# Bootstrap $p$-values

BIOS 6618

CU Anschutz

# Bootstrap Review

# Refresher: Two-Sample Bootstrap (Case Resampling)

Bootstrap sampling *mimics how the data were obtained*. For an experiment designed to compare two populations, we randomly take a sample *from each*. Hence, the bootstrap sample will mimic this process:

Given independent samples of sizes $m$ and $n$ from two populations,

1. Draw a resample of size $m$ with replacement from the first sample and a separate resample of size $n$ with replacement from the second sample. Compute a statistic that compares the two groups, such as the difference between the two sample means.
2. Repeat this resampling process many times, say 10,000.
3. Construct the bootstrap distribution of the statistic. Inspect its spread, bias, and shape.

# What Bootstraps Can Do

In our past lectures we discussed how bootstraps can be used to:

1. Estimate the standard errors for our estimators
2. Construct confidence intervals (e.g., bootstrap percentile or normal percentile) for unknown parameters

Another potential use of bootstraps is to:

3. Calculate $p$-values for test statistics *under a null hypothesis* (i.e., a new assumption we didn't need to make for 1/2 above)

# Resampling Under the Null Hypothesis

In order to estimate a $p$-value (i.e., the probability of observing something as or more extreme than what we observe assuming the null hypothesis is true), we need to resample under a null distribution.

Producing a null distribution with bootstrap testing can be somewhat nuanced and depends on the test statistic of interest. This will be discussed in terms of comparing two group means in the next section.

# Comparison with Permutation Testing

If you have a slight case of déjà vu with regards to "resampling" and "null distribution", you may be thinking about a **permutation test**.

Permutation tests sample *without replacement* from the pooled dataset (e.g., combining all $X_m$ and $Y_n$) and randomly assigning labels for belonging to $X$ or $Y$ and this process always generates the null distribution.

However, permutation tests are limited to contexts where groups under $H_0$ have the same distribution (i.e., are exchangeable). This may be true in randomized studies, but may require stronger assumptions in other settings.

# Bootstrap $p$-value Approaches

# Motivating Context

As a motivating context of introducing $p$-value definitions for bootstraps, assume our test statistic is the difference in group means such that $T = \mu_X - \mu_Y$.

Suppose that $T_0 = \bar{X} - \bar{Y}$ is the value of a test statistic $T$ computed for a particular sample. Then let $T_1^*, \cdots, T_B^*$ represent estimates from $B$ bootstrap resamples from the null distribution.

# Estimating One-Tailed Bootstrap $p$-value

If $H_0: \ \mu_X - \mu_Y \leq 0$ and $H_1: \ \mu_X - \mu_Y > 0$ (i.e., large values of $T$ support the alternative), then our one-sided $p$-value is:

$$P(T \geq T_0 | H_0) = p_B = \frac{\{\# \text{ of } T_i^* \geq T_0\}}{B}$$

Equivalently, if $H_0: \ \mu_X - \mu_Y \geq 0$ and $H_1: \ \mu_X - \mu_Y < 0$, then our one-sided $p$-value is:

$$P(T \leq T_0 | H_0) = p_B = \frac{\{\# \text{ of } T_i^* \leq T_0\}}{B}$$

Note, some add 1 to the numerator and denominator to avoid the possibility of having a $p$-value of 0.

# Estimating Two-Tailed Bootstrap $p$-value

Assume $T_0 > 0$, then we can calculate the probability of observing something as *or more extreme* in each tail as $P(T \geq T_0|H_0)$ (i.e., what we observed) and $P(T \leq -T_0|H_0)$ (i.e., as or more extreme in the other tail).

For two-sided $p$-values, multiple strategies exist in our bootstrap distribution (in order from most to least conservative):

1. Multiply max by 2: $p_B = 2 \times \max[P(T \geq T_0|H_0), P(T \leq -T_0|H_0)]$
2. Add the two tails together: $p_B = P(T \geq T_0|H_0) + P(T \leq -T_0|H_0)$
3. Multiply min by 2: $p_B = 2 \times \min[P(T \geq T_0|H_0), P(T \leq -T_0|H_0)]$

Another strategy based on a one-sided $p$-value would be:

4. Multiply the min one-sided $p$-value by 2: $p_B = 2 \times \min[P(T \geq T_0|H_0), P(T < T_0|H_0)]$

Options 1/3/4 assume symmetry of our distribution (i.e., the probability should be the same in the upper and lower tail), whereas option 2 provides the most flexibility.

