Resampling Techniques and their Application

-Class 7-

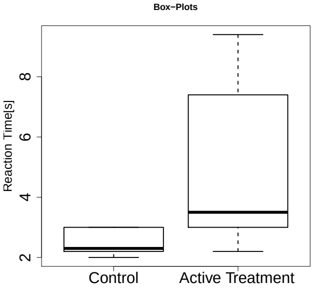
Frank Konietschke

Institut für Biometrie und Klinische Epidemiologie
Charité - Universitätsmedizin Berlin, Berlin
frank.konietschke@charite.de

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

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

- $X_{ik} \sim F_i, i = 1, 2; k = 1, \dots, n_i;\; N = n_1 + n_2$
  - $E(X_{i1}) = \mu_i$; $Var(X_{i1}) = \sigma_i^2$
  - 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

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

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

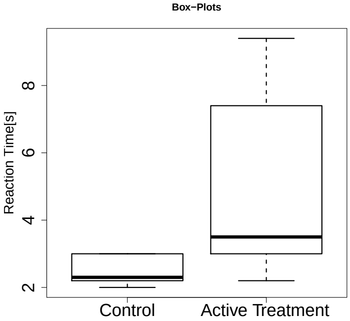
$$T = \frac{\overline{X}_{1\cdot} - \overline{X}_{2\cdot}}{\sqrt{\widehat{\sigma}_1^2/n_1 + \widehat{\sigma}_2^2/n_2}}$$

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

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

  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)
```

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

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

- 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 $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})$ (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

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

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

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**

Computation of $E(\overline{X}^*_{1\cdot} - \overline{X}^*_{2\cdot}|\mathbf{X})$

- Compute $E(\overline{X}^*_{1\cdot} - \overline{X}^*_{2\cdot}|\mathbf{X})$ in the following cases
  - Group-wise nonparametric Bootstrap
  - Nonparametric Bootstrap
  - Permutation
- Compute $Var(\overline{X}^*_{1\cdot} - \overline{X}^*_{2\cdot}|\mathbf{X})$ in the following cases
  - Group-wise nonparametric Bootstrap
  - Nonparametric Bootstrap
  - Permutation

When do Resampling Tests Work?

- Limit distribution of T: $N(0, 1)$ (Note that the t_ν distribution is the $N(0, 1)$ for large n)
- Limit Distribution of T^* given $\mathbf{X}$: $N(0, 1)$
    - Both distributions coincide and have the same limit
    - Therefore, the resampling test will work (at least for large sample sizes)
    - What means ’large’? $\hookrightarrow$ simulations necessary
    - In words
        - Resampling dist. mimics the distribution of T under H_0
        - The dist. of T departs from the resampling dist. under H_1
- References
    - Janssen (1997, 2005), Janssen and Pauls (2003)
    - Konietschke and Pauly (2012 a, b)

Confidence Intervals

- $(1-\alpha)$ confidence interval for $\delta = \mu_1 - \mu_2$
- Confidence intervals require the distribution under the alternative hypothesis
- Computation of confidence interval for δ is based on inverting

$$T = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}}}$$

- For any $\mu_1 - \mu_2$, we have $T \sim t_\nu$ (or $N(0,1)$ for large n)

$$P(t_{\alpha/2} \leq T \leq t_{1-\alpha/2}) = 1 - \alpha$$

$$CI = \left[\overline{X}_1 - \overline{X}_2 \pm t_{\nu,(1-\alpha/2)} \sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}}\right]$$

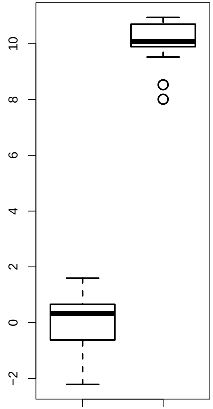
Confidence Intervals

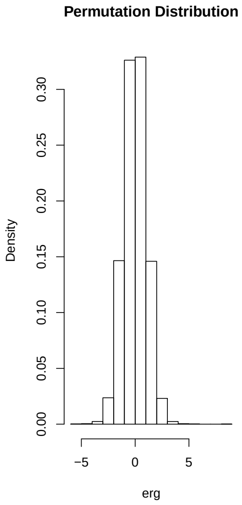
- Interpretation
  - A confidence interval is an estimator of $\mu_1 - \mu_2$
  - We estimate the difference with $(1 - \alpha)$ confidence
  - They should be compatible with the test result
  - It is false to say that *CI* covers δ with $(1 - \alpha)100\%$ probability (only holds for random prior observation)

Resampling Based Confidence Intervals

- Can we use the resampling distribution to compute confidence intervals?
- Did we ever assume that H_0 holds when we computed the resampling distribution? NO
- Illustrate with 2 samples with large effect

