Resampling Techniques and their Application

-Class 6-

Frank Konietschke

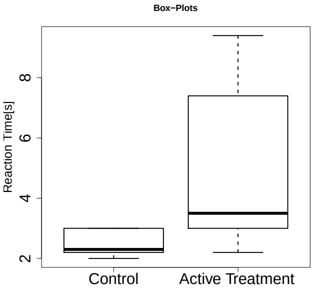
Institut für Biometrie und Klinische Epidemiologie

Charité - Universitätsmedizin Berlin, Berlin

frank.konietschke@charite.de

Motivation and Examples-III

Researchers produce a pain killer using poison from a snake. They investigate the effect of the treatment on n_1 mice in the **control** group and $n_2 = 10$ mice in the **active treatment**. The response variable is the reaction time of the mice to signal pain when a stitch is applied to their tail. Is the treatment effective? (all mice survived the dose)




```
x=c(
2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)

y=c(
2.8, 2.2, 3.8, 9.4, 8.4,
3.0, 3.2, 4.4, 3.2, 7.4)
```

- **Aim:** Test $H_0 : \mu_1 = \mu_2$ and confidence interval for $\delta = \mu_1 - \mu_2$

Statistical Model

- $X_{ik} \sim F_i, i = 1, 2; k = 1, \ldots, n_i;\ N = n_1 + n_2$
  - $E(X_{i1}) = \mu_i; Var(X_{i1}) = \sigma_i^2$
  - Asymptotics: $N \to \infty : n_i/N \to \kappa_i \in (0, 1)$
- Estimators
  - $\overline{X}_{1\cdot}$ and $\overline{X}_{2\cdot}$: means per group with

$$\overline{X}_{i\cdot} = \frac{1}{n_i} \sum_{k=1}^{n_i} X_{ik}$$

  - $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$: empirical variances per group with

$$\widehat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (X_{ik} - \overline{X}_{i\cdot})^2$$

The t-Test

- $X_{ik} \sim F_i, i = 1, 2; k = 1, \ldots, n_i;\ N = n_1 + n_2$
  - $E(X_{i1}) = \mu_i; Var(X_{i1}) = \sigma_i^2$
  - Asymptotics: $N \to \infty : n_i/N \to \kappa_i \in (0, 1)$
  - **Assume** $\sigma_1 = \sigma_2$
- **Pooled variance**

$$\widehat{\sigma}_p^2 = \frac{(n_1 - 1)\widehat{\sigma}_1^2 + (n_2 - 1)\widehat{\sigma}_2^2}{n_1 + n_2 - 2} = \frac{1}{N - 2} \sum_{i=1}^{2} \sum_{k=1}^{n_i} (X_{ik} - \overline{X}_{i\cdot})^2$$

- Test statistic

$$T = \frac{\overline{X}_{1\cdot} - \overline{X}_{2\cdot}}{\widehat{\sigma}_P \sqrt{1/n_1 + 1/n_2}}$$

- Reject H_0, if $|T| \geq t_{1-\alpha/2}(n_1 + n_2 - 2)$, the $(1 - \alpha/2)$-quantile from the t-distribution with $N - 2$ degrees of freedom

Satterthwaite-Welch t-Test

- $X_{ik} \sim F_i, i = 1, 2; k = 1, \ldots, n_i;\ N = n_1 + n_2$
  - $E(X_{i1}) = \mu_i;\ Var(X_{i1}) = \sigma_i^2$
  - Asymptotics: $N \to \infty : n_i/N \to \kappa_i \in (0, 1)$
- Test statistic

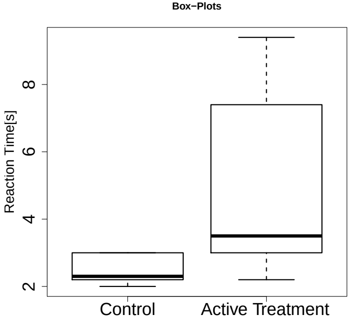
$$T = \frac{\overline{X}_{1\cdot} - \overline{X}_{2\cdot}}{\sqrt{\widehat{\sigma}_1^2/n_1 + \widehat{\sigma}_2^2/n_2}}$$

