Resampling Techniques and their Application

-Class 5-

Frank Konietschke

Institut für Biometrie und Klinische Epidemiologie

Charité - Universitätsmedizin Berlin, Berlin

frank.konietschke@charite.de

Types of Error in Hypothesis Tests

- When we carry out a test, what types of errors we can make?

	Decision	
Truth (unknown) $\downarrow$	Reject H_0	Do not reject H_0
H_0	Type-1 error	correct
H_1	correct	Type II

- **Type I error:** Reject H_0 when H_0 is actually true.
- **Type II error:** Not reject H_0 when H_1 is actually true.
- **Power** of a test: 1-Type-II error = "correct decision to reject H_0"
- These errors are defined conditional on the true status (H_0 or H_1).

Power of a Test

- Based on data, we either reject or not reject the hypothesis
- In simulation, we condition on H_0 or H_1

- **Type-1 error**
  - Assume H_0 is true
  - All operations in first row of the table
  - Data generations always under H_0

- **Type-II error**
  - Assume H_1 is true
  - All operations in second row of the table
  - Data generations under H_1

Power of a Test

- Let $X_1, \ldots, X_n$ be a sample from F with $E(X_k) = \mu$ and $Var(X_k) = \sigma^2$. We test the null hypothesis $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ at level $\alpha = 5\%$. Simulate the power of the t-statistic

$$T = \sqrt{n}\frac{\overline{X}_{\cdot}}{\widehat{\sigma}}$$

  to detect the alternative $H_1 : \mu = \delta$.
- Use $n = 10, 20, 30$ and $\delta = 0, 0.1, 0.2, \ldots, 2$

Power of a Test

```
set.seed(1)
myPower<-function(n,nsim,Distribution,delta){
erg=c()
if(Distribution=="Normal"){
x<-matrix(rnorm(nsim*n),ncol=nsim)}
if(Distribution=="Exp"){
x<-matrix(rexp(nsim*n)-1,ncol=nsim)}
x<-x+delta #Expectation of x is delta
```

```
mx<-colMeans(x)
sdx<-sqrt((colSums(x^2)-n*mx^2)/(n-1))
T<-sqrt(n)*mx/sdx
result<-data.frame(n=n,Dist=Distribution,delta=delta,
tTest=mean(abs(T)>=qt(0.975,n-1)))
result}
myPower(10,10000,"Exp",0.5)
```

Power Curve

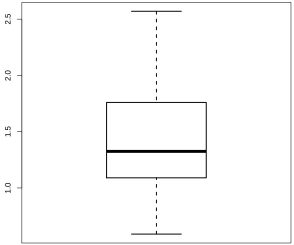
- **Power curve**: plot the power to detect δ

```
delta <- seq(0,2,0.1)
power <-c()

for(h in 1:length(delta)){
power[h]<-myPower(10,10000,"Normal",delta[h])[4]
}
plot(delta,power,type="l",lwd=3,col="blue",cex.lab=1.7,cex.axis=1.7)
abline(h=0.05,lwd=2,col="red")
```

Motivation and Examples-II

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of **n=36** bottles and obtains:




```
X=c(
0.59, 1.23, 1.00, 0.84, 0.88, 1.71,
1.81, 1.84, 2.03, 1.39, 1.30, 1.31,
1.96, 1.33, 2.57, 1.19, 1.01, 2.06,
1.32, 1.55, 1.28, 0.93, 1.63, 1.24,
1.83, 1.81, 0.94, 1.46, 1.25, 1.56,
0.61, 0.83, 1.17, 2.24, 1.68, 1.51)
```

- Estimate the mean and **the median** and their characteristics

Parameter Estimation

- Parameter estimation is key feature in statistical sciences
- Often, pretty involved
- Think about the following cases
    - Mean
    - Variance
    - Variance of an estimator (e.g. mean)
    - Variance of correlation coefficients
    - Overdispersion parameters, variance of their estimators,...

Parameter Estimation - II

- Statistical model

$$X_1, \ldots, X_n \sim F(\theta)$$

- F is a distribution
- θ are parameters of this distribution
- How to estimate $f(\theta)$?

