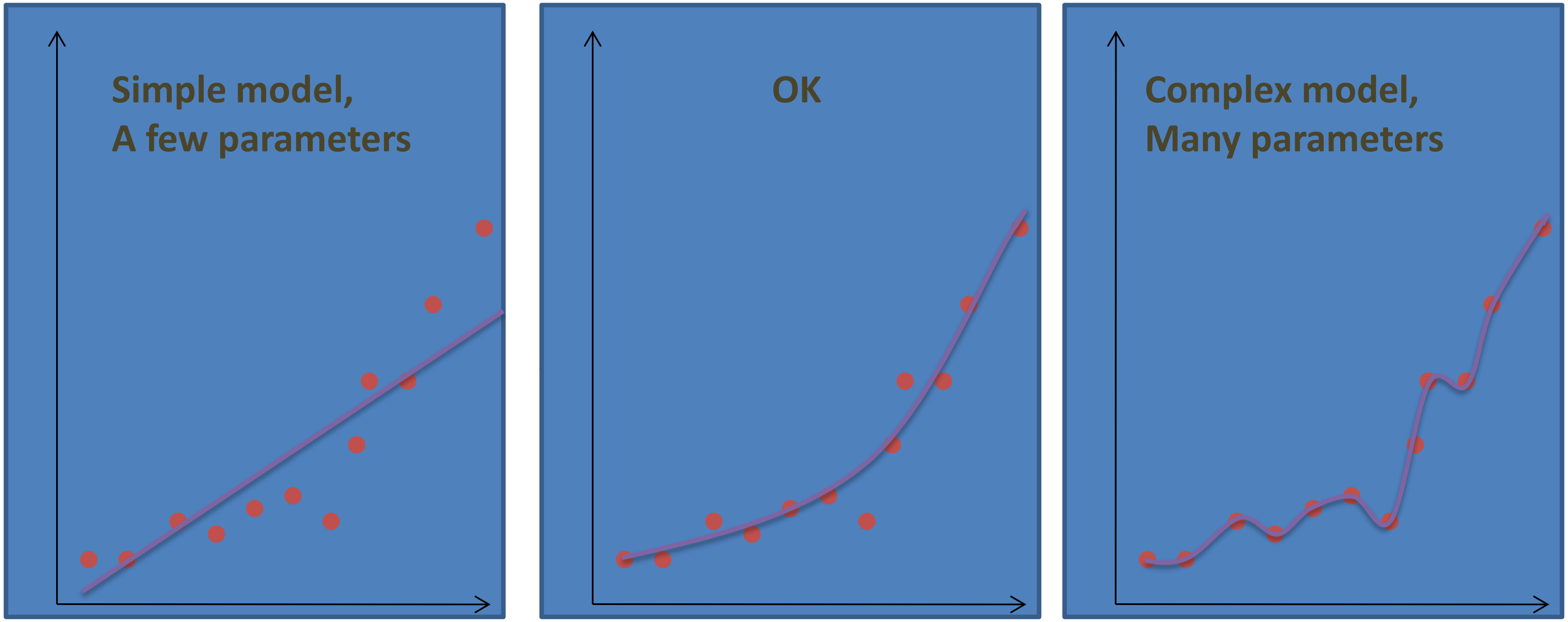
Lecture 5: Numerical model selection and hypothesis testing

Model selection

Overfitting



Model selection

- Necessary tools for selection:
  - Comparison between models
    - Cross-validation
    - Hypothesis testing
  - Uncertainty estimation
    - Confidence intervals

Hypothesis testing

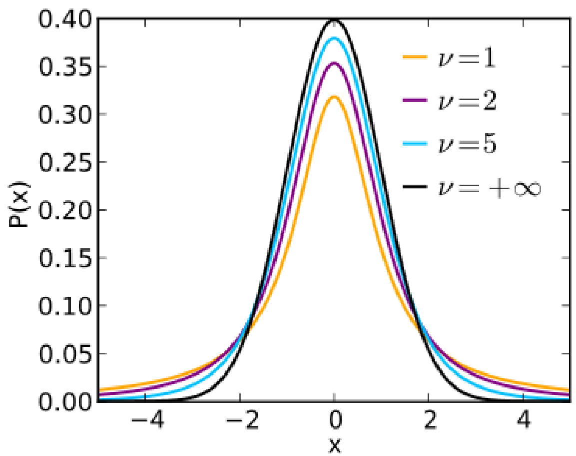
- Given data X
- Null-hypothesis and alternative hypothesis
- Test statistics
  - Some function of a sample
  - Various test statistics have various efficiency (power)
- Distribution of test statistics under H_0
- Decision making: unusual values of test statistics→H_0 is rejected.
  - Two-sided and one-sided tests