# 99 Rule (How Many Bootstraps)

When determining the number of bootstrap resamples, $B$, to use for estimating a $p$-value, Boos & Stefanski[1] recommend following the "99 Rule" of resampling $B = 19$, $B = 99$, $B = 999$, etc. *Why 99?*

Under $H_0$, we know $p$-values are uniformly distributed. Therefore, we have $B + 1$ possibilities (e.g., anywhere from 0 to all $B$ resamples meet our criteria to be counted). If we have 99 or 100 $B$, then $\frac{1}{B+1}$:

```
##          0          1          2          3          4          5          6
## B=99   0 0.01010101 0.02020202 0.03030303 0.04040404 0.05050505 0.06060606
## B=100  0 0.01000000 0.02000000 0.03000000 0.04000000 0.05000000 0.06000000
```

For $B = 99$, we see that $P(p_B \leq 0.05) = \frac{5}{100} = 0.05$, but for $B = 100$ we have $P(p_B \leq 0.05) = \frac{6}{101} = 0.059$.

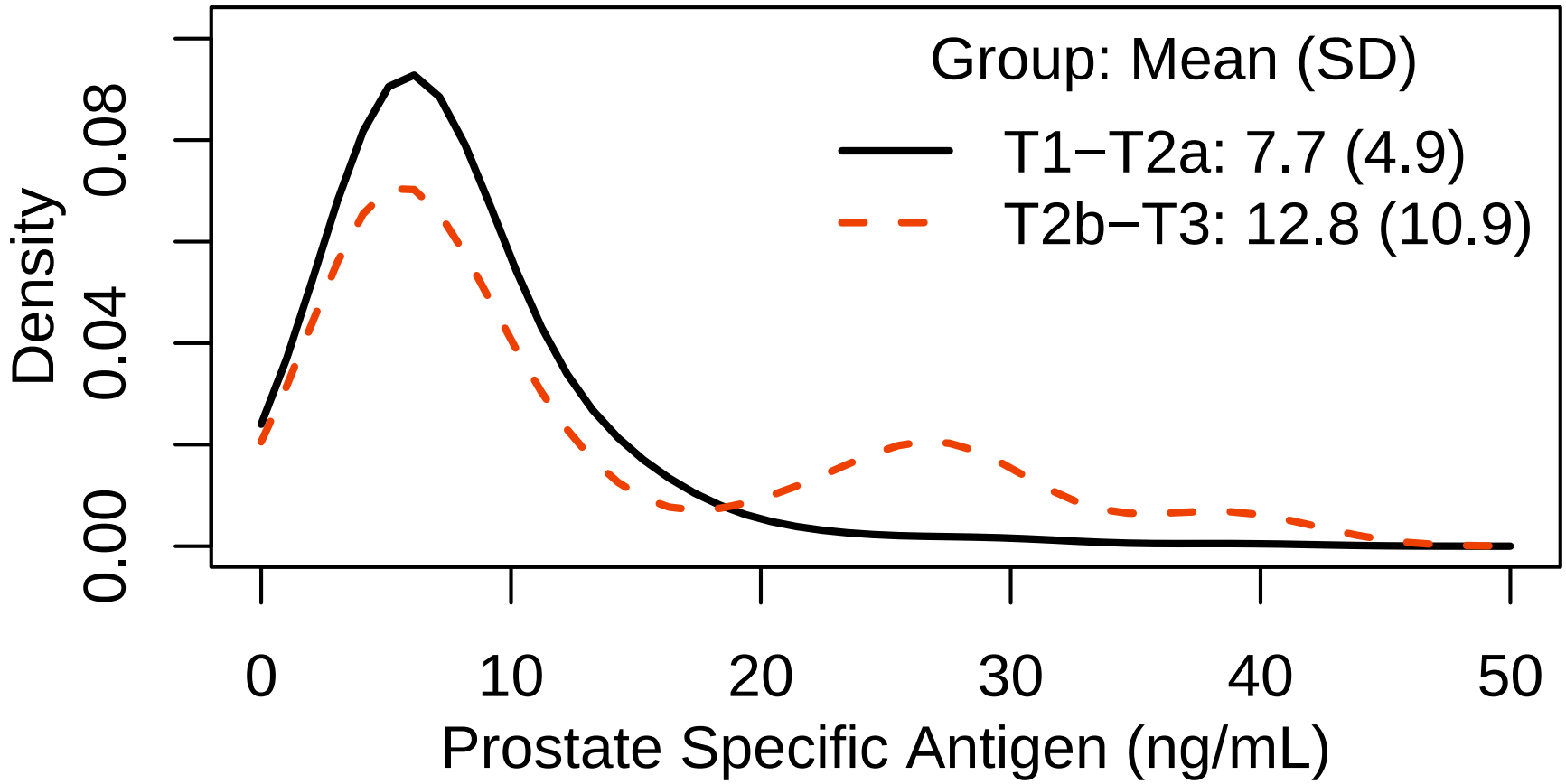
So, to maintain $\alpha$, especially with small $B$, we should follow the 99 rule.

