刘程华 2018011687 计91

Prob 1

证明经验分布函数是非参数的最大似然估计

> 我们先证明随机变量 X 只取离散值 则经验分布是最大似然估计量。

设 $\{X_i\}_{i=1}^n$ 是只能取离散值 $\{x_s\}$ 的 iid 的随机变量. 设 n_s 为 x_s 出现的次数。

$P(X=x_s)=\pi_s$, $\sum_{s=1}\pi_s=1$ $\{X_i\}$ 的似然函数取对数

$$\log L_n(\pi_1 \cdots)=\sum_{i=1}^n \log \pi_{x_i}=\sum_{s=1} n_s \log \pi_s = \sum_{s=1} n_s \log \pi_s .$$

其 Lagrange 函数 $\mathcal{L}(\pi_1 \cdots \lambda)=\sum_{s=1} n_s \log \pi_s+\lambda(\sum_{s=1}\pi_s-1)$

$$\begin{cases} \frac{n_s}{\pi_s}+\lambda=0 \quad s=1,2\cdots \\ \sum \pi_s-1=0 \end{cases} \Rightarrow \hat{\pi}_s=\frac{n_s}{n} \text{ 此时 } L_n(\pi_1\cdots) \text{ 取极大}$$

由 $\pi_s=\frac{n_s}{n}$ 可得到 $F_n(x)=\frac{1}{n}\sum_{i=1}^n I(X_i\le x)$, 证毕

> 下面我们来考虑连续型随机变量, 我们引入 X 的分布函数 $G(x)$

$$P(X=x)=G(x)-G(x_-), \quad G(x_-)=G(X<x)$$

设 $z_1<z_2<\cdots<z_m$ 是样本 $\{X_i\}_{i=1}^n$ 的不同值. n_s 是 z_s 的发生次数.

$\{X_i\}$ 的似然函数 $L_n(F)=\prod_{i=1}^n [G(X_i)-G(X_{i-})]$

令 $\hat{F}_n(x)=\frac{1}{n}\sum_{i=1}^n I(X_i\le x)$, $p_s=G(X_s)-G(X_{s-})$, $\hat{p}_s=\frac{n_s}{n}$

$$\log L_n(G)-\log L_n(\hat{F}_n)=\sum_{s=1}^m n_s \log\frac{p_s}{\hat{p}_s}=n\sum_{s=1}^m \hat{p}_s \log\frac{p_s}{\hat{p}_s}$$

$$=nE\left[\log\frac{p_s}{\hat{p}_s}\right] \le n\log\left(E\left[\frac{p_s}{\hat{p}_s}\right]\right)=n\log\left(\sum p_s\right)=0$$

$P(S=z_s)=\hat{p}_s$, Jensen 不等式

$\Rightarrow L_n(G) \le L_n(\hat{F}_n)$ 当且仅当 $p_s = \frac{n_s}{n}$ 时等号成立，证毕！

Prob 2、

Dvoretzky-Kiefer-Wolfowitz inequality

$P(\sup_x |F(x) - \hat{F}_n(x)| > \varepsilon) \le 2\exp(-2n\varepsilon^2)$

其中 $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$

$\varepsilon_n^2 = \frac{\ln(\frac{2}{\alpha})}{2n}$ ， $L(x) = \max\{\hat{F}_n(x) - \varepsilon_n, 0\}$

$U(x) = \min\{\hat{F}_n(x) + \varepsilon_n, 1\}$

$P(F \in C_n) \ge 1-\alpha$ 置信带为 $(L(x), U(x))$

通过查看R中ks.test函数实现可以发现它也是用DKW来控制误差的

因此题目中的模拟与ks.test是等价的

(1) $F = N(0,1)$ 借用R中函数 ks.test 我们很容易进行模拟

95%置信带包含真实分布函数的数据集比例为 $\frac{956}{1000}$

(2) F是标准 Cauchy 分布，比例为 $\frac{951}{1000}$

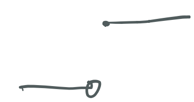
```
> in_ci_num<-0
> for(k in 1:1000){
+     n <- 100
+     x <- rnorm(n)
+     tmp<-ks.test(x,"pnorm")
+     if(tmp["statistic"]>0.05)
+         in_ci_num=in_ci_num+1
+ }
> in_ci_num
[1] 956
```

```
> in_ci_num<-0
> for(k in 1:1000){
+     n <- 100
+     x <- rcauchy(n)
+     tmp<-ks.test(x,"pcauchy")
+     if(tmp["statistic"]>0.05)
+         in_ci_num=in_ci_num+1
+ }
> in_ci_num
[1] 951
```

Prob 3

$$F_n^{-1}(p)=\inf\{y: F_n(y)\geq p\}$$

F_n:



F_n是右连续的，如果 F 在 a 处有间断

$F_n^{-1}(p)=a \Rightarrow F_n(a)\geq p \quad F_n(a-)<p$

$\Rightarrow \quad p\in(F_n(a-),F_n(a)]$

$\Rightarrow \quad \lim_{p\to F_n(a)} F_n^{-1}(p)=\lim_{p\to F_n(a-)} F_n^{-1}(p)=F_n^{-1}(F_n(a))=a$

F_n^{-1} 是左连续的

Prob 4.