Hypothesis testing

- Data

```
> x
 [1] 6.204793 5.868617 5.021237 3.179392 3.577037 4.862277 5.642055 4.007396
 [9] 5.540461 5.596270
```

- Hypotheses:
  - $H_0: \mu = 4, X \sim N(\mu, \sigma^2)$
  - $H_A: \mu \neq 4, X \sim N(\mu, \sigma^2)$



- Test statistics
  - $t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \in t(n-1)$
- Evaluate t for our sample → t_0
- Check if t_0 is in the critical area → reject H_0

Hypothesis testing

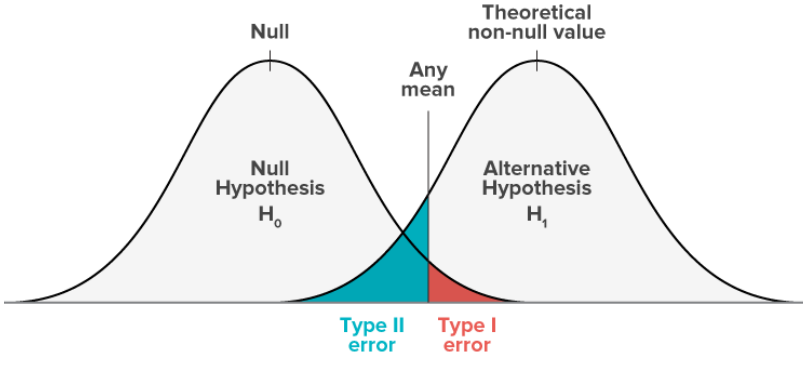
- Monte Carlo Hypothesis testing
  - Use any test statistics
  - We do not need to know how it is distributed

- Hypotheses:
  - $H_0: \mu = 4, X \sim N(\mu, \sigma^2)$
  - $H_A: \mu \neq 4, X \sim N(\mu, \sigma^2)$
- Assume $t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$

1. For $i = 1 \; to \; B$
   1. Generate from $Y \sim N(\mu, \sigma^2)$ → get $Y_1, \ldots Y_n$
   2. Compute t_i from Y
2. Use $t_1, .. t_B$ →build a histogram.
3. Use the histogram as the distribution of t under H_0

Hypothesis testing

- How good is the test statistics? Power!
- $Power = 1 - Type\ II\ error$

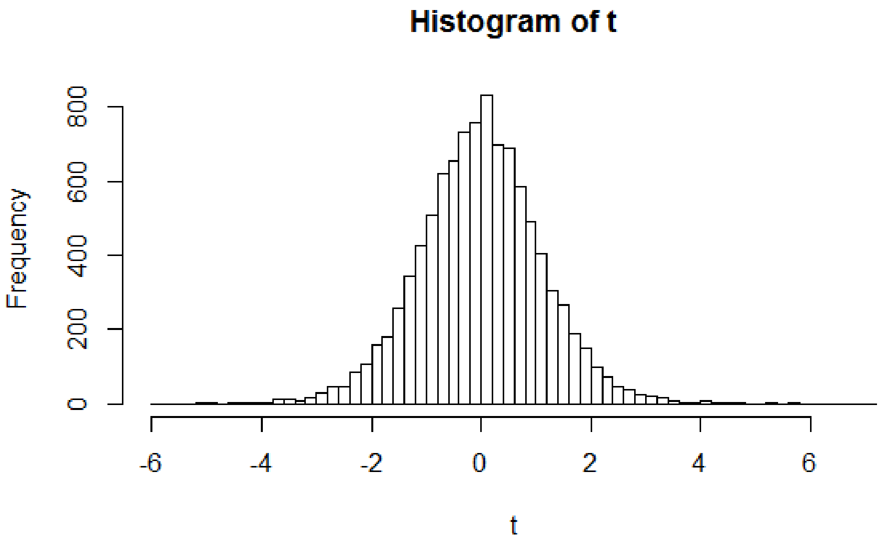


Source: grasshopper.com

- How to compute Power?
  - Generate data samples that satisfy H_a
  - Compute percent of correct rejections