[1] Boos, D. D., & Stefanski, L. A. (2013). *Essential statistical inference: theory and methods*, Ch. 11. New York: Springer.

# Bootstrap $p$-value Example

# Example Introduction

From the TSHS[2] Blood Storage dataset, we will explore if there is a mean difference in preoperative prostate specific antigen (PSA; ng/mL) between tumor stages (T1-T2a [$X$, $m$=267] versus T2b-T3 [$Y$, $n$=34]):




[2] https://www.causeweb.org/tshs/category/dataset/

# Test Statistics Evaluated

We will consider three different test statistics to estimate through our bootstraps:

1. $d = \bar{X} - \bar{Y}$
   - Most similar to what we may sample for bootstraps to estimate SE or CI
   - Doesn't account for potential differences in SEs between groups
   - Doesn't align with test statistics we typically think of with calculating $p$-values
2. $t_p = \frac{\bar{X}-\bar{Y}}{\sqrt{s_p^2(\frac{1}{m}+\frac{1}{n})}}$ where $s_p^2 = \frac{(m-1)s_X^2+(n-1)s_Y^2}{m+n-2}$
   - Reflects a two-sample t-test assuming equal variance
   - Doesn't account for potential differences in SEs between groups
3. $t_w = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{s_X^2}{m}+\frac{s_Y^2}{n}}}$
   - Reflects a two-sample t-test allowing unequal variances
   - Most flexible while accounting for SEs

# Null Distribution Approaches

We will explore two approaches for bootstrapping a null distribution proposed by Boos and Stefanski[3]:

1. Resampling with replacement $m$ cases for $X$ and $n$ cases for $Y$ from a pooled set of $X_1, \cdots, X_m, Y_1, \cdots, Y_n$.
   - Creates a common overall mean but assumes identical variance.
   - Can test $H_0$: $F(t) = G(t)$, where $F$ and $G$ are the distribution functions of $X$ and $Y$, respectively.
2. Resampling with replacement $m$ cases for $X$ from $X_1 - \bar{X}, \cdots, X_m - \bar{X}$ and $n$ cases for $Y$ from $Y_1 - \bar{Y}, \cdots, Y_n - \bar{Y}$.
   - Centers each group at mean 0, but allows different variances.
   - Can test $H_0$: $\mu_X = \mu_Y$ (i.e., more general).

In general, for each statistic, the appropriate null distribution may vary and will need to be thoughtfully considered.

[3] Boos, D. D., & Stefanski, L. A. (2013). *Essential statistical inference: theory and methods*, Ch. 11.6. New York: Springer.

# Two-Sided Test Bootstrap Code in R - I

Using $B = 999$ (i.e., 99 Rule):

```
set.seed(6618) # set seed for reproducibility
B <- 9999 # set number of bootstraps

# create objects with relevant data
X <- dat$PreopPSA[ which(dat$T.Stage==1) ]
m <- length(X)
Y <- dat$PreopPSA[ which(dat$T.Stage==2) ]
n <- length(Y)

# calculate observed test statistics
T_0_d <- mean(X) - mean(Y) # difference
T_0_tp <- t.test(X, Y, var.equal=T)$statistic # t_p
T_0_tw <- t.test(X, Y, var.equal=F)$statistic # t_w
```

# Two-Sided Test Bootstrap Code in R - II

The following code implements strategy 1 using the pooled null:

```
## resampling strategy 1: null from pooled data
pool <- c(X,Y) # create pooled object
diff_pool_d <- diff_pool_tp <- diff_pool_tw <- rep(NA,B)

for (i in 1:B){
  X_boot <- sample(pool, size=m, replace=T)
  Y_boot <- sample(pool, size=n, replace=T)

  # calculate & save test statistics
  diff_pool_d[i] <- mean(X_boot)-mean(Y_boot)
  diff_pool_tp[i] <- t.test(X_boot,Y_boot,var.equal=T)$statistic
  diff_pool_tw[i] <- t.test(X_boot,Y_boot,var.equal=F)$statistic
}
```

