实验三 统计绘图

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

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

一 目的要求：

1. 加深对描述统计的统计图意义所在的理解；

2. 熟悉并掌握R语言中各种统计图绘制函数的初步运用；

3. 理解各种统计图的含义。

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、集中趋势测度：

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。

2.1、加载数据

```
library(GEOquery)
gds4794 <- getGEO(filename='GDS4794.soft.gz')
```

2.2、提取数据表

```
Console D:/RFile/实验三/
> gds4794 <- getGEO(filename='GDS4794.soft.gz')
> data<-Table(gds4794)
> ncol(data)
[1] 67
> nrow(data)
[1] 54675
> data2<-data[,3:67]
> n=1
> col.name=colnames(data2)
> #按列随机抽样
> sam.col.name = sample(col.name,n,replace=F)
> sam.col.name
[1] "GSM1060787"
> #提取子数据集
> sub.data <- data2[, sam.col.name]
> a <-sub.data[sub.data<500] # 只取表达水平低于 500 的
数据
```

2.3、计算各种统计指标并绘制统计图：

代码如下：

```
> x1<-min(a,na.rm=TRUE) # 计算最小值
> x2<-max(a,na.rm=TRUE) # 计算最大值
> ave<-mean(a,na.rm=TRUE) # 计算均值
> med<-median(a,na.rm=TRUE) # 计算中位数
> # 连续分布的众数定义为其分布的密度函数峰值对应的取值
> ds=density(a,na.rm=TRUE)
> mode <- ds$x[which.max(ds$y)]
> quan<-quantile(a,na.rm=TRUE) # 计算四分位数 (0%,25%,
50%,75%,100%)
> dnorm_png<-png("d1-means-medium-mode.png") # 定义图
片文档
> hist(a, freq = F, breaks = 100) # 绘制频率分布直方图
> curve(dnorm(x,mean(a,na.rm=TRUE), sd(a,na.rm=TRUE)),
 xlim=c(x1,x2),
+       col="blue", lwd=3, add=TRUE) # 绘制概率分布曲
线
> abline(v=ave,lty=3,lwd=3,col="red") # 增加均值线
> abline(v=med,lty=3,lwd=3,col="purple") # 增加中位数
线
> abline(v=mode,lty=3,lwd=3,col="green") # 增加众数线
> abline(v=quan,lty=3,lwd=3,col="blue") # 增加四分位数
线
> dev.off() # 保存图片文档
null device
          1
```

2.4、log转换后计算各种统计指标并绘制统计图：

```
Console D:/RFile/实验三/
> b <- log(a)
> x1<-min(b, na.rm=TRUE) # 计算最小值
> x2<-max(b, na.rm=TRUE) # 计算最大值
> ave<-mean(b, na.rm=TRUE) # 计算均值
> med<-median(b, na.rm=TRUE) # 计算中位数
> # 连续分布的众数定义为其分布的密度函数峰值对应的取值
> ds=density(b, na.rm=TRUE)
> mode <- ds$x[which.max(ds$y)]
> quan<-quantile(b, na.rm=TRUE) # 计算四分位数 (0%, 25%,
50%, 75%, 100%)
> dnorm_png<-png("d2-means-medium-mode.png") # 定义图
片文档
> hist(b, freq = F, breaks = 100) # 绘制频率分布直方图
> curve(dnorm(x, mean(b, na.rm=TRUE), sd(b, na.rm=TRUE)),
 xlim=c(x1, x2),
+       col="blue", lwd=3, add=TRUE) # 绘制概率分布曲
线
> abline(v=ave, lty=3, lwd=3, col="red") # 增加均值线
> abline(v=med, lty=3, lwd=3, col="purple") # 增加中位数
线
> abline(v=mode, lty=3, lwd=3, col="green") # 增加众数线
> abline(v=quan, lty=3, lwd=3, col="blue") # 增加四分位数
线
> dev.off() # 保存图片文档
null device
          1
```

2.5 对比分析：

对2.3和2.4步的统计图进行对比分析，对两者之间的异同之处加以讨论。

3、离散测度

对上述 gds4794的整个数据表，在进行对数（log）转换前后分别绘制一个箱形图，然后对两者之间的异同之处加以分析讨论。

```
> png(file = "boxplot-all.png")
> boxplot(data[,3:67])
> dev.off()
null device
          1
> png(file = "boxplot-all-log2.png")
> boxplot(log(data[,3:67]))
> dev.off()
null device
          1
```

4、条形图的绘制和分析

从数据矩阵（data2）中随机选取5列数据，然后对其进行条形图的绘制和对比分析。

```
> a <- data2[, sam.col.name]
> freq = matrix(rep(0,50),10,5)
> for(i in 1:5){
+     x <-table(as.numeric(cut(a[,i],10)))
+     y <- as.data.frame(x)
+     freq[,i] <- y[,2]
+ }
> colnames(freq)<-colnames(a) # 列名
> png(file = "barplot.png")
> barplot(t(freq),beside=T,col=rainbow(5))
> dev.off()
null device
          1
> # 堆积
> png(file = "barplot2.png")
> barplot(freq,col=rainbow(10))
> dev.off()
null device
          1
>
```

```
> b<-log(a)
> # 把b 从大到小分成 10 个区间进行频数统计
> freq2 = matrix(rep(0,50),10,5) # 初始化频数矩阵
> for(i in 1:5){
+     x <-table(as.numeric(cut(b[,i],10)))
+     y <- as.data.frame(x)
+     freq2[,i] <- y[,2]
+ }
> colnames(freq2)<-colnames(b) # 列名
> png(file = "barplot-log.png")
> barplot(t(freq2),beside=T,col=rainbow(5))
> dev.off()
null device
          1
> # 堆积
> png(file = "barplot2-log.png")
> barplot(freq2,col=rainbow(10))
> dev.off()
null device
          1
```