Hypothesis testing

```
s=var(X)
B=10000
n=10
t=numeric(B)
for (i in 1:B) {
  Y=rnorm(10,4,s)
  t[i]=(mean(Y)-4)/(sd(Y)/sqrt(10))
}
hist(t,50)
```



What to do if we don't know the distribution of the data? →
permutation tests or bootstrap tests!

Permutation tests

- Introduced by Fisher 1930's → not used in practice because computationally expensive
- Applicable to certain types of hypothesis testing
  - Equality of models, populations,…
- No assumptions on distributions

- Two-sample problem:
  - Two samples coming from distributions F and G
  - $H_0: F = G$
  - $H_a: F \neq G$

Permutation tests

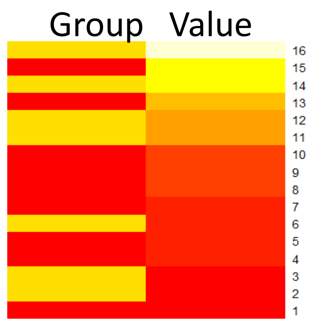
- Example: mouse data
  - Control group
  - Treatment group
    - Group variable g
    - Values variable v

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

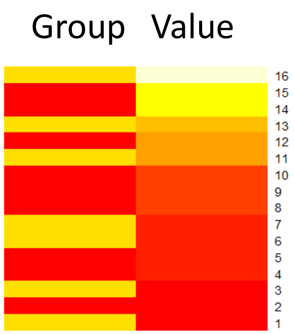
  - Does the value differ in control and treatment groups?

Permutation tests

- Main idea: if $F = G$ →group label does not matter→ we can permute those and still get a valid sample from F (or G)



Before permutation



After permutation

- Suggest test statistics $T = S(g, v)$
  - For example $T = mean(v_i | g_i = z) - mean(v_i | g_i = y)$

Permutation tests

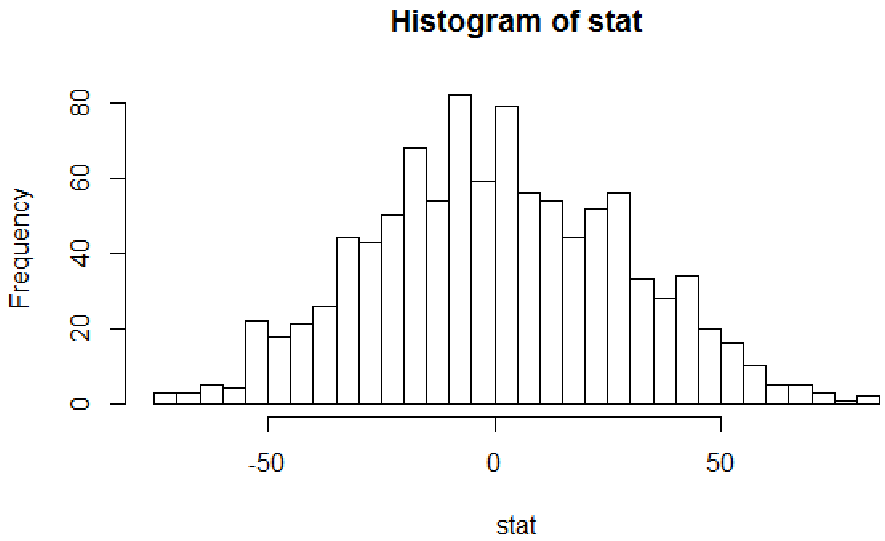
Algorithm

1. Create permutations $g_1^*, \dots g_B$ of group variable
2. Evaluate test statistics on each permutation
   – All permutations are too many? → Sample *n* elements **without replacement** from g
3. Evaluate p-value $\hat{p} = \#\{T(b) \geq T\}/B$
   – In two-sided test, $\hat{p} = \#\{|T(b)| \geq |T|\}/B$