# Two-Sided Test Bootstrap Code in R - III

The following code implements strategy 2 using the centered groups:

```
## resampling strategy 2: null from centered groups
X_center <- X - mean(X) # center X for mean 0
Y_center <- Y - mean(Y) # center Y for mean 0
diff_center_d <- diff_center_tp <- diff_center_tw <- rep(NA,B)

for (i in 1:B){
  X_boot <- sample(X_center, size=m, replace=T)
  Y_boot <- sample(Y_center, size=n, replace=T)

  # calculate & save test statistics
  diff_center_d[i] <- mean(X_boot)-mean(Y_boot)
  diff_center_tp[i] <- t.test(X_boot,Y_boot,var.equal=T)$statistic
  diff_center_tw[i] <- t.test(X_boot,Y_boot,var.equal=F)$statistic
}
```

# Two-Sided Test Bootstrap Code in R - IV

The following code provides examples for each of the 4 $p$-value estimation strategies assuming the pooled data null (resampling strategy 1) for the $t_w$ test statistic with observed value in our sample of $T_{0,t_w}$ =-2.7:

```
# calculate lower and upper tail pooled
pr_pool_tw_low <- mean(diff_pool_tw <= T_0_tw)
pr_pool_tw_up <- mean(diff_pool_tw >= -T_0_tw)
c(pr_pool_tw_low, pr_pool_tw_up) # print results
```

```
## [1] 0.00040004 0.03350335
```

```
# p-value estimates
p_s1_pool_tw <- 2*max(pr_pool_tw_low, pr_pool_tw_up) #strategy 1
p_s2_pool_tw <- pr_pool_tw_low + pr_pool_tw_up #strategy 2
p_s3_pool_tw <- 2*min(pr_pool_tw_low, pr_pool_tw_up) #strategy 3
p_s4_pool_tw <- 2*min(pr_pool_tw_low, 1-pr_pool_tw_low) #strategy 4

c(p_s1_pool_tw, p_s2_pool_tw, p_s3_pool_tw, p_s4_pool_tw)
```

```
## [1] 0.06700670 0.03390339 0.00080008 0.00080008
```

# Two-Sided Test Bootstrap in R - Results

The following table summarizes the overall results for each statistic and sampling strategy. For comparison, the two-sample t-test p-value assuming equal variances is $<0.001$ and assuming unequal variance is 0.011.

| | *Pooled Null* | | | *Centered Null* | | |
|---|---|---|---|---|---|---|
| *p-value Approach* | $d$ | $t_p$ | $t_w$ | $d$ | $t_p$ | $t_w$ |
| 1 (2 x max): | 0 | 0 | 0.067 | 0.01 | 0.009 | 0.037 |
| 2 (add tails): | 0 | 0 | 0.034 | 0.006 | 0.008 | 0.021 |
| 3 (2 x min): | 0 | 0 | 0.001 | 0.002 | 0.008 | 0.004 |
| 4 (2 x one-sided): | 0 | 0 | 0.001 | 0.01 | 0.009 | 0.004 |

# Two-Sided Test Bootstrap in R - Results

| | *Pooled Null* | | | *Centered Null* | | |
|---|---|---|---|---|---|---|
| *p-value Approach* | $d$ | $t_p$ | $t_w$ | $d$ | $t_p$ | $t_w$ |
| 1 (2 x max): | 0 | 0 | 0.067 | 0.01 | 0.009 | 0.037 |
| 2 (add tails): | 0 | 0 | 0.034 | 0.006 | 0.008 | 0.021 |
| 3 (2 x min): | 0 | 0 | 0.001 | 0.002 | 0.008 | 0.004 |
| 4 (2 x one-sided): | 0 | 0 | 0.001 | 0.01 | 0.009 | 0.004 |

- In general, results are all $<0.05$, except for $p$-value approach 1 for $t_w$ under the pooled null.
- The unequal variance t-test may not be conservative enough ($p$=0.011) compared to bootstrap $p$-values from approaches 1 and 2.
- Given the data did not appear to have equal variance in each tumor stage group, the **centered null** approach may be more appropriate.
- Similarly, it may make more sense to evaluate $d$ or $t_w$ to avoid making assumptions about equal variances in our test statistic.
- Finally, we may wish to use approach 1 to be most conservative or approach 2 to avoid assuming symmetry in the tails.

# Bootstrap $p$-value Summary

This lecture expanded our use of the bootstrap from estimating the sampling distribution of a test statistic to calculate the SE or CI, to calculating $p$-values from a bootstrapped null distribution.

As we saw in the lecture, there are many nuanced decisions to consider with selecting an appropriate null distribution, how to calculate the $p$-value, and what statistic to use. While there is no one "correct" answer, it may be better to select more conservative $p$-value calculations to avoid type I errors (i.e., false positives).

These approaches are general and can be used to estimate $p$-values for other bootstrapping approaches. Additionally, we can always interpret the bootstrap CI to evaluate if our null hypothesis value falls within the interval to make a decision.