Resampling Techniques and their Application

-Class 9-

Frank Konietschke

Institut für Biometrie und Klinische Epidemiologie

Charité - Universitätsmedizin Berlin, Berlin

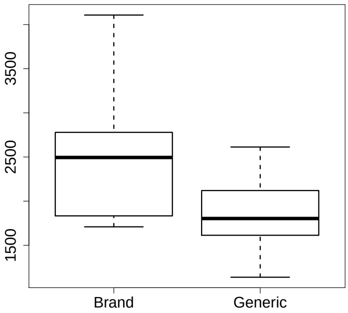
frank.konietschke@charite.de

Paired Observations

- Before and after measures
- E.g. blood pressure before and after surgery
- Measurements on the same subject
- Advantages
  - Every subject (patient) is his/her own control
  - Reducation of subjects
  - Less costs (potentially)
- Measurements from the same subject **are not necessarily independent**

Example

- Drug absorbtion study: n=10 patients received brand and generic drug (after wash out period)
- Response: Absorption of the drug in the blood



ID	Brand	Generic
1	4108	1755
2	2526	1138
3	2779	1613
4	3852	2254
5	1833	1310
6	2463	2120
7	2059	1851
8	1709	1878
9	1829	1682
10	2594	2613

- Aim: $H_0 : \mu_1 = \mu_2$ and confidence interval

Statistical Model

- $\mathbf{X}_k = (X_k, Y_k)', k = 1, \ldots, n$
  - $E(X_k) = \mu_1, E(Y_k) = \mu_2; Var(\mathbf{X}_k) = \mathbf{\Sigma}$
  - What is $\mathbf{\Sigma}$?
    - $\mathbf{\Sigma}$ is the covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma \\ \sigma & \sigma_2^2 \end{pmatrix}$$

    - σ : Covariance of X_k and Y_k
    - $\sigma = E((X_k - \mu_1)(Y_k - \mu_2))$
    - Measures the degree of the (linear) relationship between X_k and Y_k
    - On average, $(\underbrace{(X_k - \mu_1)}_{\lesseqgtr 0}\underbrace{(Y_k - \mu_2)}_{\lesseqgtr 0}) \lesseqgtr 0$

Paired t-Test

- $\mathbf{X}_k = (X_k, Y_k)', k = 1, \ldots, n$
  - $E(X_k) = \mu_i, E(Y_k) = \mu_2; Var(\mathbf{X}_k) = \mathbf{\Sigma}$
- Aim: $H_0 : \mu_1 = \mu_2 \Rightarrow t$-test
  - $D_k = X_k - Y_k$
  - $\overline{D}.$ mean of the differences
  - $\widehat{\sigma}_D^2$ empirical variance of the differences

$$T = \sqrt{n} \cdot \frac{\overline{D}.}{\widehat{\sigma}_D}$$

  - $T \xrightarrow{\mathcal{D}} N(0, 1)$ or $T \approx T_{n-1}$ (under H_0)
  - Reject H_0, if $|T| \geq t_{1-\alpha/2}(n-1)$

Example Evaluation

```
brand=c(4108,2526,2779,3852,1833, 2463,2059,1709,1829,2594)
generic=c(1755,1138,1613,2254,1310,2120,1851,1878,1682,2613)
plot(brand,generic,pch=19,cex=1.3)
n=length(brand)
x=cbind(brand,generic)
var(x)

diff=brand-generic
mD=mean(diff)
vd=var(diff)

T=sqrt(n)*mD/sqrt(vd)
pvalue=2*min(pt(T,n-1), 1-pt(T,n-1))

t.test(brand,generic,paired=TRUE)
```

Paired t-Test- Properties

- Valid if differences D_k are normally distributed (small samples)
- Valid for large sample sizes
- Test is liberal/ conservative under non-normality
- Idea: Resample the distribution of T
- But how? Differences?

Resampling the t-Test

- Resampling variables: $\mathbf{X}^* = (X_{11}^*, \ldots, X_{2n}^*)'$
  - $X_{11}^*, \ldots, X_{1n}^*$: condition 1
  - $X_{21}^*, \ldots, X_{2n}^*$: condition 2
  - $D_k^* = X_{1k}^* - X_{2k}^*$; $\overline{D}_{\cdot}^*$: mean
  - $\widehat{\sigma}_D^{2*}$ empirical variances

$$T^* = \sqrt{n} \cdot \frac{\overline{D}_{\cdot}^*}{\widehat{\sigma}_D^*}$$

- Repeat these steps n_{boot}-times
- Reject H_0, if $T < c_{\alpha/2}^*$ or $T > c_{1-\alpha/2}^*$

Generation of Resampling Variables

- Observations on the same subject are not necessarily independent
- Can we resample despite the dependencies?
  - We will study methods that **keep and ignore the dependencies**
    - Resampling the differences
    - Resampling from all data and thus ignoring dependencies

Resampling Using the Differences

- **Differences** $\mathbf{D} = (D_1, \ldots, D_n)$ (**fixed values**)
- **Drawing with Replacement:** randomly draw n observations D_k^* from **D** with replacement such that

$$P(D_1^* = D_1) = \frac{1}{n}$$

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

Resampling Using the Differences

- **Differences D** $= (D_1, \ldots, D_n)$ (**fixed values**)
- **Resampling** randomly draw n observations D_k^* from

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

- In R: $rnorm(n, 0, sd(x))$
- Also known as **Parametric Bootstrap** (Useful?)

