The bootstrap

Patrick Breheny

December 6

The empirical distribution function

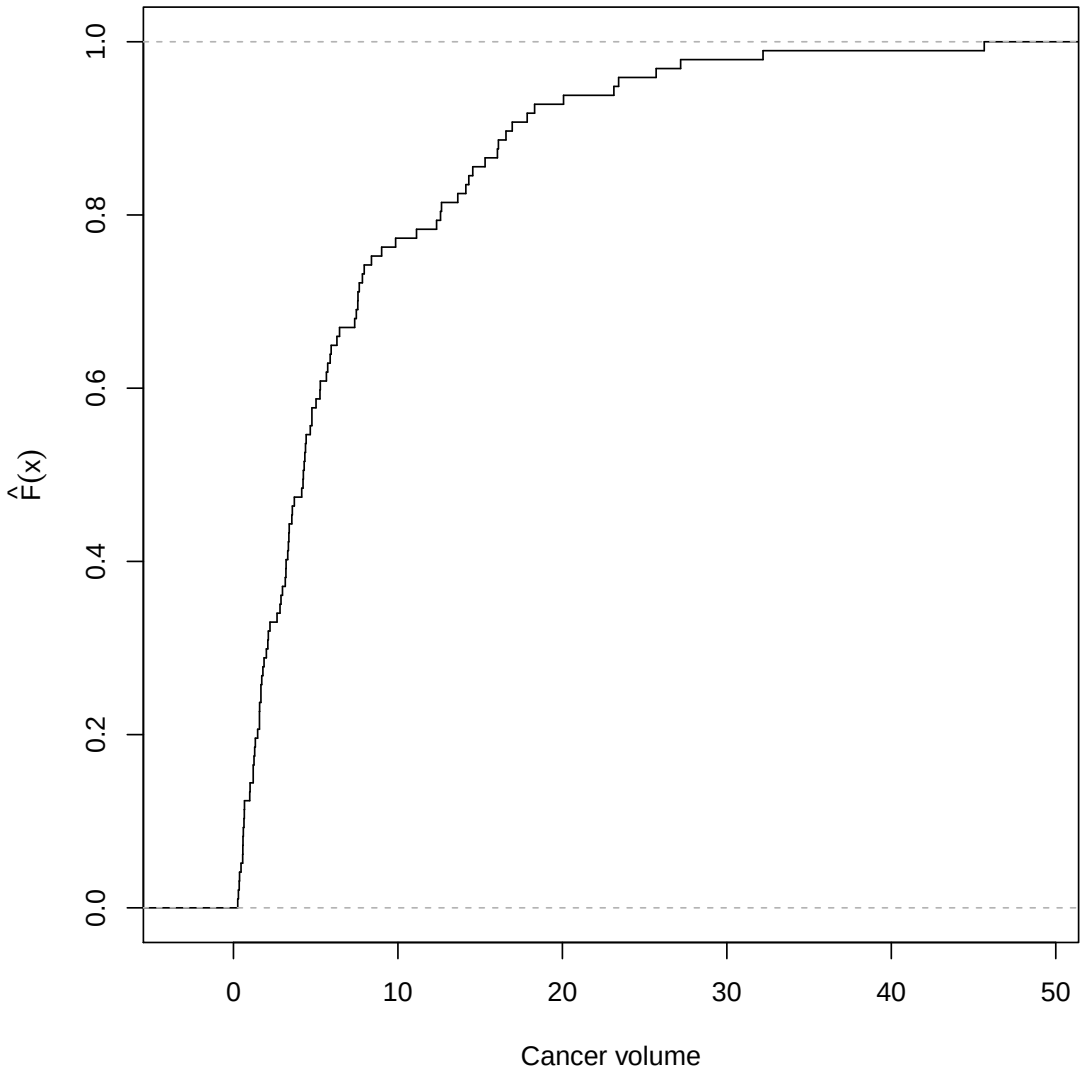
- Suppose $X \sim F$, where $F(x) = \Pr(X \leq x)$ is a distribution function, and we wish to estimate some aspect of F (for example, $\mathrm{E}\{X\}$ or $\Pr\{X > 1\}$)
- The parametric approach is to assume that F has some specific form, then estimate its parameters
- The nonparametric alternative is the *empirical distribution function*, $\hat{F}$, the CDF that puts mass $1/n$ at each data point x_i:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

where I is the indicator function

Empirical CDF: Cancer volume

From prostate study:



Properties of $\hat{F}$

- At any fixed value of x,

$$\mathrm{E}\{\hat{F}(x)\} = F(x)$$
$$\mathrm{Var}\{\hat{F}(x)\} = \frac{1}{n}F(x)(1 - F(x))$$