Permutation tests

- Code

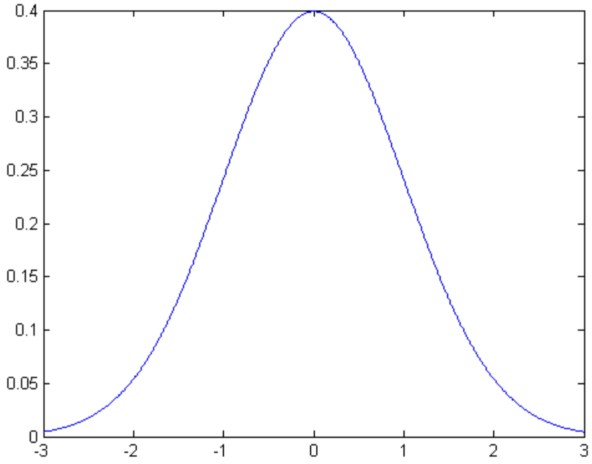
```
B=1000
stat=numeric(B)
n=dim(mouse)[1]
for(b in 1:B){
  Gb=sample(mouse$Group, n)
  stat[b]=mean(mouse$Value[Gb=='z'])-mean(
(mouse$Value[Gb=='y']))
}
hist(stat,50)

stat0=mean(mouse$Value[mouse$Group=='z'])-
mean(mean(mouse$Value[mouse$Group=='y']))

mean(stat>stat0)
```

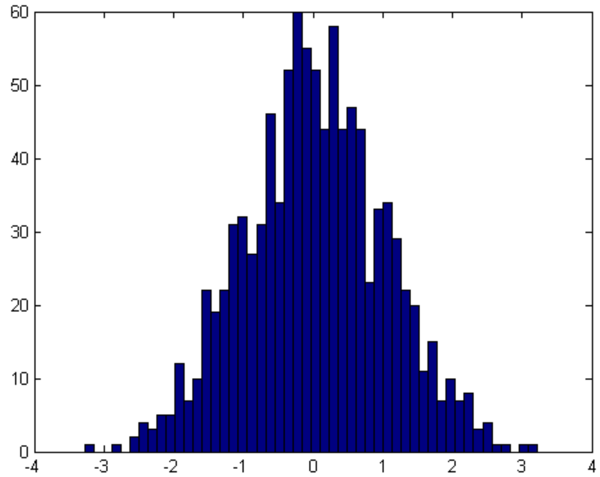



```
> stat0
[1] 30.63492
> mean(stat>stat0)
[1] 0.154
>
```

Do we reject null hypothesis?

