Bootstrapping

(Givens/Hoeting Cht. 9 & Gentle Ch 4

<u>Motivation</u>: The motivation for the bootstrap method is the same as that for the jackknife. We have an estimator T for some parameter θ of a distribution F and we would like to know characteristics (bias, variance, conf. intv) of the distribution of the estimator T.

Ideally we would have many samples from F allowing us to generate multiple estimates T_i & use these to obtain this information. However this is unrealistic and so again we turn to clever ways of utilizing a single sample.

For the jackknife the basic idea was to remove portions of the sample & then recalc the est, eventually looking at the mean of the weighted differences of the samples. For the bootstrap the main idea is that of resampling.

In general, Resampling methods involve the use of many samples each drawn from a single sample from F. Then the conditional distribution of the new sample is used for statistical inference. Since the sample set is finite, it is often easy to compute statistical functions on the sample, and thus gain info even when very little is known about F.

For the bootstrap, the basic idea is that an observ. sample should contain all of the information about the underlying population & so the observed

Sample is considered to be the population. Hence, the distribution of the test statistic T can be simulated by using random samples from the "population" (ie observed & original sample).

<u>More Formally</u>. (using the text's notation).

Sec 9.1

- Let $\theta = T(F)$ be our parameter of interest of a distribution F, expressed as a functional of F. Usually $T(F) = \int g(z)\, dF(z)$.
  For ex. $T(F) = \int z\, dF(z)$ is the mean of the dist.

$\sim F$ means $\sim f$ where F is CDF of f.

- Let $x_1, \ldots, x_n$ be data observed as a real: of the r.v. $X_1, \ldots, X_n \sim$ i.i.d. F. (F is c.d.f.) and $\mathcal{X} = \{x_1, \ldots, x_n\}$ denote the entire data set.

discrete uniform.

- Let $\hat{F}$ be the empirical dist of the observed data. (Basical weights each obs. x_i w prob $1/n$).

(as usual)

- An est. of θ, $\hat{\theta} = T(\hat{F})$.
  Ex if θ is pop. mean $\hat{\theta} = \int z\, d\hat{F}(z) = \sum_{i=1}^{n} x_i/n$.

Now, we want to know about the dist of our estimator $T(\hat{F})$. And so we ask ques. about $T(\hat{F})$ or some $R(\mathcal{X}, F)$, a statistical function of the data & their unknown dist. F.

Ex/

$$R(\mathcal{X}, F) = \frac{T(\hat{F}) - T(F)}{S(\hat{F})}$$ where $S(\hat{F})$ est. s.d. of $T(\hat{F})$

$R(\mathcal{X}, F)$ could be the bias of $T(\hat{F})$, Var, etc.. It is the function of interest. (More general than $T(\hat{F})$ and allows for easy know. of data set dependency).

As we mentioned before, the dist of $R(\mathcal{X}, F)$ is prob. unknown & may be intractable. So we use the emp. dist of the obs. data (from F) to approx the dist of $R(\mathcal{X}, F)$.

- The bootstrap Method:

sampled w/replacement.

Let $\mathcal{X}^* = \{X_1^*, \dots, X_n^*\}$ be i.i.d r.v. drawn from dist. $\hat{F}$. Then $\mathcal{X}^*$ is called the bootstrap sample of pseudodata or pseudodataset

The bootstrap stategy is to examine the dist. of $R(\mathcal{X}^*, \hat{F})$ and use it to make inf. about $R(\mathcal{X}, F)$. (Easy to resample $\mathcal{X}^*$ from $\mathcal{X}$).

Comment: in some special cases we can derive analytical results about the dist. of $R(\mathcal{X}^*, \hat{F})$ however, usually it is done through sim.

9.1 pg 254 Ex/ Suppose $\mathcal{X} = \{x_1, x_2, x_3\} = \{1, 2, 6\}$ is obs from F. and we wish to est. the mean of F, θ.

