6 The Parametric Bootstraps and R Implementations

6.1.1 Setup

We have samples $y_1,\ldots,y_n$

The sample values are viewed as the outcomes of random variables $Y_1,\ldots,Y_n$ ~ iid PDF f, CDF F.

Want to use the sample to make inference about a population characteristic θ using a statistic T whose value in the sample is t.

Question: what's the probability distribution of T? For example,

1. What're the bias, SD, and quantiles?
2. What's the distribution under certain null hypothesis?
3. How do we calculate CI for θ using T?

Two cases exist:

- Parametric: parameters ψ fully determines f. (θ is a function of ψ) example
- Non-parametric: uses only the fact that Y_j are iid.

Empirical distribution: puts equal probability n^{-1} at each sample value y_j.

Empirical distribution function (EDF): $\hat{F}(y) = \frac{\#\{y_j \le y\}}{n} = \frac{1}{n}\sum_{j=1}^{n} H(y - y_j)$

Many statistics can be though of as functions of EDF.

e.g. y-bar = $\sum_{j=1,\ldots,n} y_j$ is the mean of EDF.

They can be written as t= $t(\hat{F})$.

t(.) is called statistical function (the algorithm for computing t from Fhat)

It also defines the parameter or characteristic of interest by θ=t(F)

The mean and var are good examples of such functions:

$t_{mean}(F) = \int y dF(y)$, $\underline{t_{\mathrm{var}}(F) = \int y^2 dF(y) - \left\{\int y dF(y)\right\}^2}$

For parametric problems, θ is usually more explicit as one of the model parameters ψ.

The key parallel:

$\theta = t(F)$ and $t = t(\hat{F})$.

Let's pause and not be confused.

Ex 2.1: $\mu = \int y dF(y) \equiv t(F)$ and $\bar{y} = t(\hat{F})$ (why?)

Idea: $\hat{F} \to F$.(why?) If t(.) is 'continuous', (in some sense) $T \equiv t(\hat{F}) \to t(F) \equiv \theta$.

Also, small differences $O(n^{-1})$ don't matter…

6.1.2 Objectives

e.g. Suppose we want to calculate a $(1-2\alpha)$ CI for θ. Often,

$T \sim N(\theta + \beta, v^2)$ approximately, where β is 'bias of T'. If both bias and var are known, we can write

$\Pr(T \le t \mid F) \approx \Phi\left(\frac{t-(\theta+\beta)}{v^{1/2}}\right)$.

If $z_\alpha = \Phi^{-1}(\alpha)$, then an approximate ($1-2\alpha$) CI for θ is given by

$[t-\beta-v^{1/2}z_{1-\alpha}, t-\beta+v^{1/2}z_{1-\alpha}]$

In practice, the bias and var are unknown. What to do? Replace them with estimates!

Recall: $\beta = b(F) = E(T \mid F) - t(F), v = v(F) = \mathrm{var}(T \mid F)$ all depend on unknown F.

(Note $E(T \mid F) = E\{t(\hat{F}) \mid Y_1,...,Y_n \sim iid, F\}$)

Estimate F by $\hat{F}$, EDF or a fitted parametric distribution.

Substituting Fhat in place of F, we have estimates of bias and var:

$B = b(\hat{F}) = E(T \mid \hat{F}) - t(\hat{F}), V = v(\hat{F}) = \mathrm{var}(T \mid \hat{F})$

Use them in the above CI etc.

These are bootstrap estimates! (here used in conjunction with normal approximation but can be used to improve upon the normal approximation)

7 Parametric Simulation (for Basic Bootstrap)

Suppose we have a parametric model for the distribution of the data $y_1,..., y_n$.

Let the CDF and PDF be $F_\psi(y), f_\psi(y)$.

Estimate ψ by $\hat{\psi}$, usually by MLE. Substituting, one gets fitted model

$\hat{F}(y) = F_{\hat{\psi}}(y)$.

Let Y* denote the random variable distributed according to the fitted model.

Superscripts * will used for E*, var*, etc when these moments are calculated according to the fitted distribution.

Sometimes, we will also write $\hat{\psi} \equiv \psi^*$.

Suppose the theoretical calculation with the fitted model is too complex and either or both

1) approximations are not available
2) approximation is not trustworthy because the sample size is too small

Instead, we simulate datasets from $\hat{F}$ and estimate the moments from the simulated datasets.

