Resampling and the Bootstrap

Axel Benner
Biostatistics, German Cancer Research Center
INF 280, D-69120 Heidelberg
benner@dkfz.de

Topics

Estimation and Statistical Testing

- The Bootstrap
- Permutation Analysis

Predictive Ability using Resampling

- Data-Splitting/Cross-Validation
- The Bootstrap

Resampling methods

- Methods in which the observed data are used repeatedly, in a computer-intensive simulation analysis, to provide inferences.
- The original test statistic is considered unusual if it is unusual compared to the resampling distribution.
- Resampling methods considered here include the bootstrap method and permutation analysis.
- Other sample re-use methods such as jackknifing and data-splitting are used for sensitivity analysis or measuring predictive ability.

Simulation

- Approximations obtained by random sampling or simulation are called **Monte Carlo** estimates.

  Assume: Random variable Y has a certain distribution
  $\rightarrow$
  Use simulation or analytic derivations to study how an estimator, computed from samples from this distribution, behaves.

  e.g. Y has lognormal distribution $\Rightarrow$ standard error of the median?

  1. Analytical solution?

  2. Computational solution:
     Simulate B samples of size n from the lognormal distribution, compute the sample median for each sample, and then compute the sample variance of the 500 sample medians.

Example of 100 random deviates

```
n <- 100
set.seed(12345)
# one sample:
rlnorm(n,meanlog=0,sdlog=1)

# 500 samples:
y <- matrix(rlnorm(500*n,meanlog=0,sdlog=1),nrow=n,ncol=500)

ym <- apply(y, 2, median)
summary(ym)  # Distribution of 500 median values
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7446  0.9246  1.0060  1.0130  1.0910  1.4140
```

```
ys <- sd(ym)
print(ys) # standard error(median)
```

```
[1] 0.1222032
```

Example: Assume that we have a sample set y, given by

82	79	81	79	77	79	79	78	79	82	76	73	64

What is the standard error of the median of y?

Use the bootstrap:

```
library(boot)
y <- scan()
82 79 81 79 77 79 79 78 79 82 76 73 64

med <- function(d, i)  median(d[i])
b <- boot(y, med, R=999, stype="i")
```

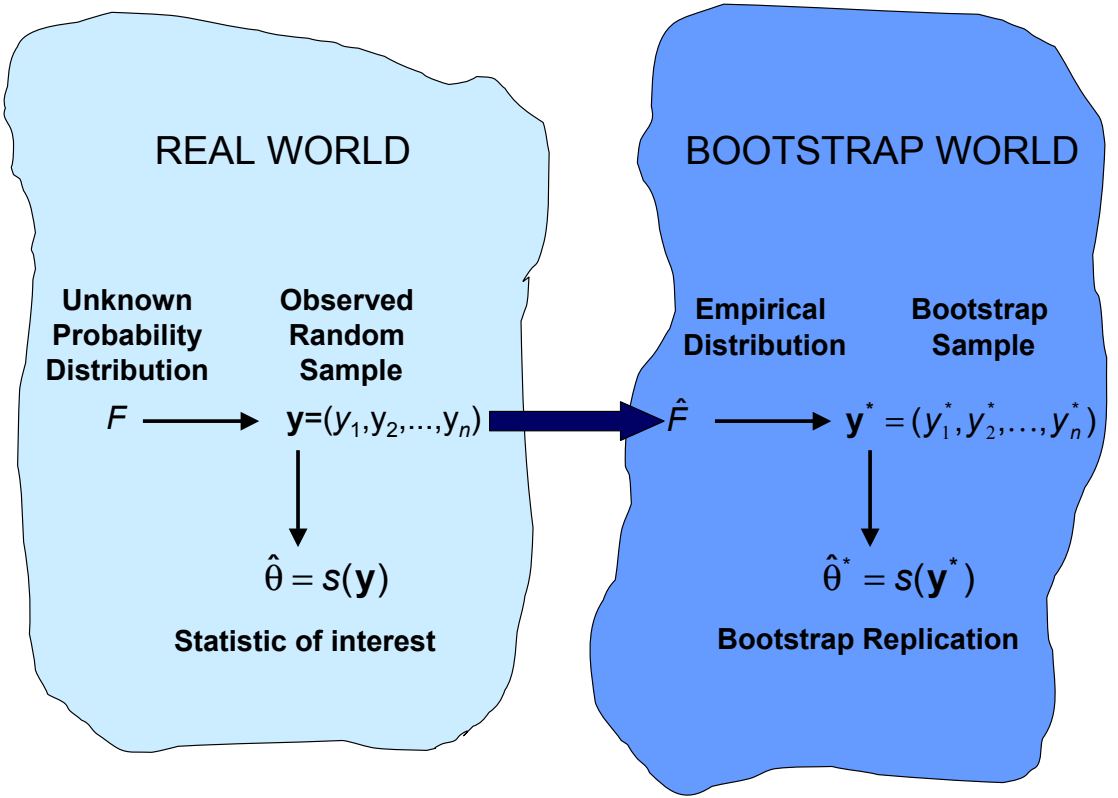
```
Bootstrap Statistics :
    original  bias    std. error
t1*       79  -0.226   0.7010103
```