5、频率分布图：

```
> data<-Table(gds4794)
> data2 <- log(data[,3:67])
> x1<-min(data2,na.rm=TRUE)
> x2<-max(data2,na.rm=TRUE)
> y_max<-0.25
> dnorm_png<-png("all-hist.png")
> curve(dnorm(x,mean(data2[,1],na.rm=TRUE), sd(data2[,
1],na.rm=TRUE)), xlim=c(x1,x2),ylim=c(0,y_max), col=1,
 lwd=3)
> for (i in 2:ncol(data2))
+ {
+     curve(dnorm(x,mean(data2[,i], na.rm=TRUE), sd(da
ta2[,i], na.rm=TRUE)), add=TRUE ,xlim=c(x1,x2), ylim=c
(0,y_max), col=i, lwd=3)
+ }
> dev.off()
null device
          1
```

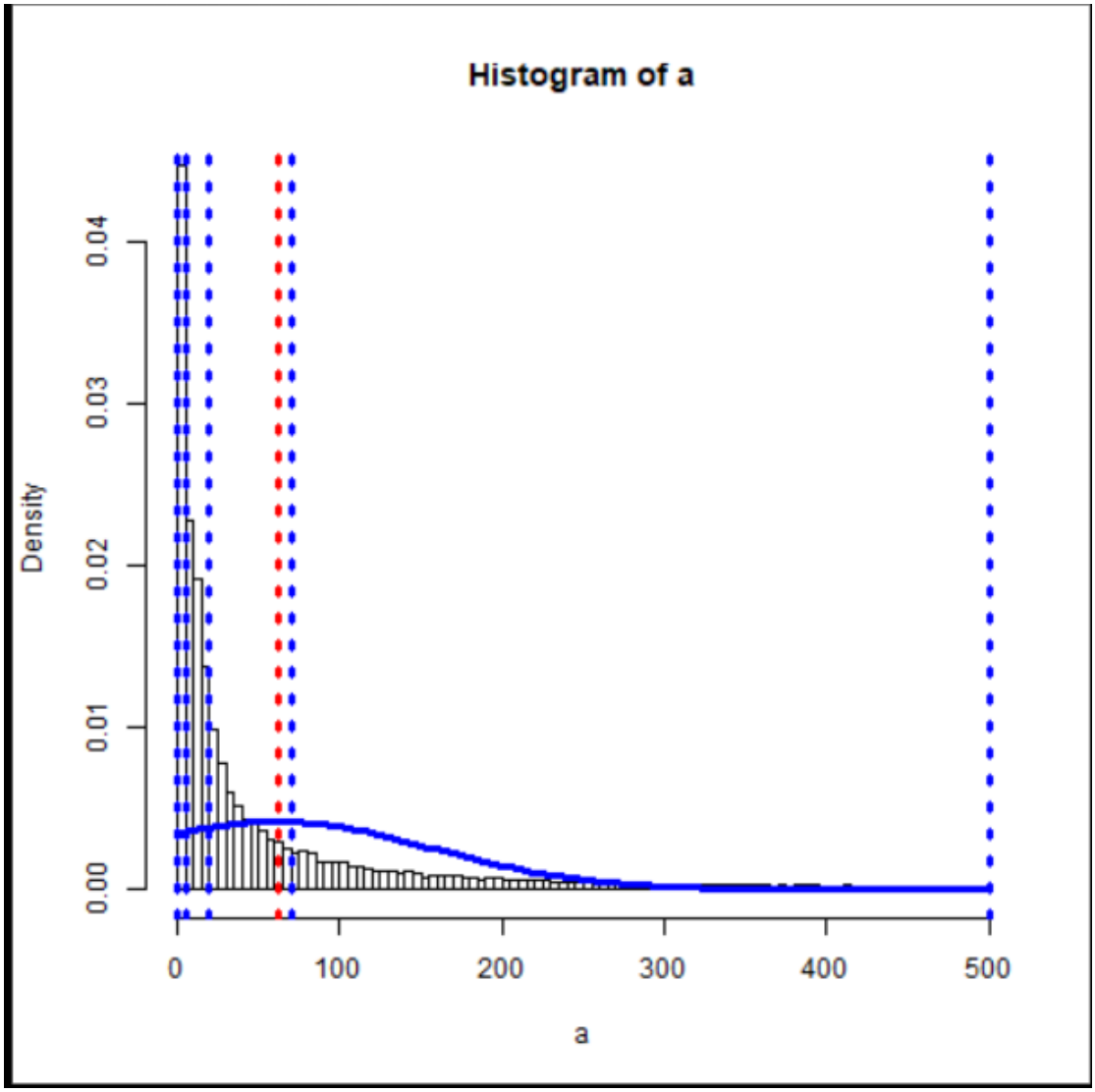