- Reject H_0, if $|T| \geq t_{1-\alpha/2}(\nu)$,

$$\nu = \frac{\left(\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}\right)^2}{\frac{\widehat{\sigma}_1^4}{n_1^2(n_1-1)} + \frac{\widehat{\sigma}_2^4}{n_2^2(n_2-1)}}$$

  degrees of freedom **(Satterthwaite's approximation).**

Researchers produce pain killer using poison from a cobra. The investigate the effect of the treatment on n_1 mice in the **control** group and $n_2 = 10$ mice in the **active treatment**. The response variable is the reaction time of the mice to signal pain when a stitch is applied to their tail. Is the treatment effective?




```
react <- data.frame(resp=c(x,y),
grp=factor(c(rep(1,10),rep(2,10))))

t.test(resp~grp,data=react,
var.equal=TRUE)

t.test(resp~grp,data=react,
var.equal=FALSE)
```

Properties of the t-Tests

- **t-Test**
- Strictly: Only valid if data is normally distributed
- Method is valid for large sample sizes, in general
- Assumes equal variances
- Simulations show bad performance if data is skewed
- Can we do better? ($\Rightarrow$ Resampling)

- **Satterthwaite t-Test**
- Strictly: Only valid if data is normally distributed
- Method is valid for large sample sizes, in general
- Allows for unequal variances
- Simulations show bad performance if data is skewed
- Can we do better? ($\Rightarrow$ Resampling)

First Attempts: Permutation Tests (1990)

- Original permutation test (Romano, 1990)
- Basically the t-test without studentization (no denominator)
- Test statistic

$$T_{ns} = \sqrt{\frac{n_1 n_2}{N}}(\overline{X}_{1\cdot} - \overline{X}_{2\cdot})$$

- Idea: Use of **permutations** to find its distribution
- Compute critical and p-values from the permutation distribution of T_{ns}

First Attempts: Permutation Tests (1990) -II

- **Workflow**
  - Collect **all** data in $\mathbf{X} = (X_{11}, \dots, X_{2n_2})'$
  - Randomly **permute** the values in $\mathbf{X}$ and obtain $\mathbf{X}^* = (X^*_{11}, \dots, X^*_{2n_2})'$
  - Re-assign $X^*_{11}, \dots, X^*_{1n_1}$: group 1
  - Re-assign $X^*_{21}, \dots, X^*_{2n_2}$: group 2
  - Compute $\overline{X}^*_{1\cdot}$ and $\overline{X}^*_{2\cdot}$
  - Compute $T^*_{ns} = \sqrt{\frac{n_1 n_2}{N}}(\overline{X}^*_{1\cdot} - \overline{X}^*_{2\cdot})$ and safe this value
  - Repeat the above a large number of times (n_{perm}) and estimate the p-value

First Attempts: Permutation Tests (1990) -III

- **When will the test work?**
- Impact of n_1 and n_2
- Impact of the variances
- Impact of the joint distribution $F(X_{11}, X_{21})$
- Write a simulation program ($\alpha = 5\%$)

- **Implementation**
- Write the sample as $X_1, \ldots, X_N$
- $\sqrt{\frac{n_1 n_2}{N}}(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) = \sum_{\ell=1}^{N} c_\ell X_\ell$
- Permutation version

$$\sqrt{\frac{n_1 n_2}{N}}\,(\overline{X}^*_{1\cdot} - \overline{X}^*_{2\cdot}) = \sum_{\ell=1}^{N} c_\ell X^*_\ell = \sum_{\ell=1}^{N} c^*_\ell X_\ell$$

- Permute the “coefficients” c_ℓ
- Use matrix technique to multiply a permutation matrix with the sample

Implementation

- Write the **pooled** sample $X_{11}, \ldots, X_{2n2}$ as $X_1, \ldots, X_N$
- Index $(11) \hookrightarrow 1; (12) \hookrightarrow 2; \ldots; (2n_2) \hookrightarrow N$
- Test statistic

$$\sqrt{\frac{n_1 n_2}{N}} \left(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}\right) = \sum_{\ell=1}^{N} c_\ell X_\ell, \;\; c_\ell = \begin{cases} 1 \leq \ell \leq n_1 : & \sqrt{\frac{n_1 n_2}{N}} \frac{1}{n_1} \\ n_1 + 1 \leq \ell \leq N : & -\sqrt{\frac{n_1 n_2}{N}} \frac{1}{n_2} \end{cases}$$

- Permutation version

$$\sqrt{\frac{n_1 n_2}{N}} \left(\overline{X}^*_{1\cdot} - \overline{X}^*_{2\cdot}\right) = \sum_{\ell=1}^{N} c_\ell X^*_\ell = \sum_{\ell=1}^{N} c^*_\ell X_\ell$$

