Resampling

STK473 – Praktikum 7

MRJ

Simulasi Monte Carlo

- Simulasi yang memanfaatkan informasi mengenai sebaran data yang **diketahui (dihipotesiskan, dianggap tahu)** dengan pasti.

- Pendugaan Peluang

$$P(Y > a) \; atau \; P(Y < a)$$

dengan $Y = g(x),\ X \sim f(x)$

  – Menghitung peluang dari suatu peubah acak
  – Pengujian hipotesis

Simulasi Monte Carlo

– Pengujian Hipotesis

Pada kasus contoh acak yang diambil berasal dari populasi yang menyebar normal, nilai statistik uji t akan menyebar $t_{db=(n-1)}$, sementara untuk populasi dengan sebaran lainnya, perlu dikaji ulang sebaran dari nilai statistik uji t

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Simulasi dilakukan untuk melihat hasil dari sebaran hipotetik populasi ketika menghitung nilai statistik uji t

Ilustrasi 1

Misal X~$N(0,1)$, jika diketahui Y=2X, berapa $P(Y$>2)?

Algoritme :

1. Bangkitkan X~$N(0,1)$
2. Hitung Y=2X
3. Ulangi langkah 1 dan 2 sebanyak 100000 kali
4. Hitung presentasi nilai Y>2

Simulasi

```
n<-100000
x<-rnorm(n)
y<-2*x
y1<-ifelse(y>2,1,0)
mean(y1) #P(Y>2) secara empirik
```

$$X \sim N(0,1)\ ;Y = 2X$$
$$E(Y) = 2E(X) = 0\ ;\ V(Y) = 2^2 V(X) = 4$$
$$Y \sim N(0,4)$$

```
pnorm(2,0,2,lower.tail=F) #P(Y>2) dari sebaran hipotetik
```

Ilustrasi 2

- Misal peubah acak $X \sim \exp(\lambda)$, kita ingin menguji hipotesis

$$H_0 : \lambda = 2$$
$$H_1 : \lambda = 4$$

Jika kita menolak H_0 ketika $x > 1$.

Hitung α dan kuasa ujinya !

Jawaban

- $\alpha = P(tolak\ H_0 |\ H_0\ benar)$
  $= P(x > 1 | \lambda = 2)$
  $= \int_1^\infty \frac{1}{2} e^{-x/2}\, dx = 0.135$
- Kuasa uji $= 1 - \beta$
  $= P(tolak\ H_0 |\ H_0\ salah)$
  $= P(x > 1 | \lambda = 4)$
  $= \int_1^\infty \frac{1}{4} e^{-x/4}\, dx = 0.018$

Simulasi

```
lambda<-2
N<-100000
k<-1
x<-rexp(N,lambda)
y<-ifelse(x>k,1,0)
hasil<-mean(y) #alpha

lambda<-4
N<-100000
k<-1
x<-rexp(N,lambda)
y<-ifelse(x>k,1,0)
hasil<-mean(y) #kuasa uji
```

Ilustrasi 3

- Berikut tersedia data dari sebaran eksponensial:

1.68760	0.03120
0.03037	0.61068
0.91673	0.63169
1.34939	2.99986
0.08164	2.70955

- $H_0: \lambda = 1$
- Apa kesimpulan yang bisa diambil?

Ilustrasi 3

- Algoritme :

1. Hitung $t_{data} = \left|\frac{\bar{x}-\mu_0}{s/\sqrt{n}}\right|$ dari data
2. Bangkitkan 10 contoh acak~$\exp(\lambda = 1)$
3. Hitung $t_{dist} = \left|\frac{\bar{x}-\mu_0}{s/\sqrt{n}}\right|$ dari sebaran hipotetik
4. Ulangi langkah 2 dan 3 sebanyak 10000 kali
5. Hitung nilai-p=$P(t_{dist}>t_{data})$

