实验二 统计基础练习 2

年级： 15级　　专业：生信　　学号：1513401013　　姓名：郑磊

编号　　一　　二　　三　　四　　总分　　评阅人

得分

一、软硬件平台：

1. 硬件平台：（硬件配置）i5，2.9HZ处理器，16G内存，64位操作系统

2. 系统平台：（操作系统及其版本号）Windows10 企业版

3. 软件平台：（软件系统及其版本号，若是在线分析平台，还需要提供URL地址）R3.4.1 ，Rstudio

二、实验内容：

1、准备工作:

进入R语言环境后，进行以下准备工作：（1)安装GEOquery包；（2)从Genbank的GEO Datasets数据库中下载制定ID的表达谱数据。当然，你也可以使用其他方式，如Web网页的在线下载模式先下载一个数据集；也可以使用授课教师提供的数据集。

2、二项分布模拟：

2.1、相关函数：

dbinom(x, size, prob)#该函数给出了每个点的概率密度分布。

pbinom(x, size, prob)#该函数给出事件的累积概率，它用于表示概率的单个值。

qbinom(p, size, prob)#该函数采用概率值，并给出其累积值与概率值匹配的数字。

rbinom(n, size, prob)#该函数从给定样本生成所需数量的给定概率的随机值。

参数的描述：

x - 是数字的向量，p - 是概率向量，n - 是观察次数，size - 是试验的次数，prob-是每次试验成功的概率。

2.2、药物有效性评估：

临床数据表明某种药物治疗某种非传染性疾病的有效率为0.88，无效率为0.12。（1）今用该药治疗该疾病患者100人，试分别计算这100人中有80人、90人、100人有效的概率。（**2**）如果治疗**100**该疾病患者，试评估不同人数的治疗有效率。

3、中心极限定理验证：

该环节需要大家提前准备好一个基因表达谱数据，如果没有，则有授课教师提供。以下示例以教师提供的一个来自于Genbank的GEO Datasets数据的GDS-format数据进行分析的。

3.1、加载数据

```
#加载本地的数据
gds4794 <- getGEO(filename='GDS4794.soft.gz')
#查看数据类型
mode(gds4794)
#查看注释信息
Meta(gds4794)$channel_count
Meta(gds4794)$feature_count
```

Meta(gds4794)$platform

Meta(gds4794)$sample_count

Meta(gds4794)$sample_organism

Meta(gds4794)$sample_type

Meta(gds4794)$title

Meta(gds4794)$type

#查看数据表的列名

colnames(Table(gds4794))

#查看部分数据标内容，前10行，前6列; #从第三列开始是数据列

Table(gds4794)[1:10,1:6]

3.2、提取数据表

#从S4数据类中提取所需数据表

data<-Table(gds4794)

#第一列设定为行标题【本次实验不需要】

#rownames(data)<-data[,1]

#查看数据表的行、列数【实验结果中需要记录】

ncol(data)

#[1] 67

nrow(data)

#[1] 54675

#去除标题列的干扰【前两列】

data2<-data[,3:67]

#随机抽取至少5列数据

n=5

#得到列名称【标题行】

col.name=colnames(data2)

#按列随机抽样

sam.col.name = sample(col.name,n,replace=F)

#查看抽样结果【实验结果中需要记录】

sam.col.name

#按行随机抽样【本次实验不需要】

#row.name=rownames(data2)

#sam.row.name = sample(row.name,n,replace=F)

#提取子数据集

sub.data <- data2[, sam.col.name]

3.3、绘制概率密度分布图，查看基因表达谱的数据分布规律

#计算数据子集的最大、最小值，用作限制横坐标范围

x1 <- min(sub.data, na.rm=TRUE)

x2 <- max(sub.data, na.rm=TRUE)

#定义纵坐标最大值，根据绘图结果自行调整，最小值固定为0

y_max = 7e-4

#绘制概率分布图，不同曲线使用不同颜色

dnorm_png<-png("dnorm.png")

for (i in 1:ncol(sub.data))

```
{ curve(dnorm(x,mean(sub.data[,i], na.rm=TRUE), sd(sub.data[,i],
na.rm=TRUE)),
add=TRUE , xlim=c(x1,x2), ylim=c(0,y_max), col=rev(rainbow(i)), lwd=3)
}
#保存图片
dev.off()
```

