STAT 641: BOOTSTRAPPING METHODS

Jiyoun Myung

Department of Statistics and Biostatistics
California State University, East Bay

Spring 2021, Day 12

Review

Permutation test: The two-sample problem:

We observe two independent random samples such as

$$F \to \boldsymbol{z} = (z_1, z_2, \ldots, z_n) \text{ independent of}$$
$$G \to \boldsymbol{y} = (y_1, y_2, \ldots, y_m).$$

Having observed $\boldsymbol{z}$ and $\boldsymbol{y}$, we wish to test

$$H_0 : F = G.$$

The equality $F = G$ means that F and G assign equal probabilities to all sets, $\text{Prob}_F\{A\} = \text{Prob}_G\{A\}$ for A any subset of the common sample space of the $\boldsymbol{z}$'s and $\boldsymbol{y}$'s.

Reveiw

Algorithm 15.1

Computation of the two-sample permutation test statistic

1. Choose B independent vectors $\mathbf{g}^*(1)$, $\mathbf{g}^*(2), \cdots, \mathbf{g}^*(B)$, each consisting of n z's and m y's and each being randomly selected from the set of all $\binom{N}{n}$ possible such vectors. [B will usually be at least 1000; see Table (15.3).]

2. Evaluate the permutation replications of $\hat{\theta}$ corresponding to each permutation vector,

$$\hat{\theta}^*(b) = S(\mathbf{g}^*(b), \mathbf{v}), \qquad b = 1, 2, \cdots, B. \tag{15.17}$$

3. Approximate $\mathrm{ASL}_{\mathrm{perm}}$ by

$$\widehat{\mathrm{ASL}}_{\mathrm{perm}} = \#\{\hat{\theta}^*(b) \geq \hat{\theta}\}/B. \tag{15.18}$$

Mouse Data

Mean differences

Student's two-sample t test

```
trt <- c(94,197,16,38,99,141,23)
ctrl <- c(52,104,146,10,51,30,40,27,46)

t.test(trt, ctrl, alternative="greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  trt and ctrl
## t = 1.1208, df = 14, p-value = 0.1406
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -17.50517       Inf
## sample estimates:
## mean of x mean of y
##  86.85714  56.22222
```

Mouse Data

Mean differences

Permutation Test

```
obs.diff.means <- mean(trt) - mean(ctrl)

sur <- c(trt, ctrl)
diff.means <- numeric()
set.seed(1234)
for (i in 1:5000) {
  pm <- sample(sur, 16, replace = FALSE)
  diff.means[i] <- mean(pm[1:7]) - mean(pm[8:16])
}
length(diff.means[ diff.means >= obs.diff.means])/5000
```

```
## [1] 0.1444
```

Your turn

Quesition 1

Suppose you conduct an experiment and inject a drug into three mice. Their times for running a maze are 8, 10, and 15 seconds; the times for two control mice are 5 and 9 seconds.

- Compute the difference in mean times between the treatment group and the control group.
- Write out all possible permutations of these times to the two groups and calculate the difference in means.
- What proportion of the differences are as large or larger than the observed difference in mean times?
- For each permutation, calculate the mean of the treatment group only. What proportion of these means are as large or larger than the observed mean of the treatment group?

Your turn

Quesition 2

The file **Phillies2009** contains data from the 2009 season for the baseball team the Philadelphia Phillies.

```
library(resampledata)
str(Phillies2009)
```

```
## 'data.frame':    162 obs. of  7 variables:
##  $ Date      : Factor w/ 162 levels "1-Aug","1-Jul",..: 130 141 147 7 12 17 23 33 38 43 ...
##  $ Location  : Factor w/ 2 levels "Away","Home": 2 2 2 1 1 1 1 1 2 2 ...
##  $ Outcome   : Factor w/ 2 levels "Lose","Win": 1 1 2 1 2 2 2 1 1 1 ...
##  $ Hits      : int  4 6 11 7 15 13 10 5 14 8 ...
##  $ Doubles   : int  2 1 3 2 3 3 3 1 3 2 ...
##  $ HomeRuns  : int  0 0 1 1 1 2 3 0 1 3 ...
##  $ StrikeOuts: int  6 3 6 3 6 4 7 3 5 7 ...
```

- Find the mean number of strike outs per game (**StrikeOuts**) for the home and the away games (**Location**).
- Perform a permutation test to see if the difference in means is statistically significant.

