The Bootstrap and Jackknife

Bootstrap & Jackknife
Motivation

In scientific research

• Interest often focuses upon the estimation of some unknown parameter, θ. The parameter θ can represent for example, mean weight of a certain strain of mice, heritability index, a genetic component of variation, a mutation rate, etc.

• Two key questions need to be addressed:

1. How do we estimate θ?
2. Given an estimator for θ, how do we estimate its precision/accuracy?

• We assume Question 1 can be reasonably well specified by the researcher

• Question 2, for our purposes, will be addressed via the estimation of the estimator's standard error

Bootstrap Motivation

Challenges

• Answering Question 2, even for relatively simple estimators (e.g., ratios and other non-linear functions of estimators) can be quite challenging

  • Solutions to most estimators are mathematically intractable or too complicated to develop (with or without advanced training in statistical inference)

• However

  • Great strides in computing, particularly in the last 25 years, have made computational intensive calculations feasible.

• We will investigate how the bootstrap allows us to obtain robust estimates of precision for our estimator, θ, with a simple example…

Bootstrap Estimation

Estimating the precision of the sample mean

• A dataset of *n* observation provides more than an estimate of the population mean (denoted here as $\overline{X}$), where

- $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

• It gives an estimate of the precision of $\overline{X}$, namely

- $$\hat{se}\left[\overline{X}\right] = \sqrt{\hat{\sigma}^2 / n}\,,$$

  where $$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}\,,$$

  is an estimate of the population variance.

• The problem with this standard error estimate is that it does not extend to estimators other than $\overline{X}$ in an obvious way.

Bootstrap Estimation

Estimating the precision of the sample mean

• From the formulas on the previous page, we can obtain an estimate of precision for $\overline{X}$ by estimating the population variance and "plugging" it into the formula for the standard error estimate.

• **Question**: What IF you did not know the formula for the standard error of the sample mean, BUT you had access to modern PC. How might you obtain an estimate of precision?

• **Answer**: The bootstrap!

Bootstrap Algorithm

Bootstrapping

• Assuming the sample accurately reflects the population from which it is drawn $\overline{X}$

• Generate a large number of "bootstrap" samples by resampling (with replacement) from the dataset

• Resample with the same structure (dependence, sample sizes) as used in the original sample

• Compute your estimator, θ , (here, $\theta = \overline{X}$), for each of the bootstrap samples

• Compute the "standard deviation" from the statistics calculated above.

Bootstrap Algorithm

Bootstrap sample | Bootstrap estimates

$$1: \quad (X_1^{(1)}, X_2^{(1)}, \ldots, X_n^{(1)}) \rightarrow \theta(X_1^{(1)}, X_2^{(1)}, \ldots, X_n^{(1)}) = \overline{X}^{(1)}$$

$$2: \quad (X_1^{(2)}, X_2^{(2)}, \ldots, X_n^{(2)}) \rightarrow \theta(X_1^{(2)}, X_2^{(2)}, \ldots, X_n^{(2)}) = \overline{X}^{(2)}$$

$$\vdots \qquad\qquad \vdots$$

$$B: \quad (X_1^{(B)}, X_2^{(B)}, \ldots, X_n^{(B)}) \rightarrow \theta(X_1^{(B)}, X_2^{(B)}, \ldots, X_n^{(B)}) = \overline{X}^{(B)}$$

Compute $\hat{\sigma}_b^2$, where $\hat{\sigma}_b^2 = \dfrac{\sum_{j=1}^{B} (\overline{X}^{(j)} - \overline{\overline{X}}^{(.)})^2}{B-1}$, and

$\overline{\overline{X}}^{(.)} = \dfrac{1}{B} \sum_{j=1}^{B} \overline{X}^{(j)}$.

The bootstrap standard error is $\hat{se}_{boot}\left[\overline{X}\right] = \sqrt{\hat{\sigma}_b^2}$.

For other estimators, simply replace $\overline{X}$ with the θ of your choice.

Bootstrap Estimation Examples

Estimating the precision of the sample mean

• Example: Generated a sample of size n=49 observations with the following summary statistics:

- $\overline{X} = \frac{1}{49}\sum_{i=1}^{n} X_i = 49.71$
- $\hat{se}\left[\overline{X}\right] = \sqrt{\hat{\sigma}^2 / n} = \sqrt{49.104/49} = 1.001$

• We generated B=100,000 bootstrap samples of size n=49 to obtain 100,000 bootstrap estimates of the sample mean, i.e., $\overline{X}^{(1)}, \overline{X}^{(2)}, \ldots, \overline{X}^{(100,000)}$.

• The bootstrap standard error was

- $\hat{se}_{boot}\left[\overline{X}\right] = \sqrt{\hat{\sigma}_b^2} = \sqrt{0.982} = 0.991$

• A reasonably close estimate to the "true" standard error estimate of 1.001

Bootstrap Estimation
Examples

Confidence Intervals on the Sample Median

• Approximate confidence intervals for the median can be obtained using asymptotic theory

  • The sample median is asymptotically normally distributed
  • The formula for the standard error is difficult to use

$$X_m \sim N\left(mdn(X), \frac{1}{4n(f(mdn(X)))^2}\right)$$

  where *f* is the density function of the true median.