- Arrange the coefficients c^*_ℓ in a $n_{perm} \times N$ matrix $\mathbf{P}$ and multiply with $\mathbf{X}$. This operation computes all permuted values T^*_{ns}

```
mypermu<-function(nsim,nperm,n1,n2,s1,s2,Distribution){
N<-n1+n2;erg<-c()
#-----------Permutation Matrices--------------#
i1<-sqrt(n1*n2/N)*c(rep(1/n1,n1),rep(-1/n2,n2))
P<-t(apply(matrix(i1,nrow=nperm,ncol=N,byrow=TRUE),1,sample))
```

```
if(Distribution=="Normal"){
x1<-matrix(rnorm(n1*nsim)*sqrt(s1),ncol=nperm,nrow=n1)
x2 <-matrix(rnorm(n2*nsim)*sqrt(s2),ncol=nperm,nrow=n2)}
x<-rbind(x1,x2)
mx1<-colMeans(x1); mx2<-colMeans(x2)
Tns <-sqrt(n1*n2/N)*(mx1-mx2)
```

```
for (i in 1:nsim){
X<-x[,i]
#---------------Permutations--------------------#
TnsP <- P%*%X #actual samole
pvalue<-2*min(mean(TnsP<=Tns[i]),mean(TnsP>=Tns[i]))
```

```
erg[i] <-(pvalue<0.05)}
result<-data.frame(nsim=nsim,nperm=nperm,n1=n1,n2=n2,s1=s1,s2=s2,Dist=Distribution,
Permu=mean(erg))
result}
mypermu(10000,10000,10,10,1,1,"Normal")
```

General Validity of the Permutation Test

- Type-I error simulation (10K simulations, 10K n_{perm} $\alpha = 5\%$); Test $H_0 : \mu_1 = \mu_2$
- Sample sizes $(n_1, n_2) = (10, 10), (10, 20), (20, 10)$ and one with $(50, 100)$
- Variances $(\sigma_1^2, \sigma_2^2) = (1, 1), (1, 3)$
- Fill the table: (*set.seed(1)*; most important is the last column)

Sample Size	Variances	Shape	Emp. Type-I	Accurate (yes/no)
(10,10)	(1,1)	Symmetric		
(10,20)	(1,1)	Symmetric		
(10,10)	(1,3)	Symmetric		
(10,20)	(1,3)	Symmetric		
(20,10)	(1,3)	Symmetric		
(50,100)	(1,3)	Symmetric		
(100,50)	(1,3)	Symmetric		

- Any issues? Observations?

Theoretical Investigations

- The above can only work if the distribution of T_{ns} and T_{ns}^* coincide
- Let us compute

$$\begin{aligned}
E(T_{ns}) &= 0 \\
E(T_{ns}^*|\mathbf{X}) &= E\left(\sqrt{\frac{n_1 n_2}{N}}\left(\overline{X}_{1\cdot}^* - \overline{X}_{2\cdot}^*\right)|\mathbf{X}\right) = 0
\end{aligned}$$

$$\begin{aligned}
Var(T_{ns}) &= \frac{n_1 n_2}{N}\frac{\sigma_1^2}{n_1} + \frac{n_1 n_2}{N}\frac{\sigma_2^2}{n_2} \to \kappa_2\sigma_1^2 + \kappa_1\sigma_2^2 \\
Var(T_{ns}^*|\mathbf{X}) &= \frac{1}{N-1}\sum_{i=1}^{2}\sum_{k=1}^{n_i}(X_{ik} - \overline{X}_{\cdot\cdot})^2 \to \kappa_1\sigma_1^2 + \kappa_2\sigma_2^2
\end{aligned}$$

- $\overline{X}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i} X_{ik}$
- Only valid, if... $n_1 = n_2$ or $\sigma_1^2 = \sigma_2^2$