```
set.seed(1)
n1<-15;n2<-15; N<-n1+n2
x<-rnorm(n1,0)
y<-rnorm(n2,10)
boxplot(x,y)
erg<-c()
for(i in 1:100000){
xx<-sample(c(x,y))
xx1<-xx[1:n1]
xx2<-xx[(n1+1):(N)]
mxx1<-mean(xx1); mxx2<-mean(xx2)
vxx1<-var(xx1); vxx2<-var(xx2)
Tstar<-(mxx1-mxx2)/sqrt(vxx1/n1 + vxx2/n2)
erg[i]<-Tstar}
hist(erg,freq=F)
```





Resampling Based Confidence Intervals

- Both the distribution of T and its resampling distribution coincide
- We can use the distribution of T^* for the computation of $(1-\alpha)100\%$- confidence intervals

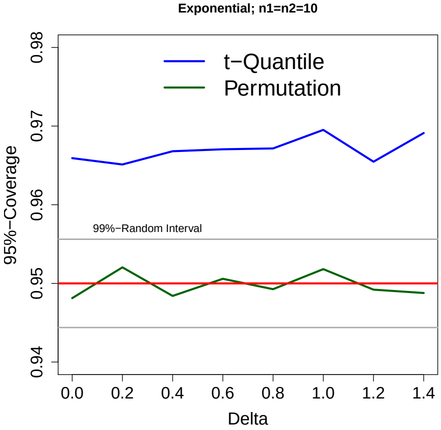
$$P(c^*_{\alpha/2} \leq T \leq c^*_{1-\alpha/2}) \approx 1-\alpha$$

$$CI_p = \left[\overline{X}_{1\cdot} - \overline{X}_{2\cdot} - c^*_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}}; \overline{X}_{1\cdot} - \overline{X}_{2\cdot} - c^*_{\alpha/2} \cdot \sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}}\right]$$

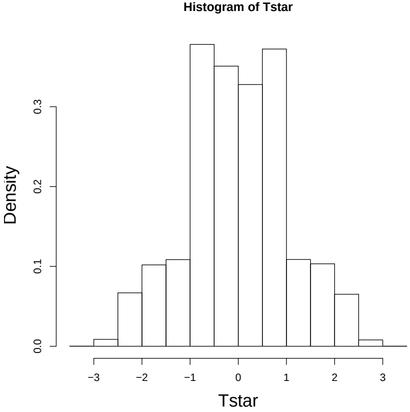
- Studentization „deletes” the shift; dist. is invariant
- References
  - Pauly, Asendorf and Konietschke (2016)

Confidence Intervals - II

- $X_{11}, \ldots, X_{1n_1} \sim Exp(1); X_{21}, \ldots, X_{2n_2} \sim Exp(1) + \delta$
- $nsim = nperm = 10,000$



```
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)
n1<-length(x)
n2<-length(y)
N<-n1+n2
mx<-mean(x);my<-mean(y)
vx<-var(x);vy<-var(y)
T<-(mx-my)/sqrt(vx/n1 + vy/n2)
erg<-c()
for(i in 1:100000){
xx<-sample(c(x,y))
xx1<-xx[1:n1]
xx2<-xx[(n1+1):(N)]
mxx1<-mean(xx1); mxx2<-mean(xx2)
vxx1<-var(xx1); vxx2<-var(xx2)
Tstar<-(mxx1-mxx2)/sqrt(vxx1/n1 + vxx2/n2)
erg[i]<-Tstar}
hist(erg)
c1star<-quantile(erg,0.025)
c2star<-quantile(erg,0.975)
```



Implementation

- Either in the same as in the 1-sample case, or
- Writing

$$\overline{X}_{1\cdot} - \overline{X}_{2\cdot} = \sum_{\ell=1}^{N} c_\ell X_\ell$$

- Permutation version

$$\overline{X}^*_{1\cdot} - \overline{X}^*_{2\cdot} = \sum_{\ell=1}^{N} c_\ell X^*_\ell = \sum_{\ell=1}^{N} c^*_\ell X_\ell$$

- Permute the “coefficients” c_ℓ
- Same strategy for the variance estimator
- Example next slide

