Bootstrap

Outline

- Introduction to Bootstrap
- Bootstrap Procedures
- Bootstrap Confidence Intervals
- Bootstrap to Estimate Bias
- Bias of Variance Estimator

Study

Conisder the following controlled, randomized, double-blind study showing that the risk of heart attack could be reduced by taking aspirin.

- half the subjects are randomly assigned to take aspirin half a placebo
- subjects and physicians were blinded to the assignments
- tablets given every other day

The data

Group	Heart Attacks	Subjects
Aspirin	104	11037
Placebo	189	11034

The Estimated Rates

Given that a person was an aspirin taker, the fraction of people that had a heart attack was

104/11037.

Given that a person was a placebo taker, the fraction of people that had a heart attack was

189/11034.

The Estimation

The Ratio of the Two Rates: $\hat{\theta} = \frac{104/11037}{189/11034} = .55026$

Aspirin takers only have 55% as many heart attacks as placebo-takers.

The problems

We are not interested in $\hat{\theta}$ the estimated ratio in the sample. We are interested in θ the true ratio, that is the ratio we would see if we could treat all subjects, and not just a sample of them.

The tough question is how do we know that $\hat{\theta}$ might not come out much less favorably if the experiment were run again?

Revisit Aspirin example

This is where statistical inference comes in. Statistical theory allows us to make the following inference: the true value of θ lies in the interval a $< \theta <$ b with 95% confidence.

In the aspirin study, it also track strokes. The results are presented as the following:

Group	Strokes	Subjects
Aspirin	119	11037
Placebo	98	11034

What's the conclusion?

For strokes, the ratio of the two rates is

$$\hat{\theta} = \frac{119/11037}{98/11034} = 1.21.$$

It now looks like taking aspirin is actually harmful. Our question is how do we know that $\hat{\theta}$ might come out beneficial if the experiment were run again.

The task of data analyst is to find the sampling distribution of the chosen estimator $\hat{\theta}$.

If we can construct the interval for the true stroke ratio θ with 95% confidence, and if this includes the neutral value $\theta = 1$, at which aspirin would be no better or worse than placebo.

Repetition

Suppose we were to repeat the study B times. We would estimate $\hat{\theta}_1, \ldots, \hat{\theta}_B$

Each estimate would certainly not be 55%. The cause of the variation is sampling error. How much would they vary?

Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed with cdf F.

Let $T = T(X_1, X_2, \ldots, X_n)$ be a statistic. We often are faced with the problem of determining some aspect of the sampling distribution of T. However, in many cases there may be no easy way to derive it.

The **bootstrap** is a technique proposed by Bradley Efron in 1979 that cleverly resamples from the dataset to learn about this sampling distribution. The basic idea of the bootsrapping method is that, in absence of any other information about the distribution, the observed sample contains all the available information about the underlying distribution, and hence resampling is the best guide to what can be expected from resampling from the distribution.

Let F_T denote the cdf of T. Because T is a function of the data, F_T is determined by F.

Thus, if we could repeatedly draw samples of size n from F forever, we could obtain independent draws of T and estimate F_T to any desired accuracy.

However, we only have the one dataset. Efron's idea was to repeatedly draw datasets from the empirical cdf $\hat{F}$.

Recall the definition of the empirical cdf $\hat{F}$.

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

The bootstrap procedure for approximating the cdf F_T is as follows:

1.

For $b = 1, 2, \ldots, B$

- Generate the bth bootstrap sample $x^*_{1b}, x^*_{2b}, \ldots, x^*_{nb}$, from the empirical cdf $\hat{F}$.
- Compute the bth bootstrap replicate $T^*_b = T(x^*_{1b}, x^*_{2b}, \ldots, x^*_{nb})$

2. The bootstrap estimate of F_T is the empirical cdf of the bootstrap replicates

$$\hat{F}_T(t) = \frac{1}{B}\sum_{b=1}^{B} I[T_b^* \leq t]$$

From $\hat{F}_T$ we can estimate any parameter of the sampling distribution of T that we wish by finding the corresponding parameter determined by $\hat{F}_T$.

Sampling from $\hat{F}$ is as easy as obtaining samples of size n by drawing with replacement from the original dataset.

We sometimes wish to compute a parameter of F_T such as the standard deviation, and we can simply do this by computing the parameter with the bootstrap replicates without actually constructing $\hat{F}_T$.

The theory for the bootstrap relies on the fact F determines F_T and $\hat{F}$ determines $\hat{F}_T$.

So as n becomes large and $\hat{F}$ converges to F, then $\hat{F}_T$ will converge to F_T if B is large.

Example (Sampling distribution of sample median)

Let T be the sample median. We'll take the definition that T is the smallest real number with the property that at least half of the data are less than or equal to it.

Suppose that X has density $f(x) = e^{-x}$ for $x > 0$ with corresponding cdf $F(x) = 1 - e^{-x}$, and suppose the sample size is $n = 25$. Let's compute $\hat{F}_T$.