Clearly, $\hat{\theta} = T(\hat{F})$ or $R(\mathcal{X}, F)$ is the sample mean $\hat{\theta} = 9/3$. we would like to have more info about the dist. of $\hat{\theta}$.

The empirical dist of F is $\hat{F}$ which places mass $1/3$ at each observed value.

A bootstrap sample $\mathcal{X}^* = \{X_1^*, X_2^*, X_3^*\}$ will be elements drawn i.i.d. from $\hat{F}$. There are $3^3 = 27$ possibilities for $\mathcal{X}^*$, each with prob $1/27$.

Ex ($\mathcal{X}^* = \{1, 1, 6\}$)

- Let $\hat{F}^*$ denote the empirical dist func. of $\mathcal{X}^*$
ex ($\hat{F}^*$ puts 2/3 at 1 & 1/3 at 6).

- and we have the new corresp. bootstrap est
$\hat{\theta}^* = T(\hat{F}^*)$ ex ($\hat{\theta}^* = 2/3 + 6(1/3) = 8/3$).

The bootstrap strategy is to examine (analy. if poss. o.w. via simulation the dist of $\hat{\theta}^*$).

For this case $\hat{\theta}^*$ can take on 10 dist. values since the order of the data does. matter.

$\mathcal{X}^*$	$\hat{\theta}^*$	$P^*[\hat{\theta}^*]$
1, 1, 1	3/3	1/27
1, 1, 2	4/3	3/27
1, 2, 2	5/3	3/27
2, 2, 2	6/3	1/27
1, 1, 6	8/3	3/27
→ 1, 2, 6	9/3	6/27
2, 2, 6	10/3	3/27
1, 6, 6	13/3	3/27
2, 6, 6	14/3	3/27
6, 6, 6	18/3	1/27
		sum to 1.

- For each $\mathcal{X}^*$ we calc. $\hat{\theta}^*$
- For each $\hat{\theta}^*$ we can write down its prob $P^*[\hat{\theta}^*]$ w.r.t. the bootstrap exp of drawing $\mathcal{X}^*$ conditional on ~~from~~ $\mathcal{X}$. ($R(\mathcal{X}^*, \hat{F})$)

- The bootstrap principle is to equate the dist of $R(\mathcal{X}, F)$ w/ that of $R(\mathcal{X}^*, \hat{F})$. and make our inference. (based on dist of $\hat{\theta}^*$).

So

$$P^*[\hat{\theta}^* \leq \tfrac{6}{3}] = 8/27.$$

or a simple $25/27$ ($\sim 93\%$) conf. interval for θ is $(4/3, 14/3)$.

Sec 9.2. Basic Methods.

Sec. 9.2.1 Nonparametric bootstrap.

For realistic sample sizes (n), the number of potential bootstrap ~~da~~ pseudo-data sets is very large (n^n) and so complete enumer. is not possible. Instead, B ind. rand. bpd. are drawn from the empirical dist. $\hat{F}$ of the observed data.

- i.e. draw $\mathcal{X}_i^* = \{X_{i1}^*, \ldots, X_{in}^*\}$ for $i = 1, \ldots, B$ and then use $R(\mathcal{X}_i^*, \hat{F})$ $(i = 1, \ldots, B)$ to approx. dist of $R(\mathcal{X}, F)$.

Comments: - no parametric assumptions are required (don't need any info about unknown dist)
- can answer wider range of ques.
- often more accurate than standard parametric theory.
- can make simulation error small by increasing B.

Ex/ Previous ex. non-parametric bootstrap.

$$\hat{F} = \begin{cases} 1 & 1/3 \\ 2 & 1/3 \\ 6 & 1/3 \end{cases}$$

Generate $\mathcal{X}_i^*$ by sampling $X_{i1}^*, X_{i2}^*, X_{i3}^*$ w/ replacement from $\{1, 2, 6\}$. Each $\mathcal{X}^*$ yields a $\hat{\theta}^*$. Ex pg 254 shows freq of $\hat{\theta}^*$ when $B = 1000$. These freq approx $P[\hat{\theta}^*]$.