Write such a dataset as $Y_1^*,...,Y_n^*$ where Y_j^* are iid sampled from $\hat{F}$.

When the statistic of interest is calculated from a simulated dataset, we denote it by T*.

From R repetitions of the data simulation, we obtain $T_1^*,..., T_R^*$.

Properties of $T-\theta$ are estimated from $T_1^*,..., T_R^*$.

e.g.
$$\begin{aligned} B &= b(\hat{F}) = E(T \mid \hat{F}) - t(\hat{F}) = E^*(T^*) - t \\ V &= v(\hat{F}) = \mathrm{var}(T \mid \hat{F}) = \mathrm{var}^*(T^*) \end{aligned}$$

are estimated by

$B_R = \frac{1}{R}\sum_{r=1}^{R} T_r^* - t = \overline{T}^* - t$ (a simulation analog of $T-\theta$)

$V_R = \frac{1}{R-1}\sum_{r=1}^{R}(T_r^* - \overline{T}^*)^2$

As R increases, B_R->B and V_R->V. (Why? LLN)

* We focus on the property of T relative to θ rather than absolute property of T.

7.1.1 Distribution and quantile estimates

The normal approximation is inaccurate for many cases. (can be seen from Q-Q plot of the simulated values t_1*, …,t_R*. Use empirical distribution of these simulated values to approximate the distribution of T!

<u>Goal</u>: approximate the dist(T-θ) by dist(T*-t).

<u>Procedure</u>: approximate

G(u) = Pr(T-θ ≤u)

by the <u>simulation estimate</u>

$\hat{G}_R(u) = \frac{\#\{t_r^* - t \le u\}}{R} = \frac{1}{R}\sum_{r=1}^{R} I(t_r^* - t \le u)$.

As R increases, this will converge to

$\hat{G}(u)$, the exact CDF of dist(T*-t) under sampling from the fitted model.

Two sources of error in approximation $\hat{G}_R$ to G:

1) that b/w $\hat{G}$ and G (data variability)
2) that b/w $\hat{G}_R$ and $\hat{G}$ (finite simulation)

Among others, <u>quantiles of dist(T-θ)</u> are approximated by <u>those of by dist(T*-t)</u>. The p quantile of the latter can be approximated by sample quantiles of t*-t. (R+1)p the ordered values of t*-t, i.e. $t_{((R+1)p)}$*-t .

* The simulated approximation $\hat{G}_R$ and the corresponding quantiles are in principle <u>BETTER than results obtained from normal approximation</u> since it doesn't assume dist(T*-t) has a particular form! (R needs to be large enough)

* It is useful to be able to look at the density of T. A rough idea of the density g(u) of U=T-θ can be had from a histogram of the values of t*-t. Use <u>kernel density estimates</u> on t_r*-t! R needs to be even larger (R=1,000 or more)

8 Nonparametric Simulation (for Basic Bootstrap)

No parametric model assumption, except Y_1,…,Y_n are iid with unknown F.

=> Estimate F by the EDF $\hat{F}$. Then?

1) Theoretical computation if possible,
2) o.w., simulation of datasets and empirical calculation of required properties.

e.g. $E^*(\overline{Y}^*) = \overline{y}, \mathrm{var}^*(\overline{Y}^*) = \frac{(n-1)}{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}(y_j - \overline{y})^2$

* Simulating from EDF: sample each Y* independently at random from the original data values $y_1,\ldots,y_n$. RSWR from the data. This is called 'nonparametric bootstrap'.

* Need more simulation (larger R) than parametric bootstrap to get reliable estimates of quantiles etc.

* There are up to $m_n = \binom{2n-1}{n-1}$ possible values of t*. For very small n, it is very large…

```
n <- 1:100
plot(n, choose(2*n-1, n-1), log='y')
```

So the effects of discreteness in the distribution of T* is fairly minimal.
Still, outliers are problems.

8.1.1 Simple confidence intervals

1) Use normal approximation to the distribution of T. Use bootstrap to estimate only bias and variance.
2) Equitailed $(1-2\alpha)$ CI intervals called the basic bootstrap confidence limits. (use R>1000) $t-(t^*_{((R+1)(1-\alpha))}-t)$, $t-(t^*_{((R+1)\alpha)}-t)$.
    a. The accuracy depends upon the extent to which the dist(T*-t) agrees with dist(T-θ). Using studentized bootstrap may help sometimes.