三、实验结果：

2.2：

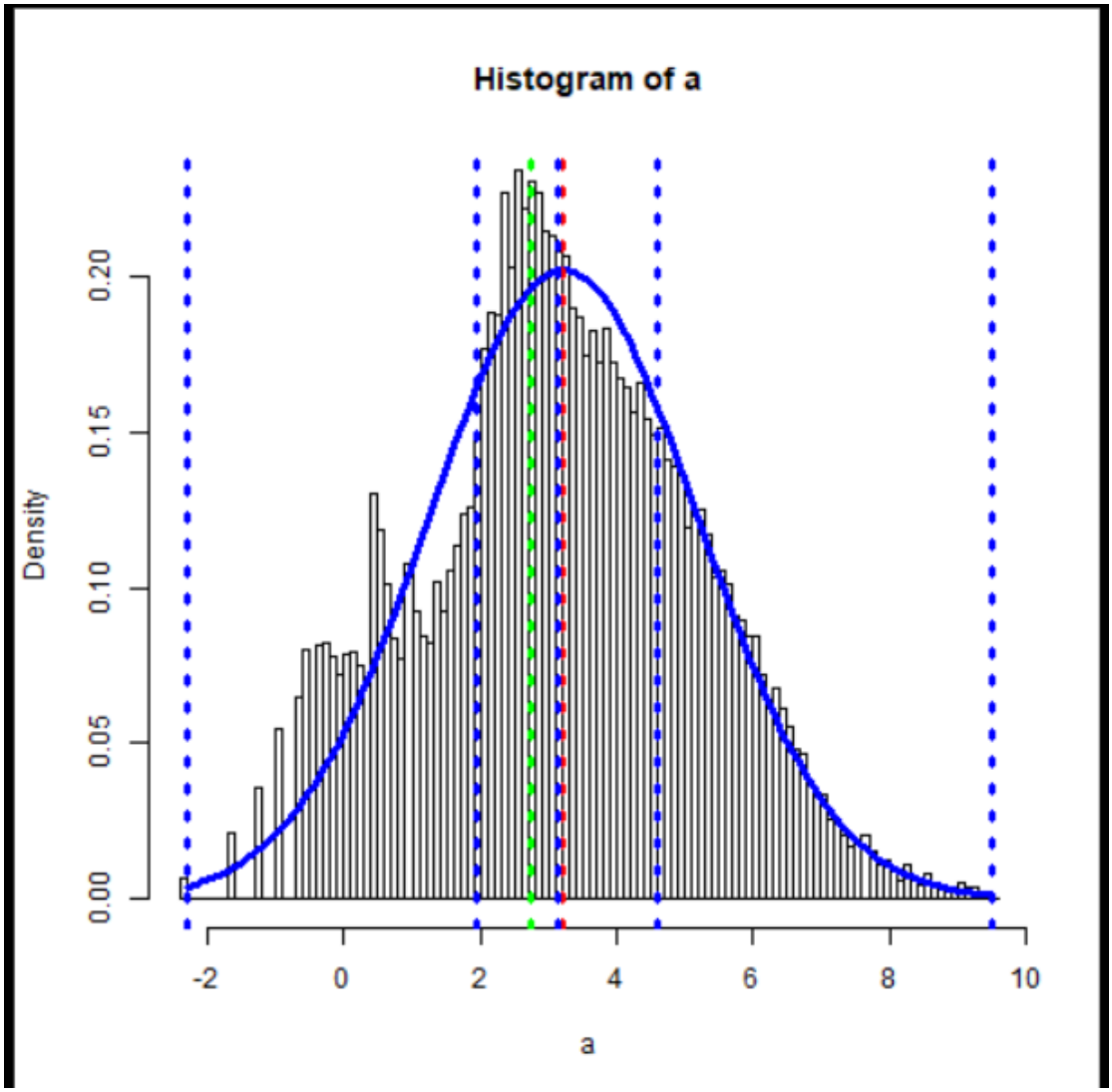
数据有67行，54675列

抽样结果为：GSM1060787

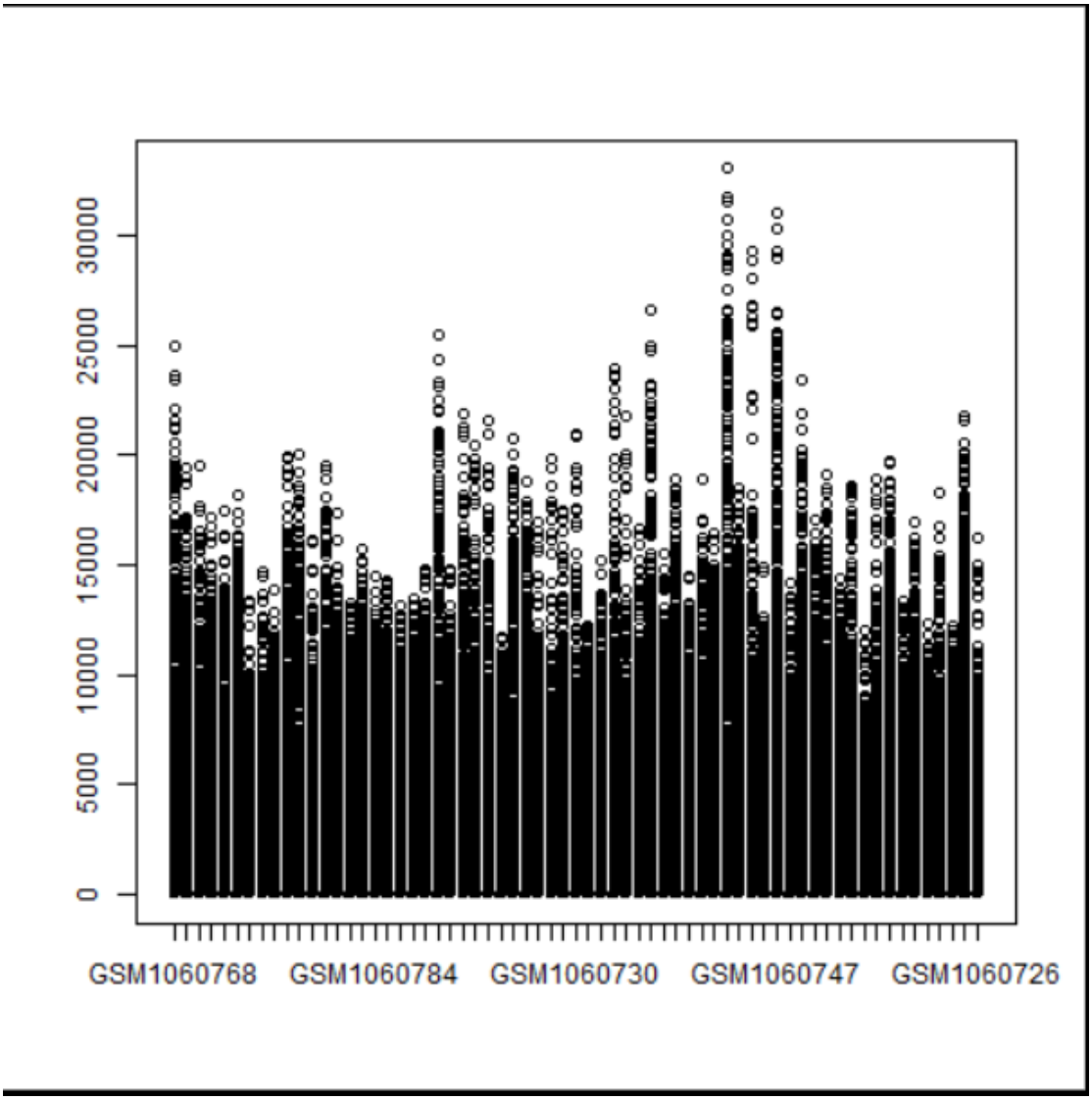
2.3


Histogram of a
Density
0.00
0.01
0.02
0.03
0.04
0
100
200
300
400
500
a

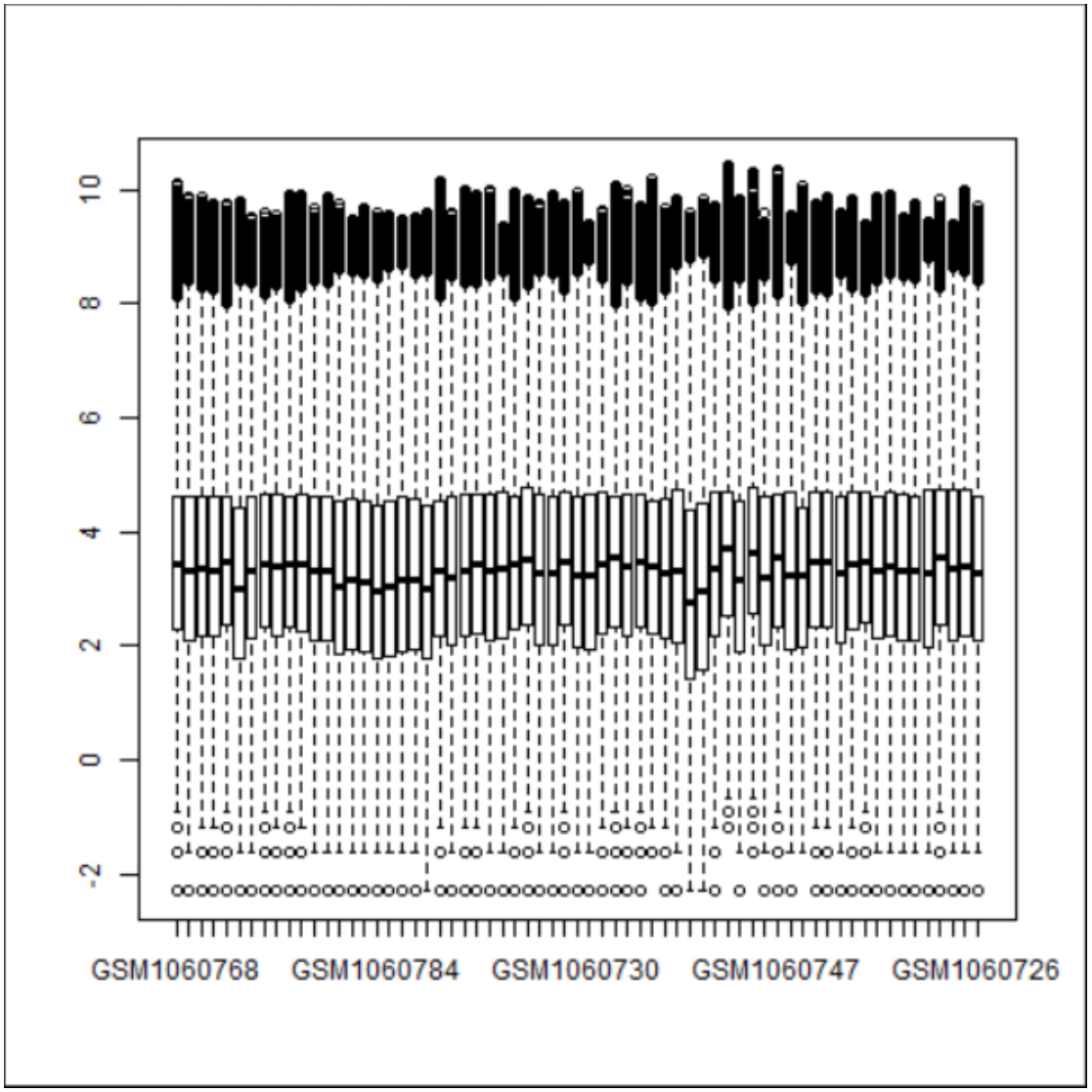
2.4 :