The bootstrap: general principle



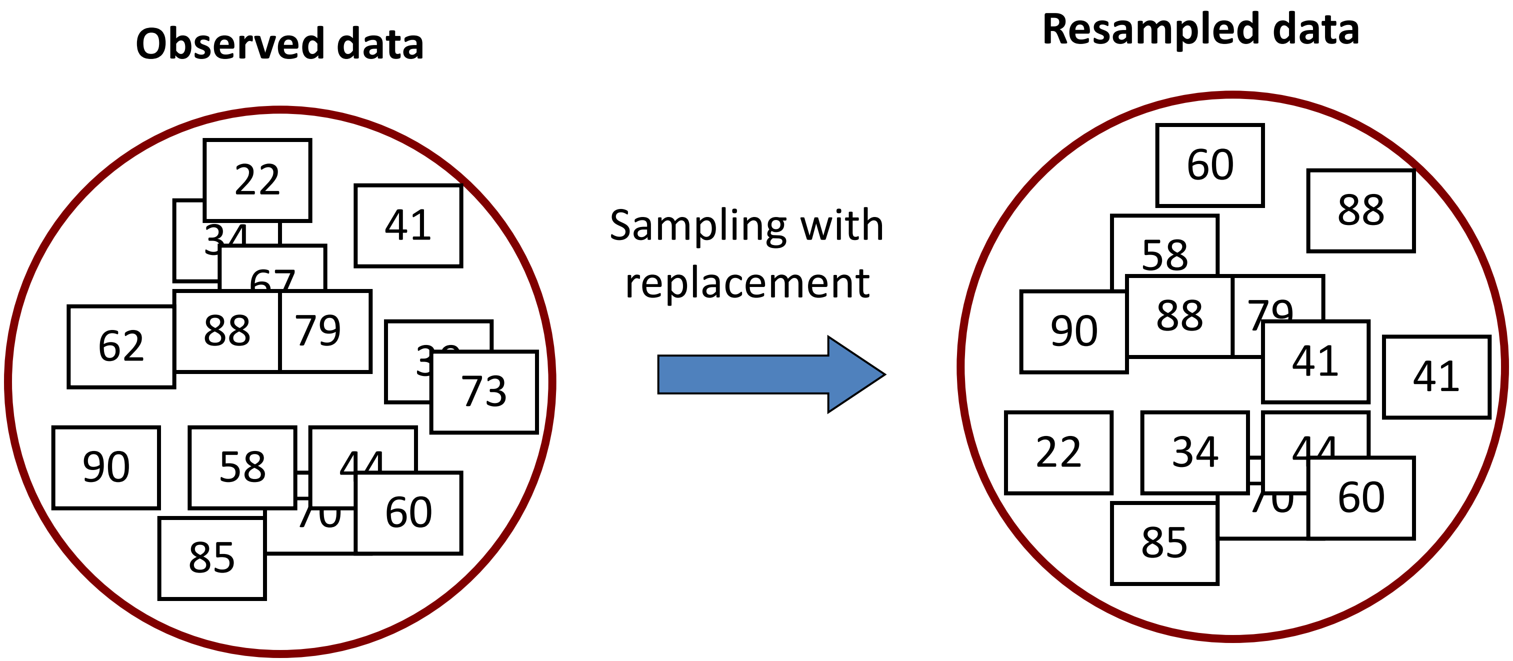


We want to determine uncertainty of $T(D)$

1. Generate many different D_i from their distribution
2. Use histogram of $T(D_i)$ to determine confidence limits→ unfortunately can not be done (distr of *D is often unknown)*

Instead: Generate many different D_i^* from the empirical distribution (histogram)

Nonparametric bootstrap



$\bar{x}$

$\bar{x}_1^*, \bar{x}_2^*, \ldots, \bar{x}_N^*$

Nonparametric bootstrap

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

Assume $X \sim F(X, w)$, F and w are unknown

1. Estimate $\widehat{w}$ from data **D**=(X_1,...X_n)
2. Generate $\mathbf{D_1}$ =(X^*_1,...X^*_n) by sampling with replacement
3. Repeat step 2 B times
4. The distribution of w is given by $T(D_1), \dots T(D_B)$

Nonparametric bootstrap can be applied to any deterministic estimator, distribution-free

Parametric bootstrap

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

Assume $X \sim F(X, w)$, F is known and w is unknown

1. Estimate $\widehat{w}$ from data **D=**$(X_1,\ldots X_n)$
2. Generate $\mathbf{D_1}$ **=**$(X^*_1,\ldots X^*_n)$ by generating from $F(X, \widehat{w})$
3. Repeat step 2 B times
4. The distribution of θ is given by $T(D_1), \ldots T(D_B)$

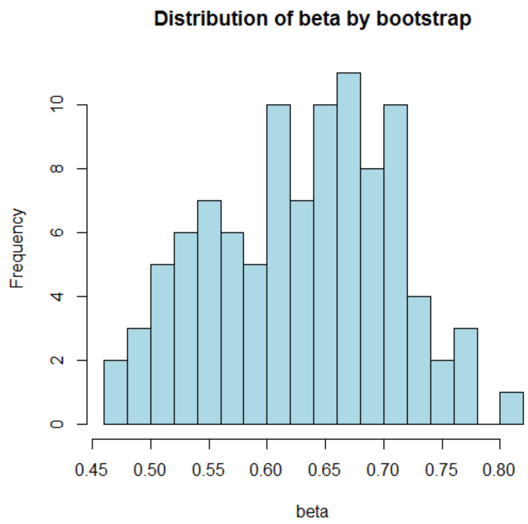
Parametric bootstrap is **more** precise if the distribution form is correct