The Bootstrap

Efron's bootstrap is a general purpose technique for obtaining estimates of properties of statistical estimators without making assumptions about the distribution of the data.

Often used to find

1. standard errors of estimates
2. confidence intervals for unknown parameters
3. p values for test statistics under a null hypothesis


REAL WORLD
Unknown Probability Distribution
Observed Random Sample
F
y=(y1,y2,...,yn)
θ̂ = s(y)
Statistic of interest
BOOTSTRAP WORLD
Empirical Distribution
Bootstrap Sample
F̂
y* = (y1*, y2*, ..., yn*)
θ̂* = s(y*)
Bootstrap Replication

Bootstrap Simulation

Suppose Y has a cumulative distribution function (cdf)
$F(y) = P(Y \leq y)$

We have a sample of size n from $F(y)$, $y_1, y_2, \ldots, y_n$

Steps:

1. Repeatedly simulate sample of size n from F
2. Compute statistic of interest
3. Study behavior of statistic over B repetitions

Bootstrap Simulation (again)

- $y_1, ..., y_n$ random sample from F.
- Estimation function T is given by $T = t(y_1, ..., y_n) \equiv t(\hat{F})$.
- Think of $t(\cdot)$ as an algorithm
  - applied to F gives parameter $\theta = t(F)$.
  - applied to $\hat{F}$ gives estimate $t = t(\hat{F})$.
- **Resample**:
  $y_1^*, ..., y_n^* \overset{i.i.d.}{\sim} \hat{F}$ giving $T^* = t(y_1^*, ..., y_n^*) \equiv t(\hat{F}^*)$.
- Repeat B times to get $t_1^*, ..., t_B^*$

Comments

- Without knowledge of F we use the empirical cdf $F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \leq y)$ as an estimate of F.

- Pretend that $F_n(y)$ is the original distribution $F(y)$.

- Sampling from $F_n(y)$ is equivalent to sampling with replacement from originally observed $y_1, ..., y_n$

- For large n the expected fraction of original data points that are selected for each bootstrap sample is 0.632

$$\begin{aligned} P(\text{obs. } i \in \text{ bootstrap sample } b) &= 1-(1-\frac{1}{n})^n \\ &\approx 1-e^{-1} \\ &= 0.632 \end{aligned}$$

Note: $1-\frac{1}{n}$ is probability for not being selected at a specific drawing; with n drawings we get that $(1-\frac{1}{n})^n$ is probability of not being selected at least once.

- From bootstrap sampling we can estimate any aspect of the distribution of $s(y)$ [which is any quantity computed from the data y], for example its standard error

$$\widehat{se}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^*(b) - \hat{\theta}^*(\cdot))^2 \right\}^{1/2}$$

where $\hat{\theta}^*(b) = s(y^{*b})$ is the bootstrap replication of $s(y)$ and $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$.

The Jackknife

- We have a sample $y = (y_1, ..., y_n)$ and estimator $\hat{\theta} = s(y)$.
- Target: Estimate the bias and standard error of $\hat{\theta}$.
- The leave-one-out observation samples

$$y_{(i)} = (y_1, ..., y_{i-1}, \quad y_{i+1}, ..., y_n)$$

  for $i = 1, ..., n$ are called **jackknife samples**.

- Jackknife estimators are $\hat{\theta}_{(i)} = s(y_{(i)})$.

- The jackknife estimate of bias is

$$\widehat{bias}_J \;=\; (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$$

  where $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$

- The jackknife estimate of the standard error is

$$\hat{se}_J \;=\; \sqrt{\frac{n-1}{n}\sum(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2}$$

- Use jackknife to measure the uncertainty of a bootstrap estimate of a statistic $s(y)$: *jackknife-after-bootstrap*
  $\rightarrow$ cp. function `jack.after.boot` from R package `boot`.

