interval for $\hat{\theta}$.

Really, to do pretty much *anything* we could do for $\mathbf{x}$, but for $\hat{\theta}$.

Correlation

We still have a similar problem to what we saw with the jackknife: bootstrapped estimates of the statistic of interest are different, but not *that* different.

This is because bootstrap realizations of the sample statistic of interest are made with partially the same data, and so are correlated.

Consider the dataset $\{x_i\}_{i=1}^n$. What is the probability of any particular point x_i being included in any given bootstrap sample?

Correlation

$$P(x_i \text{ selected in any given bootstrap replication}) = 1/n$$
$$P(x_i \text{ not selected in any given bootstrap replication}) = (1 - 1/n)$$
$$P(x_i \text{ not selected in bootstrap sample}) = (1 - 1/n)^n \equiv p_n$$
$$\lim_{n\to\infty} p_n = e^{-1} \approx 0.368$$

So as the sample size gets large, there is approximately a

$$1 - \lim_{n\to\infty} p_n \approx 0.632$$

or nearly $2/3$ chance of each datapoint being included in the bootstrap sample.

Correlation

The expected number of points that are shared between two bootstrap samples, then, is not trivial. What is the significance of this?

Consider, for example, the sample means generated from two bootstrap samples, $\bar{x}_1$ and $\bar{x}_2$, where the underlying data is $X_i \sim N(\mu, \sigma^2)$.

Can we evaluate $Cov(\bar{x}_1, \bar{x}_2)$? Covariance will be induced by points which are shared between these two bootstrapped realizations of $\bar{X}$.

Without loss of generality, take $\mu = 0$ and $\sigma^2 = 1$.

Correlation

This is actually tougher than it sounds. There are two sources of variation: the original dataset $\mathbf{x}$, which is a realization of the joint distribution of $\mathbf{X}$, and the variability introduced by our resampling scheme.

Recall the conditonal expectation and variance identities from STA257 (chapter 4 of the textbook). For any random variables (X, Y),

$$E_X(X) = E_Y E_{X|Y}(X|Y)$$
$$Var_X(X) = Var_X(E_{X|Y}(X|Y)) + E_X(Var_{X|Y}(X|Y))$$

Correlation

Let $x_i^{(k)}$ be the i^{th} datapoint in the k^{th} bootstrap sample. Then

$$
\begin{aligned}
E(x_i^{(k)}) &= E_{\mathbf{X}} E(x_i^{(k)}|\mathbf{x}) = E_{\mathbf{X}}(\bar{X}) = 0 \\
Var(x_i^{(k)}) &= Var_{\mathbf{X}}(E(x_i^{(k)})|\mathbf{x}) + E_{\mathbf{x}}(Var(x_i^{(k)}|\mathbf{x})) \\
&= Var(\bar{X}) + E(\overline{X^2} - (\bar{X})^2) \\
&= \frac{1}{n} + 1 - \frac{1}{n} = 1
\end{aligned}
$$

Correlation

Because the sampling is done with replacement, **conditional on $\mathbf{x}$**, datapoints in the same bootstrap sample or across multiple bootstrap samples are independent. We can use this to evaluate the unconditional covariance of $x_i^{(k)}, x_j^{(m)}$, for $k = m$ or $k \neq m$ and $i \neq j$:

$$\begin{aligned} Cov(x_i^{(k)}, x_j^{(m)}) &= E(x_i^{(k)} x_j^{(m)}) \\ &= E\left(E(x_i^{(k)} x_j^{(m)} | \mathbf{x})\right) \\ &= E\left(E(x_i^{(k)} | \mathbf{x}) E(x_j^{(m)} | \mathbf{x})\right) \\ &= E(\bar{X}^2) \\ &= \frac{1}{n} \end{aligned}$$

Correlation

Hence,

$$Cov(\bar{x}_1, \bar{x}_2) = \frac{1}{n^2} \sum_{i,j} Cov(x_i^{(1)}, x_j^{(2)}) = \frac{1}{n}$$

and

$$\begin{aligned} Var(\bar{x}_k) &= \frac{1}{n^2} \left(\sum_i Var(x_i^{(k)}) + \sum_{i \neq j} Cov(x_i^{(k)}, x_j^{(k)}) \right) \\ &= \frac{1}{n^2} \left(n \times (1) + n(n-1) \times \left(\frac{1}{n} \right) \right) \\ &= \frac{2n-1}{n^2} \end{aligned}$$