8.1.2 Why the above form?

Let A and B be αth and $(1-\alpha)$th quantiles of $\hat{G}_R(u) = P(t^* - t \le u \mid \hat{F})$.
Recall $G(u) = P(T - \theta \le u \mid F)$. Thus, the true $(1-2\alpha)$ CI is specified by

$$1 - 2\alpha = P(G^{-1}(\alpha) \le T - \theta \le G^{-1}(1-\alpha))$$
$$= P(T - G^{-1}(1-\alpha) \le \theta \le T - G^{-1}(\alpha))$$

Now we approximate $G^{-1}(\alpha) \approx \hat{G}_R^{-1}(\alpha) = t^*_{((R+1)\alpha)} - t$ and similarly for $G^{-1}(1-\alpha)$. Replacing them into the above 'true' boundaries of CI, we obtain simple confidence intervals.

8.2 Statistical Issues

It's important to ask the following questions:

1) When does the bootstrap work? Practical and theoretical aspects.
    a. Consistency: some mathematical conditions. Things like sample maximum doesn't work.
    b. Asymptotic accuracy: for many common statistics, theory says (studentized) bootstrap CI has advantage over normal approximations etc.
2) When might the bootstrap fail?
    a. Incomplete data
    b. Dependent data
    c. Dirty data

8.2.1 Air-conditioning data example

For n=12 observations of between-failure times, we wish to estimate the underlying <u>mean</u> μ or the failure <u>rate</u> μ^{-1}. The density of Exp(1/μ) is given by

$$f(y) = \frac{1}{\mu}\exp\left(-\frac{y}{\mu}\right), y > 0.$$

R representation is {r,p,q,d}exp(…,rate=1). Note it is parameterized by rate, not mean.
Exponential distribution is a special case of Gamma(μ,κ) distribution with density

$$f_{\mu,\kappa}(y) = \frac{1}{\Gamma(\kappa)}\left(\frac{\kappa}{\mu}\right)^{\kappa} y^{\kappa-1}\exp(-\kappa y/\mu), y > 0, \mu, \kappa > 0$$

Again, R representation is {r,p,q,d}gamma(…,shape, rate = 1, scale = 1/rate).
Under the exponential model, the estimator $T = \overline{Y}$ for μ is

1) unbiased,
2) has variance μ^2/n.

since $T = \overline{Y} \sim Gamma(n, n/\mu)$ (for certain parametrization)

* Some exponential and Gamma distribution backgrounds:

1) For U~Gamma(α,1/β), E(U)= $\alpha\beta$ and var(U)= $\alpha\beta^2$
2) V~Exp(1/μ)=Gamma(1,1/μ). Naturally, E(V)=μ and var(V)= μ^2.
3) If $U_1,\ldots,U_n$~iid gamma(α,1/β), their sum ~ Gamma(nα,1/β)
4) IF U~Gamma(α,1/β), then for a constant a>0, U/a~Gamma(α,a/β).
5) Thus, if $V_1,\ldots,V_n$ ~ iid Exp(1/μ), their average ~ Gamma(n, n/μ)

Example 2.3: Approximate CI for μ can be calculated from these + normal approximation for dist(T).

Example 2.4: $Y^* \sim Exp(1/\bar{y})$. Thus,
$E^*(\bar{Y}^*) = \bar{y}, \mathrm{var}^*(\bar{Y}^*) = \bar{y}^2/n$.
The estimated bias of and variance are those given in Example 2.3. Trivial.

But if we want to calculate the bias and variance of $T = \log \bar{Y}$ under the fitted model, $E^*(\log \bar{Y}^*) - \log \bar{y}, \mathrm{var}^*(\log \bar{Y}^*)$, exact calculation is more difficult and theory is not that helpful--- still, we can approximate from simulated samples of Y*s.

Example 2.5. $Y^* \sim Exp(1/\bar{y}) = Exp(1/108.083)$. We know the answer but simulate them and estimate the bias and variance nonetheless for illustration. Figure 2.1 shows the bias and variance. As R increases, the closer the simulation calculation is to the right answer. How large do they have to be? --- It depends. Not trivial.