3：


0
5000
10000
15000
20000
25000
30000
GSM1060768
GSM1060784
GSM1060730
GSM1060747
GSM1060726

没有经过log2处理



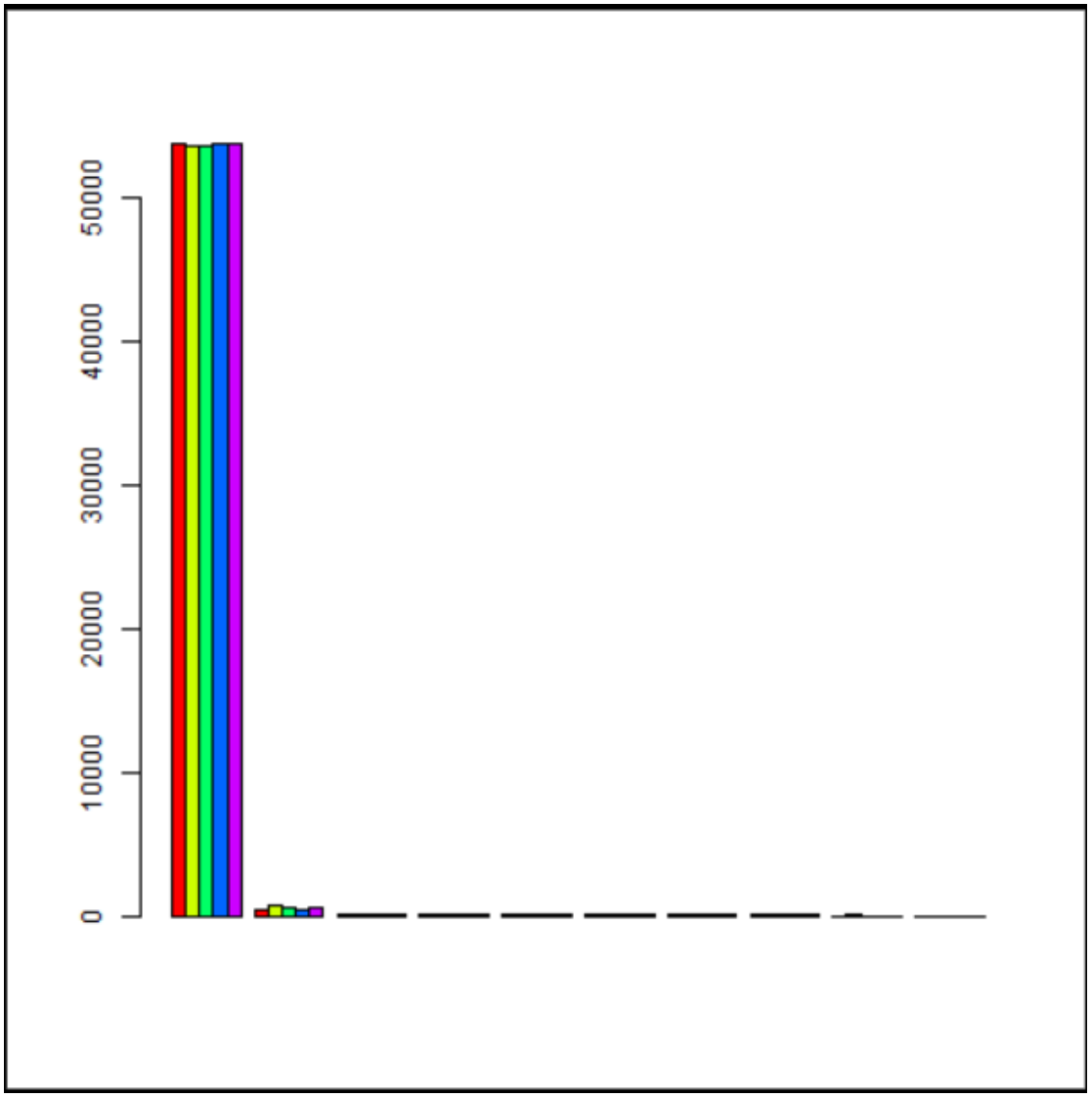
Log2处理过后

4.