Hypothesis Tests

$H_0 : \mu = \mu_0$ vs $H_A : \mu > \mu_0$

$H_0 : \mu \leq \mu_0$ vs $H_A : \mu > \mu_0$

$H_0 : \mu = \mu_0$ vs $H_A : \mu < \mu_0$

$H_0 : \mu \geq \mu_0$ vs $H_A : \mu < \mu_0$

$H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$

$H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$

Chapter 16: Hypothesis testing with the bootstrap

Introduction

- In addition to providing standard errors and confidence intervals, the bootstrap can also be used to test statistical hypothesis.
- The bootstrap tests give similar results to permutation tests when both are available. The bootstrap hypothesis tests are more widely applicable then the permutation tests though less accurate.
- The bootstrap tests of hypotheses can be done in some situations for which a permutation test doesn't exist.

The two sample problem

Suppose we have two samples $\boldsymbol{z}$ and $\boldsymbol{y}$ from possibly different probability distributions F and G, and we wish to test the null hypothesis $H_0 : F = G$.

Denote the combined sample by $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{y})$ and let its empirical distribution be $\hat{F}_0$, putting probability $1/(n+m)$ on each member of $\boldsymbol{x}$. Under H_0, $\hat{F}_0$ provides a nonparametric estimate of the common population that gave rise to both $\boldsymbol{z}$ and $\boldsymbol{y}$. An achieved significance level is calculated

$$\text{ASL} = \text{Prob}_{H_0}\{t(\boldsymbol{x}^*) \geq t(\boldsymbol{x})\}$$

where $t(\boldsymbol{x})$ is a test statistic. The quantity $t(x)$ is fixed at its observed value and the random variable x^* has a distribution specified by the null hypothesis H_0.

HT with the bootstrap

Computing the bootstrap test statistic for $H_0 : F = G$

Algorithm 16.1

Computation of the bootstrap test statistic for testing $F = G$

1. Draw B samples of size $n + m$ with replacement from $\mathbf{x}$. Call the first n observations $\mathbf{z}^*$ and the remaining m observations $\mathbf{y}^*$.
2. Evaluate $t(\cdot)$ on each sample,

$$t(\mathbf{x}^{*b}) = \bar{\mathbf{z}}^* - \bar{\mathbf{y}}^*, \quad b = 1, 2, \cdots B. \tag{16.2}$$

3. Approximate ASL_{boot} by

$$\widehat{\text{ASL}}_{\text{boot}} = \#\{t(\mathbf{x}^{*b}) \geq t_{obs}\}/B, \tag{16.3}$$

where $t_{obs} = t(\mathbf{x})$ the observed value of the statistic.

HT with the bootstrap

Mouse data

```
trt <- c(94,197,16,38,99,141,23)
ctrl <- c(52,104,146,10,51,30,40,27,46)
sur <- c(trt, ctrl)

obs.diff.means <- mean(sur[1:7]) - mean(sur[8:16])

set.seed(1234)
boot_sam <- matrix(sample(sur, 16*5000, replace = TRUE), nrow = 5000 )

diff.means <- apply(boot_sam, 1,
                    function(x) {mean(x[1:7]) - mean(x[8:16])})

# (bootstrap based on difference in sample means):
length(diff.means[diff.means >= obs.diff.means])/5000
```

```
## [1] 0.1244
```

HT with the bootstrap

An alternative bootstrap test

More accurate testing can be obtained through the use of a studentized statistic. Instead of $t(\boldsymbol{x}) = \bar{\boldsymbol{z}} - \bar{\boldsymbol{y}}$, one can use

$$t(\boldsymbol{x}) = \frac{\bar{\boldsymbol{z}} - \bar{\boldsymbol{y}}}{\sqrt{\bar{\sigma}^2(1/n + 1/m)}}$$

where $\bar{\sigma}^2 = \frac{\sum_{i=1}^{n}(z_i - \bar{\boldsymbol{z}})^2 + \sum_{i=1}^{m}(y_i - \bar{\boldsymbol{y}})^2}{n+m-2}$.

HT with the bootstrap

With studentized statistics

```
stud.t <- function(x) {
    num <- mean(x[1:7]) - mean(x[8:16])
    pool.var <- (var(x[1:7])*6 + var(x[8:16])*8)/14
    den <- sqrt(pool.var*(1/7 + 1/9))
    return(num/den)
}
# observed Student's t statistic:
obs.stud.t <- stud.t(sur)

new.t <- apply(boot_sam, 1, stud.t)
# one-sided test p-value (bootstrap based on Student's t):
length(new.t[new.t >= obs.stud.t])/5000
```

```
## [1] 0.14
```

In this calculation, we used exactly the same set of bootstrap samples that gave the value 0.126. Unlike in the permutation test that studentization does not affect the answer, studentization does produce a different value for $\widehat{\mathsf{ASL}}_{\mathsf{boot}}$.

