Model Selection and Hypothesis Testing

732A90
Computational Statistics

Maryna Prus
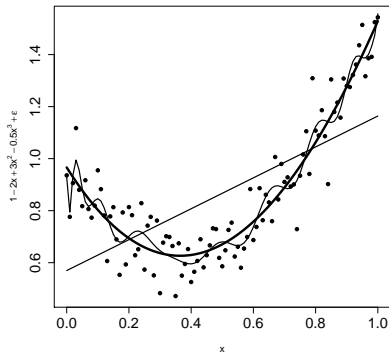(maryna.prus@liu.se)

Slides originally by Krzysztof Bartoszek

November 23, 2021
Department of Computer and Information Science
Linköping University

Tools for model selection

- Comparing different models
  - Hypothesis testing

- Uncertainty estimation
  - Confidence intervals

Hypothesis Testing

1. State

   H_0 - null hypothesis (e.g. no difference between groups)

   H_1 - alternative (e.g. there is difference)

2. Observe data

3. Calculate test statistic

4. Under H_0 test statistic has "known" distribution

5. Decision:

   Is value of test statistic *surprising* (in *critical region*)?

   If so reject H_0 in favour of H_1

6. Two-sided vs. one-sided tests

Hypothesis Testing: Example

$x_1, \ldots, x_n$ - data from $N(\mu, \sigma^2)$

Hypotheses:

$H_0 : \mu = 4$
$H_1 : \mu \neq 4$

Test statistic

$T(x) = \frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$

```
tx<-(mean(x)-4)/(sqrt(var(x)/length(x)))
t0<-qt(0.975,df=length(x)-1)
(tx>t0)||(tx<(-t0)) ## reject if TRUE
```
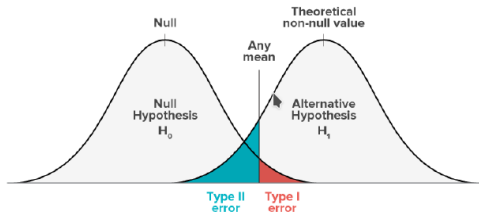
Type II error - probability not to reject H_0 if it is "false"

Power $= 1 -$ Type II error

- ability to correctly identify
surprise,
i.e. indicate H_1

Computing power

- Analytically ?

- Generate data samples that satisfy H_1
  Compute percent of correct rejections

Theoretical
non-null value

Null

Any
mean

Null
Hypothesis
H_0

Alternative
Hypothesis
H_1

Type II   Type I
error   error

Source: grasshopper.com

Monte Carlo Hypothesis Testing

Knowledge of test statistic distribution not needed!

$x_1, \ldots, x_n$ - data from $N(\mu, \sigma^2)$

$H_0 : \mu = 4$
$H_1 : \mu \neq 4$

Test statistic

$$T(x) = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

1: **for** $i = 1$ to B **do**
2:     Generate $Y_1, \ldots, Y_n$ i.i.d. from H_0, i.e. $N(\mathbf{4}, \sigma^2)$, σ^2 est.
3:     Compute t_i from $Y_1, \ldots, Y_n$
4: **end for**
5: Use $t_1, \ldots, t_B$ to construct a histogram
6: Use the histogram as the distribution of $T(x)$ under H_0

Monte Carlo Hypothesis Testing


```
x<-rnorm(10,4,1)
s<-var(x)
B<-10000
n<-length(x)
tsamp<-rep(NA,B)
for (i in 1:B){
    Y<-rnorm(n,4,s)
    tsamp[i]<-(mean(Y)-4)/(sd(Y)/sqrt(length(Y)))
}
hist(tsamp,breaks=50,col=gray(0.8),main="",xlab="t"
    ,ylab="",freq=FALSE,cex.axis=1.5,cex.lab=1.5)
```
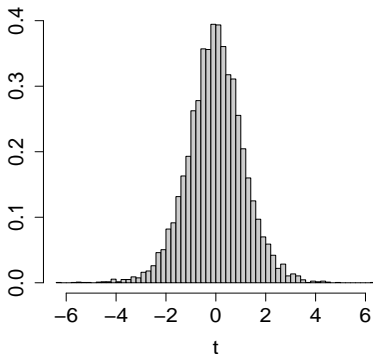
- Introduced by Fisher 1930's

- Computationally expensive

- Applicable to certain types of hypothesis testing:
  - Equality of models, populations, etc.

- Two samples problem:
  - Population 1 distributed as F
  - Population 2 distributed as G

    $H_0 : F = G$
    $H_1 : F \neq G$

Comparing two groups:

- Control - "z"
- Treatment - "y"
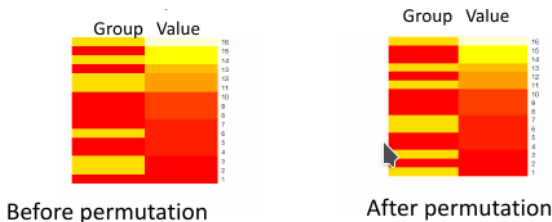
Data:

```
> t(mouse)
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]  [,10] [,11] [,12]
Group "y"   "z"   "z"   "y"   "y"   "z"   "y"   "y"   "y"   "y"   "z"   "z"
Value " 10" " 16" " 23" " 27" " 31" " 38" " 40" " 46" " 50" " 52" " 94" " 99"
      [,13] [,14] [,15] [,16]
Group "y"   "z"   "y"   "z"
Value "104" "141" "146" "197"
```

Do the values differ significantly between control and treatment groups?