Sec 9.2.2. Parametric Bootstrap.

In ordinary bootstrap $\mathcal{X}^*$ gener. by drawing $X_1^*, .. X_n^*$ iid. from $\hat{F}$. (The empirical dist w/ no other assumptions).
However, if it is believed that F is a parametric distribution $F(x,\theta)$ another method may be employed.

Parametric Bootstrap.

- Draw $X_1, \dots, X_n \sim$ iid $F(x,\theta)$. and use $\mathcal{X}$ to est. θ, $\hat{\theta}$.
- Each parametric bootstrap pseudo-data ~~set~~ $\mathcal{X}^*$ is generated ~~from~~ by drawing $X_1^*, \dots, X_n^* \sim$ iid. $F(x, \hat{\theta})$.

Ex/ Assume $X_1, \dots, X_n$ are Normal$(\theta, 1)$.
compute $\hat{\theta} = \Sigma X_i / n$
draw $X_1^*, \dots, X_n^*$ Normal $(\hat{\theta}, 1)$.

If the model is known or believed to be a good fit, this is a powerful tool, allowing inference & producing more accurate confidence intervals than standard asymptotic theory.

Sec 9.2.4
Sec 4.1 Bootstrap Bias Correction

We want to know the bias of $T(F) = \theta$.
$\therefore$ we are interested in using bootstrap analysis on $R(\mathcal{X}, F) = T(\hat{F}) - T(F) = \hat{\theta} - \theta$.

We want to know $\text{Exp}[\text{Bias}] = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta$
exp w.r.t. F.

So using the bootstrap principle we calc.

$$E^*[\cancel{\text{Bias}}\ \hat{\theta}^* - \hat{\theta}] = E^*[\hat{\theta}^*] - \hat{\theta} = \bar{\theta}^* - \hat{\theta}$$

where

$$\bar{\theta}^* = \sum_{j=1}^{B} \hat{\theta}_j^* / B .$$

each $\hat{\theta}_j^*$ based on x_j^* drawn from $\hat{F}$.

And so our new bias corrected estimator is

$$\theta_{Bias} = \hat{\theta} - (\bar{\theta}^* - \hat{\theta}) = 2\hat{\theta} - \bar{\theta}^* .$$

and should have less bias than $\hat{\theta}$.

Sec 4.2

Bootstrap Estimation of Variance

Suppose now we wish to estimate the variance of $T(\hat{F}) = \hat{\theta}$.

We create this estimate by calculating the sample variance of our bootstrap estimates $\hat{\theta}_j^*$.

$$\hat{V}(\hat{\theta}) = \hat{V}(\theta^*) = \frac{1}{B-1} \sum_{j=1}^{B} (\hat{\theta}_j^* - \bar{\theta}^*)^2$$

Sec 9.3.1
Sec 4.3

Bootstrap Confidence Intervals – Percentile Method

The simplest method for drawing inference about a univariate para. θ using bootstrap simulations is to construct a confidence interval using the percentile method. This method amounts to reading percentiles off the histogram of $\hat{\theta}^*$ values produced by bootstrapping.

We had ex. earlier, look pg 258 Fig 9.1 we can find a bootstrap est of $1-\alpha$ conf. interval based on $((1-\alpha/2)100^{th})$ & $(\alpha/2\ 100^{th})$ emp. percentiles of the histogram.

Thus 95% conf. int fo θ is $(-.0.205, -.174)$.

In practice if we wish to use B bootstrap samples to est. our confidence interval of size $(1-\alpha)100$th we obtain the inteval

$$\left(t^*_{(\alpha/2)},\ t^*_{(1-\alpha/2)} \right)$$

where $t^*_{(\pi)}$ is the $[\pi B]^{th}$ order statistic of our B sized bootstrap samples $\hat{\theta}^*$.