3.4、数据频率分布直方图的绘制，直接看看数据自身大小分布规律

```
a<- sub.data[,1]
#频率频率直方图，分100个bins
#count图
png(file = "gds4794-hist1.png")
hist(a, freq = T, breaks = 100)
dev.off()
#Frequency图
png(file = "gds4794-hist2.png")
hist(a, freq = F, breaks = 100)
dev.off()
```

3.5、抽样评估验证中心极限定理

对下面三种图进行对比分析和讨论。

```
#随机抽样1次
png(file = "gds4794-hist-sample1.png")
hist(a[sample(a, 100)], freq = F, breaks = 100)
```

```
dev.off()
#重复抽样100次
png(file = "gds4794-hist-sample100.png")
x <- replicate(100, sample(a, size=100, replace = FALSE))
hist(x, freq = F, breaks = 100)
dev.off()
#重复抽样100次，绘制均值分布图
png(file = "gds4794-hist-sample100-mean.png")
x<-replicate(100, mean(a[sample(a, 100)]))
hist(x, freq = F, breaks = 100)
dev.off()
```

4、数据转换的重要性

通过第3.3步的绘图结果，我们可以了解到数据的概率密度分布规律【本例中是严重的偏态分布】；再第根据3.4步的绘图结果，我们可以发现数据自身隐含的变化规律【本例中是指数型】；故而，可以采用对数（log）转换后，再次进行分析。

```
#对数转换（log）
b<-log(a)
x1<-min(b,na.rm=TRUE)
x2<-max(b,na.rm=TRUE)
#Frequency图
hist(b, freq = F, breaks = 100)
```

#概率密度分布图

curve(dnorm(x,mean(b,na.rm=TRUE),sd(b,na.rm=TRUE)),xlim=c(x1,x2),co

l="red",lwd=3, add=TRUE)

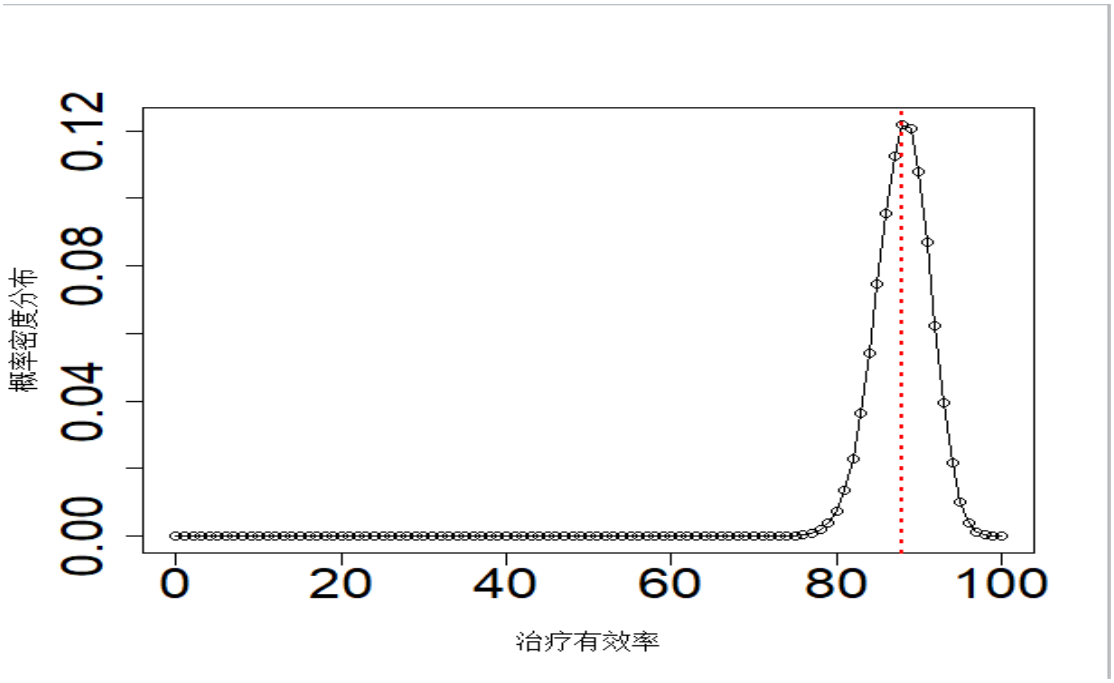
三、实验结果：

2.2. 药物有效性评估：

(1)

```
Console D:/RFile/实验二/
> setwd("D:/RFile/实验二")
> pbinom(80, 100, 0.88)
[1] 0.0147063
> pbinom(90, 100, 0.88)
[1] 0.7743481
> pbinom(100, 100, 0.88)
[1] 1
```