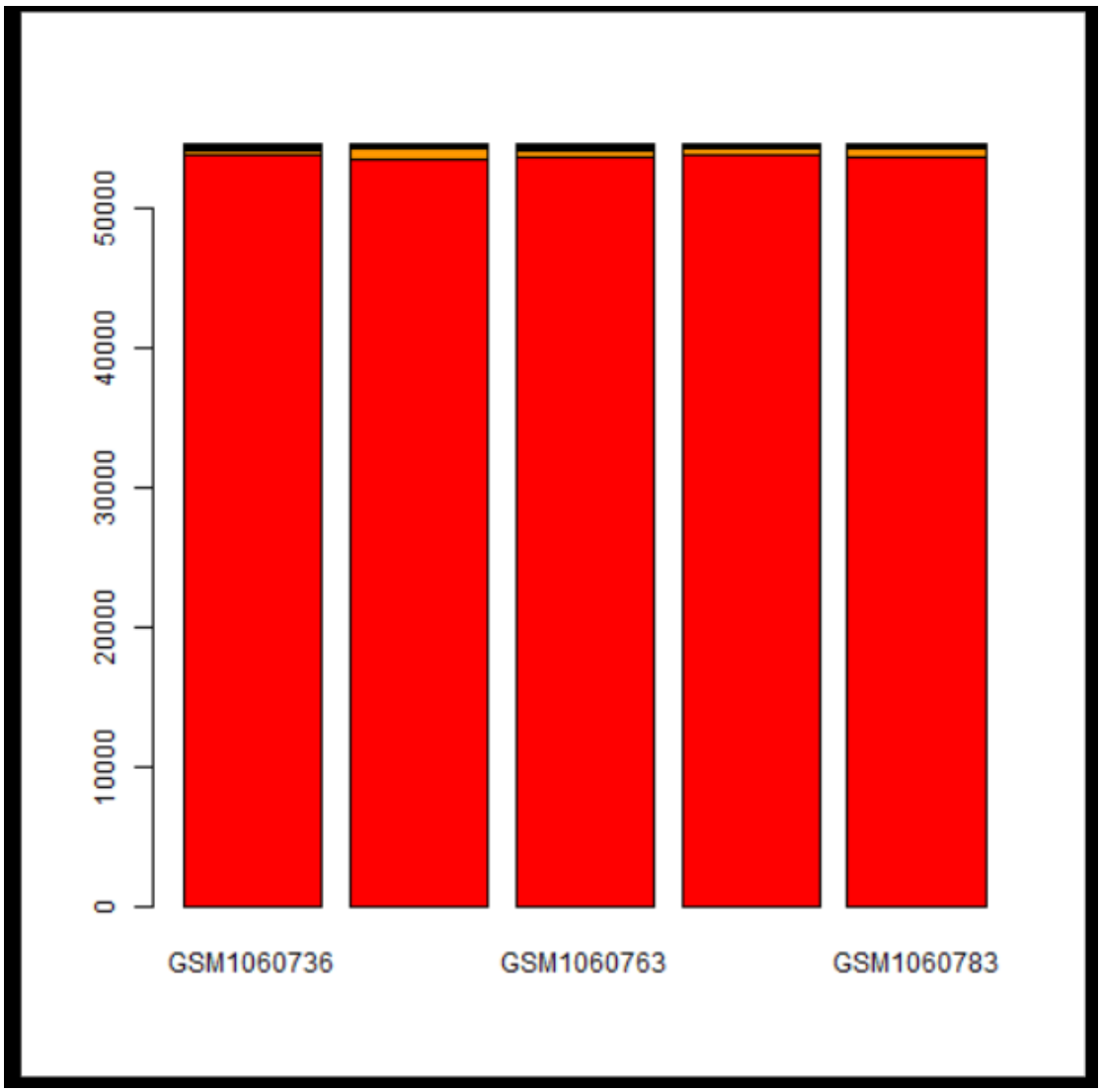
```
> n=5
> sam.col.name = sample(col.name, n, replace=F)
> sam.col.name
[1] "GSM1060736" "GSM1060748" "GSM1060763"
[4] "GSM1060733" "GSM1060783"
```