Sec 9.3.1.1 Justification for the percentile Method.

Consider a strictly increasing transformation ϕ and a distribution function H that is cont. and symmetric $(H(z) = 1 - H(-z))$. with the property that

$$(*) \quad P\left[h_{\alpha/2} \leq \phi(\hat{\theta}) - \phi(\theta) \leq h_{1-\alpha/2} \right] = 1-\alpha$$

(we do not need to know H! & ϕ). where h_α is the α quantile of H.

&/ if ϕ is the normalizing variance stabilizing transform the H is standard Normal. (doesn't matter).

Apply the bootstrap principlal to eq $(*)$.

$$1-\alpha \approx P^*\left[h_{\alpha/2} \leq \phi(\hat{\theta}^*) - \phi(\hat{\theta}) \leq h_{1-\alpha/2} \right]$$

isolate $\hat{\theta}^*$

$$= P^*\left[h_{\alpha/2} + \phi(\hat{\theta}) \leq \phi(\hat{\theta}^*) \leq h_{1-\alpha/2} + \phi(\hat{\theta}) \right]$$

$$= P^*\left[\phi^{-1}\left(h_{\alpha/2} + \phi(\hat{\theta}) \right) \leq \hat{\theta}^* \leq \phi^{-1}\left(h_{1-\alpha/2} + \phi(\hat{\theta}) \right) \right]$$

And so we have rewritten the expression in terms of the confidence interval of $\hat{\Theta}^*$ and the bootstrap dist.

Since the bootstrap dist. is observed by us, its percentiles are known quantities.

i.e.

we know ε_α the α quantile of the empirical dist of $\hat{\Theta}^*$

Hence we know

$$P^*\left[\varepsilon_{\alpha/2} \le \hat{\Theta}^* \le \varepsilon_{1-\alpha/2}\right] = 1-\alpha$$

and we can approx

$$\phi^{-1}(h_{\alpha/2} + \phi(\hat{\Theta})) \approx \varepsilon_{\alpha/2}$$
$$\phi^{-1}(h_{1-\alpha/2} + \phi(\hat{\Theta})) \approx \varepsilon_{1-\alpha/2}.$$

Now, going back to the original eq (*) we have

$$1-\alpha = P\left[h_{\alpha/2} - \phi(\hat{\Theta}) \le -\phi(\Theta) \le h_{1-\alpha/2} - \phi(\hat{\Theta})\right]$$
$$= P\left[-h_{1-\alpha/2} + \phi(\hat{\Theta}) \le \phi(\Theta) \le -h_{\alpha/2} + \phi(\hat{\Theta})\right]$$
$$= P\left[\phi^{-1}\left[-h_{1-\alpha/2} + \phi(\hat{\Theta})\right) \le \Theta \le \phi^{-1}\left(-h_{\alpha/2} + \phi(\hat{\Theta})\right)\right]$$
$$= P\left[\phi^{-1}\left[h_{\alpha/2} + \phi(\hat{\Theta})\right) \le \Theta \le \phi^{-1}\left(h_{1-\alpha/2} + \phi(\hat{\Theta})\right)\right]$$

since by symm. $h_{\alpha/2} = -h_{1-\alpha/2}$ $(H(z) = 1-H(-z))$.

$$\approx P\left[\varepsilon_{\alpha/2} \le \hat{\Theta}^* \le \varepsilon_{1-\alpha/2}\right]$$

and so the conf. limits happily coincide w/ those for $\hat{\Theta}^*$ we read off the quantiles for $\hat{\Theta}^*$ from the bootstrap dist as our confidence limits for Θ.

Sec 9.3.2 Pivoting

Although simple, the percentile method is prone to bias & inaccurate coverage prob. To ensure best performance, the bootstrapped statistic should be approx. pivitol. (that is its dist. should not depend on the unknown parameter θ).