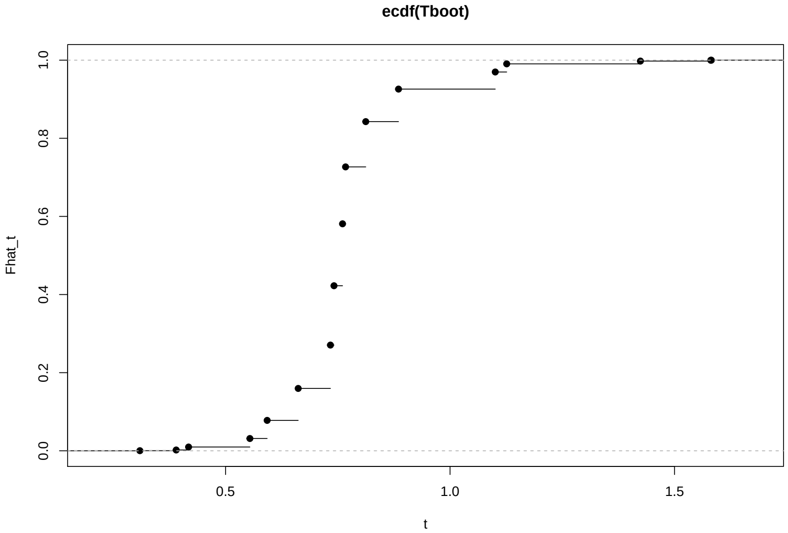
```r
n=25
x=rexp(n)
T=median(x)

B=10000
Tboot=numeric(B)

for(b in 1:B){
 xb=sample(x,25,replace=TRUE)
 Tboot[b]=median(xb)
}
```

Plot the empirical CDF of bootstrap replicates

```
plot(ecdf(Tboot),xlab="t", ylab="Fhat_t")
```



A common problem is to find the standard deviation of the sampling distribution of a statistic T.

When T is an estimator $\hat{\theta}$ of a parameter θ, such as the sample median estimator of the population median, this standard deviation is known as the standard error.

When F isn't known, the standard error of the sample median is esentially impossible to write down analytically.

It has been shown that the variance of the sample median is asymptotically equivalent to $\frac{1}{4nf^2(m)}$ where f is the density function of X and m is the population median. However, this requires knowing f.

We can obtain a bootstrap estimate of the standard error by simply computing the standard deviation of our bootstrap replicates,

$$\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(T_b^* - \bar{T}^*)^2}$$

In our particular sample we have

```r
sd(Tboot)
```

```
## [1] 0.1349032
```

Bootstrap Confidence Intervals

Standard Normal Bootstrap Interval

Finding the standard error in this way can be very helpful in constructing confidence intervals.

Consider a statistic $\hat{\theta}$ that is an estimator of a parameter θ determined by F. In many cases large sample theory tells us

$$\frac{\hat{\theta} - \theta}{se(\hat{\theta})} \xrightarrow{D} Z$$

where $Z \sim N(0, 1)$.

For example, that would typically be the case if $\hat{\theta}$ is a maximum likelihood estimator, or a weighted or unweighted average of independent random variables.

When $\hat{\theta}$ is *asymptotically normal* in this way, we can easily derive an asymptotic $(1-\alpha)100\%$ confidence interval for θ by

$$(\hat{\theta} - z_{\alpha/2} se(\hat{\theta}), \hat{\theta} + z_{\alpha/2} se(\hat{\theta}))$$

and use the bootstrap to find $se(\hat{\theta})$ as discussed above.

Percentile Bootstrap Confidence Intervals

An alternative method for using bootstrap to construct confidence intervals that does not rely on the asymptotic normality of $\hat{\theta}$ is to simply refer to the empirical distribution of the bootstrap replicates.

Suppose that B bootstrap samples are drawn, each time computing $\hat{\theta}^*_b$ for $b = 1, 2, ..., B$.

Order the values,

$$\hat{\theta}^*_{(1)} \leq \hat{\theta}^*_{(2)} \leq \cdots \leq \hat{\theta}^*_{(B)}.$$

Then the percentile bootstrap confidence interval is just

$$(\hat{\theta}^*_{(B(\alpha/2)}, \hat{\theta}^*_{(B(1-\alpha/2))})$$

In other words, it's just the middle $(1-\alpha)100\%$ of the bootstrap replicates.

For example, if we had $B = 10000$ and wished to compute a 95 percent confidence interval for θ, we'd let the left endpoint be the 250th value and the right endpoint be the 9750th value in the ordered $\hat{\theta}^*$s.

The sample median is an example of an asymptotically normal estimator of the population median. Let's try an example drawing from a chi-squared distribution with 1 degree of freedom. We'll look at both the normal interval and the percentile interval.

```r
## First let's see true median
qchisq(.5,df=1)
```

```
## [1] 0.4549364
```

```r
## try sample size=100 and 10000 replications
n=100
B=10000
x=rchisq(n,df=1)
thetahat=median(x)
thetahatboot=numeric(B)

for(b in 1:B){
  xb=sample(x,n,replace=TRUE)
  thetahatboot[b]=median(xb)
  }

sethetahat=sd(thetahatboot)
```

```
#Normal interval
zval=qnorm(.975,0,1)
c(thetahat-zval*sethetahat,thetahat+zval*sethetahat)
```

```
## [1] 0.2230095 0.5309382
```

```
## the percentile inteval
thetahatboot=sort(thetahatboot)
thetahatboot[c(250,9750)]
```

```
## [1] 0.2541737 0.5841492
```