Dvoretzky–Kiefer–Wolfowitz inequality

$$P(\sup_x |F(x)-\hat{F}_n(x)|>\varepsilon)\leq 2\exp(-2n\varepsilon^2)$$

其中 $\hat{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n} I(X_i\leq x)$

$\varepsilon_n^2=\dfrac{\ln(\frac{2}{\alpha})}{2n}$, $L(x)=\max\{\hat{F}_n(x)-\varepsilon_n,0\}$

$U(x)=\min\{\hat{F}_n(x)+\varepsilon_n,1\}$

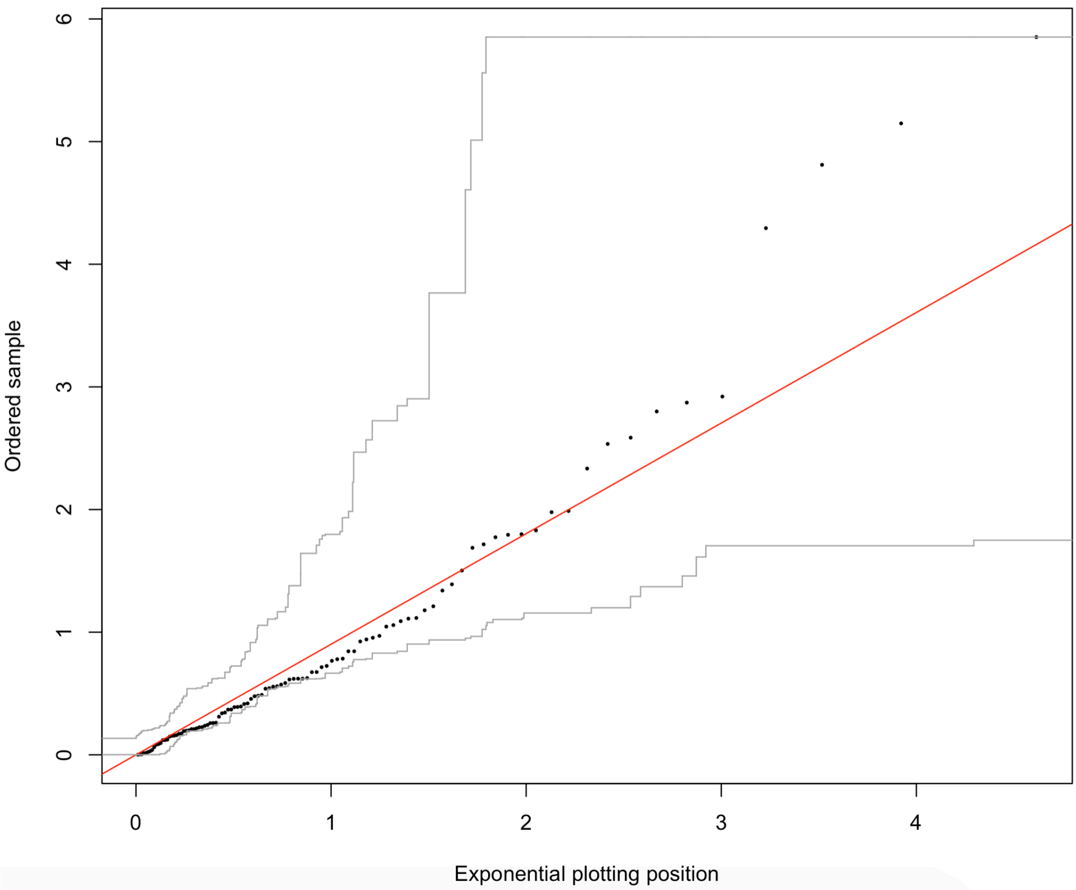
$P(F\in C_n)\geq 1-\alpha$ 置信带为 $(L(x),U(x))$

Q-Q 图中

上界 $(x, F^{-1}(U(\hat{F}_n(x))))$ 其中 F^{-1}的定义同 Prob 3

下界 $(x, F^{-1}(L(\hat{F}_n(x))))$

图和代码见下页




```
qqexp_cb <-  function(y, line=FALSE) {
  y <- y[!is.na(y)]
  Fn<-ecdf(y)
  n <- length(y)
  x <- qexp(c(1:n)/(n+1))
  m <- mean(y)
  if (any(range(y)<0)) stop("Data contains negative values")
  ylim <- c(0,max(y))
  qqplot(x, y, xlab="Exponential plotting position",ylim=ylim,ylab="Ordered
sample",col="black", pch=16, cex=0.4)
  if(line)abline(0,m,col="red",lty=1)
  y<-sort(y)
  epsilon<-sqrt((1/(2*n))*log(2/0.05))
  for(i in 1:n+1){
    l[i]<-quantile(y,max(Fn(y[i])-epsilon,0))
    u[i]<-quantile(y,min(Fn(y[i])+epsilon,1))
  }
  plot(stepfun(y,l),add=T,do.points=F,col="darkgrey")
  plot(stepfun(y,u),add=T,do.points=F,col="darkgrey")
  invisible()
}
```