Example 2.6. Normal approximation to T-θ vs. empirical distribution of t*-t. The letter is better…

Example 2.9. distribution of t* approximated from nonparametric bootstrap matches gamma model better than exponential model.

8.2.2 Air conditioning data example code

```
                                          #
                                          # Generate data
                                          #
y <- c(3, 5, 7, 18, 43, 85, 91, 98, 100, 130, 230, 487)
n <- length(y)

                                          # Ex. 1.1
                                          # Is Exponential assumption reasonable?
par(mfrow=c(2,1))
                                          # draw EDF and parametric fit
plot(ecdf(y), main='Empirical Distribution of Y')
curve(pexp(x, 1/mean(y)), add=TRUE, col='red')
legend(200, 0.6, c('EDF','Fitted model (Exponential)'),
       lty=c(1,1), col=c('black','red'))
                                          # Q-Q plot of EDF vs parametric fit
plot(qexp(ppoints(y), 1/mean(y)), sort(y),
     xlab='Quantiles of Exponential Distribution',
     ylab='Quantiles of Y')
abline(0,1)

                                          # Ex 2.1. Sample mean is 108.0833
mean(y)

                                          # Ex 2.3. approximate normal CI
                                          # 46.93055 169.23611
c(mean(y)-qnorm(.975)*mean(y)/sqrt(n),
```

```
  mean(y)+qnorm(.975)*mean(y)/sqrt(n))

                                        # Ex 2.5. Simulation vs Theory

R <- 500                                # Try different R values, like 10, 100

t.stars <- numeric(R)
for(r in 1:R){
  y.stars <- rexp(n, 1/mean(y))
  t.stars[r] <- mean(y.stars)
}

mean(t.stars)-mean(y)                   # compare this with 'correct' bias=0
var(t.stars)                            # compare this with 'correct' var=973.5
## Repeat the above many times. Do we get consistent results?
mean(y)^2/n                             # 'correct' var (y-bar)^2/n


                                        # Ex 2.5. (complete)

n.trials <- 7
Rs <-seq(10, 500, length=10)
boot.bias <- matrix(NA, nrow=length(Rs),ncol=n.trials )
boot.var <- matrix(NA, nrow=length(Rs),ncol=n.trials )

for(trial in 1:n.trials){
  R <- max(Rs)
  t.stars <- numeric(R)
  for(r in 1:R){
    y.stars <- rexp(n, 1/mean(y))
    t.stars[r] <- mean(y.stars)
  }
  for(i in 1:length(Rs)){
    boot.bias[i, trial] <- mean(t.stars[1:Rs[i]])-mean(y)
    boot.var[i, trial] <- var(t.stars[1:Rs[i]])
  }
}

par(mfrow=c(1,2))
matplot(Rs, boot.bias, log='x', type='b', pch=1)
abline(h=0)
matplot(Rs, boot.var, log='x', type='b', pch=1)
abline(h=mean(y)^2/n)


                                        # Ex 2.6. Normal approximation
                                        #  vs G_R estimate

R <- 999                                # Also try R=99
t.stars <- numeric(R)
for(r in 1:R){
  y.stars <- rexp(n, 1/mean(y))
  t.stars[r] <- mean(y.stars)
}
                                        # Normal and Gamma Q-Q plot
par(mfrow=c(1,2))
qqnorm(t.stars)
qqline(t.stars)
plot(qgamma(ppoints(t.stars), n, n/mean(y)), sort(t.stars),
```

```
     xlab='Exponential model quantiles', ylab='t*')
abline(0,1)

                                        # Distribution G_R of t*
par(mfrow=c(1,1))
hist(t.stars, nclass=20, prob=TRUE, col='gray', border='white')
lines(density(t.stars))


                                        #
                                        # Ex 2.9. Nonparametric bootstrap
                                        #

R <- 999
t.stars <- numeric(R)
set.seed(101)
for(r in 1:R){
  y.stars <- sample(y, replace=TRUE)
  t.stars[r] <- mean(y.stars)
}
                                        # Exp and Gamma Q-Q plot
par(mfrow=c(1,2))
plot(qgamma(ppoints(t.stars), n, n/mean(y)), sort(t.stars),
     xlab='Exponential model quantiles', ylab='t*')
abline(0,1)
plot(qgamma(ppoints(t.stars), n*0.71, 0.71*n/mean(y)), sort(t.stars),
     xlab='Gamma model quantiles', ylab='t*')
abline(0,1)


                                        #
                                        # Simple 95% confidence intervals
                                        #
                                        # 25.1625 171.3375

alpha <- .025
par(mfrow=c(1,1))
plot(ecdf(t.stars))
abline(h=c(alpha, 1-alpha))
abline(v=quantile(t.stars, prob=c(alpha, 1-alpha)))
c(mean(y)-(quantile(t.stars, 1-alpha) - mean(y)),
  mean(y)-(quantile(t.stars, alpha) - mean(y)))

par(mfrow=c(1,1))
hist(t.stars, nclass=50, col='gray')
abline(v=c(mean(y)-(quantile(t.stars, 1-alpha) - mean(y)),
         mean(y)-(quantile(t.stars, alpha) - mean(y))), col='red')
abline(v=c(mean(y)-qnorm(.975)*mean(y)/sqrt(n),
         mean(y)+qnorm(.975)*mean(y)/sqrt(n)), col='blue')
```