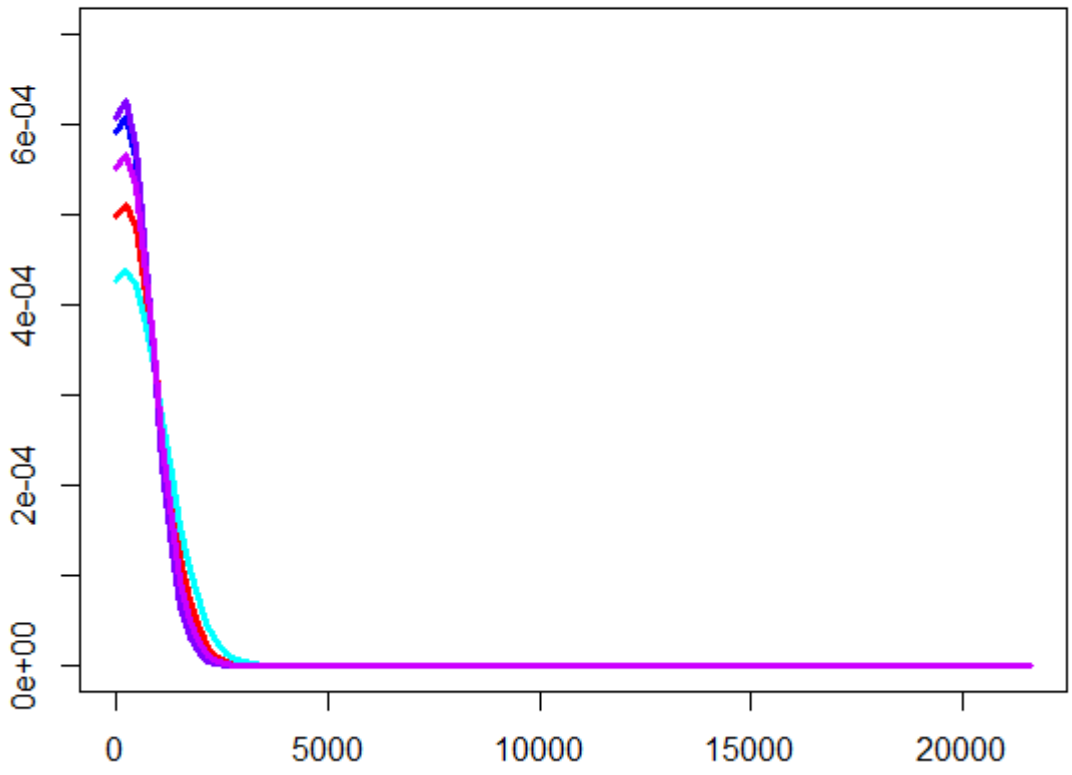
(2)