Parameter Estimation - III

Estimation of θ	**Properties**
Maximum-Likelihood	F must be known
	Algorithm can be difficult
	Algorithm might not converge
	Large sample for distribution
Moment based	F can be unknown
	Computation usually feasible
	Usually exist (no converging issues)
	Small sample approximations
Resampling Methods	

Parameter Estimation - Resampling. But How?

- Data $\mathbf{X} = (X_1, \ldots, X_n)'$
- **Draw observations with replacement** from $\mathbf{X}$:

$$X_1^*, \ldots, X_n^*$$

  - Distribution of $X_1^*, \ldots, X_n^*$: $\widehat{F}_n$ (empirical distribution)
  - We draw observations from $\widehat{F}_n$
  - We know

$$\widehat{F}_n \to F, n \to \infty$$

  - Basically, we simulate data from $\widehat{F}_n$

Parameter Estimation - Resampling -II

- Fix the data **X**
  - Generate a bootstrap sample (drawing with replacement from **X**): $X_1^*, \ldots, X_n^*$
  - Compute the estimator $\widehat{\theta}^*$ and safe this value
  - Repeat the previous steps n_{boot} times
  - Estimate the parameter of interest using the values of $\widehat{\theta}_1, \ldots, \widehat{\theta}_{n_{boot}}$

Parameter Estimation - Resampling -III

- Model: $X_1, \ldots, X_n \sim F(\theta)$ (iid)
- Parameters $E(X_k) = \mu$ and $Var(X_k) = \sigma^2$
- Task: Estimate $f(\theta) = \tau^2 = Var(\overline{X}.)$

- Data $\mathbf{X}$
- Generate $X_1^*, \ldots, X_n^*$
- Compute $\overline{X}^*_.$ (safe in $\widehat{\theta}_\ell$)
- Repeated the steps $\ell = 1, \ldots, n_{boot}$ times
- Estimator of $\tau^2 = Var(\overline{X}.)$ is

$$\widehat{\tau}^2 = \frac{1}{n_{boot} - 1} \sum_{\ell=1}^{n_{boot}} (\widehat{\theta}_\ell - \overline{\widehat{\theta}}.)^2, \quad \overline{\widehat{\theta}}. = \frac{1}{n_{boot}} \sum_{\ell=1}^{n_{boot}} \widehat{\theta}_\ell$$

```
X#cork diameter data
n <- 36
nboot <-10000
B<- apply(matrix(1:n,
ncol=nboot,nrow=n),
2,sample,replace=TRUE)
xstar <- matrix(X[B],
ncol=nboot,nrow=n)
mxstar <- colMeans(xstar)
tauhat2 <- var(mxstar)
```

Project: How Good is the Estimator?

- Use computer simulations to assess the quality of the estimator $\widehat{\tau}^2$
- Compare with $\widehat{\tau}^2_{emp} = \widehat{\sigma}^2 / n$
- Compare the bias and MSE of the estimators
  - Generate n_{sim} random samples from different distributions
  - Compute the estimator upon the sample and safe the value in $\widehat{\theta}_s$
  - Assess the bias and MSE

$$Bias = \frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} (\widehat{\theta}_s - \theta) \text{ and } MSE = \frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} (\widehat{\theta}_s - \theta)^2$$

Project

- Task: Estimate the variance of **correlation coefficients** using resampling strategies
- How to get an idea about the true variance?
- Is bootstrap a good way?