IDEA: If $F = G$ then *group label does not matter*

  $\rightarrow$ permute labels and still have sample from F (or G)

Before permutation — After permutation

Test statistic:

$$T = \text{mean}(\text{values}|\text{group} = z) - \text{mean}(\text{values}|\text{group} = y)$$

1: $T(x)$ value of statistic from observed data

2: Create permutations $g_1^*, \ldots, g_B^*$ of group variable

If number of (all) permutations too large
$\rightarrow$ sample B^* of them randomly *without replacement*
further B^* instead of B

3: Evaluate test statistic on each permutation

4: Estimate p-value:
One-sided test: $\hat{p} = \#\{T(X_{g_b^*}) \geq T(X)\}/B$
Two-sided test: $\hat{p} = \#\{|T(X_{g_b^*})| \geq |T(X)|\}/B$

5: p-value law $\Rightarrow$ difference significant $\Rightarrow$ reject H_0

Data	Group	A	B	A	B	B	A	$\overline{A} = 5/3$
	Value	1	2	2	1	1	2	$\overline{B} = 4/3$
Set 1	Group	A	A	B	A	B	B	$\overline{A} = 4/3$
	Value	1	2	2	1	1	2	$\overline{B} = 5/3$
Set 2	Group	B	B	A	B	A	A	$\overline{A} = 5/3$
	Value	1	2	2	1	1	2	$\overline{B} = 4/3$
Set 3	Group	B	B	B	A	A	A	$\overline{A} = 4/3$
	Value	1	2	2	1	1	2	$\overline{B} = 5/3$

Observed difference: $\overline{A} - \overline{B} = 5/3 - 4/3 = 1/3$

From permuted data: $\overline{A} - \overline{B} \in \{-1/3,\ 1/3,\ -1/3\}$

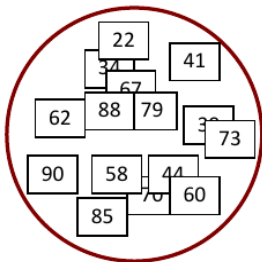Difference between two groups significant?

Do we reject H_0?


```
B=1000
stat=numeric(B)
n=dim(mouse)[1]
for(b in 1:B){
  Gb=sample(mouse$Group, n)
  stat[b]=mean(mouse$Value[Gb=='z'])-mean(mouse$
      Value[Gb=='y'])
}
stat0=mean(mouse$Value[mouse$Group=='z'])-mean(
    mouse$Value[mouse$Group=='y'])
print(c(stat0,mean(stat>stat0)))
## [1] 30.63492   0.12700
```
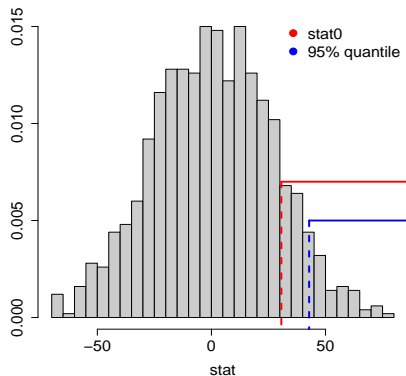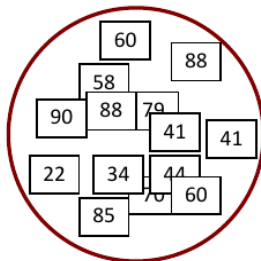
Resampling Methods

Observed data

22
34
67
41
62
88
79
39
73
90
58
44
70 60
85

$\overline{x}$

Sampling with replacement

Sampling without replacement

Resampled data

60
88
58
90
88
79
41
41
22
34
44
70 60
85

$\overline{x}_1^*, \overline{x}_2^*, ..., \overline{x}_N^*$

Jackknife and Bootstrap

Theory *different* **but** coding *similar*

$D = (X_1, \ldots, X_n)$ - data from $F(\cdot, w)$

$\widehat{w} = T(D)$ - estimator

1: **for** $i = 1, \ldots, B$ **do**

2: Generate

   $D_i^* = (X_1^*, \ldots, X_n^*)$ by sampling *with replacement* {*Nonparametric Bootstrap*, F unknown}

   $D_i^* = X_{[-i]}$ {*Jackknife*, F unknown} $\rightarrow$ *see next slide*

   $D_i^* = (X_1^*, \ldots, X_n^*)$ by generating from $F(\cdot, \hat{w})$ {*Parametric Bootstrap*, F known}

3: **end for**

4: Use histogram by $T(D_1^*), \ldots, T(D_B^*)$ as distribution of $\widehat{w}$

- Partition sample into r groups

- Group size $k \geq 1$

- For simplicity: $n = r \cdot k$

- $X_{[-i]}$ sample without group i, $i = 1, \ldots, r$
    - $\rightarrow$ size of $X_{[-i]}$: $n - k$