Which gives

$$Cor(\bar{x}_1, \bar{x}_2) = \frac{Cov(\bar{x}_1, \bar{x}_2)}{\sqrt{Var(\bar{x}_1) Var(\bar{x}_2)}} = \frac{2n-1}{n} \approx 50\%$$

Correlation

This example illustrates that we might not expect the correlation between any two bootstrapped realizations of $\hat{\theta}$ to be small. This lowers the quality of the bootstrapped distribution of $\hat{\theta}$, from the perspective of making inferences about θ.

Nonetheless, the non-parametric bootstrap is an incredibly powerful tool, which lets you compute complicated statistics of data with complicated underlying distributions.

What if we were willing (as we have been in this course) to make assumptions about the parametric family of distributions that generated the original observed data $\mathbf{x}$? Could we do better?

Parametric Bootstrap

The **Parametric Bootstrap** is similar to the non-parametric bootstrap, except samples are drawn from the distribution that generated the data, with parameters replaced by their sample estimates.

Recap: the original dataset was generated according to

$$F_\theta \to \mathbf{x} \to \hat{\theta}(\mathbf{x})$$

The non-parametric bootstrap generated samples according to

$$\hat{F}_\theta \to \{\mathbf{x}_b\}_{b=1}^B \to \left\{\hat{\theta}\right\}_{b=1}^B$$

The parametric bootstrap generates samples according to

$$F_{\hat{\theta}} \to \{\mathbf{x}_b\}_{b=1}^B \to \left\{\hat{\theta}\right\}_{b=1}^B$$

The hat jumps off the F and lands on the θ.

Parametric Bootstrap

The algorithm is then:

- Choose $B \in \mathbb{N}$
- For b in $1 \dots B$:
  - Obtain the bootstrap sample $\mathbf{x}_b$ by samping n points from $F_{\hat{\theta}}$
  - Compute $\hat{\theta}_b = \hat{\theta}(\mathbf{x}_b)$

The Parametric Bootstrap for Hypothesis Tests

What about for testing a null hypothesis, $H_0 : \theta = \theta_0$?

We saw with both the parametric and non-parametric bootstraps that the distribution of $\hat{\theta}_b$ was centered on $\hat{\theta}$, the observed estimate from our original sample, not on the true value of θ that generated $\mathbf{x}$.

So using it for a pivot might not be the best idea.

How to proceed?

The Parametric Bootstrap for Hypothesis Tests

For problems where we know nothing about the sampling distribution of the quantity we wish to analyze, we sample from our original sample in order to estimate variability.

But when testing $H_0 : \theta = \theta_0$, what we are really testing is whether $X \sim F_{\theta_0}$ or not.

So instead of sampling from our original sample, which gave us samples from $\hat{F}_\theta$, in the **parametric bootstrap for hypothesis tests** we sample directly from F_{θ_0}:

$$F_{\theta_0} \rightarrow \{\mathbf{x}_b\}_{b=1}^{B} \rightarrow \left\{\hat{\theta}\right\}_{b=1}^{B}$$

The Parametric Bootstrap for Hypothesis Tests

The algorithm is then:

- Choose $B \in \mathbb{N}$
- For b in $1 \ldots B$:
  - Obtain the bootstrap sample $\mathbf{x}_b$ by samping n points from F_{θ_0}
  - Compute $\hat{\theta}_b = \hat{\theta}(\mathbf{x}_b)$

The Parametric Bootstrap for Hypothesis Tests

We then have a sample from the sampling distribution of $\hat{\theta}$, under $H_0 : \theta = \theta_0$.

Recall that we defined a p-value as being the probability of observing a test statistic giving at least as much evidence against H_0 as what we observed, if H_0 is true.

We can directly evaluate our parametric bootstrap p-value:

$$p_{boot} = \frac{1}{B} \sum_{b=1}^{B} I\left(\left| \hat{\theta}_b - \theta_0 \right| > \left| \hat{\theta} - \theta_0 \right| \right)$$

which is just the frequency with which $\hat{\theta}_b$ is farther away from θ_0 than the $\hat{\theta}$ from our sample.