The Bootstrap

November 8, 2012

Motivation

- Suppose we take a sample of 1,000 people from a large population. We are interested in estimating, say, the average height of the people in the population.
- Suppose population heights have mean μ and standard deviation σ. We are interested in estimating μ.
- We know (by the CLT) that the sample average should be approximately normal with mean μ and variance $\sigma^2/1000$. We can use this fact to obtain standard errors and form confidence intervals or perform hypothesis tests. This is an easy inference problem.

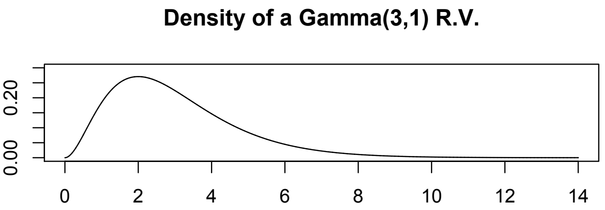
Motivation

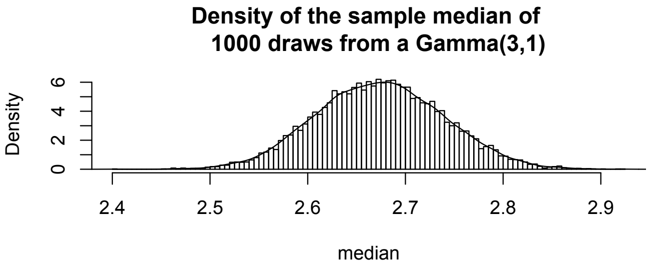
- Suppose we take a sample of 1,000 people from a large population. We are interested in estimating, say, the average height of the people in the population.
- Suppose population heights have mean μ and standard deviation σ. We are interested in estimating μ.
- We know (by the CLT) that the sample average should be approximately normal with mean μ and variance $\sigma^2/1000$. We can use this fact to obtain standard errors and form confidence intervals or perform hypothesis tests.
  This is an easy inference problem.
- What if we weren't interested in the population mean, but the population median?
- Depending on the population, there may not be a "nice formula" for the distribution sample median.
  This is a harder inference problem.

Ideal Scenario

- To estimate the distribution of the sample median, we could take samples of 1,000 people over and over and over and over again.
- For each sample of 1,000 people, find the sample median.
- Draw a histogram of these sample medians: should be close to the true distribution.

Gamma distribution:





Bootstrap Idea:

- In practice, we only have one sample. Impossible to sample many times to obtain a sampling distribution.

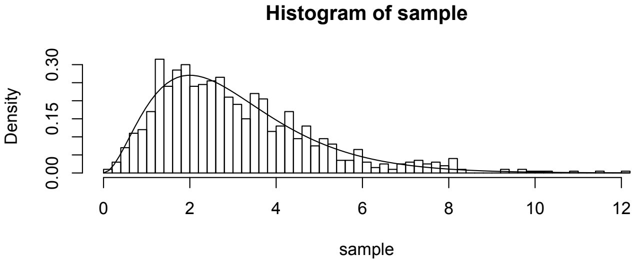
Bootstrap Idea:

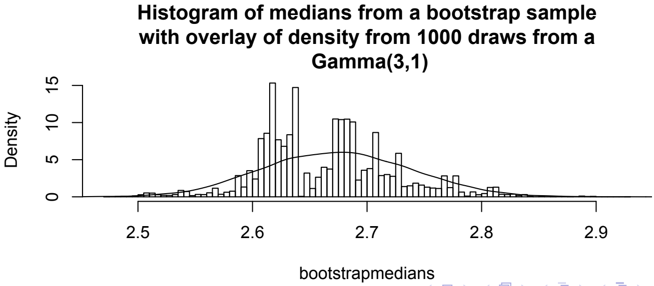
- In practice, we only have one sample. Impossible to sample many times to obtain a sampling distribution.
- HOWEVER, if we sampled well, the data from our sample should be close in distribution to the data from the population. (Key idea: Empirical distribution obtained by the sample converges to the true distribution)
- Resampling (with replacement) from our sample many, many times is ALMOST like resampling from the entire population.

Bootstrap Idea:

- In practice, we only have one sample. Impossible to sample many times to obtain a sampling distribution.
- HOWEVER, if we sampled well, the data from our sample should be close in distribution to the data from the population. (Key idea: Empirical distribution obtained by the sample converges to the true distribution)
- Resampling (with replacement) from our sample many, many times is ALMOST like resampling from the entire population.
- For many statistics, we can get close to the sampling distribution this way.