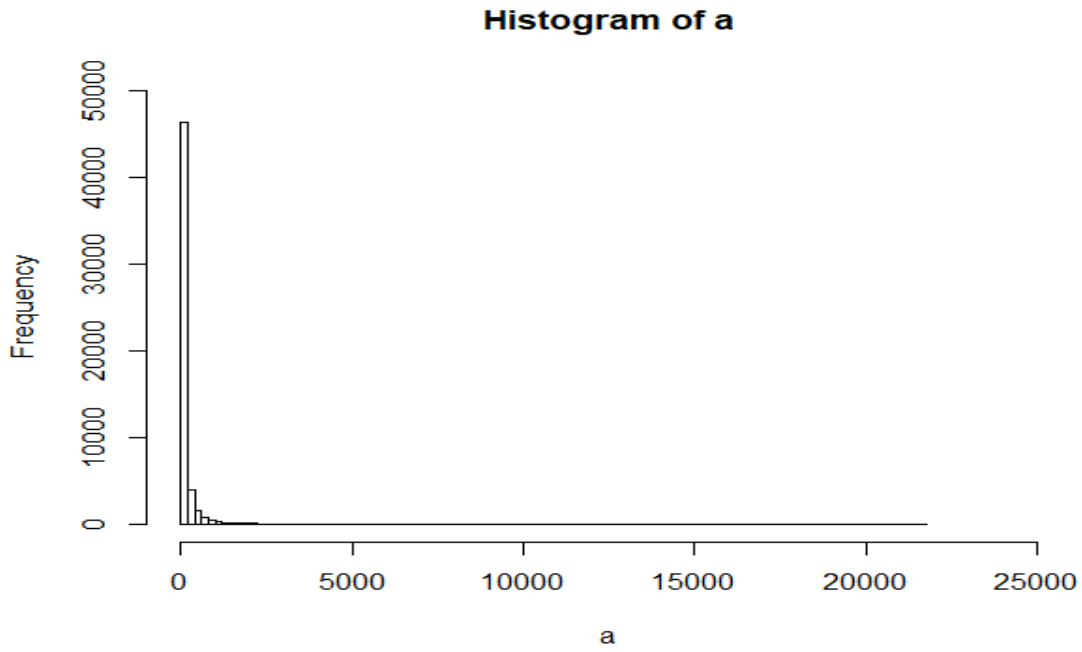
3.2

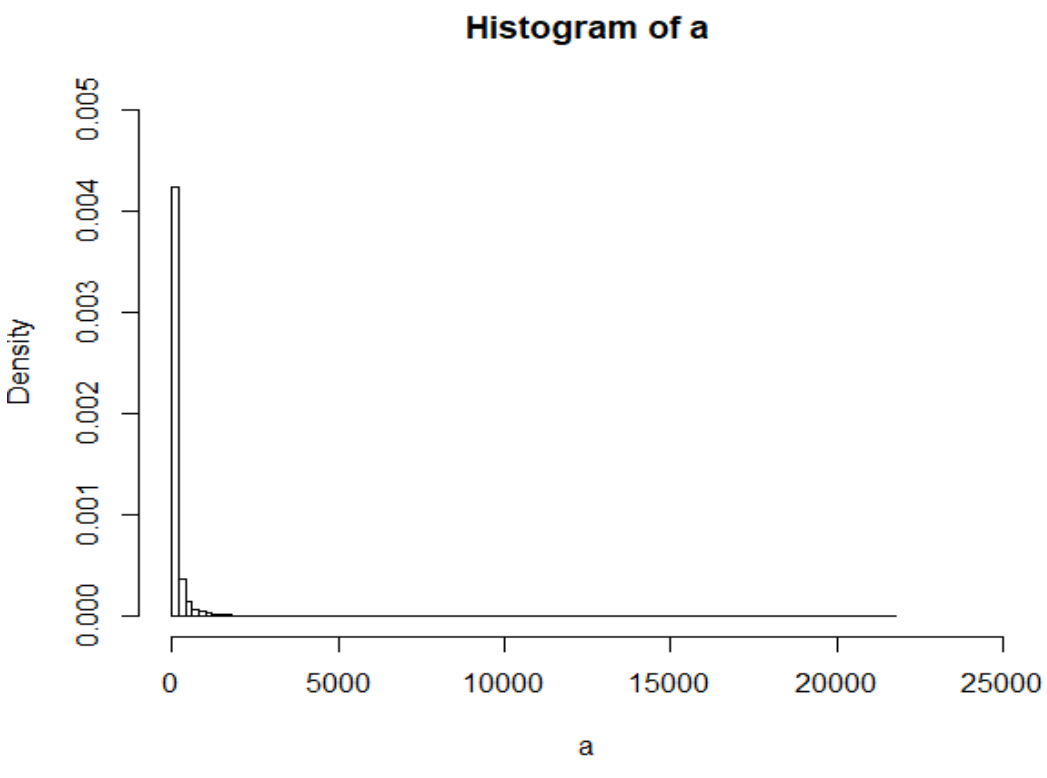
```
Console D:/RFile/实验二/
> library(GEOquery)
> gds4794 <- getGEO(filename='GDS4794.soft.gz')
> data<-Table(gds4794)
> ncol(data)
[1] 67
> nrow(data)
[1] 54675
> data2<-data[,3:67]
> n=5
> col.name=colnames(data2)
> sam.col.name = sample(col.name,n,replace=F)
> sam.col.name
[1] "GSM1060756" "GSM1060758" "GSM1060783"
[4] "GSM1060790" "GSM1060752"
> sub.data <- data2[, sam.col.name]
>
```

