STAT 641: BOOTSTRAPPING METHODS

Jiyoun Myung

Department of Statistics and Biostatistics
California State University, East Bay

Spring 2021, Day 11

Chapter 15: Permutation tests

Introduction

Permutation tests

- It is a computer-intensive statistical technique that predates computers.
- The idea was introduced by R.A. Fisher in the 1930's, more as a theoretical argument supporting Student's t-test than as a useful statistical method in its own right.
- Simple, intuitive, widely applicable, and free (light) of mathematical assumptions.
- There is a close connection with the bootstrap.

The two-sample problem

The main application of permutation tests is to the two-sample problem. (But, it can be used for a one-sample test of location.)

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.

The two-sample problem

A Traditional Hypothesis Test

Data: Mouse Data

Group	*Survival time (in days)*									Mean
Treatment	94	197	16	38	99	141	23			86.86
Control	52	104	146	10	51	30	40	27	46	56.22

Type I error & Type II error

		Decision	
		Reject H_0	Fail to Reject H_0
TRUTH	H_0 is True	Erroneous Decision (Type I error, α)	Correct Decision
	H_0 is False	Correct Decision	Erroneous Decision (Type II error, β)

- α: the level of the test.
- $1-\beta$: the power of the test.

p-value

Achieved significance level (ASL): p-value

Having observed test statistic $\hat{\theta}$, the achieved significance level of the test, abbreviated ASL, is defined to be the probability of observing a more extreme value for the test statistic than the one actually observed when the null hypothesis is true,

$$\text{ASL} = \text{Prob}_{H_0}\{\hat{\theta}^* \geq \hat{\theta}\}.$$

- The random variable $\hat{\theta}^*$ has the null hypothesis distribution.
- The smaller the value of ASL, the stronger the evidence against H_0.
- We reject H_0 if ASL is less than α.
- Less formally, we observe ASL and rate the evidence against H_0 according to the following rough conventions:

ASL $< .10$	borderline evidence against H_0.
ASL $< .05$	reasonably strong evidence against H_0.
ASL $< .025$	strong evidence against H_0.
ASL $< .01$	very strong evidence against H_0.

Traditional hypothesis tests

The main practical difficulty with traditional hypothesis tests

- Calculating the ASL, $\text{Prob}_{H_0}\{\hat{\theta}^* \geq \hat{\theta}\}$ as if the null hypothesis H_0 specifies a single distribution, but not.
- In order to calculate the ASL, we had to either approximate the null hypothesis variance, or use Student's t method.

Permutation Test

Fisher's permutation test

- If the null hypothesis is correct, any of the survival times for any of the mice could have come equally well from either of the treatments.
- Combine all the $m + n$ observations from both groups together, then take a sample of size m **without replacement** to represent the first group; the remaining n observations constitute the second group.
- Compute the difference between group means and then repeat this process a large number of times.
- If the original difference in sample means falls outside the middle 95% of the distribution of differences, the two-sided permutation test rejects the null hypothesis at a 5% level.

Permutation Test

Fisher's permutation test

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

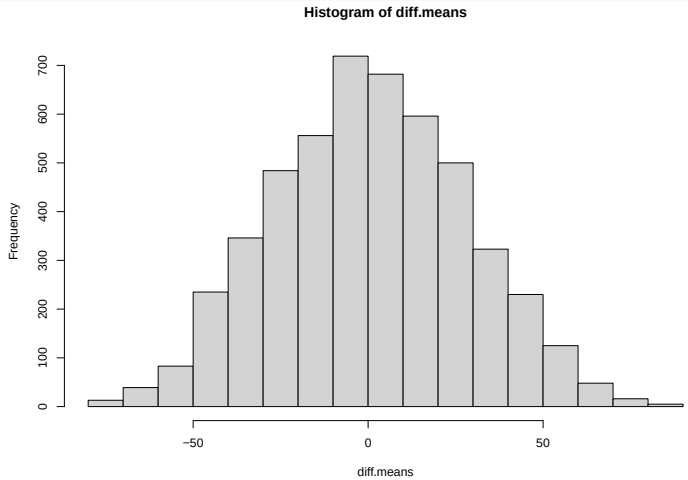
```
obs.diff.means <- mean(trt) - mean(ctrl)

surv <- c(trt, ctrl)
diff.means <- numeric()
set.seed(123)
for ( i in 1:5000 ) {
  pm <- sample(surv, 16, replace = FALSE)
  diff.means[i] <- mean(pm[1:7]) - mean(pm[8:16])
}
```