Parameter Estimation: Variance of the Median

- $X_1, \ldots, X_n \sim F$
- Median $\nu = F^{-1}(\frac{1}{2})$ of the population (50% quantile)
- Empirical Median: $\widehat{\nu} = \widehat{F}^{-1}(\frac{1}{2})$
  - If n is uneven: middle value of sorted list $X_{[1]}, \ldots, X_{[n]}$
  - If n is even: $\frac{X_{[n/2]} + X_{[n/2+1]}}{2}$
- Compute the variance of $\widehat{\nu}$
- Sampling strategies

Parameter Estimation: Uncertainty of Median

- Sample from known distributions to get an idea about the true value
- Set sample size $n = 10, 50, 500, 50000$, $n_{sample} = 10K$

- $N(0, 1)$ **population** ($\nu = 0$)

```
set.seed(1)
n<-50
erg <-c()
for(i in 1:10000){
x <-rnorm(n)
erg[i]<-median(x)}
hist(erg)
mean(erg)
```

- $Exp(1)$ **population** ($\nu = 0.693$)

```
set.seed(1)
n<-50
erg <-c()
for(i in 1:10000){
x <-rexp(n)
erg[i]<-median(x)}
hist(erg)
mean(erg)
```

Percentile Method: Confidence Interval

- Compute a confidence interval for ν
    1. Sample $X_1^*, \ldots, X_n^*$ with replacement from $\mathbf{X}$
    2. Compute $\widehat{\nu^*}$ from $X_1^*, \ldots, X_n^*$ and safe this value in $\widehat{\nu}_\ell^*$
    3. Repeat the above n_{boot} times and obtain $\widehat{\nu}_1^*, \ldots, \widehat{\nu}_{n_{boot}}^*$
    4. Sort the values from smallest to largest: $\widehat{\nu}_{[1]}^*, \ldots, \widehat{\nu}_{[n_{boot}]}^*$
    5. Estimate the $(1-\alpha)$ confidence interval for ν by

$$CI_\nu = \left[\widehat{\nu}^*_{[0.025*nboot]}, \widehat{\nu}^*_{[0.975*nboot]}\right]$$

Simulation Study: Coverage Probability

- The CI should cover the true ν in $(1-\alpha)100\%$
- Simulation with $n = 50$, $n_{sim} = 1K$, $n_{boot} = 1K$
- Normal and Exponential distributions

Simulation Study: Coverage Probability

```
myCI<-function(n,nsim,nboot,Distribution){
erg<-c()
B<- apply(matrix(1:n,ncol=nboot,nrow=n),2,sample,replace=TRUE)
if(Distribution=="Normal"){
x<-matrix(rnorm(n*nsim),ncol=nsim)
nu<-0}
if(Distribution=="Exp"){
x<-matrix(rexp(n*nsim),ncol=nsim)
nu<-0.693}
for(i in 1:nsim){
xstar = matrix(x[,i][B],ncol=nboot,nrow=n)
nustar<-apply(xstar,2,median)
nustarS<-sort(nustar)
lower<-nustarS[0.025*nboot]; upper<-nustarS[0.975*nboot]
erg[i]<-(lower <nu && upper>nu)}
result <- data.frame(nsim=nsim, nboot=nboot,nu=nu,
CI=mean(erg))
result}
myCI(50,1000,1000,"Exp")
```

Parameter Estimation: Uncertainty of Median

- Revise the cork diameter example
- $n_{boot} = 100K$

```
hist(X,freq=F)
nustar <-c()
nboot<-100000
set.seed(1)
for(i in 1:nboot){
xB<- sample(X,36,replace=TRUE)
nustar[i]<-median(xB)}
hist(nustar,freq=F)
nustarS<-sort(nustar)
lower<-nustarS[0.025*nboot]; upper<-nustarS[0.975*nboot]
c(lower,upper)
```

Project: Width of a Distribution

- In statistics, describing the width of a distribution is key and indeed a rather challenging task
- Often, the width of a distribution is described by the Interquartile Range (IQR), which is defined as the IQR=75%-quantile - 25% quantile
- Using sampling, find the true IQR of $N(0,1)$ and $Exp(1)$ distributions and illustrate its variability. Check using the functions $qnorm()$ and $qexp()$
- Using resampling, find a 95% confidence interval for the IQR
- Investigate its coverage probability in a simulation study with $n_{sim} = 1000$ and $n_{boot} = 1000$ runs for normal and exponential distributions with sample sizes $n \in 10, 20, 30, 40$
- Compute the precision interval for the empirical coverage probability and state your conclusion