- Note that these two facts imply that

$$\hat{F}(x) \xrightarrow{\mathsf{P}} F(x)$$

- An even stronger proof of convergence is given by the *Glivenko-Cantelli Theorem*:

$$\sup_x \left|\hat{F}(x) - F(x)\right| \xrightarrow{\text{a.s.}} 0$$

Introduction

- The empirical CDF is interesting in its own right, but also as the fundamental component of a statistical approach called the *bootstrap*
- The bootstrap is an extremely important idea in modern nonparametric statistics; indeed, Casella & Berger call it "perhaps the single most important development in statistical methodology in recent times"

Derivation of bootstrap

- Suppose we are interested in assessing the variance of an estimate $\hat{\theta} = \theta(\mathbf{x})$
- Its actual variance is given by

$$\mathrm{Var}(\hat{\theta}) = \int \cdots \int \{\theta(x_1, \ldots, x_n) - \mathrm{E}(\hat{\theta})\}^2 dF(x_1) \cdots dF(x_n)$$

  where $\mathrm{E}(\hat{\theta}) = \int \cdots \int \theta(x_1, \ldots, x_n) dF(x_1) \cdots dF(x_n)$
- There are two problems with evaluating this expression directly

The ideal bootstrap

- The first is that we do not know F
- The natural solution is to plug in the empirical cdf, $\hat{F}$:

$$\hat{\mathrm{Var}}(\hat{\theta}) = \int \cdots \int \{\theta(x_1, \ldots, x_n) - \hat{\mathrm{E}}(\hat{\theta})\}^2 d\hat{F}(x_1) \cdots d\hat{F}(x_n)$$

- For reasons that will become clear, we will call this the *ideal bootstrap* estimate

The ideal bootstrap (cont'd)

- The second problem, however, is that this integral is difficult to evaluate
- Because $\hat{F}$ is discrete,

$$\hat{\mathrm{Var}}(\hat{\theta}) = \sum_j \frac{1}{n^n} \{\theta(\mathbf{x}_j) - \hat{\mathrm{E}}(\hat{\theta})\}^2$$

  where $\mathbf{x}_j$ ranges over all n^n possible combinations of the observed data points $\{x_i\}$ (not all of which are distinct)
- Unless n is small, this may take a long time to evaluate

Monte Carlo approach

- However, we can approximate this answer instead using Monte Carlo integration – the technique we have been using all semester long in our simulations
- Instead of actually evaluating the integral, we approximate it numerically by drawing random samples of size n from $\hat{F}$ and finding the sample average of the integrand
- This approach gives us the bootstrap – an approximation to the ideal bootstrap
- By the law of large numbers, this approximation will converge to the ideal bootstrap as the number of random samples that we draw goes to infinity

Bootstrap estimate of variance

The procedure for finding the bootstrap estimate of the variance (or "bootstrapping the variance") is as follows:

(1) Draw $\mathbf{x}_1^*, \ldots, \mathbf{x}_B^*$ from $\hat{F}$, where each *bootstrap sample* $\mathbf{x}_b^*$ is a random sample of n data points drawn iid from $\hat{F}$

(2) Calculate $\hat{\theta}_b^*$, where $\hat{\theta}_b^* = \theta(\mathbf{x}_b)$; these are called the *bootstrap replications*

(3) Let

$$v_{boot} = \frac{1}{B} \sum_{b=1}^{B} \left\{ \hat{\theta}_b^* - \bar{\theta}^* \right\}^2,$$

where $\bar{\theta}^* = B^{-1} \sum_b \hat{\theta}_b^*$

Resampling

- What does a random sample drawn from $\hat{F}$ look like?
- Because $\hat{F}$ places equal mass at every observed value x_i, drawing a random sample from $\hat{F}$ is equivalent to drawing n values, with replacement, from $\{x_i\}$
- In practice, this is how the $\mathbf{x}_b^*$'s from step 1 on the previous page are generated
- This somewhat curious phenomenon in which we draw new samples by sampling our original sample is called *resampling*

Bootstrapping statistical models

- The same approach can be applied to multivariate data and statistical models, provided that the observations are still independent and identically distributed
- In this case, we resample elements of $\{\mathbf{x}_i, y_i\}_{i=1}^n$ (*i.e.*, rows of our data set) and apply the model to $\{\mathbf{X}_b^*, \mathbf{y}_b^*\}$
- It should be noted, however, that this approach treats $\mathbf{X}$ as random
- In certain cases where $\mathbf{X}$ is fixed, this approach does not make sense; it is still possible to use the bootstrap for these situations, but this involves bootstrapping the residuals of the model and is somewhat more complicated

Bootstrap estimation of the CDF of $\hat{\theta}$

- The bootstrap is not limited to the variance – we can use it to estimate any aspect of the sampling distribution of $\hat{\theta}$
- In particular, we can calculate the 2.5th and 97.5th percentiles of $\hat{\theta}^*$, thereby obtaining a 95% confidence interval for $\hat{\theta}$ that, unlike $\hat{\theta} \pm 1.96\,\mathrm{SE}$, can be asymmetric and thereby reflect the skewness of an estimate
- This confidence interval is called the bootstrap *percentile interval*
- The percentile approach is not the only way of creating bootstrap confidence intervals; actually, a considerable body of research has been devoted to this problem, although it is very simple, has nice properties, and is widely used.

