2017－2018 年第一学期

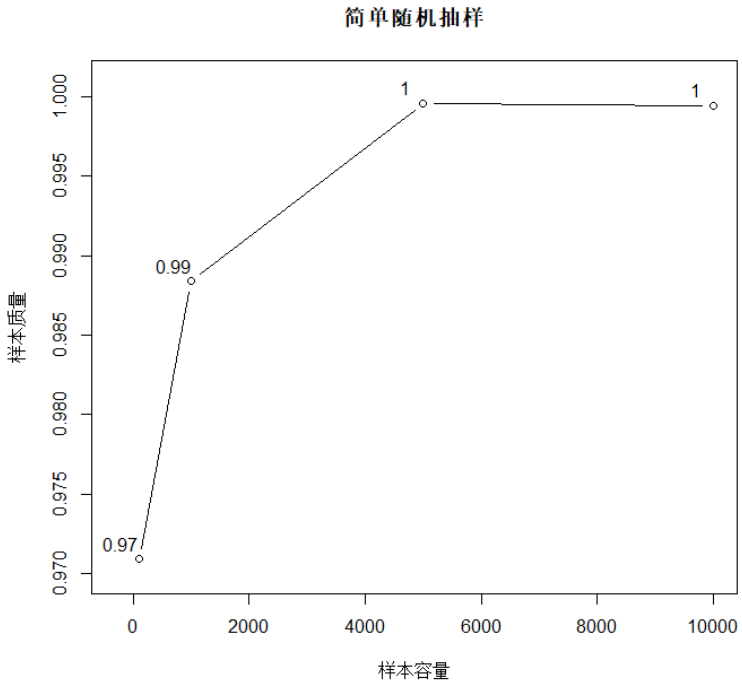
《大数据统计基础》试题　答题纸

学校：中央财经大学　　学号：2017210785　　姓名：司徒雪颖　　成绩_____

一、抽样

1. 抽取样本容量为 100、1000、5000、10000 的样本质量

如图所示，样本质量分别是 0.97,0.99,1,1



2. 求样本质量为 95%时最优样本容量

代码输出结果为 180

```
> s = smooth.spline(q11,samp1) #对曲线进行拟合
> pr = predict(s,0.95);pr #当样本质量为95%时的样本容量
$x
[1] 0.95

$y
[1] 180.0419
```

附：抽样代码

```
setwd("E:/graduate/class/2017《大数据统计基础》考试题/")
mydata = read.csv("LoanStats3c.csv",header = T,skip=1)
# head(mydata)
data0 = na.omit(mydata$loan_amnt)
# hist(data0)
N = length(data0)
data1 = cut(data0,breaks = c(0,5000*(1:6),max(data0)))
pd = table(data1)/N

samp = c(100,1000,5000,10000) #抽样
n  = length(samp);n
fun1 = function(i,data0,maxsamp)
{
  p = sample(data0,i)#p 抽到的样本编号，一共抽了 i 个样本
  p = c(p,matrix(NA,1,maxsamp-length(p)))
  return(p)#把抽到的样本编号存到 p 里，p 的长度是 samp[n]-length(p)，不够的 NA 补齐
}
samp = as.matrix(samp)
ma = apply(samp,1,fun1,data0,samp[n])
# head(ma)
# dim(ma)

fun2 = function(datasam1,pd)
{
  datasam11 = cut(na.omit(datasam1),breaks = c(c(0,5000*(1:6),max(data0))))
  ps = table(datasam11)/length(na.omit(datasam1))+0.00001#????
  j = sum((ps-pd)*(log(ps/pd)))
  q = exp(-j)
  return(q)
}
q1 = apply(ma,2,fun2,pd);q1
# length(q1);dim(samp)
par(mfrow = c(1,1))
plot(samp,q1,xlab = "样本容量",ylab = "样本质量",main = "简单随机抽样",
     type = "b",xlim = c(-300,10000),ylim = c(0.97,1.001))
text(samp-300,q1+0.001,round(q1,2))

#求最优样本容量
x = seq(6,20,0.1)
y = 2^x
plot(x,y)
```

```
samp1 = round(y)[y>100 & y<35000];samp1 #去头去尾
samp1 = as.matrix(samp1)
n1  = length(samp1);n1
ma1 = apply(samp1,1,fun1,data0,samp1[n1])
q11 = apply(ma1,2,fun2,pd)
plot(samp1,q11,xlab = "样本容量",ylab = "样本质量",main = "简单随机抽样",type =
"l")

s = smooth.spline(q11,samp1) #对曲线进行拟合
pr = predict(s,0.95);pr #当样本质量为95%时的样本容量
```