Bootstrap for Gamma:



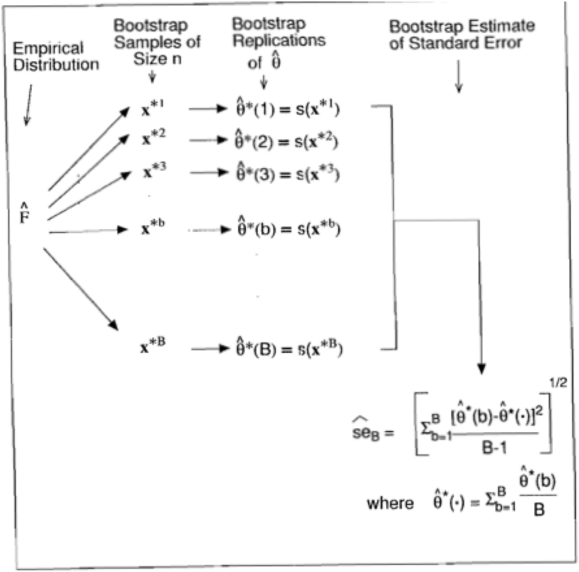


Estimation of standard errors:

- Let **x** denote the original sample of n units. Let $\hat{\beta}$ denote the median (or any other parameter of interest) of the sample
- Select (Large) B independent bootstrap samples $\mathbf{x}^{*\mathbf{1}}, \mathbf{x}^{*\mathbf{2}}, \ldots, \mathbf{x}^{*\mathbf{B}}$, each consisting of n data values draw **with replacement** from **x**.
- Compute the median for each sample. Let $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ denote these medians. Let $\bar{\theta}^* = \frac{1}{B}\sum \hat{\theta}_i^*$ denote the average of these medians.
- Estimate the standard error for the sample median by taking the standard deviation of the B bootstrap medians.

$$\widehat{\mathrm{se}}_B = \left\{ \sum_{i=1}^{B} [\hat{\theta}_i^* - \bar{\theta}^*]^2 / (B-1) \right\}^{1/2}.$$

The Bootstrap Algorithm for SE



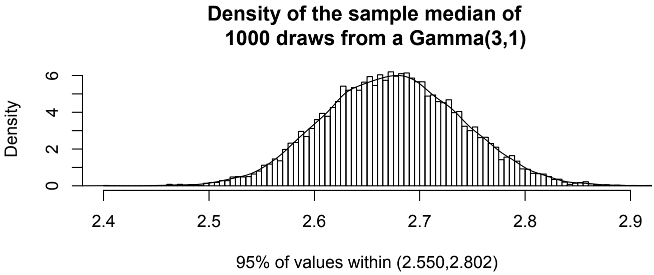
Bootstrap confidence intervals:

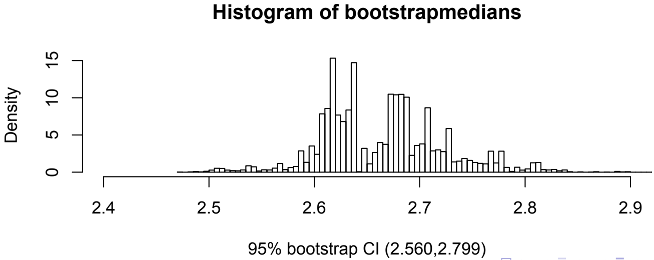
- Bootstrap confidence intervals are easy too!
- Suppose we want to find a $1 - \alpha$ confidence interval.
- We can form a bootstrap confidence interval by finding the $\alpha/2$ and the $(1 - \alpha/2)$ percentile of the bootstrap medians $(\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*)$.

Bootstrap confidence intervals:

- Bootstrap confidence intervals are easy too!
- Suppose we want to find a $1-\alpha$ confidence interval.
- We can form a bootstrap confidence interval by finding the $\alpha/2$ and the $(1-\alpha/2)$ percentile of the bootstrap medians $(\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B)$.
- For example, if we took 10,000 bootstrap samples, the if we denote $\hat{\theta}^*_{(i)}$ as the i^{th} largest bootstrap median, a 95% bootstrap confidence interval would be $[\hat{\theta}^*_{(251)}, \hat{\theta}^*_{(9750)}]$.

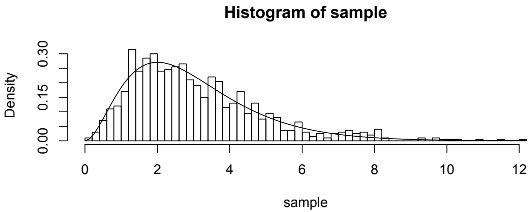
Bootstrap CI:





Warnings:

- Note: bootstrap estimation is only as good as the data you begin with.
    - Median of the distribution: 2.674
    - Median of sample: 2.668



- If sample does not look like original distribution, then bootstrapping may fail (think Type I Errors). There's no good way to check this unless you make an assumption about the distribution of the population.