HT with the bootstrap

What if we wanted to test only whether their means were equal?

Use bootstrapping to do a test about the means without assuming that both distributions have the same shape.

- To proceed we need estimates of F and G that use only the assumption of a common mean.
- We can shift the two empirical distributions (based on the two observed samples) so that they each have a mean equal to the sample mean of the combined sample (so that the two shifted empirical distributions have the same mean).
- Resample from these shifted empirical distributions.
- We can either do a bootstrap test based on the difference in two sample means, or we can do a test based on Welch's statistic (since there is no reason to assume that the two underlying distributions have the same variance).

HT with the bootstrap

Algorithm 16.2

Computation of the bootstrap test statistic for testing equality of means

1. Let $\hat{F}$ put equal probability on the points $\tilde{z}_i = z_i - \bar{z} + \bar{x}, i = 1, 2, \ldots n$, and $\hat{G}$ put equal probability on the points $\tilde{y}_i = y_i - \bar{y} + \bar{x}$, $i = 1, 2, \ldots m$, where $\bar{z}$ and $\bar{y}$ are the group means and $\bar{x}$ is the mean of the combined sample.
2. Form B bootstrap data sets $(\mathbf{z}^*, \mathbf{y}^*)$ where $\mathbf{z}^*$ is sampled with replacement from $\tilde{z}_1, \tilde{z}_2, \cdots \tilde{z}_n$ and $\mathbf{y}^*$ is sampled with replacement from $\tilde{y}_1, \tilde{y}_2, \ldots \tilde{y}_m$.
3. Evaluate $t(\cdot)$ defined by (16.5) on each data set,

$$t(\mathbf{x}^{*b}) = \frac{\bar{z}^* - \bar{y}^*}{\sqrt{\bar{\sigma}_1^{2*}/n + \bar{\sigma}_2^{2*}/m}}, \quad b = 1, 2, \cdots B. \qquad (16.6)$$

4. Approximate $\mathrm{ASL}_{\mathrm{boot}}$ by

$$\widehat{\mathrm{ASL}}_{\mathrm{boot}} = \#\{t(\mathbf{x}^{*b}) \geq t_{obs}\}/B, \qquad (16.7)$$

where $t_{obs} = t(\mathbf{x})$ is the observed value of the statistic.

Testing equality of means

```
t.shift <- trt - mean(trt) + mean(sur)
c.shift <- ctrl - mean(ctrl) + mean(sur)

bt <- matrix(sample(t.shift, 7*5000, replace = TRUE), nrow = 5000)
bc <- matrix(sample(c.shift, 9*5000, replace = TRUE ), nrow = 5000)
bsample <- cbind(bt, bc)

diff.var <- function(d){
   num <- mean(d[1:7]) - mean(d[8:16])
   denom <- sqrt( var(d[1:7])/7 + var(d[8:16])/9)
   return(num/denom)
}

sams <- apply(bsample , 1, diff.var)
obs.sams <- diff.var(sur)
# one-sided test p-value
length(sams[sams  >= obs.sams])/5000
```

```
## [1] 0.1514
```

Testing equality of means

For comparison purposes

```
# the Welch (or Satterthwaite) approximation to the degrees of freedom is used.
t.test(trt, ctrl, alternative = "greater", var.equal = FALSE )
```

```
##
##  Welch Two Sample t-test
##
## data:  trt and ctrl
## t = 1.0587, df = 9.6545, p-value = 0.1578
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -22.00168       Inf
## sample estimates:
## mean of x mean of y
##  86.85714  56.22222
```

Permutation test vs Bootstrap

- Permutation methods tend to apply to only a narrow range of problems. However when they apply, as in testing $F = G$ in a two-sample problem, they give gratifyingly exact answers without parametric assumptions.
- The bootstrap distribution was originally called the "combination distribution". It was designed to extend the virtues of permutation testing to the great majority of statistical problems where there is nothing to permute.
- The permutation ASL is exact, while the bootstrap ASL is approximated.

Read section 16.4 for more information.