Resampling Using the Differences

- **Differences D** $= (D_1, \dots, D_n)$ (**fixed values**)
- Generate random weights $W_1, \dots, W_n$ with $E(W_1) = 0$ and $Var(W_1) = 1$
  - Random signs $P(W_1 = 1) = P(W_1 = -1) = 1/2$
  - Asymmetric signs $P(W_1 = \frac{1+\sqrt{5}}{2}) = \frac{\sqrt{5}-1}{2\sqrt{5}}$ and $P(W_1 = \frac{1-\sqrt{5}}{2}) = \frac{\sqrt{5}+1}{2\sqrt{5}}$
- Wild-Bootstrap Method
- Note that centering is not necessary (why?)

Resampling Using the Differences

- **Data $\mathbf{X}_k = (X_{1k}, X_{2k})'$ (fixed values)**
- Randomly permute the data within each pair: $\mathbf{X}_k^* = (X_{1k}^*, X_{2k}^*)'$
- This method is equivalent to....

Resampling Using Original Data

- **Data X** $= (X_{11}, \ldots, X_{2n})$ (**fixed values**)
- **Permutation** randomly permute the $2n$ observations X_{ik}^* in **X**
- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}^* = (2, 2, 4, 3, 2)$
  $\mathbf{X}^* = (1, 1, 2, 3, 3)$
  $\mathbf{X}^* = (2, 5, 5, 3, 3)$
  ...
- In R: sample(x)
- **Ignoring the dependency**

Resampling Using Original Data

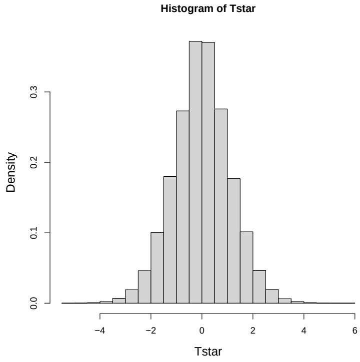
- **Data X** $= (X_{11}, \ldots, X_{2n})$ (**fixed values**)
- **Nonparametric Bootstrap** randomly draw with replacement $2n$ observations X^*_{ik} from **X**
- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}^* = (2, 2, 4, 3, 2)$
  $\mathbf{X}^* = (1, 1, 2, 3, 3)$
  $\mathbf{X}^* = (2, 5, 5, 3, 3)$
  ...
- In R: sample(x,replace=TRUE)
- Also known as **Nonparametric Bootstrap II**
- **Ignoring the dependency**

Permuting all variables despite dependencies

- Permuting (or drawing with replacement) all data is not intuitive
- Observations on same subject might be dependent
- Reason: The permutation distribution mimics the distribution of T
- Reference: Konietschke and Pauly (2015)

Illustration

```
x=brand
y=generic
plot(x,y,pch=19,cex=1.3)
n=10
d=x-y
T=sqrt(n)*mean(d)/sd(d)
pvalue=2*min(pt(T,n-1),1-pt(T,n-1))
pvalue
Tstar=c()
xy=c(x,y)
for(i in 1:100000){
xstar=sample(xy) #permutation overall
dstar=xstar[1:n]-xstar[(n+1):(2*n)]
Tstar[i]= sqrt(n)*mean(dstar)/sd(dstar)
}
pstar= 2*min(mean(Tstar<=T),mean(Tstar>=T))
pstar
```



Data Generation

- Data generation
- Different methods are possible

$$
\begin{aligned}
\mathbf{X}_k &= \boldsymbol{\mu} + \boldsymbol{\Sigma}^{-1/2}\mathbf{Z}_k \;\; E(\mathbf{Z}_k) = \mathbf{0}, \;\; Var(\mathbf{Z}_k) = \mathbf{I} \text{ or} \\
\mathbf{X}_k &= (F_1^{-1}(\Phi(Z_{1k})), F_2^{-1}(\Phi(Z_{2k}))), \;\; \mathbf{Z}_k \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ (quantile method)}, \;\text{ or} \\
X_{0k} &\sim E(X_{0k}) = 0 \text{ and } Var(X_{0k}) = 1 \\
X_{1k} &\sim E(X_{1k}) = 0 \text{ and } Var(X_{1k}) = 1 \\
X_{2k} &= \rho X_{1k} + \sqrt{1-\rho^2} X_{0k}
\end{aligned}
$$

- Elegant way: Copula (not covered in this class)
- Note: Most distributions cannot have a perfect correlation and most have bounded possible correlations within $[-1, 1]$.
- Multivariate normal distribution in *R*: packages *mvtnorm* or *multcomp*
- $\Phi(x)$: CDF of $N(0, 1)$

Project

- In a paired data setting, permuting data overall and thus ignoring the dependency is somewhat counter intuitive. Verify the validity of the method for the paired t-test in a simulation study at 5% level of significance. Use $n_{sim} = 10,000$ and $n_{perm} = 10,000$ permutation runs. Generate bivariate normal data with variance $\sigma_i^2 = 1$ and different covariances $\sigma \in \{-0.95, -0.5, 0, 0.5, 0.95\}$ and sample sizes $n \in 10, 20$.