Implementation in R and SAS

- Although it is not difficult to write your own for loop to conduct bootstrapping (easier in R than SAS), both platforms provide tools to do most of the work for you
- In both platforms, bootstrapping is fairly simple, although you must write your own function to analyze the data and return $\hat{\theta}^*$

R

- Bootstrapping in R can be accomplished via the boot package (by default, installed but not loaded)
- The function you write to calculate $\hat{\theta}^*$ must be a function of two arguments: the first is the original data and the second is a vector of indices specific to the bootstrap sample
- Thus, in order to use the bootstrap to, say, estimate the standard error of the variance, you will need to define a function like the following:

```
var.boot <- function(x,ind){var(x[ind])}
```

boot example

- Once you have defined such a function, its usage is straightforward:

```
> boot(cavol,var.boot,1000)
...
Bootstrap Statistics :
    original    bias    std. error
t1* 62.17922 -0.951639    17.24447
```

- We can obtain bootstrap confidence intervals via:

```
> out <- boot(cavol,var.boot,1000)
> boot.ci(out,type="perc")
...
Level     Percentile
95%   (32.55, 98.88 )
```

How big should B be?

- What is a good value for B?
- On the one hand, computing time increases linearly with B, so we would like to get by with a small B
- This desire is particularly acute if θ is complicated and time-consuming to calculate
- On the other hand, the lower the value of B, the less accurate and more variable our estimated standard error is

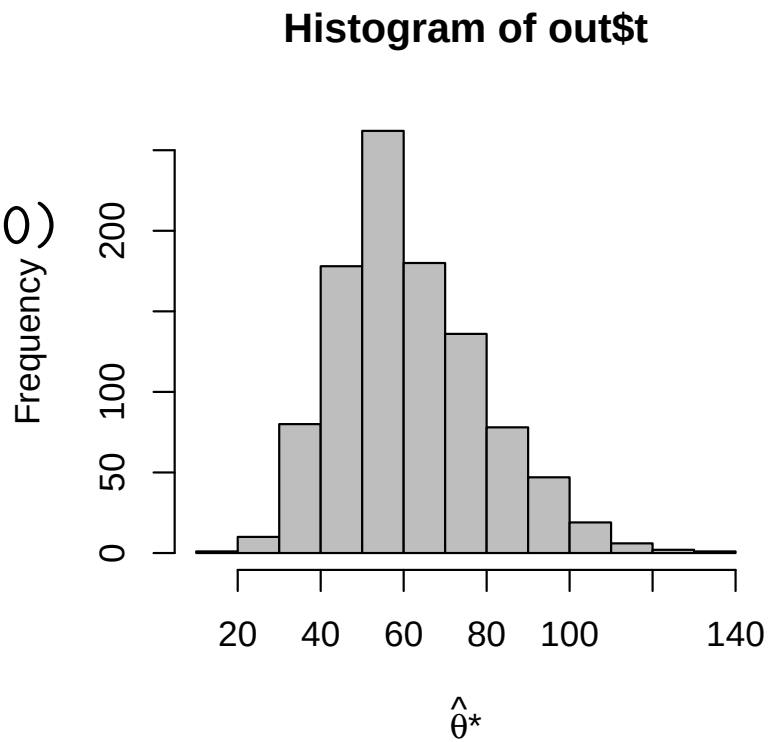
How big should B be? (cont'd)

- How much accuracy do we lose by stopping at B bootstrap samples instead of going to ∞?
- This can be assessed by standard statistical methods: $\{\hat{\theta}_b^*\}$ are iid and our SE estimate is a standard deviation
- Generally speaking, published articles in recent years tend to use 1000 bootstrap replications; however, for highly computer intensive statistics, 100 or even 50 may be acceptable
- However, each application is different – bootstrap data, just like real data, often deserves a closer look: in the words of Brad Efron, "it is almost never a waste of time to display a histogram of the bootstrap replications"

Histogram of bootstrap replications

```
out <- boot(cavol,var.boot,1000)
hist(out$t)
```

95% CI for bootstrap SE:
(16.8,18.4)



SAS implementation

- In SAS, your function that calculates $\hat{\theta}^*$ must be a macro called %analyze which takes two arguments: data and out
- So, in order to use the bootstrap to study the sampling distribution of the variance, we could write:

```
%macro analyze(data,out);
  proc means noprint data=&data;
    output out=&out(drop=_freq_ _type_) var=var_cavol;
    var cavol;
  run;
%mend;
```

SAS implementation (cont'd)

- To use the %boot and %bootci macros, you also need to download the macro definitions (available on support.sas.com or the course website) and source them:

```
%inc "jackboot.sas.txt";
```

- You can then obtain bootstrap summaries and bootstrap confidence intervals via:

```
%BOOT(data=prostate);
%BOOTCI(percentile);
```