Sensitivity analysis (Jackknife after Bootstrap)

How different would the results have been if an observation y_j has been absent from the original data?

Measure effect of y_j on calculations by comparing full simulation with the subset of statistics $t_1^*, ..., t_B^*$ obtained from bootstrap samples without y_j.

Using frequencies f_{bj}^* counting the number of times y_j appears in the bth simulation we restrict to replicates with $f_{bj}^* = 0$.

$\Rightarrow$ Measure effect of y_j on the bias by scaled difference

$$n(bias_{-j} - bias) = \left\{ \frac{1}{B_{-j}} \sum_{b:f_{bj}^*=0} (t_b^* - t_{-j}) - \frac{1}{B} \sum (t_b^* - t) \right\}$$

t_{-j} is the value of t when y_j is excluded from the original data.

Hypothesis testing

- Null hypothesis (H_0), absence of some effect, vs. alternative hypothesis (H_1)
- Perform a level α test
  - within the context of a parametric model,
  - without a model
    (1) non-parametric test
    (2) permutation test
    (3) bootstrap test

The significance level of a test and the p-value

- The *significance level* (or size) α of a test is the probability of making a Type I error; that is, α is the probability of deciding erroneously on the alternative when, in fact, the hypothesis is true.
- The p-value is the chance of obtaining a test statistic as or more extreme (as far away from what we expected or even farther in the direction of the alternative) than the one we got, assuming the null hypothesis is true.
- This chance is called the observed significance level, or p-value.
- A test statistic with a p-value less than some prespecified false positive level (or size) α is said to be 'statistically significant' at that level.
- The p-value represents the probability that we would observe a difference as large as we saw (or larger) if there were really nothing happening other than chance variability.

Type I and type II error; power of a test

- Two types of error

		Decision	
		H_0	H_1
The Facts	H_0	-	Type I error α
	H_1	Type II error β	-

- The *power* $1-\beta$ of a test is 1 minus the probability of making a type II error; that is, $1-\beta$ is the probability of deciding on the alternative when the alternative is the correct choice.
- The ideal statistical test would have a significance level α of zero and a power of 1, but this ideal can not be realized.
- In practice, we will fix a significance level $\alpha > 0$ (usually this will be 0.05), and choose a statistic that maximizes or comes closest to maximizing the power $1-\beta$.

Exact test

- Often variance or distribution of a variable are unknown. Therefore we usually test a *compound hypothesis* such as $H_0 : \ mean(Y) = 0$, (Y can be normal with mean 0 and variance 0.5 or with variance 1, or has a gamma distribution with mean 0 and 4 degrees of freedom).
- A test is said to be *exact* with respect to a compound hypothesis if the probability of making a type I error is exactly α for each of the possibilities that make up the hypothesis.
- A test is said to be *conservative*, if the type I error never exceeds α.

Permutation tests

- Also called *randomization tests, rerandomization tests, exact tests.*
- Introduced by Fisher and Pitman in the 1930s.
- Usually require only a few weak assumptions.

  – underlying distributions are symmetric
  – the alternatives are shifts in value

  A preliminary rank transformation often can ensure that the tests are distribution-free.

Five steps to a Permutation Test

1. Choose a test statistic $s(y)$ which will distinguish the hypothesis from the alternative.

2. Compute the test statistic for the original set (labelling) of the observations.

3. Obtain the permutation distribution of s by rearranging observations.

4. Compute the test statistic for all possible rearrangements (permutations) of the observations.

5. Make a decision:
   Reject the null hypothesis if the value of the test statistic for the original labelling (original data) is an extreme value in the permutation distribution of the statistic. Otherwise, accept the null hypothesis.

Example: t test vs. permutation test

- data

X			Y		
A	B	C	D	E	F
121	118	110	34	12	22
$\overline{x}_n = 116.33$			$\overline{y}_n = 22.67$		

  t test statistic: $t = 13.0875$, two-sided p-value: $p = 0.0002$

- after one permutation:

X			Y		
A	B	D	C	E	F
121	118	34	110	12	22
$\overline{x}_n = 91$			$\overline{y}_n = 48$		

- how many permutations exist?