Mouse Data

Mean differences

Permutation Test

```
hist(diff.means)
```




```
length(diff.means[ diff.means >= obs.diff.means])/5000
```

```
## [1] 0.1434
```

Infer package

specify()

```
library(tidyverse)
library(infer)
library(moderndive)

id <- as.factor(c(rep("x", length(trt)), rep("y", length(ctrl))))
mouse <- tibble(id, surv)

null_dist <- mouse %>%
  specify(formula = surv ~ id)
null_dist
```

```
## Response: surv (numeric)
## Explanatory: id (factor)
## # A tibble: 16 x 2
##     surv id
##    <dbl> <fct>
##  1    94 x
##  2   197 x
##  3    16 x
##  4    38 x
##  5    99 x
##  6   141 x
##  7    23 x
##  8    52 y
##  9   104 y
## 10   146 y
## 11    10 y
## 12    51 y
## 13    30 y
## 14    40 y
```

Infer package

hypothesize()

- hypothesize(null = "point") : for hypotheses involving a single sample
- hypothesize(null = "independence") : for hypotheses involving two samples

```
null_dist <- mouse %>%
  specify(formula = surv ~ id) %>%
  hypothesize(null = "independence") # for hypotheses involving two samples
null_dist
```

```
## Response: surv (numeric)
## Explanatory: id (factor)
## Null Hypothesis: independence
## # A tibble: 16 x 2
##     surv id
##    <dbl> <fct>
##  1    94 x
##  2   197 x
##  3    16 x
##  4    38 x
##  5    99 x
##  6   141 x
##  7    23 x
##  8    52 y
##  9   104 y
## 10   146 y
## 11    10 y
## 12    51 y
## 13    30 y
## 14    40 y
## 15    27 y
## 16    46 y
```

Infer package

generate()

```
null_dist <- mouse %>%
  specify(formula = surv ~ id) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 5000, type = "permute") # w/o replacement
null_dist
```

```
## Response: surv (numeric)
## Explanatory: id (factor)
## Null Hypothesis: independence
## # A tibble: 80,000 x 3
## # Groups:   replicate [5,000]
##      surv id    replicate
##     <dbl> <fct>     <int>
##  1    23 x             1
##  2    52 x             1
##  3    30 x             1
##  4    40 x             1
##  5    51 x             1
##  6    94 x             1
##  7    99 x             1
##  8   197 y             1
##  9    27 y             1
## 10    38 y             1
## # ... with 79,990 more rows
```

Infer package

calculate()