Simulasi

```
lambda<-1
k<-10000
data1<-c(1.6876,0.03037,0.91673,1.34939,0.08164,
         0.0312,0.61068,0.63169,2.99986,2.70955)
n<-length(data1)
m<-mean(data1)
s<-sd(data1)
tdata<-abs((m-1/lambda)/(s/sqrt(n)))

data2<-matrix(rexp(n*k,lambda),k)
m1<-apply(data2,1,mean)
s1<-apply(data2,1,sd)
tdist<-abs((m1-1/lambda)/(s1/sqrt(n)))

y<-ifelse(tdist>tdata,1,0)
pvalue<-mean(y)
kesimpulan<-ifelse(pvalue<0.05,"Tolak H0","Tak Tolak H0")
pvalue;kesimpulan
```

Simulasi

- Cara lain menghitung nilai-p. Ulangi 9999 kali. Gabungkan nilai t dari data asli dengan nilai t dari simulasi. Urutkan nilai t, dan perhatikan pada persentil keberapa nilai t hitung yang kita miliki.

```
y1<-c(tdata,tdist[-k])
y1<-sort(y1)
pvalue1<-1-which(y1==tdata)/k
kesimpulan1<-ifelse(pvalue1<0.05,"Tolak H0",
    "Tak Tolak H0")
pvalue1;kesimpulan1
```

Ilustrasi 3

- Bagaimana jika contoh acak tersebut diasumsikan berasal dari sebaran normal, lalu hipotesis yang ingin diuji adalah:

- $H_0: \mu = 2$

- Kesimpulan apa yang bisa diambil? Dari program sebelumnya, apa saja yang perlu diubah?
- Hint: untuk kesederhanaan gunakan $\sigma^2 = 1$

Case

Suppose we interest to estimate the ratio of Y/X, or Y/(X+Y), etc

So we decide to take a sample of X and Y

The question is “What is the estimated value of the ratio” and also “How we can estimate the SE of estimated ratio” ?

Case

The sample

$X = \{x_1, x_2, \ldots., x_n\}$

$Y = \{y_1, y_2, \ldots., y_n\}$

Ratio Y/X

Estimated by : $r = \frac{\bar{y}}{\bar{x}}$

Estimated SE : $\sqrt{\hat{V}(r)} = \left(1 - \frac{n}{N}\right)\left(\frac{1}{{\mu_X}^2}\right)\frac{{s_r}^2}{n}$

Ratio Y/(X+Y)

?

MRJ

Case

Ratio Y/(X+Y)

Estimator : $r = \frac{\bar{y}}{\bar{x}+\bar{y}}$

Estimator of SE ?

→ **Resampling** For estimating SE

Jackknife

- We have a sample $y = (y_1, \dots, y_n)$ to estimate θ with the estimator $\hat{\theta} = f(y)$
- Target : estimate standard error of $\hat{\theta}$
- The leave-one-out observation samples
  $$y_{(i)} = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$$
  for $i = 1, \dots, n$ are called **jacknife samples**
- Jacknife statistics are $\hat{\theta}_{(i)} = f(y_{(i)})$

Jacknife estimators

$$\hat{\theta}_{jack} = n\hat{\theta} - (n-1)\hat{\theta}_{(.)} \qquad \hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)}$$

$$V_{jack}(\hat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(.)}\right)^2$$

Example 1

Data

Branch store	Income	
	Item1	Item2
A	1363	1087
B	670	571
C	761	518
D	746	612
E	991	770
F	798	655

Estimate the ratio of Item1/total item using jacknife!