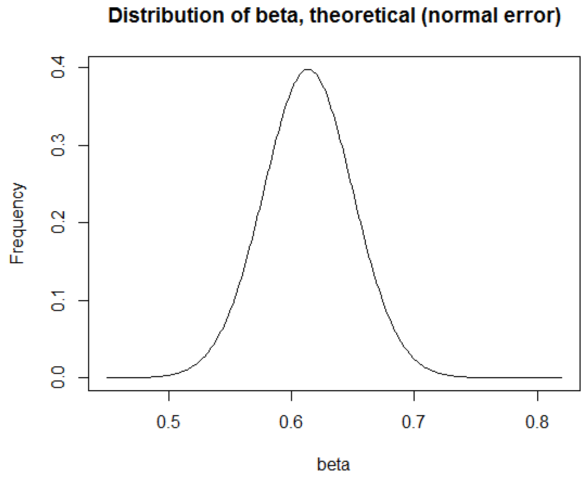
Example

- Distribution of regression coefficient

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

```
stat1<-function(data,n){
  data1=data[n,];
  res<-lm(Price~Area, data1)
  ret=res$coefficients[2]
  return(ret)
}
res=boot(data,stat1,R=100)
hist(res$t,20)
```





Bootstrap confidence intervals

- obtained from the distribution given by the bootstrap
  - **R**: boot.ci() for one variable, envelope() for many variables

```
boot.ci(res)
```

```
> boot.ci(res)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = res)

Intervals :
Level      Normal              Basic
95%   ( 0.4569,  0.7595 )   ( 0.4665,  0.7631 )

Level     Percentile            BCa
95%   ( 0.4642,  0.7609 )   ( 0.4289,  0.7346 )
Calculations and Intervals on Original Scale
```

Uncertainty estimation

1. Get $D_1, \dots D_B$ by bootstrap
2. Use $T(D_1), \dots T(D_B)$ to estimate the uncertainty
   - Boostrap percentile
   - Bootstrap-t
   - Bootstrap Bca
   - …

- Bootstrap works for all distribution types but approximate
- Can be bad accuracy for small data sets $n < 40$ (empirical is far from true)
- Parametric bootstrap works even for small samples

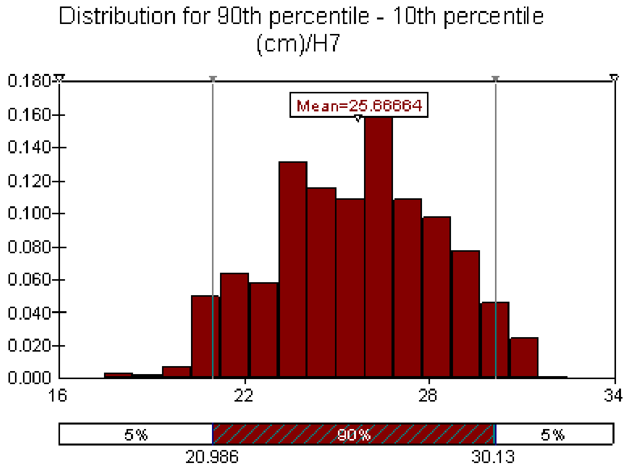
Bootstrap confidence intervals

- To estimate 100(1-α) confidence interval for w

Bootstrap percentile method

1. Using bootstrap, compute $T(D_1), \dots T(D_B)$, sort in ascending order, get $w_1 \dots w_B$
2. Define A_1=ceil(B α/2), A_2=floor(B-B α/2)
3. Confidence interval is given by

$$\left(w_{A_1}, w_{A_2}\right)$$



Look at the plot…

Bootstrap confidence intervals

Bootstrap-t method

- Done by analogy with t test

1. Using bootstrap, compute $T^{*1}=T(\mathbf{D_1})\ldots T^{*B}=T(\mathbf{D_B})$
2. Compute $$t_j=\frac{T^{*j}-T(\mathbf{D})}{se(T^{*j})},\ j=1\ldots B$$
3. Define A_1=ceil(B α/2), A_2=floor(B-B α/2)
4. Confidence interval is $\left(T(D)-se(T)\cdot t_{A_2},T(D)-se(T)\cdot t_{A_1}\right)$