3.3 概率密度分布图



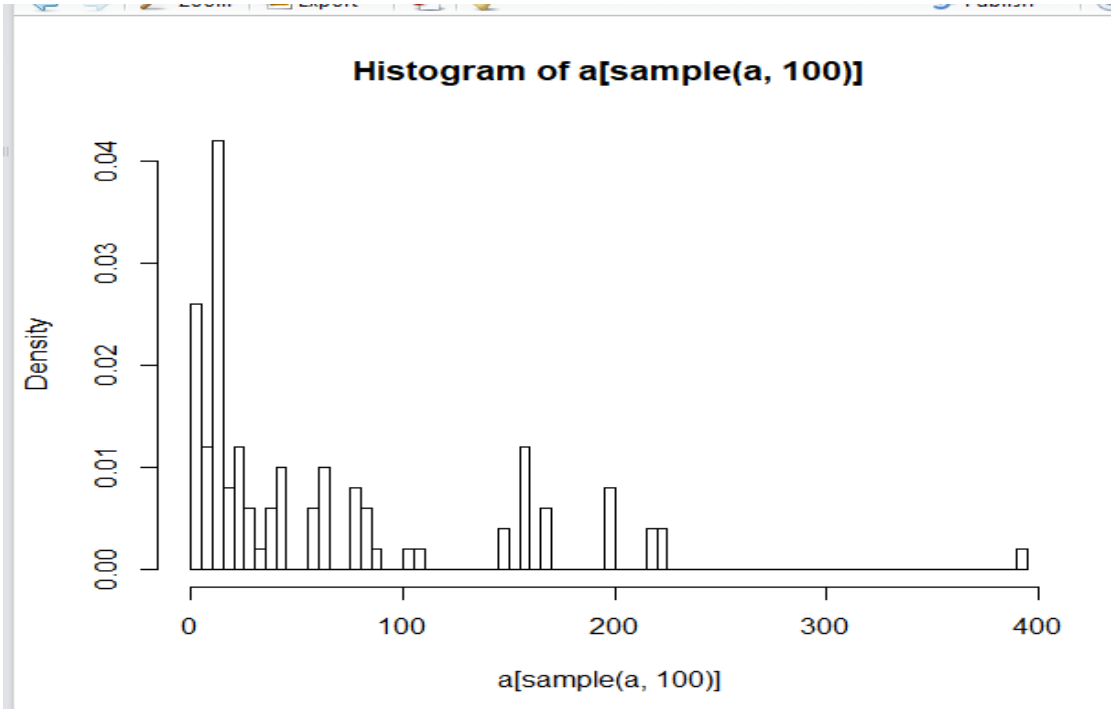
3.4 频率分布直方图