Theoretical Investigations-II

- Derivation of $Var(T^*_{ns}|\mathbf{X})$

  We compute the expectation and variance of

$$\sqrt{\frac{n_1 n_2}{N}}\,(\overline{X}^*_{1\cdot} - \overline{X}^*_{2\cdot}) = \sum_{\ell=1}^{N} c_\ell X^*_\ell$$

$$\begin{aligned} E(X^*_\ell|\mathbf{X}) &= \frac{1}{N}\sum_{\ell=1}^{N} X_\ell = \overline{X}_{\cdot\cdot} \\ E\left(\sum_{\ell=1}^{N} c_\ell X^*_\ell|\mathbf{X}\right) &= \sum_{\ell=1}^{N} c_\ell E(X^*_\ell|\mathbf{X}) \\ &= \overline{X}_{\cdot\cdot}\sum_{\ell=1}^{N} c_\ell = 0 \end{aligned}$$

$$\begin{aligned} E(X^{2*}_i|\mathbf{X}) &= \sum_{\ell=1}^{N} X^{2*}_\ell P(X^*_i = X_\ell) = \frac{1}{N}\sum_{\ell=1}^{N} X^2_\ell \\ Var(X^*_i|\mathbf{X}) &= E(X^{2*}_i|\mathbf{X}) - E^{2*}(X_i|\mathbf{X}) = \frac{1}{N}\sum_{\ell=1}^{N}(X_\ell - \overline{X}_{\cdot\cdot})^2 \\ E(X^*_i X^*_j|\mathbf{X}) &= \sum_{\ell\neq k} X_\ell X_k P(X^*_i = X_\ell, X^*_j = X_k) = \frac{1}{N(N-1)}\sum_{\ell\neq k} X_\ell X_k \\ Cov(X^*_i, X^*_j|\mathbf{X}) &= E(X^*_i X^*_j|\mathbf{X}) - E(X^*_i|\mathbf{X})E(X^*_j|\mathbf{X}) \\ &= -\frac{1}{N(N-1)}\sum_{\ell=1}^{N}(X_\ell - \overline{X}_{\cdot\cdot})^2 \end{aligned}$$

  Now plug everything together

-

Repairment: Studentize the Statistic

- Instead of using the numerator of the statistic only, investigate the permutation distribution of

$$T = \frac{\overline{X}_{1\cdot} - \overline{X}_{2\cdot}}{\sqrt{\widehat{\sigma}_1^2/n_1 + \widehat{\sigma}_2^2/n_2}}$$

- Verify in a simulation study
- We call these methods **Studentized Permutation Tests**
    - Permuted values $\mathbf{X}^* = (X_{11}^*, \ldots, X_{2n_2}^*)'$
    - $X_{11}^*, \ldots, X_{1n_1}^*$: group 1
    - $X_{21}^*, \ldots, X_{2n_2}^*$: group 2
    - $\overline{X}_{1\cdot}^*$ and $\overline{X}_{2\cdot}^*$: means
    - $\widehat{\sigma}_1^{2*}$ and $\widehat{\sigma}_2^{2*}$: empirical variances

$$T^* = \frac{\overline{X}_{1\cdot}^* - \overline{X}_{2\cdot}^* - E(\overline{X}_{1\cdot}^* - \overline{X}_{2\cdot}^* | \mathbf{X})}{\sqrt{\widehat{\sigma}_1^{2*}/n_1 + \widehat{\sigma}_2^{2*}/n_2}}$$

Project: Validity of Studentized Permutation Test

- Type-I error simulation (10K simulations, 10K n_{perm} $\alpha = 5\%$); Test $H_0 : \mu_1 = \mu_2$
- Sample sizes $(n_1, n_2) = (10, 10), (10, 20), (20, 10)$ and one with $(50, 100)$
- Variances $(\sigma_1^2, \sigma_2^2) = (1, 1), (1, 3)$
- Fill the table: (*set.seed(1)*; most important is the last column)

Sample Size	Variances	Shape	Emp. Type-I	**Accurate (yes/no)**
(10,10)	(1,1)	Symmetric		
(10,20)	(1,1)	Symmetric		
(10,10)	(1,3)	Symmetric		
(10,20)	(1,3)	Symmetric		
(20,10)	(1,3)	Symmetric		
(50,100)	(1,3)	Symmetric		
(100,50)	(1,3)	Symmetric		

- Any issues? Observations?