One sample Hypothesis Testing

Suppose that other investigators have run experiments similar to ours but with many more mice, and they observed a mean lifetime of 129.0 days for treated mice. We might want to test whether the mean of the treatment group was 129.0 as well

$$H_0 : \mu_{\boldsymbol{z}} = 129.0.$$

- Assuming a normal population, under the null hypothesis, we would have

$$\bar{z} \sim N\left(129.0, \frac{\sigma^2}{n}\right).$$

- If σ^2 is unknown, we can estimate

$$\bar{\sigma} = \left\{ \sum_{i=1}^{n} (z_i - \bar{z})^2 / (n-1) \right\}^{1/2} = 66.8$$

- We base the bootstrap hypothesis test on the distribution of the test statistic

$$t(\boldsymbol{z}) = \frac{\bar{z} - 129.0}{\bar{\sigma}/\sqrt{7}}$$

under the null hypothesis.

What is the appropriate null distribution?

- We need a distribution F that estimates the population of treatment times under H_0.
- Note that the empirical distribution $\hat{F}$ is not an appropriate estimate for F because it does not obey H_0. That is, the mean of F is not equal to the null value of 129.0. Somehow we need to obtain an estimate of the population that has mean 129.0.
- We use as our estimated null distribution the empirical distribution on the values

$$\tilde{z}_i = z_i - \bar{z} + 129.0 = z_i - 86.9 + 129.0 = z_i + 42.1.$$

- We sample $\tilde{z}_1^*, \ldots, \tilde{z}_7^*$ with replacement from $\tilde{z}_1, \ldots, \tilde{z}_7$ and for each bootstrap sample compute the statistic

$$t(\tilde{\boldsymbol{z}}^*) = \frac{\bar{\tilde{z}}^* - 129.0}{\bar{\tilde{\sigma}}^*/\sqrt{7}},$$

  where $\bar{\tilde{\sigma}}^*$ is the standard deviation of the bootstrap sample.

- No assumption of normality. And the sampling distribution of $t(\tilde{\boldsymbol{z}}^*)$ does not generally follow a t-distribution.

Mouse Data

```
set.seed(1234)
mean(trt)
```

```
## [1] 86.85714
```

```
# t - test for comparison
t.test(trt, mu = 129)
```

```
##
##  One Sample t-test
##
## data:  trt
## t = -1.67, df = 6, p-value = 0.146
## alternative hypothesis: true mean is not equal to 129
## 95 percent confidence interval:
##   25.10812 148.60616
## sample estimates:
## mean of x
##  86.85714
```

```
# test statistic using the observed sample.
t.stat <- sqrt(7)*(mean(trt) - 129)/sd(trt)

t.shift <- trt - mean(trt) + 129
bootsample <- matrix( sample(t.shift, 7*5000, replace = TRUE ), nrow = 5000)
mean.boot <- apply(bootsample, 1, mean)
sd.boot <- apply(bootsample, 1, sd)
t.boot <- sqrt(7)*(mean.boot - 129)/sd.boot

lwd <- length(t.boot[t.boot <= t.stat])/5000
upr <- length(t.boot[t.boot >= t.stat])/5000
(p.value <- 2*min(lwd, upr))
```

```
## [1] 0.1912
```

Mouse Data

```r
set.seed(1234)

# t - test for comparison
t.test(trt, mu = 129, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  trt
## t = -1.67, df = 6, p-value = 0.07298
## alternative hypothesis: true mean is less than 129
## 95 percent confidence interval:
##      -Inf 135.8942
## sample estimates:
## mean of x
##  86.85714
```

```r
# test statistic using the observed sample.
t.stat <- sqrt(7)*(mean(trt) - 129)/sd(trt)

t.shift <- trt - mean(trt) + 129
bootsample <- matrix( sample(t.shift, 7*5000, replace = TRUE ), nrow = 5000)
mean.boot <- apply(bootsample, 1, mean)
sd.boot <- apply(bootsample, 1, sd)
t.boot <- sqrt(7)*(mean.boot - 129)/sd.boot

lwd <- length(t.boot[t.boot <= t.stat])/5000
upr <- length(t.boot[t.boot >= t.stat])/5000
(p.value <- min(lwd, upr))
```

```
## [1] 0.0956
```

Your turn

Recall the old faithful data set available in **MASS** package. This is a classic data set containing the time between 299 eruptions of the Old Faithful geyser in Yellowstone, and the length of the subsequent eruptions.

To check if the waiting time is 1 hour $= 60$ minutes: $H_0 : \mu = 60$ vs $H_A : \mu \neq 60$.

NOTE: The mean of waiting variable should be exactly 60.

```
library(MASS)
data(geyser)
```