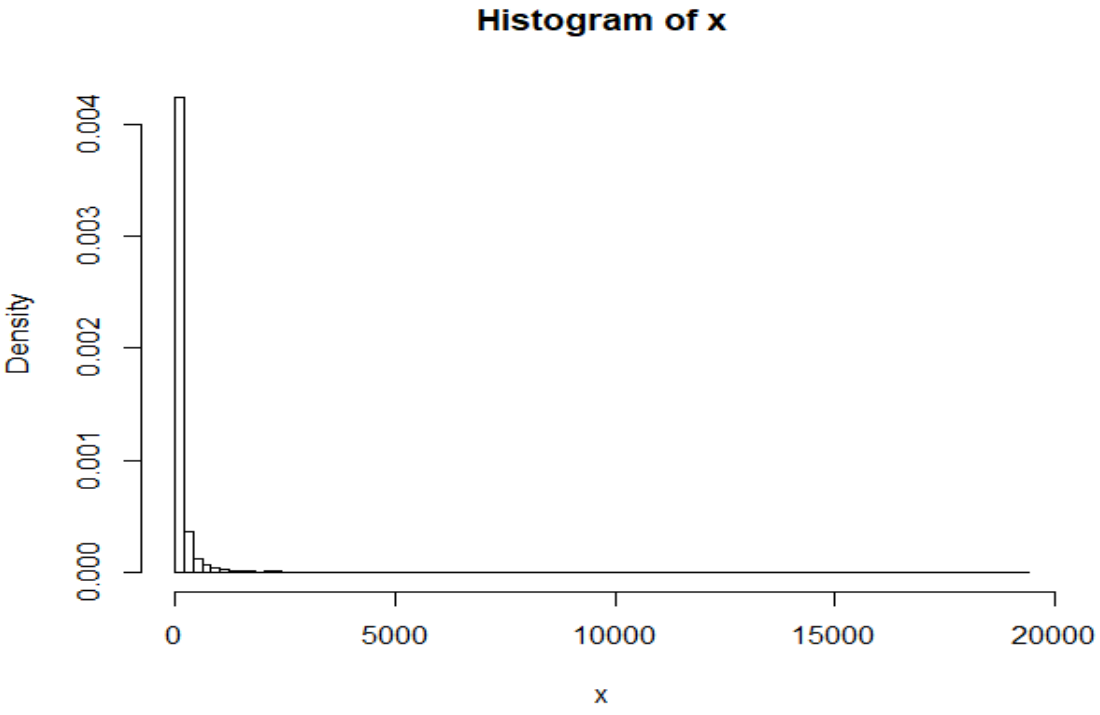


3.5 抽样评估

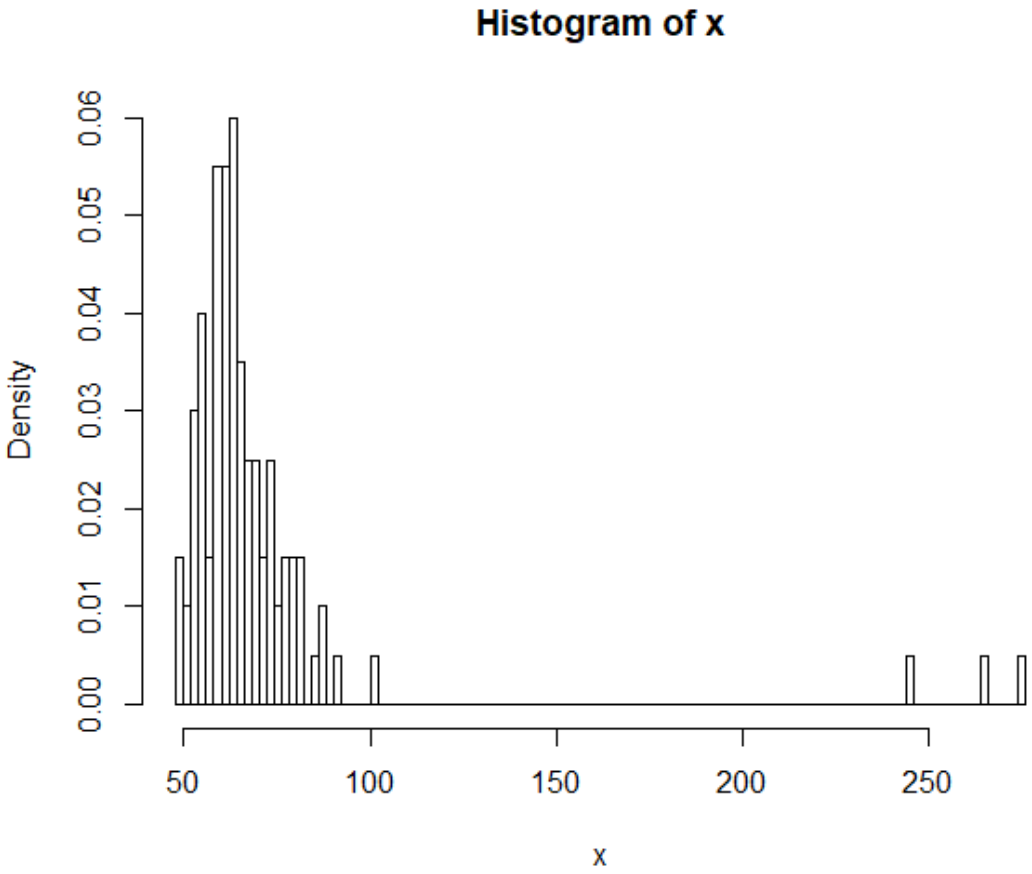
1.