Ex/ if g is our variance stabalizing transformation. then the variance of $g(\hat{\theta})$ is ind. of $\hat{\theta}$, and is a good pivot.

Sec 9.3.2 discusses several pivoting tech.

Sec 9.3.22 The bootstrap t.

An approx. pivot that is quite easy to implement is given by the bootstrap t method, also called the studentized bootstrap.

Suppose we wish to est $\theta = T(F)$ by $\hat{\theta} = T(\hat{F})$. We can est. the variance of $\hat{\theta}$ by $V(\hat{F})$. Doing so, it is reasonable to hope that

$$R(X,F) = \frac{T(\hat{F}) - T(F)}{\sqrt{V(\hat{F})}} = \frac{\hat{\theta} - \theta}{\sqrt{V(\hat{F})}}$$

is roughly pivitol.

And so we bootstrap $R(X,F)$ obtaining a collection of $R(X^*, \hat{F})$.

Let $\hat{G}$ be the dist of $R(X,F)$ and $\hat{G}^*$ be the dist of $R(X^*, \hat{F})$. unknown observed.

Then by def a $1-\alpha$ confidence interval for θ can be obtained from

$$1-\alpha = P\left[\xi_{\alpha/2}(\hat{G}) \le R(X,F) \le \xi_{1-\alpha/2}(\hat{G})\right]$$
$$= P\left[\xi_{\alpha/2}(\hat{G}) \le \frac{T(\hat{F})-T(F)}{\sqrt{V(\hat{F})}} \le \xi_{1-\alpha/2}(\hat{G})\right]$$
$$= P\left[\hat{\theta} - \sqrt{V(\hat{F})}\,\xi_{1-\alpha/2}(\hat{G}) \le \theta \le \hat{\theta} - \sqrt{V(\hat{F})}\,\xi_{\alpha/2}\right]$$

where $\xi_\alpha(\hat{G})$ is the α quantile of $\hat{G}$. However F (and hence $\hat{G}$) is unknown.
So we use the bootstrap priciple to imply that $\hat{G}$ is roughly equal to $\hat{G}^*$,
$\therefore$ $\xi_\alpha(\hat{G}) \approx \xi_\alpha(\hat{G}^*)$ $\forall \alpha$. And since we can obtain $\xi_\alpha(\hat{G}^*)$ from the histogram of bootstrap values $R(X^*, \hat{F})$ we have the bootstrap confidence interval for θ:

$$\left(T(\hat{F}) - \sqrt{V(\hat{F})}\,\xi_{1-\alpha/2}(\hat{G}^*),\ T(\hat{F}) - \sqrt{V(\hat{F})}\,\xi_{\alpha/2}(\hat{G}^*)\right)$$

Comment: these are percentiles of the tails & thus.
B should be very large (several thousand).
-Can use bootstrap variance est. for $V(\hat{F})$.

Bootstrap-t usually provide confidence interal coverage rates that closely approx. the nominal conf. level. They are most reliable when $T(\hat{F})$ is approx a location statistic $\rightarrow$ a constant shift in all the data values will induce the same shift in $T(\hat{F})$. Also, can be sensitive to the presence of outliers & should be used with caution in these cases. Also, it performs poorly when the underlying dist. is heavy tailed.

Se. 9.3.3 pg 268 Bootstrap for hypothesis testing.

Conducting a hyp. test is closely related to est. conf int. The simplest approach for bootstrap hyp. testing is to base the p-value on a bootstrap conf. int. Specifically, consider a null hyp based on a para whose est. can be bootstrapped.

- Obtain conf. int by percentile or other method
- if the $(1-\alpha)100\%$ b.c.i does not cover the null value, reject w/ p-value no greater than α.

- This is a simple method & works better if the bootstrap sampling is done in a manner that reflects the null hyp.

