抽样作业

司徒雪颖
中央财经大学

一、 简单随机抽样

样本容量 samp 从 64 到 32768，一共 91 种

```
setwd("E:/graduate/class/抽样/")
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

x = seq(6,20,0.1)
y = 2^x
plot(x,y)
samp = round(y)[y>50 & y<35000];samp #去头去尾
n  = length(samp);n
```

```
#简单随机抽样--------------------------
fun1 = function(i,data0,sampmax)
{
  set.seed(1995)
  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,sampmax-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)
length(q1);dim(samp)
plot(samp,q1,xlab = "样本容量",ylab = "样本质量",main = "简单随机抽样
")
```

由图 1-1 可以看出，当样本容量少于 100 或 100 左右时，样本质量低于 90%，此后随着样本容量上升，样本质量迅速趋近于 1，当样本容量为 832 时样本质量已达 99%。

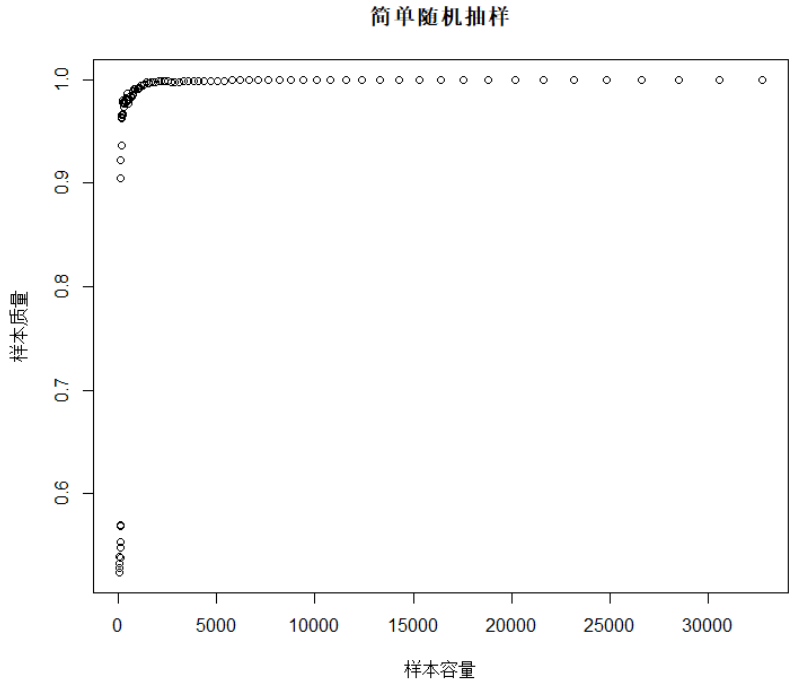

图 1- 1 简单随机抽样的样本容量与样本质量的散点图

二、 分层抽样

```
#分层抽样--------------------
str = length(levels(data1))
data2 = cbind(data0,data1)
fun3 = function(s,pd,data0)
{
  p = NULL
  N = length(data0)
  for(j in 1:str)
  {
    set.seed(1995)
    samp2 = NULL
    samp2 = sample((1:N)[data2[,2]==j],round(s*pd[j]))
    p = c(p,samp2)
  }
  res = c(data0[p],matrix(NA,1,samp[n]+5-length(p)))
  return(res)
```

```
}
mb = apply(samp,1,fun3,pd,data0)
q2 = apply(mb,2,fun2,pd)
plot(samp,q2,xlab = "样本容量",ylab = "样本质量",main = "分层抽样")
```

从图 1-2 可以看出，当样本容量为最小值 64 时，样本质量已达 99%。说明分层抽样小样本即可取得良好的抽样精度。

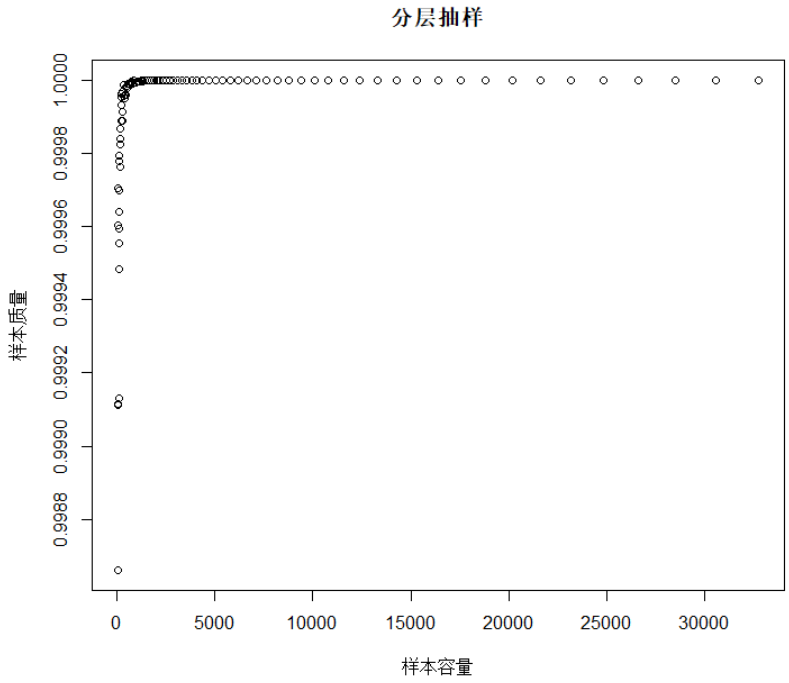

图 1- 2 分层抽样的样本容量与样本质量的散点图

三、当要求样本质量为 99%时，两种抽样方式的最优样本容量

从输出结果可以看出，当要求样本质量为 99%时，简单随机抽样的最优样本容量为 827，而分层抽样只需要 45 个样本，说明在样本质量一定时，分层抽样所需的样本远远少于简单随机抽样，在样本容量一定时，分层抽样的精度远远好于简单随机抽样。

```
> s1 = smooth.spline(q1,samp) #对曲线进行拟合
> pr1 = predict(s1,0.99);pr1#预测当样本质量为99%时的随机抽样样本容量
$x
[1] 0.99

$y
[1] 826.551
```

```
> s2 = smooth.spline(samp,q2) #对曲线进行拟合
> pr2 = predict(s2,45);pr2 #预测当样本质量为99%时的分层抽样样本容量45
$x
[1] 45

$y
[1] 0.9900626
```