9 Further Ideas on Bootstrap

9.1 Bootstrapping Several Samples (Review of ANOVA)

Suppose we're interested in a parameter thatdepends on the populations $F_1, \ldots, F_k$.
The data consist of independent random samples form these populations.
The i'th sample $y_{i1},\ldots,y_{in_i}$ arises from population F_i for i=1,…,k.
The nonparametric estimate of F_i is the edf of the ith sample.
Nonparametric simulation leads to stratified sample $y_{i1}^*,\ldots,y_{in_i}^*$.
Proceed as One-sample case.
Example (Difference of population means)
Interested in the difference of two population means,

$$\theta = t(F_1, F_2) = \int y dF_1(y) - \int y dF_2(y)$$

The estimate is

$$t(\hat{F}_1, \hat{F}_2) = \bar{y}_1 - \bar{y}_2 .$$

Simulate $t^* = \bar{y}_1^* - \bar{y}_2^*$ to obtain CI etc.

9.2 Smooth Estimate of F

If EDF is discrete and the effects of discreteness are severe (D&H 2.3.2, sample medians and other quantiles), a smooth estimate of F may be better.
One possibility is to use kernel density estimation.

$$\hat{f}_h(y) = \frac{1}{nh}\sum_{j=1}^{n} w\left(\frac{y - y_j}{h}\right)$$

and do calculations or simulations based on the corresponding CDF $\hat{F}_h$ rather than EDF. This corresponds to simulation by setting

$Y_j^* = y_j^* + h\varepsilon_j,\ j = 1,..,n$ (why?)

which is called the smoothed bootstrap.

10 Tests

Simple null hypothesis H_0: $y_1,\ldots,y_n$ ~ known F_0.
Composite null hypothesis: some aspects of F are not determined even when H_0 is true. (normal with mean 1)
A test statistic T: measures the discrepancy b/w the data and the null hypothesis.
We can make large values of T are evidence against H_0 in most cases. (why? Example?)
Let t be the observed value of T.
The level of evidence against H0 is measured by the significance probability

$p = \Pr(T \geq t \mid H_0)$, called the P-value.
A critical value t_p for t, associated with testing at level p or 100p%:
$\Pr(T \geq t_p \mid H_0) = p$.(error rate or the size of the test)
Level p critical region of the test is {(y1,…,yn): $t \geq t_p$}
Dist(T|H_0) = null distribution of T.
Under H0, the P-value ~ Uniform(0,1).
How to pick test statistic: consider 'alternative' hypothesis. Likelihood ratio test statistic or goodness of fit test statistic.
In the nonparametric setting, no particular forms are specified for distributions and choice of T should be based on a qualitative notion of what is of concern should H_0 not be true.
e.g. Suppose we wish to test H_0: X and Y are independent, given the random sample $(X_1,Y_1),\ldots,(X_n,Y_n)$. The correlation is a convenient measure of dependence and is zero under H_0. (necessary? Sufficient?) T=(correlation) or T=(correlation)^2 can be used depending on situations. (What kind of ?)
In most parametric problems and all nonparametric problems, the null hypothesis is composite. Some parameters are unknown. $p = \Pr(T \geq t \mid F)$ depends on various F satisfying H_0. Neat, theoretical solutions are there. A less satisfactory appoach is to estimate F by a CDF $\hat{F}_0$ which satisfies H0 and calculate
$p = \Pr(T \geq t \mid \hat{F}_0)$.