没有log处理：

频数图：

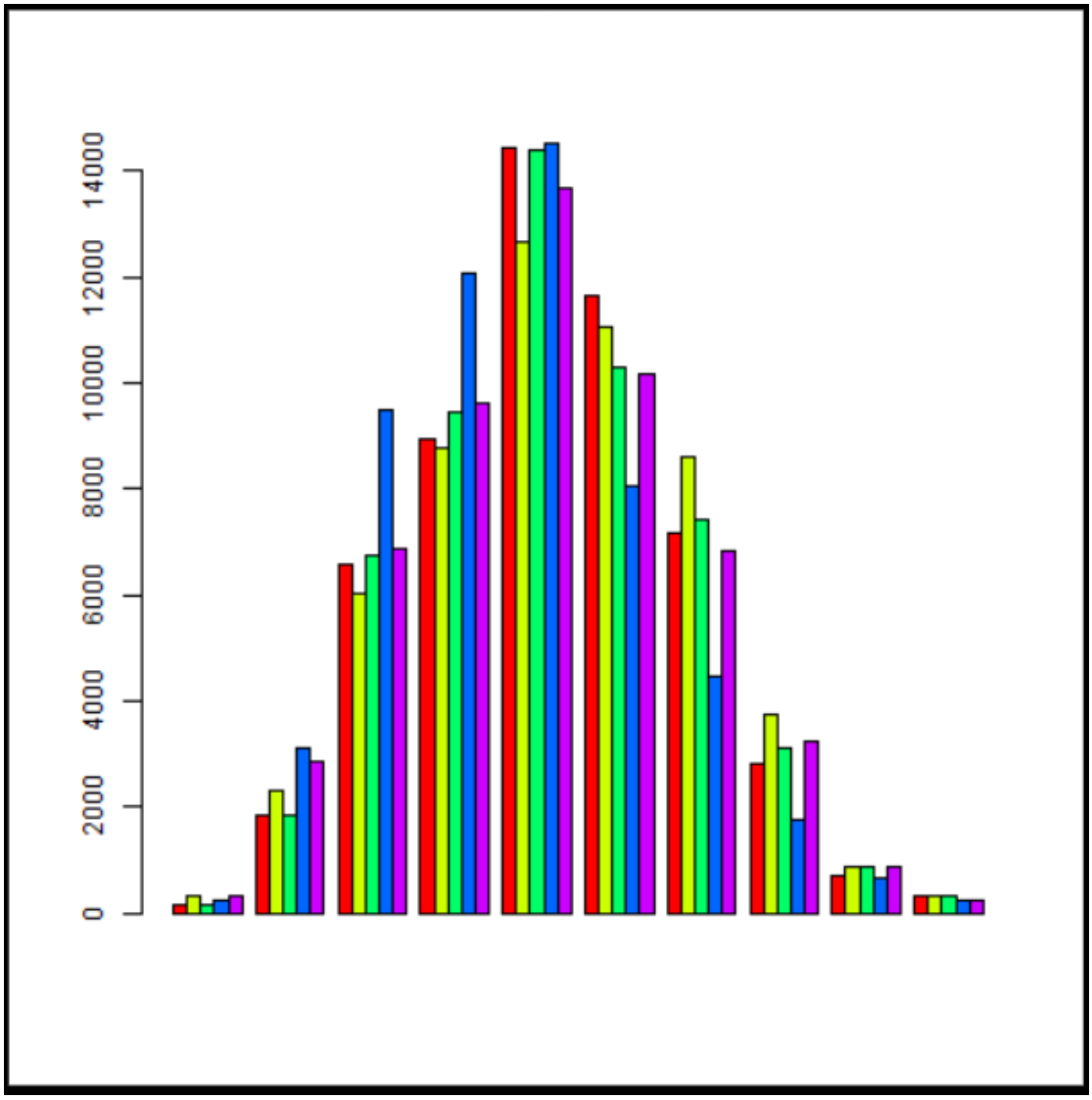


堆积图：

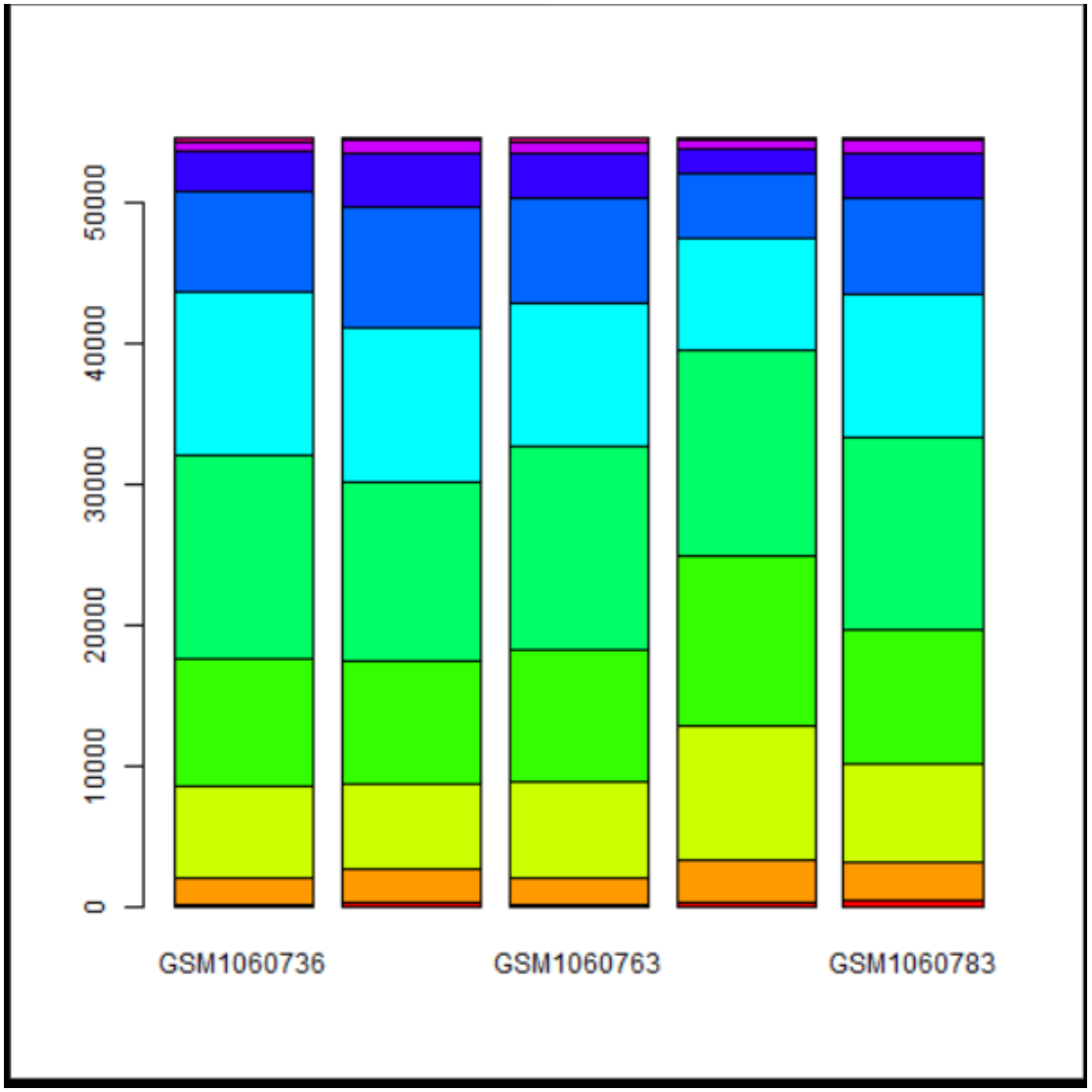

0
10000
20000
30000
40000
50000
GSM1060736
GSM1060763
GSM1060783

Log处理后：

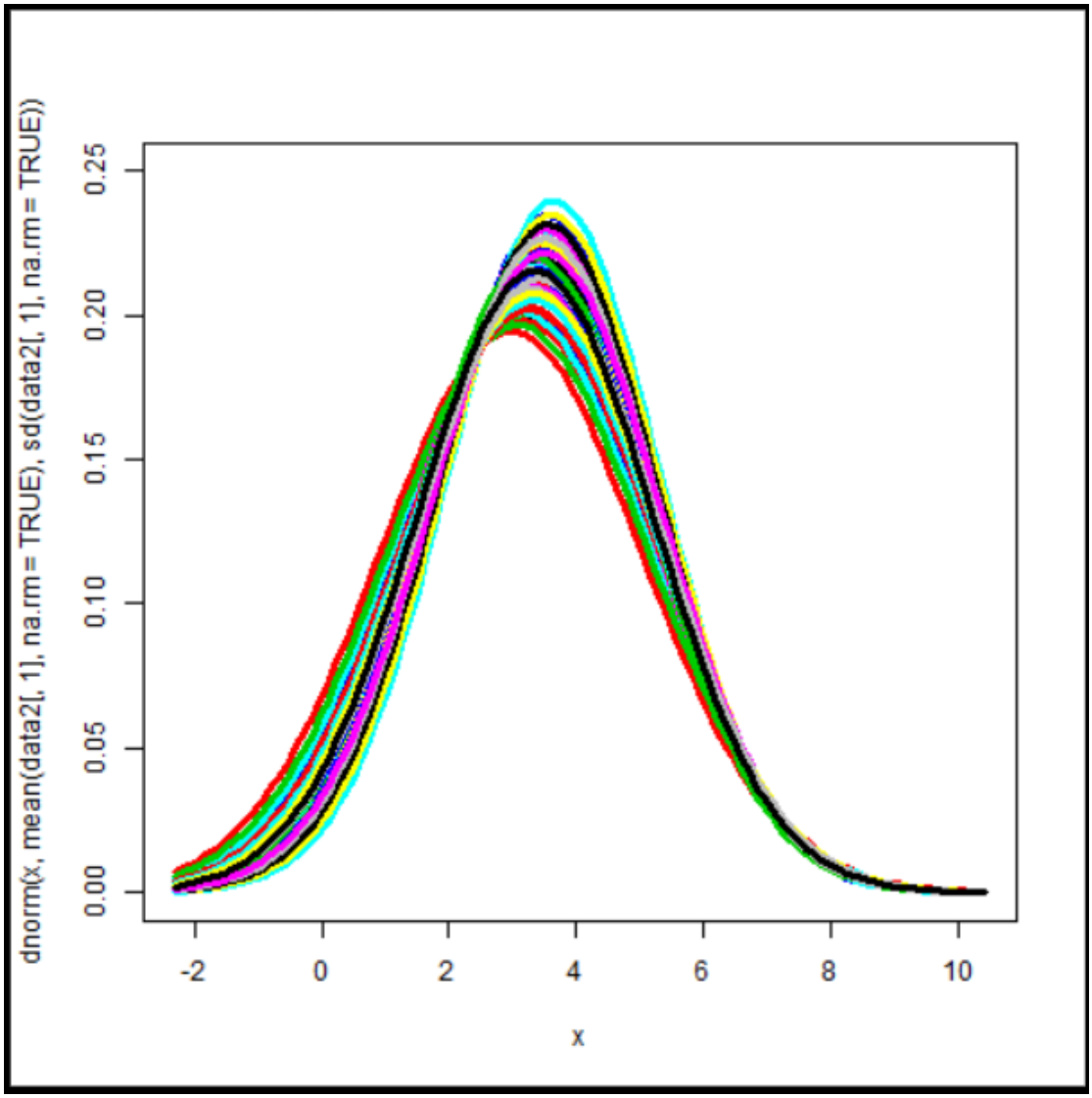
频数图：



堆积图：


0
10000
20000
30000
40000
50000
GSM1060736
GSM1060763
GSM1060783

5.



四、讨论：

2.5.

没有经过log处理的数据，中位数，众数，平均数较为分散，都不适合代表这组数据的特征。

Log2处理修正之后，其近似于正太分布，三条线较为集中，四分位数线分布也较为合理，可以较好的代表这组数据的特征。

3.

没有log2处理的数据画出的箱型图分布和合理，离群点较多，且不能直观的得到有用的信息。

经过log2处理的数据可以直观地看出各组数据的分布情况，但都不能做组与组之间的分析比较。

4.

频数图：

未处理前数据偏向偏态分布，且组与组之间比较不明显，处理后，近似于与正太分布组与组之间的差别比较明显。

堆积图：

处理前全部积压在一起，没有比较的意义。处理后可以直观的看出不同样本内不同的区间内探针的数量。

5.

所有样本数据经log2处理之后都可以近似的转化为正太分布。