- In step i compute estimation T from $X_{[-i]}$
    - $\rightarrow B \leq n$ (or $B \leq r$ if $k > 1$)

Nonparametric Bootstrap: Example

Data: 1, 2, 2, 1, 3, 4, 1 $\overline{X} = 14/7 = 2$
Set 1: 1, 2, 2, 2, 3, 1, 1 $\overline{X}_1^* = 12/7$
Set 2: 1, 1, 1, 3, 3, 2, 2 $\overline{X}_2^* = 13/7$
Set 3: 4, 4, 4, 3, 1, 2, 1 $\overline{X}_3^* = 19/7$
Set 4: 2, 1, 1, 2, 1, 1, 4 $\overline{X}_4^* = 12/7$
Set 5: 3, 4, 4, 3, 2, 2, 4 $\overline{X}_5^* = 22/7$

Boostrap sample: $\{12/7, 13/7, 19/7, 12/7, 22/7\}$

$\text{Var}\left[\overline{X}^*\right] = \text{Var}\left[\{12/7, 13/7, 19/7, 12/7, 22/7\}\right]$

Quantiles: $P(\overline{X} \leq 3) \approx 4/5$
$\qquad\qquad P(\overline{X} \leq 2) \approx 3/5$

IDEA: Instead of working with true (unknown) distribution underlying $\overline{X}$ we take the observed sample as substitute

Bootstrap Confidence Intervals

Estimate $100(1 - \alpha)\%$ percentile confidence interval for w

1: Compute $T(D_1^*), \ldots, T(D_B^*)$ using Bootstrap

2: *Bootstrap percentile method*:
   Sort $T(D_1^*), \ldots, T(D_B^*)$ in ascending order obtaining $y_1, \ldots, y_B$

   Bootstrap t-method:
   Compute $y_i = (T(D_i^*) - T(D))/(se(T(D_i^*)))$, $i = 1, \ldots, B$

3: Define $A_1 = \lceil (B \cdot \alpha/2) \rceil$, $A_2 = \lfloor (B - B \cdot \alpha/2) \rfloor$

4: Confidence interval:
   (y_{A_1}, y_{A_2}) {*percentile method*} **or**
   $(T(D) - se(T(D^*)) \cdot y_{A_1}, T(D) + se(T(D^*)) \cdot y_{A_2})$ {*t-method*}

Statistic from observed data not in CI $\rightarrow$ reject H_0

Bootstrap

$$\widehat{\text{Var}[T(\cdot)]} = \frac{1}{B-1} \sum_{i=1}^{B} \left(T(D_i^*) - \overline{T(D^*)} \right)^2$$
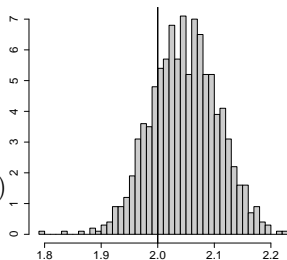
Jackknife $(n = B)$

$$\widehat{\text{Var}[T(\cdot)]} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(T_i^* \right) - J(T))^2 \, ,$$

where

$$T_i^* = nT(D) - (n-1)T(D_i^*) \quad J(T) = \frac{1}{n} \sum_{i=1}^{n} T_i^*$$

Bootstrap in R


```r
library("boot")
stat1<-function(data,vn){
    data<-as.data.frame(data[vn,])
    res<-lm(Response~Predictor,data)
    res$coefficients[2]
}
x<-rnorm(100);data<-cbind(Predictor=x,Response=3+2*
    x+rnorm(length(x),sd=0.5))
res<-boot(data,stat1,R=1000)
print(boot.ci(res))
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
##Based on 1000 bootstrap replicates
#Intervals :
#Level          Normal               Basic
#95%   ( 1.933,  2.164 )   ( 1.935,  2.162 )
# Level       Percentile            BCa
#95%   ( 1.934,  2.161 )   ( 1.936,  2.166 )
```

Bootstrap Bias Correction

1: Observed data: $D = (X_1, \ldots, X_n)$, estimator $\widehat{w} = T(D)$

2: **for** $i = 1, \ldots, B$ **do**

3:     Generate $D_i^* = (X_1^*, \ldots, X_n^*)$ by sampling with replacement

4:     Calculate $T_i^* = T(D_i^*)$

5: **end for**

6: Bias corrected estimator is

$$T_1 := 2 \cdot T(D) - \frac{1}{B} \sum_{i=1}^{B} T_i^*$$

Jackknife also has a bias correction method

- Bootstrap–t method is more accurate than percentile

- Permutations: sampling *without* replacement
  bootstrap: sampling *with* replacement

- Permutation: p-value *exact* if all permutations used
  bootstrap: p-value always *approximate*

- Bootstrap may be used for a wider class of problems

- Methods do not require distributional assumptions

Summary

- Hypothesis testing

- Monte Carlo hypothesis testing

- Resampling methods
    - permutations
    - jackknife
    - nonparametric bootstrap

- Simulation methods
    - parametric bootstrap