$$C_3^6 = \binom{6}{3} = \frac{6!}{3! \cdot 3!} = \frac{6 \cdot 5 \cdot 4}{1 \cdot 2 \cdot 3} = 20$$

permutation	X	Y	$\overline{x}_n$	$\overline{y}_n$	$\overline{x}_n - \overline{y}_n$	t
1	ABC	DEF	116.33	22.67	93.67	13.087
2	ABD	CEF	91.00	48.00	43.00	1.019
3	ABE	CDF	87.00	52.00	35.00	0.795
4	ABF	CDE	83.67	55.33	28.33	0.627
5	ACD	BEF	88.33	50.67	37.67	0.866
6	ACE	BDF	84.33	54.67	29.67	0.659
7	ACF	BDE	81.00	58.00	23.00	0.500
8	ADE	BCF	59.00	80.00	-21.00	-0.455
9	ADF	BCE	55.67	83.33	-27.67	-0.611
10	AEF	BCD	51.67	87.33	-35.67	-0.813
11	BCD	AEF	87.33	51.67	35.67	0.813
12	BCE	ADF	83.33	55.67	27.67	0.611
13	BCF	ADE	80.00	59.00	21.00	0.455
14	BDE	ACF	58.00	81.00	-23.00	-0.500
15	BDF	ACE	54.67	84.33	-29.67	-0.659
16	BEF	ACD	50.67	88.33	-37.67	-0.866
17	CDE	ABF	55.33	83.67	-28.33	-0.627
18	CDF	ABE	52.00	87.00	-35.00	-0.795
19	CEF	ABD	48.00	91.00	-43.00	-1.019
20	DEF	ABC	22.67	116.33	-93.67	-13.087

- Test decision: In two of 20 cases overall the absolute value of the test statistic t is greater than or equal to the absolute value of $t = 13.087$ we obtained for the original labelling.

  Therefore we obtain the exact p value $p = 2/20 = 0.1$.

- Note: 0.1 is the smallest p value you can get for comparing two groups of size 3.
- Note: If both groups have equal size only half of permutations is really needed (symmetry)
- Note: The number of permutations for comparing two groups of size m and $n - m$ is

$$C_m^n = \binom{n}{m} = \frac{n!}{m! \cdot (n-m)!}$$

  e.g. for $n = 52$ and $m = 18$

$$C_{18}^{52} = \binom{52}{18} = \frac{52!}{18! \cdot 34!} = 4.27 \times 10^{13}$$

- It may be necessary to use Monte Carlo sampling to approximate the permutation test

Microarray Data: Resampling in multiple testing

Estimate the joint distribution of the test statistics $T_1, \ldots, T_G$ under the complete null hypothesis H_0^C by permuting the columns of the $(G \times n)$ gene expression data matrix $\mathbf{X}$.

Permutation algorithm for non-adjusted p-values

- For the b-th permutation, $b = 1, \ldots, B$
  1. Permute the n columns of the data matrix $\mathbf{X}$.
  2. Compute test statistics $t_{1,b}, \ldots, t_{G,b}$ for each hypothesis.
- The permutation distribution of the test statistic T_g for hypothesis H_g, $g = 1, \ldots, G$, is given by the empirical distribution of $t_{g,1}, \ldots, t_{g,B}$. For two-sided alternative hypotheses, the permutation p-value for hypothesis H_g is

$$p_g^* = \frac{1}{B} \sum_{b=1}^{B} I(|t_{g,b}| \geq |t_g|)$$

where $I(.)$ is the indicator function, equaling 1 if the condition in parenthesis is true, and 0 otherwise.

Permutation algorithm of Westfall & Young (maxT)

- Order observed test statistics: $|t_{r_1}| \geq |t_{r_2}| \geq \ldots \geq |t_{r_G}|$.
- For the b-th permutation of the data ($b = 1, \ldots, B$):
  - divide the data into its artificial control and treatment group
  - compute test statistics $t_{1,b}, \ldots, t_{G,b}$
  - compute successive maxima of the test statistics

$$\begin{aligned} u_{G,b} &= |t_{r_G,b}| \\ u_{g,b} &= \max\{u_{g+1,b}, |t_{r_g,b}|\} \text{ for } g = G-1, \ldots, 1 \end{aligned}$$

- compute adjusted p-values:

$$\tilde{p}^*_{r_g} = \frac{1}{B} \sum_{b=1}^{B} I(u_{g,b} \geq |t_{r_g}|)$$

Permutation algorithm of Westfall & Young – Example

gene	|t|	
1	0.1	t_{r_G}
4	0.2	$t_{r_{G-1}}$
5	2.8	:
2	3.4	t_{r_2}
3	7.1	t_{r_1}

sort observed values

gene	$\|t_b\|$	u_b	$I(u_b>\|t\|)$
1	1.3	1.3	1
4	0.8	1.3	1
5	3.0	3.0	1
2	2.1	3.0	0
3	1.8	3.0	0