Other applications of Bootstrap:

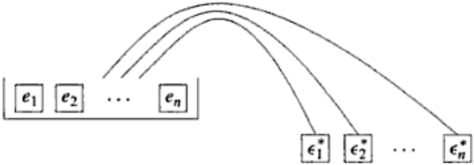
- Many possible applications for bootstrap, not just finding sampling distributions.
- Find estimates, standard errors, and bias in complicated models fitted to data. (See *Statistical Models* by David Freedman for some examples)
- Can also be used for testing.

Other applications of Bootstrap:

- Many possible applications for bootstrap, not just finding sampling distributions.
- Find estimates, standard errors, and bias in complicated models fitted to data. (See *Statistical Models* by David Freedman for some examples)
- Can also be used for testing.
- Key idea: mechanism for resampling has to preserve original structure of data.
- For example: If a set of data points is assumed to be i.i.d., we can mimic their distribution by resampling from the data points with replacement.

Example: Regression Models

- We know the formulas for finding standard errors in in regression, but suppose we forgot.
- Suppose we assume the model $Y_i = X_i\beta + \epsilon_i$, where the design matrix X is fixed and has full rank and the errors $\epsilon_1, \ldots, \epsilon_n$ are IID with mean 0 and variance σ^2.
- Now, Y_i's are not i.i.d., but the ϵ_i are. If the Y_i are linear in X, the residuals $e_i = Y_i - X_i\hat{\beta}$ should be close to the actual errors ϵ_i.
- By resampling from the residuals we preserve the randomness structure.



Example: Regression Models

- Draw n times at random with replacement from this population to get bootstrap errors $\epsilon_1^*, \ldots \epsilon_n^*$. These are i.i.d. (because you sample them that way).
- Generate the Y_i^*:

$$Y_i^* = X_i\hat{\beta} + \epsilon_i^*$$

- Given Y^* and X, can then get the regression estimate $\hat{\beta}^* = (X'X)^{-1}X'Y^*$.

Example: Regression Models

- Draw n times at random with replacement from this population to get bootstrap errors $\epsilon_1^*, \ldots \epsilon_n^*$. These are i.i.d. (because you sample them that way).
- Generate the Y_i^*:

$$Y_i^* = X_i\hat{\beta} + \epsilon_i^*$$

- Given Y^* and X, can then get the regression estimate $\hat{\beta}^* = (X'X)^{-1}X'Y^*$.
- Do this over and over to get many, many $\hat{\beta}^*$.
- Distribution of $\hat{\beta}^* - \hat{\beta}$ is a good approximation for the distribution of $\hat{\beta} - \beta$.
- The empirical covariance matrix of the $\hat{\beta}^*$ (computed by actually taking variances and correlations of $\hat{\beta}^*$ terms) should be close to the thoretical covariance matrix of $\hat{\beta}$.

Example: Kolmogorov-Smirnov

- Here is the procedure for computing bootstrap p-values for the KS test in the Matching package.
  (Very similar to permutation tests.)
- Suppose we a treatment group of m units and a control group of n. Members of each group are selected i.i.d, and units in the treatment group are selected independently from the control group.
- Let $\widehat{KS}$ denote the value of the KS statistic for these groups.

Example: Kolmogorov-Smirnov

- Here is the procedure for computing bootstrap p-values for the KS test in the Matching package.
  (Very similar to permutation tests.)
- Suppose we a treatment group of m units and a control group of n. Members of each group are selected i.i.d, and units in the treatment group are selected independently from the control group.
- Let $\widehat{KS}$ denote the value of the KS statistic for these groups.
- Under the null hypothesis of a KS test: both groups have the same distribution.
- Let $y_1, \ldots, y_m$ denote the observations from the treated group and let $y_{m+1}, \ldots, y_{m+n}$ denote the observations from the control group.
- Under null, the distribution of $(y_1, \ldots, y_m)$ is the same as of $(y_{m+1}, \ldots, y_{m+n})$ is the same as $(y_1, \ldots, y_{m+n})$.

Example: Kolmogorov-Smirnov

To get distribution of KS test statistic under the null hypothesis:

1. Draw $m + n$ observations with replacement from $(y_1, \ldots, y_{m+n})$.
2. Assign first m observations to "treatment," assign next n to "control."
3. Compute the KS statistic $\widehat{KS}^*$ for this assignment of treatment and control.
4. Do this many, many, many times to get a distribution of the KS statistic under the null hypothesis.

The KS bootstrap p-value is the proportion of bootstrap trials with a KS statistic $\widehat{KS}^* \geq \widehat{KS}$.