```
myPermuCI<-function(nsim,nperm,n1,n2,v1,v2,delta, Distribution){
PermCI=c()
N<-n1+n2
#------Data Generation-----#
vvec = sqrt(c(rep(v1,n1),rep(v2,n2)))
if (Distribution == "Normal"){
x1=matrix(rnorm(n1*nsim,delta)*sqrt(v1),ncol=nsim)
x2=matrix(rnorm(n2*nsim)*sqrt(v2),ncol=nsim)}
xy = rbind(x1,x2)
x12 = x1^2; x22=x2^2
mx = colMeans(x1); my = colMeans(x2)
vx = (colSums(x12)-n1*mx^2)/(n1-1)
vy = (colSums(x22)-n2*my^2)/(n2-1)
df=(vx/n1+vy/n2)^2/(vx^2/(n1^2*(n1-1))+vy^2/(n2^2*(n2-1)))
T.L <-mx-my-qt(0.975,df)*sqrt(vx/n1+vy/n2)
T.U <-mx-my+qt(0.975,df)*sqrt(vx/n1+vy/n2)
#-----------Permutation Matrices--------------#
P<-t(apply(matrix(1:N,nrow=nperm,ncol=N,byrow=TRUE),1,sample))
#-------Helping Variables for Permutation Distribution---#
i1<-c(rep(1/n1,n1),rep(0,n2))
i2<-c(rep(0,n1),rep(1/n2,n2))
i3<-c(rep(1/(n1*(n1-1)),n1), rep(0,n2))
i4<-c(rep(0,n1), rep(1/(n2*(n2-1)),n2))
Im1<-matrix(i1[P],nrow=nperm,ncol=N)
Im2<-matrix(i2[P],nrow=nperm,ncol=N)
Iv1<-matrix(i3[P],nrow=nperm,ncol=N)
Iv2<-matrix(i4[P],nrow=nperm,ncol=N)
#----------------Begin of Simulation----------------#
for (i in 1:nsim){
X<-xy[,i]
#---------------Permutations---------------------#
mxP <- Im1%*%X
myP = Im2%*%X
vxP <- Iv1%*%X^2 - n1/(n1*(n1-1))*mxP^2
vyP <- Iv2%*%X^2 - n2/(n2*(n2-1))*myP^2
TP = (mxP -myP )/sqrt(vxP +vyP  )
c1<-quantile(TP,0.025); c2<-quantile(TP,0.975)
lower <-mx[i]-my[i]-c2*sqrt(vx[i]/n1+vy[i]/n2)
upper <- mx[i]-my[i]-c1*sqrt(vx[i]/n1+vy[i]/n2)
PermCI[i]<-(lower<delta& upper >delta)
#-------End of Simulation-------------#
}
Result <- data.frame(nsim=nsim,nperm=nperm,delta=delta,
n1=n1,n2=n2,v1=v1,v2=v2, SW=mean(T.L <delta & T.U >delta),
PermCI=mean(PermCI),
distribution=Distribution)
print(Result)
#-------End of Function------------------------#
}
myPermuCI(1000,1000,10,20,1,3,1,"Normal")
```

Project

- Simulate the coverage probabilities of the 95%-confidence intervals CI and CI_p for $\delta \in \{0, 0.1, \ldots, 2\}$
- Investigate normal and exponential distributions
- Use $\sigma_i^2 = 1$ under all settings and varying $n_i \in \{10, 20, 30\}$
- Use $n_{sim} = 10,000$ and $n_{perm} = 10,000$ permutation runs
- Instead of using permutations, would a wild-bootstrap approach also be possible?
    - Compute $(1 - \alpha)$- confidence intervals for μ (one-sample problem) and $\delta = \mu_1 - \mu_2$ using a wild-bootstrap approach. Provide a detailed derivation (formula, no simulation)