10.1 Resampling for parametric test

10.1.1 Monte Carlo test

The basic Monte Carlo test: compares the observed t to R independent values of T which are obtained from corresponding samples independently simulated under H_0. Call them $t_1^*,\ldots,t_R^*$. Under H_0, all R+1 values are equally likely values of T, i.e. (if T is continuous)
$\Pr(T < T_{(r)}^* \mid H_0) = \frac{r}{R+1}$. Then, if exactly k of t^* exceeds t,
$p = \Pr(T \geq t \mid H_0) \approx p_{mc} = \frac{k+1}{R+1}$. (the Monte Carlo P-value)
(Slightly more complicated for discrete T.)

Example: (logistic regression) Suppose y1,…,yn are independent binary outcomes with corresponding scalar covariate values x1,…,xn. Q: Does x influence y? If we choose the model

$$\log \frac{\Pr(Y_j = 1 \mid x_j)}{\Pr(Y_j = 0 \mid x_j)} = \lambda + \psi x_j . j = 1,\ldots,n$$

then $H_0 : \psi = 0$.
Use $S = \sum Y_j$ and $T = \sum x_j Y_j$.

The null distribution of Y1,..,Yn given S=s is uniform over all (n choose s) permutations of y1,…,yn. Instead of full permutations, generate R random permutations and apply the above. A simulated samples will then be $(x_1, y_1^*),\ldots,(x_n, y_n^*)$ and the associated test statistic $t^* = \sum x_j y_j^*$.

10.1.2 Parametric bootstrap test

Fit the null model $\hat{F}_0$ and let $p = \Pr(T \ge t \mid \hat{F}_0)$.

10.2 Nonparametric Permutation Test

<u>E.g. (Correlation test)</u> Suppose Y=(U,X) is a random pair and n of them are observed.
H_0: U and X are independent.
Use T=sample correlation.
Now, under the null, S=(ordered Us and Xs) (equivalent to two marginal EDFs) is minimal sufficient for F. Conditional test can be applied. When S is constrained to equal s, the random sample of $(U_1,X_1),\ldots,(U_n,X_n)$ is equivalent to $(u_{(1)}, X_1^*),\ldots,(u_{(n)}, X_n^*)$ with X*s random permutations of ordered Xs. Under H0: all such permutations are equally likely and there are n! of them. The one-sided P-value is

$$p = \frac{\#\text{ of permutations such that } T^* \ge t}{n!}.$$

Practically, it is rarely possible or necessary to compute the permuation P-value exactly. Make use of Monte Carlo method: take a large number R of random permutations and approximate p by

$p_{mc} = \dfrac{1 + \#\{t_r^* \ge t\}}{R+1}$. 99 or 999 random permutations should suffice.

<u>Eg.</u> For comparing the means of two populations, by assuming one distribution is a shifted version of the other, we can use permutation test.

Note, this is special nonparametric resampling tests, in which resampling is doen WITHOUT replacement.

10.3 Nonparameteric Bootstrap Tests

Resapling W or W/O replacement usually makes little difference.
But bootstrap tests apply to a much wider class of testing problem.
Obtain the null resampling distribution $\hat{F}_0$, the basic bootstrap test computes the P-value as $p_{boot} = \Pr^*(T^* \ge t \mid \hat{F}_0)$ or approximately

$p = \frac{1 + \#\{t_r^* \geq t\}}{R+1}$ using the results from R bootstrap samples.

e.g. Comparison of two means : H0: $\mu_2 = \mu_1$. Let $\hat{F}_0$ be the pooled EDF. Same as permutation test above but sample WITH replacement.

10.4 And more!

Test of unimodality of a density
Bootstrap Censored Data (Survival Analysis)
Missing values
Bootstrap time series
Bootstrap Linear Regression
Bootstrap Nonparametric Regression

10.4.1 You're supposed to…

Be able to see if you see a bootstrap
Be able to see if you see a need for bootstrap
Be able to look up what methods are available
Be able to implement in your favorite software package