To illus. what we mean, consider a null hyp about a univariate para θ with null value θ_0.
Let the test statistic be $R(\mathcal{X}, F) = \hat{\theta} - \theta_0$
the null hyp would reject whenever $|\hat{\theta} - \theta_0|$ is large compared to a reference dist.

A tempting method would be to gen the ref. dist by resampling values $R(\mathcal{X}^*, F) = \hat{\theta}^* - \theta_0$ via bootstrap. However if if the null is false, then this stat. does not have the correct ref. dist. If θ_0 is far from true θ, then $|\hat{\theta} - \theta_0|$ will not seem big compared to $|\hat{\theta}^* - \theta_0|$.

Instead use $R(\mathcal{X}^*, \hat{F}) = \hat{\theta}^* - \hat{\theta}$ to generate bootstrap estimate. Then if θ_0 is far from θ, the values $|\hat{\theta}^* - \hat{\theta}|$ will be small compared to $|\hat{\theta} - \theta_0|$. Thus comparing $\hat{\theta} - \theta_0$ to $\hat{\theta}^* - \hat{\theta}$ yields greater statistical power.

Sec 9.4
Gentle 4.5

Reducing MC Error. (Variance/bias reduction).

MC $\hat{}$BS estimators have two sources of variation due to

1. the initial sampling
2. the bootstrap sampling.

We will discuss 2 methods to reduce variance.

— Jackknife after Bootstrap

Gentle.

First we need to est. the variance of the bootstrap estimator. One way to do this is to use the jackknife. The brute force way to do this is to

- do n sep. bootstraps on the original sample w/ a diff obs removed each time.
- Then use jackknife discussed earlier.

This is not comp effiecient.

Another procedure (Efron 1992) draw m bs samples and stores the indices of the original sample included in each bootstrap sample in an $n \times m$ matrix

m bootstrap samples

$$\begin{array}{ccc} 1 & \cdots & m \\ \end{array}$$
$$\begin{pmatrix} x_1 & x_2 & \\ x_4 & x_3 & \\ x_7 & x_9 & \\ \vdots & \vdots & \\ \vdots & \vdots & \end{pmatrix}$$

Then for each of the m samples not containing x_i, treat it as though it came from the orig. sample w/ x_i omitted.

Bootstrap samples obt. either way have the same dist. Then use these samples to gen. Jackknife.

- Can have problems if x_j in all samples, but large m (rel to n) has a low prob. of this.

Sec 9.4.1 **Balanced Bootstrap.**

Consider a bootstrap bias correction of the sample mean.
The bias correction should be zero, because
$\bar{X}$ is unbiased for the true mean μ.

Now

$$R(\mathcal{X}, F) = \bar{X} - \mu$$

will have bootstrapped values

$$R(\mathcal{X}_j^*, \hat{F}) = \bar{X}_j^* - \bar{X} \qquad j = 1, \ldots, B.$$

whose mean is our bias est.

Although $\bar{x}$ is unbiased it is highly unlikely that ~~This will cause bias est~~ a random select of the pseudo-data sets will produce a dist $R(\mathcal{X}_j^*, \hat{F})$ whose mean is exactly 0.
This is due to ordinary M.C. variation.

However, if each data value occurs in $\{\mathcal{X}_1^*, \ldots \mathcal{X}_B^*\}$ w/ the same rel freq as it does in $\mathcal{X}$ then the bootstrap bias est

$$\frac{1}{B}\sum_{j=1}^{B} R(\mathcal{X}_j^*, \hat{F}) \text{ must be zero.}$$

This is called balancing the bootstrap.

The easiest way to do this is to concatenate B copies of the data set $\mathcal{X}$, permute the values and then read off B blocks n values at a time.
The jth block becomes $\mathcal{X}_j^*$.

Other comments:

Sec 9.5 Fairly common idea Bootstrap aggregating (Bagging). Basically replace $\hat{\theta}$ w/ $\bar{\theta}^* = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j^*$ → model averaging. reduces var due to small chan. in data.