B=1000 permutations

$\sum$	$\tilde{p} = \sum / B$
935	0.935
876	0.876
138	0.138
145	0.145
48	**0.048**

adjusted p-values

O. Hartmann - NGFN Symposium, 19.11.2002 Berlin

Nonparametric Bootstrap Tests

- permutation tests are special nonparametric resampling tests, in which resampling is done without replacement
- the special nature of significance tests requires that probability calculations be done under a null hypothesis model, that means we must resample from a distribution $\hat{F}_0$, say, which satisfies the relevant null hypothesis H_0
- the basic bootstrap test will be to compute the p-values as

$$p_{boot} = P^*(T^* \geq t | \hat{F}_0)$$

approximated by

$$p = \frac{1}{B} \sum_{b=1}^{B} I(t_b^* \geq t)$$

using the results $t_1^*, t_2^*, \ldots, t_B^*$ from B bootstrap samples

Example: Comparison of population means

$$H_0 : \mu_x = \mu_y \textbf{ vs. } H_A : \mu_x \neq \mu_y$$

- If the shapes of the underlying distributions are identical, then the two distributions are the same under H_0.
- Choose for $\hat{F}_0$ the pooled empirical distribution function of the two samples.
- the bootstrap test will be the same as the permutation test, except that random permutations will be replaced by random samples of size $n_x + n_y$ drawn **with replacement** from the pooled data

Bootstrap test for testing $F = G$ (Efron &Tibshirani, 1993)

1. Draw B samples of size $n_x + n_y$ with replacement from the pooled data of the original sets x and y. Call the first n_x observations of the bootstrap sample x^* and the remaining y^*.
2. Evaluate $t(\cdot)$ on each sample, e.g.

$$t(x^{*b}) = \bar{x}^* - \bar{y}^*, \quad b = 1, ..., B$$

3. Compute bootstrap p value by

$$\hat{p} = \#\{t(x^{*b}) \geq t_{obs}\}/B$$

where $t_{obs} = \bar{x} - \bar{y}$ is the observed value of the test statistic on the original data sets.

(Monte Carlo) Permutation vs. Bootstrap Resampling

- In MC sampling one samples values of the test statistic from its underlying permutation distribution
- In Bootstrapping there are two sources of error:
  1. Error caused by resampling from an empirical cumulative distribution function formed from the initial data set.
  2. Error caused from by carrying out only a finite number of resamples.
- For messier problems when the test statistic has a complicated analytically intractible distribution the bootstrap can provide a reasonable answer while the permutation test may not work.
- Permutation methods only apply in a narrow range of problems. When they apply, as in testing $F = G$ in two-sample problems, they give "exact" answers without parametric assumptions.

(Monte Carlo) Permutation vs. Bootstrap Resampling (cont)

An example comparing the location of two distributions by one-sided tests:

```
x <- scan()
16 23 38 94 99 141 197

y <- scan()
10 27 31 40 46 50 52 104 146
```

The observed test statistic $\bar{x} - \bar{y}$ is

```
mean(x)-mean(y)

[1] 30.63492
```

(Monte Carlo) Permutation vs. Bootstrap Resampling (cont)

We want to compute $P(\bar{X} - \bar{Y} \geq 30.63 | F = G)$. The permutation test is done using $16!/(7!9!) = 11440$ partitions of the 16 cases into two groups of 9 and 7, respectively.

```
library(exactRankTests)
perm.test(x, y, alter="greater")

        2-sample Permutation Test

data:  x and y
T = 608, p-value = 0.1406
alternative hypothesis: true mu is greater than 0
```

A bootstrap test was done with 1000 bootstrap samples.
In 126 of these the bootstrap estimate of $\bar{x} - \bar{y}$ equalled or exceeded the original mean difference of 30.63.
Thus the bootstrap estimate of the p-value is $126/1000 = 0.126$

When does the permutation test fail?

The permutation test is exact, if:

- in the one-sample problem, the variables have a symmetric distribution
- in the two- and k-sample problem, the variables are exchangeable among the samples

A permutation test for comparing the means of two populations does not fail, if either the variances are the same, or the sample sizes are the same (cp. Romano, JASA 1990, p.686-692).

A permutation test for the difference of the medians of two distributions will not be exact, even asymptotically, unless the underlying distributions are the same. This is independent of the sample sizes (cp. Romano, JASA 1990, p.686-692).