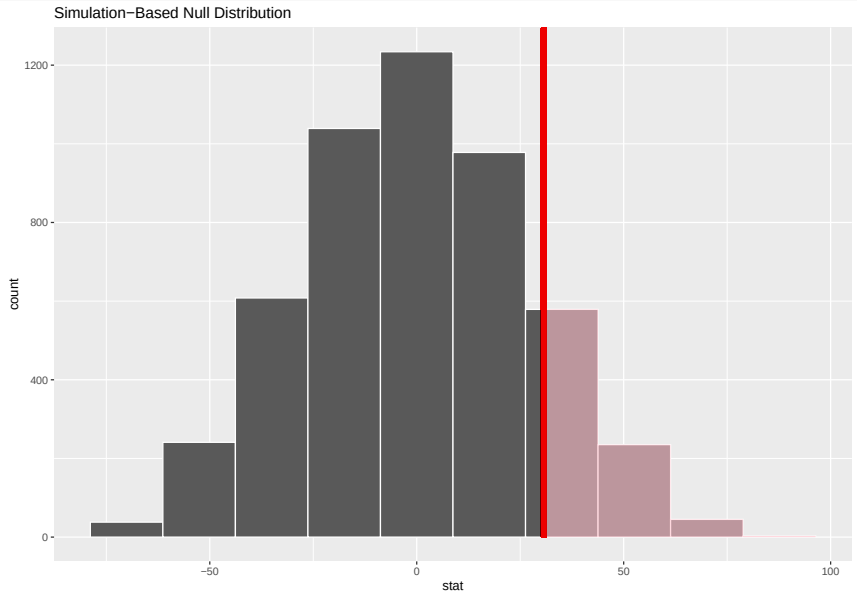
```
null_dist <- mouse %>%
  specify(formula = surv ~ id) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 5000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("x", "y"))
null_dist
```

```
## # A tibble: 5,000 x 2
##    replicate    stat
##        <int>   <dbl>
##  1         1 -46.8
##  2         2 -26.5
##  3         3 -13.8
##  4         4  13.6
##  5         5  17.2
##  6         6 -31.1
##  7         7 -25.7
##  8         8 -15.1
##  9         9   0.413
## 10        10  11.6
## # ... with 4,990 more rows
```

Infer package

visualize p-value

```
obs.diff.means <- mean(trt) - mean(ctrl)
visualize(null_dist, bins = 10) +
  shade_p_value(obs_stat = obs.diff.means, direction = "right")
```



Infer package

visualize p-value

```
null_dist %>%
  get_p_value(obs_stat = obs.diff.means, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.134
```

The permutation ASL is close to the t-test ASL, even though there are no normality assumptions. Fisher demonstrated a close theoretical connection between the permutation test based on $\bar{z} - \bar{y}$, and Student's test. (See Problem 15.9).

Permutaion Test

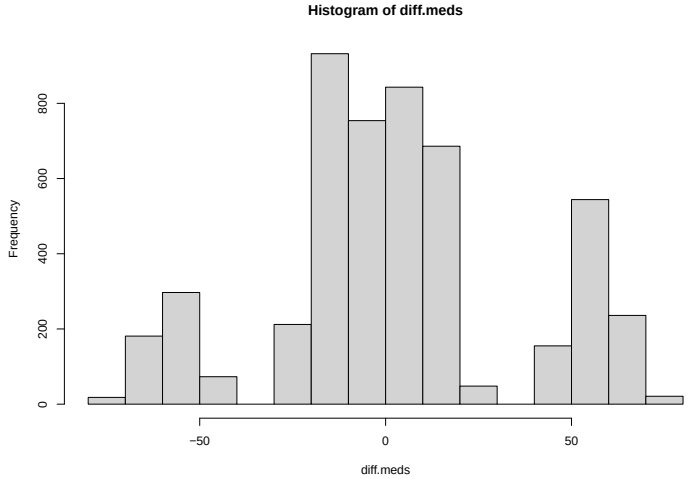
Other Statistics: Median

Median differences

```
obs.diff.meds <- median(trt) - median(ctrl)

diff.meds <- numeric()
set.seed(123)
for (i in 1:5000) {
  pm <- sample(surv, 16, replace = FALSE)
  diff.meds[i] <- median(pm[1:7]) - median(pm[8:16])
}
length(diff.meds[diff.meds >= obs.diff.meds])/5000
```

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

```
hist(diff.meds)
```



How many permutations?

- Table 15.2 in the textbook.
- Suggestions: start with 1000 permutations and continue to larger numbers only if p-value is small enough to be interesting, for example, p value < 0.1.
- Parallel computing of permutations will be useful.

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.

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

```
sur <- c(trt, ctrl)

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

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] 0.126
```

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] 0.1438
```

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.