•Approximate confidence intervals for the median can be obtained using asymptotic theory

• Bootstrapping would be easier/easiest.

Bootstrap Estimation
Examples

Bootstrapped estimates of the standard error for sample median

	Data	Median
Original sample:	{1, 5, 8, 3, 7}	5
Bootstrap 1:	{1, 7, 1, 3, 7}	3
Bootstrap 2:	{7, 3, 8, 8, 3}	7
Bootstrap 3:	{7, 3, 8, 8, 3}	7
Bootstrap 4:	{3, 5, 5, 1, 5}	5
Bootstrap 5:	{1, 1, 5, 1, 8}	1
etc.		
Bootstrap *B* (=1000)		

Bootstrap Estimation
Examples

Bootstrapped estimates of the standard error for sample median (cont.)

• Descriptive statistics for the sample medians from 1000 bootstrap samples

B	1000
Mean	4.964
Standard Deviation	**1.914**
Median	5
Minimum, Maximum	1, 8
25th, 75th percentile	3, 7

• We estimate the standard error for the sample median as 1.914

• A 95% asymptotic (with n=5?) confidence interval (using the 0.975 quantile of the standard normal distribution) is

$$5 +/- 1.96(1.914) = (1.25, 8.75)$$

Bootstrap Estimation
Examples

Confidence Intervals on the relative risk

• Approximate confidence intervals for the estimated relative risk, r = P[D|Exposed]/P[D|Not exposed] can also be obtained using asymptotic theory

  • The log[r] is asymptotically normally distributed with mean equal to the log of the true relative risk and variance

$$\text{var}\left[\log(r)\right] = \frac{1-P[D\,|\,E]}{n_1 P[D\,|\,E]} + \frac{1-P[D\,|\,\overline{E}]}{n_2 P[D\,|\,\overline{E}]}$$

  • 95% confidence intervals for the relative risk are therefore obtained by using the 0.975 quantile of the standard normal distribution (1.96) in the formula

$$\left(r \times \exp\left[-\sqrt{1.96\,\text{var}\left[\log(r)\right]}\right], r \times \exp\left[+\sqrt{1.96\,\text{var}\left[\log(r)\right]}\right]\right)$$

• We'll compare this approximation to the bootstrap in our example below

Bootstrap Estimation
Examples

Bootstrapped estimates of the standard error for sample relative risk

Cross-classification of Framingham Men by high systolic blood pressure and heart disease

	Heart Disease	
High Systol BP	No	Yes
No	915	48
Yes	322	44

The sample estimate of the relative risk is

$r = (44/366)/(48/963) = 2.412$

The asymptotic 95% confidence interval is

$(2.412*0.756, 2.412*1.322) = (1.82, 3.19)$.

Bootstrap Estimation
Examples

Bootstrapped estimates of the standard error for the relative risk (cont.)

• Descriptive statistics for the sample relative risks

B	100000
Bootstrap mean, *r*	2.464
Bootstrap Median	2.412
Standard Deviation	**0.507**

• The bootstrap standard error for the estimated relative risk is 0.507

• A 95% bootstrap confidence interval is

$$2.412 +/- 1.96(0.507) = (1.42, 3.41)$$

Bootstrap Summary

Advantages

- All purpose computer intensive method useful for statistical inference.
- Bootstrap estimates of precision do not require knowledge of the theoretical form of an estimator's standard error, no matter how complicated it is.

Disadvantages

- Typically not useful for correlated (dependent) data.
- Missing data, censoring, data with outliers are also problematic.

Jackknife Estimation

• The jackknife (or leave one out) method, invented by Quenouille (1949), is an alternative resampling method to the bootstrap.

• The method is based upon sequentially deleting one observation from the dataset, recomputing the estimator, here, $\theta_{(i)}$, n times. That is, there are exactly n jackknife estimates obtained in a sample of size n.

• Like the bootstrap, the jackknife method provides a relatively easy way to estimate the precision of an estimator, θ.

• The jackknife is generally less computationally intensive than the bootstrap

Jackknife Algorithm

Jackknifing

• For a dataset with n observations, compute n estimates by sequentally omitting each observation from the dataset and estimating θ on the remaining $n - 1$ observations.

• Using the n jackknife estimates, $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(n)}$,

we estimate the standard error of the estimator as

$$\hat{se}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\theta_{(i)} - \bar{\theta}_{(.)})^2}$$

• Unlike the bootstrap, the jackknife standard error estimate will not change for a given sample

Jackknife Summary

Advantages

- Useful method for estimating and compensating for bias in an estimator.
- Like the bootstrap, the methodology does not require knowledge of the theoretical form of an estimator's standard error.
- Is generally less computationally intensive compared to the bootstrap method.

Disadvantages

- The jackknife method is more conservative than the bootstrap method, that is, its estimated standard error tends to be slightly larger.
- Performs poorly when the the estimator is not sufficiently smooth, i.e., a non-smooth statistics for which the jackknife performs poorly is the median.