A permutation test fails, if one tests for interaction in an unbalanced design! (cp. Good, Permutation Tests, 1993).

When does the permutation test fail?

An example comparing the location of two distributions by two-sided tests, where the true means and variances as well as the two group sizes are different:

```
set.seed(34561)
x <- rnorm(25,0,1)
y <- rnorm(75,1,4)

library(exactRankTests)
perm.test(x, y, exact=T)
t.test(x, y, var.equal=T)
t.test(x, y)
wilcox.exact(x, y)
```

Two-sample permutation test: $p = 0.147$
Two-sample t-test: $p = 0.156$
Welch two-sample t-test: $p = 0.020$
Wilcoxon rank sum test: $p = 0.088$

Bootstrap Tests

```
library(boot)
# transform x and y to get common mean
x1 <- x-mean(x)+mean(c(x,y))
y1 <- y-mean(y)+mean(c(x,y))

xy <- data.frame(value=c(x1,y2),
      group=c(rep("x",length(x)), rep("y",length(y))))

diff.t <- function(d, i)
{
     n <- table(as.numeric(d[,2]))
     gp1 <- 1:n[1]
     t.test(d[i,1][gp1],d[i,1][-gp1])$statistic
}
# resample separately: use strata arg.
set.seed(12345)
b <- boot(xy, diff.t, R=1000, stype="i", strata=xy[,2])
sum(abs(b$t) >= abs(b$t0))/b$R
```

Bootstrap t test: $p = 0.042$

Bootstrap test for testing equality of means

(Efron &Tibshirani, 1993)

1. Transform x and y according $\tilde{x}_i = x_i - \bar{x} + \bar{z},\ i = 1, ..., n_x$, and $\tilde{y}_i = y_i - \bar{y} + \bar{z},\ i = 1, ..., n_y$, where $\bar{z}$ is the mean of the combined sample of x and y.

2. Build B bootstrap data sets (x^*, y^*) using samples of size n_x with replacement from $\tilde{x}_1, ..., \tilde{x}_{n_x}$ and samples of size n_y with replacement from $\tilde{y}_1, ..., \tilde{y}_{n_y}$.

3. Evaluate $t(\cdot)$ on each sample,

$$t(x^{*b}) \ = \ (\bar{\tilde{x}}^* - \bar{\tilde{y}}^*)/\sqrt{s^2_{\tilde{x}}/n_x + s^2_{\tilde{y}}/n_y}, \quad b = 1, ..., B$$

4. Compute bootstrap p value by

$$\hat{p} \ = \ \#\{t(x^{*b}) \geq t_{obs}\}/B$$

where $t_{obs} = (\bar{x} - \bar{y})/\sqrt{s^2_x/n_x + s^2_y/n_y}$ is the observed value of the test statistic on the original data sets, $s^2_x = \sum_1^{n_x}(x_i - \bar{x})^2/(n_x - 1)$ and $s^2_y = \sum_1^{n_y}(y_i - \bar{y})^2/(n_y - 1)$.

A small simulation study

$X \; sim \; \mathcal{N}(0,1); Y \; \sim \; \mathcal{N}(1,4)$

Test of the null hypothesis $H_0 : \; \mu_x \; = \mu_y$.

	median p value	
test	n1=n2=50	n1=25; n2=75
t	0.087	0.193
Welch	0.089	0.036
Wilcoxon	0.052	0.132
Permutation	0.085	0.195
Bootstrap (eq var)	0.082	0.200
Bootstrap (uneq var)	0.091	0.033

When might the bootstrap fail?

- Incomplete data
- Dependent data
- Dirty data ("outliers")

Recommendations

Which test should be used?

- Consider a permutation test before you turn to a bootstrap
  - The bootstrap is not exact except for large samples and has often low power
    (but can sometimes be applied when permutation tests fail).
  - The permutation test is exact if observations in the combined samples are exchangeable (note: i.i.d. observations are exchangeable).
- An important advantage of the permutation test over the t test is that it is exact even for small samples whether or not observations come from normal distributions.

References

- Davison AC, Hinkley DV (1997). Bootstrap Methods and Their Applications. Cambridge University Press.
- Efron B, Tibshirani RJ (1993). An Introduction to the Bootstrap. Chapman & Hall, Inc..
- Good P (2000). Permutation Tests. Springer Verlag.

R Packages

- Permutation tests: `exactRankTests`, `multtest`.
- Bootstrap: `boot`, `Design`.