Resampling the t-Test

- Goal: estimate the distribution of T via resampling
  - Data: $\mathbf{X} = (X_{11}, \ldots, X_{2n_2})'$
    - Resampling variables: $\mathbf{X}^* = (X_{11}^*, \ldots, X_{2n_2}^*)'$
    - $X_{11}^*, \ldots, X_{1n_1}^*$: group 1
    - $X_{21}^*, \ldots, X_{2n_2}^*$: group 2
    - $\overline{X}_{1\cdot}^*$ and $\overline{X}_{2\cdot}^*$: means
    - $\widehat{\sigma}_1^{2*}$ and $\widehat{\sigma}_2^{2*}$: empirical variances

$$T^* = \frac{\overline{X}_{1\cdot}^* - \overline{X}_{2\cdot}^* - E(\overline{X}_{1\cdot}^* - \overline{X}_{2\cdot}^* | \mathbf{X})}{\sqrt{\widehat{\sigma}_1^{2*}/n_1 + \widehat{\sigma}_2^{2*}/n_2}}$$

  - Repeat these steps *nboot*-times
- Reject H_0, if $T < c_{\alpha/2}^*$ or $T > c_{1-\alpha/2}^*$
- c_α^*: α- quantile from resampling distribution

Group wise Nonparametric Bootstrap

- $\mathbf{X}_1 = (X_{11}, \ldots, X_{1n_1})$ (**fixed values**)
- $\mathbf{X}_2 = (X_{21}, \ldots, X_{2n_2})$ (**fixed values**)
- **Drawing with Replacement:** randomly draw n_1 and n_2 observations from $\mathbf{X}_1$ and $\mathbf{X}_2$
- Example $\mathbf{X}_1 = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}_1^* = (2, 2, 4, 3, 2)$
  $\mathbf{X}_1^* = (1, 1, 2, 3, 3)$
  $\mathbf{X}_1^* = (2, 5, 5, 3, 3)$
  ...
- In R: sample(x1,replace=TRUE)
- Also known as **Group wise Nonparametric Bootstrap**

Nonparametric Bootstrap

- **Data X** $= (X_{11}, \ldots, X_{2n_2})$ (**fixed values**)
- **Drawing with Replacement:** randomly draw N observations X_k^* from **X** with replacement such that
$$P(X_{11}^* = X_{11}) = \frac{1}{N}$$
- In R: sample(x,replace=TRUE)
- Also known as **Nonparametric Bootstrap**

Permutation

- **Data X** $= (X_{11}, \ldots, X_{2n_2})$ (**fixed values**)
- **Drawing without Replacement:** randomly draw N observations X^*_{ik} from **X** without replacement such that

$$P(X^*_{11} = X_{11}) = \frac{1}{N}$$

- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}^* = (4, 1, 3, 2, 5)$
  $\mathbf{X}^* = (5, 1, 2, 3, 4)$
  $\mathbf{X}^* = (3, 1, 2, 5, 4)$
  ...
- In R: sample(x)
- Also known as **Permutation**

Parametric Bootstrap

- **Data** $\mathbf{X}_i = (X_{ik}, \ldots, X_{in_i})$ (**fixed values**)
- **Resampling** randomly draw n_i observations X^*_{ik} from

$$N(0, \widehat{\sigma}^2_i)$$

- In R: $rnorm(n, 0, sd(x))$
- Also known as **Parametric Bootstrap** (Why is that not equivalent to the t-approximation?)

Skewed Parametric Bootstrap

- **Data $\mathbf{X}_i = (X_{i1}, \dots, X_{in_i})$ (fixed values)**
- Estimate the skewness of each sample by

$$\widehat{\mu}_{i,3} = \frac{n_i}{(n_i - 1)(n_i - 2)} \sum_{k=1}^{n_i} \left(\frac{X_{ik} - \overline{X}_{i\cdot}}{\widehat{\sigma}_i} \right)^3$$

- **Resampling** randomly draw n_i observations X_{ik}^* from

$$sign(\widehat{\mu}_{i,3})\widehat{\sigma}_i \frac{\chi^2_{f_i} - f_i}{\sqrt{2f_i}}$$

- $f_i = 8/\widehat{\mu}^2_{i,3}$

Wild Bootstrap

- **Data $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})$ (fixed values)**
- Fix the values $Z_{ik} = X_{ik} - \overline{X}_{i\cdot}$
- **Resampling** randomly generate iid weights W_{ik} with $E(W_{ik}) = 0$ and $Var(W_{ik}) = 1$. Generate X_{ik}^* by
$$X_{ik}^* = W_{ik} * Z_{ik}$$
- Examples: $W_{ik} \sim N(0, 1)$
- Rademacher: $P(W_{ik} = 1) = P(W_{ik} = -1) = 1/2$

  ...
- Also known as **Wild-Bootstrap**

Project

- **Which resampling method is better? Permutation, Nonparametric Bootstrap or Parametric Bootstrap?**
- Write your own simulation program