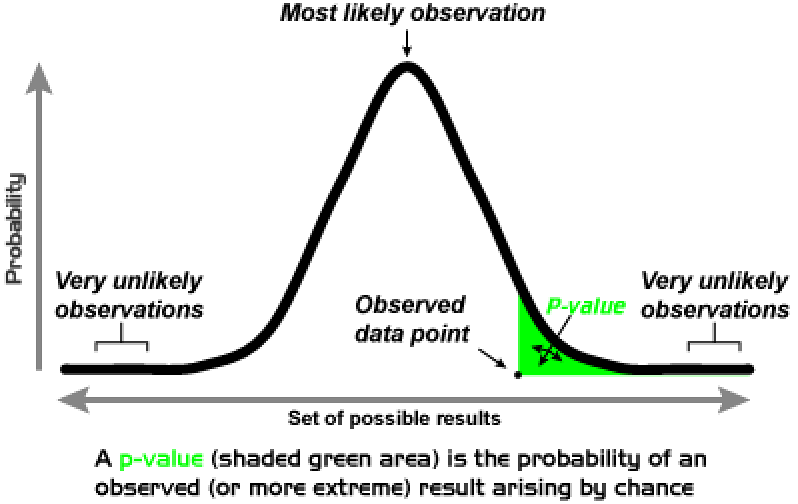
Bootstrap confidence intervals

Comments

- *se* is square root of estimated variance
- Estimation $se(T^{*j})$ typically requires second-level bootstrap -> bootstrap-t is computationally intensive
- Bootstrap-t is more accurate than percentile (coverage error)
- Bootstrap BC_a is a more advanced bootstrap CI method

Bootstrap hypothesis testing

- Bootstrap distribution $T^{*1}=T(\mathbf{D_1})\ldots T^{*B}=T(\mathbf{D_B})$
- Assume $H_0: T = T_0$
  - For ex $\beta = 0$ in regression
- Compute P-value by checking the tail corresponding T_0
- Much more complicated bootstrap hypothesis testing methods exist



Source: Wikipedia

Bootstrap tests vs permutation tests

- Permutation tests do sampling **without** replacement**,** bootstrap does sampling **with** replacement
- Permutation p-value is exact if all permutations considered, bootstrap is always approximate (becomes more exact as $n \to \infty$)
- Bootstrap histograms centered around T, permutation histograms around 0
- Bootstrap tests cover larger class of problems
- $sample\_variance(T^{*b})$ has no meaning for permutation tests, but for bootstrap it is an estimate of variance of T.
- Both methods require no distributional assumptions
- For permutation test, the accuracy of p-value depends on B
  - 10% accuracy achived for p=0.05 if $B \approx 2000$

Permutation tests for model selection

- Given $X_0 = (X_a, X_b, Y)$, model M
- Test
  - H_0: $variables\ X_b should\ not\ be\ in\ M$
  - H_a: $all\ variables\ are\ significant$
- Given test statistics $T(M)$

- Algorithm
  1. Get $\hat{X}$ by permuting columns X_a and fit model $Y = M(\hat{X}, X_b)$
  2. Compute test statistics for this model
  3. Repeat steps 1-2 B times and get a distribution of T
  4. Use it and $T(M(X_o))$ to compute p-value

Bootstrap bias corrections

- Theory shows

$$T_1 = 2T(P_n) - \mathrm{E}\Big(T\big(P_n^{(1)}\big)\;\Big|\;P_n\Big)$$

- The last term is computed by
  1. Using observation set **D=**$(X_1,\ldots X_n)$, sample with replacement and get bootstrap sample $\mathbf{D_1}$ **=**$(X^*_1,\ldots X^*_n)$,
  2. Repeat step 1 B times
  3. Take the mean of T(**D**$_{\mathbf{1}}$)… T(**D**$_{\mathbf{B}}$)