Simulation

Algorithm:

1. Compute $\hat{\theta} = r = \frac{\bar{y}}{\bar{x}+\bar{y}}$
2. Take jacknife sample $x_{(i)}$ & $y_{(i)}; i = 1, \dots, n$
3. Compute jacknife estimator

$$\hat{\theta}_{jack} = n\hat{\theta} - (n-1)\hat{\theta}_{(.)} = n\frac{\bar{y}}{\bar{x}+\bar{y}} - \frac{(n-1)}{n}\sum_{i=1}^{n}\frac{\bar{y}_{(i)}}{\bar{x}_{(i)}+\bar{y}_{(i)}}$$

$$V_{jack}(\hat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{\bar{y}_{(i)}}{\bar{x}_{(i)}+\bar{y}_{(i)}} - \frac{1}{n}\sum_{i=1}^{n}\frac{\bar{y}_{(i)}}{\bar{x}_{(i)}+\bar{y}_{(i)}}\right)^2$$

Simulation

```
store<-LETTERS[1:6]
item1<-c(1363,670,761,746,991,798)
item2<-c(1087,571,518,612,770,655)
data1<-data.frame(store,item1,item2)
n<-nrow(data1)
teta_hat<-mean(data1$item1)/
     (mean(data1$item1)+mean(data1$item2))
y<-matrix(NA,n,n-1)
x<-matrix(NA,n,n-1)
for (i in 1:n) {
  y[i,]<-data1$item1[-i]
  x[i,]<-data1$item2[-i]
}
ybar<-apply(y,1,mean)
xbar<-apply(x,1,mean)
teta_i<-ybar/(xbar+ybar)
teta_jk<-n*teta_hat-(n-1)*mean(teta_i)
var_jk<-(n-1)/n*(sum(teta_i^2)-(n*mean(teta_i)^2))
se_jk<-sqrt(var_jk)
```

Bootstrap

- We have a sample $y = (y_1, \dots, y_n)$ to estimate θ with the estimator $\hat{\theta} = f(y)$
- Steps
  - Repeatedly simulate sample of size n from $y \rightarrow {y^b}_{(i)}$
  - Compute statistic of interest ${\hat{\theta}^b}_{(i)} = f({y^b}_{(i)})$
  - Study behavior of statistic over N repetitions

Bootstrap estimators

$$\hat{\theta}_b = \frac{1}{N}\sum_{i=1}^{N} {\hat{\theta}^b}_{(i)} \qquad V_b(\hat{\theta}) = \frac{1}{N-1}\sum_{i=1}^{N}\left({\hat{\theta}^b}_{(i)} - \hat{\theta}_b\right)^2$$

Example 2

Data

Branch store	Income	
	Item1	Item2
A	1363	1087
B	670	571
C	761	518
D	746	612
E	991	770
F	798	655

Estimate the ratio of Item1/total item using bootstrap!

Simulation

Algorithm:

1. Repeatedly simulate sample of size n from $x \rightarrow x^b{}_{(i)}$ & $y \rightarrow y^b{}_{(i)}$
2. Compute statistic $\hat{\theta}^b{}_{(i)} = \frac{\bar{y}^b{}_{(i)}}{\bar{x}^b{}_{(i)} + \bar{y}^b{}_{(i)}}$
3. Compute bootstrap estimator

$$\hat{\theta}_b = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}^b{}_{(i)} = \frac{1}{N}\sum_{i=1}^{N} \frac{\bar{y}^b{}_{(i)}}{\bar{x}^b{}_{(i)} + \bar{y}^b{}_{(i)}}$$

$$V_b(\hat{\theta}) = \frac{1}{N-1}\sum_{i=1}^{N} \left(\hat{\theta}^b{}_{(i)} - \hat{\theta}_b\right)^2$$

Simulation

```
store<-LETTERS[1:6]
item1<-c(1363,670,761,746,991,798)
item2<-c(1087,571,518,612,770,655)
data1<-data.frame(store,item1,item2)
n<-nrow(data1)
b<-1000
y<-matrix(sample(data1$item1,n*b,replace=T),b)
x<-matrix(sample(data1$item2,n*b,replace=T),b)
ybar<-apply(y,1,mean)
xbar<-apply(x,1,mean)
teta_i<-ybar/(xbar+ybar)
teta_b<-mean(teta_i)
var_b<-(sum(teta_i^2)-(b*teta_b^2))/(b-1)
se_b<-sqrt(var_b)
```

thank you!

MRJ