Histogram of a[sample(a, 100)]
Density
0.00
0.01
0.02
0.03
0.04
0
100
200
300
400
a[sample(a, 100)]

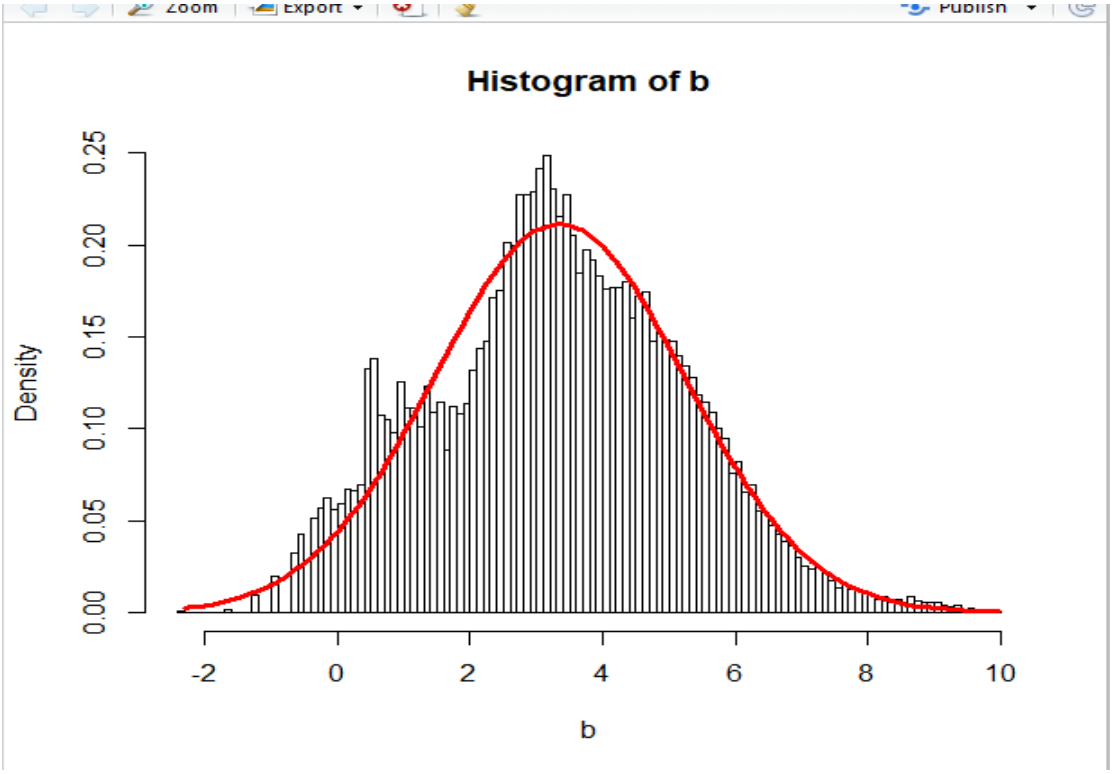
2.


Histogram of x
Density
0.000
0.001
0.002
0.003
0.004
0
5000
10000
15000
20000
x

3.


Histogram of x
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
50
100
150
200
250
X

4. 数据转换


Zoom
Export
Publish
Histogram of b
Density
0.00
0.05
0.10
0.15
0.20
0.25
-2
0
2
4
6
8
10
b

四、讨论：

3.4 绘制数据频率分布直方图时，freq参数控制纵坐标时频数还是频率，一般设置是：freq=F；

由3.5的三幅图可以看出抽样越大，越接近正态分布。即：无论总体分布如何，只要当抽取的样本容量足够大，那么样本均值的抽样分布就近似于正态分布。

R语言很容易就可以验证一些较难理解的统计学问题，比如中心极限定理，以后要勤加练习R语言。