- The first term is the 2T**(D)**

Bootstrap variance estimation

- Using bootstrap, compute $T^{*1}=T(\mathbf{D_1})\ldots T^{*m}=T(\mathbf{D_B})$

$$\widehat{\mathrm{V}}(T) = \frac{1}{m-1}\sum(T^{*j}-\overline{T}^{*})^2$$

Jackknife methods

- Idea: similar to CV, but used in statistical inference
  - Bias estimation
  - Variance estimation

”Jackknife methods make use of systematic partitions of a dataset to estimate properties of an estimator computed from the full sample”

- Suppose, we are given a random sample $\mathbf{Y}=(Y_1,\ldots,Y_n)$ and some estimator $T(\mathbf{Y})$

Jackknife methods

First-order jackknife

1. Obtain $\mathbf{Y}_{(-j)}$ by dropping group of observations j from $\mathbf{Y}$
2. For each j, compute $T_{(-j)} = T(\mathbf{Y}_{(-j)})$
3. Compute pseudovalues and J(T), called *jackknifed* T:

$$\overline{T}_{(\bullet)} = \frac{1}{r}\sum_{j=1}^{r} T_{(-j)} \qquad T_j^* = rT - (r-1)T_{(-j)}$$

$$\mathrm{J}(T) = \frac{1}{r}\sum_{j=1}^{r} T_j^* = \overline{T}^*$$

- Equivalently, $\mathrm{J}(T) = rT - (r-1)\overline{T}_{(\bullet)}$

Jackknife variance estimate

- We can use $T_{(-j)}$ or pseudovalues as estimates of T for different samples (both give equivalent expression).

- Variance becomes

$$\widehat{\mathrm{V}(T)}_{\mathrm{J}} = \frac{\sum_{j=1}^{r} \left(T_j^* - \mathrm{J}(T)\right)^2}{r(r-1)}$$

Sometimes, one takes $\frac{\sum_{j=1}^{r} (T_j^* - T)^2}{r(r-1)}$

!The variance is often overestimated

Jackknife bias correction

First-order jackknife

- The bias reduced to order n^{-1} (we take r=n)

$$\mathrm{Bias}(T) = \mathrm{E}(T) - \theta \;\; = \sum_{q=1}^{\infty} \frac{a_q}{n^q}$$

$$\begin{aligned}
\mathrm{Bias}(\mathrm{J}(T)) &= \mathrm{E}(\mathrm{J}(T)) - \theta \\
&= n(\mathrm{E}(T) - \theta) - \frac{n-1}{n} \sum_{j=1}^{n} \mathrm{E}(T_{(-j)} - \theta) \\
&= n \sum_{q=1}^{\infty} \frac{a_q}{n^q} - (n-1) \left(\sum_{q=1}^{\infty} \frac{a_q}{(n-1)^q} \right) \\
&= a_2 \left(\frac{1}{n} - \frac{1}{n-1} \right) + a_3 \left(\frac{1}{n^2} - \frac{1}{(n-1)^2} \right) + \ldots \\
&= -a_2 \left(\frac{1}{n(n-1)} \right) + a_3 \left(\frac{1}{n^2} - \frac{1}{(n-1)^2} \right) + \ldots
\end{aligned}$$

Jackknife estimation of bias

- We see that

$$\mathrm{E}(\mathrm{J}(T)) - \theta \;=\; \mathrm{E}(T) - \theta + (n-1)\left(\mathrm{E}(T) - \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}(T_{(-j)})\right)$$

- Hence, bias is $B_{\mathrm{J}} = (n-1)\left(\overline{T}_{(\bullet)} - T\right)$

Higher-order jackknife

The order of the bias can be further reduced

- Second-order jackknife

$$\mathrm{J}^2(T) = \frac{n^2 \mathrm{J}(T) - (n-1)^2 \sum_{j=1}^{n} \mathrm{J}(T)_{(-j)}/n}{n^2 - (n-1)^2}$$

- Higer order jackkifes –combining jackknifes of lower orders:

$$T_w = \frac{T_1 - wT_2}{1 - w}$$

Higher-order jackknife

Comments

- High order jackknifes reduce the bias but they increase the variance

- Delete-1 jackknife is not always appropriate (median). Use delete-k