# 实验五 t 检验

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t分布和检验的理解；

2、熟悉并掌握 t检验相关R语言函数和脚本。

## 二、实验内容：

1、等方差和异方差t检验的比较

1.1、随机生成两组、每组100个0~100之间的数值，然后分别进行等方差和异方差t检验；

代码：

```
#随机数生成
set.seed(1)
a<-seq(0,100,length.out=100)
set.seed(2)
b<-seq(0,100,length.out=100)
```

```
#t检验
t.test(a,b,var.equal=TRUE) #等方差t检验
t.test(a,b) #异方差t检验
#概率密度分布图
#a
png(file = "t_test.png")
curve(dnorm(x,mean(a,na.rm=TRUE),sd(a,na.rm=TRUE)),xlim=c(0,100),ylim=c(0,0.04),col="blue",lwd=3)
abline(v=mean(a,na.rm=TRUE),lty=3,lwd=3,col="blue") #  增加均值线
abline(v=mean(a,na.rm=TRUE)+sd(a,na.rm=TRUE),lty=3,lwd=3,col="blue") #增加标准差线
abline(v=mean(a,na.rm=TRUE)-sd(a,na.rm=TRUE),lty=3,lwd=3,col="blue") #增加标准差线
#b
curve(dnorm(x,mean(b,na.rm=TRUE),sd(b,na.rm=TRUE)),add=TRUE,xlim=c(0,100),ylim=c(0,0.04),col="red",lwd=3)
abline(v=mean(b,na.rm=TRUE),lty=3,lwd=3,col="red") #  增加均值线
abline(v=mean(b,na.rm=TRUE)+sd(b,na.rm=TRUE),lty=3,lwd=3,col="red") #增加标准差线
abline(v=mean(b,na.rm=TRUE)-sd(b,na.rm=TRUE),lty=3,lwd=3,col="red") #增加标准差线
dev.off()
```

1.2、随机生成两组0~100之间的数值（seq函数），第一组100个数值，第二组50个数值，然后分别进行等方差和异方差t检验；

将1.1中的b<-seq(0,100,length.out=100)改为b<-seq(0,100,length.out=50)即可

1.3、 随机生成两组、每组100个的数值，第一组60~90之间，第二组70~80之间，然后分别进行等方差和异方差t检验；

将1.1中的a<-seq(0,100,length.out=100)改为a<-seq(60,90,length.out=100)

将1.1中的b<-seq(0,100,length.out=100)改为b<-seq(70,80,length.out=100)即可

1.4、 随机生成两组、每组100个的数值，第一组50~80之间，第二组70~100之间，然后分别进行等方差和异方差t检验；

将1.1中的a<-seq(0,100,length.out=100)改为a<-seq(50,80,length.out=100)

将1.1中的b<-seq(0,100,length.out=100)改为b<-seq(70,100,length.out=100)即可

1.5、对上述统计分析结果进行分析讨论。

2、t检验在基因表达水平分析中的应用

2.1、加载数据

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

2.2、查看样本信息

```
#查看样本数量
Meta(gds4794)$sample_count
#查看列注释信息
Columns(gds4794)
```

2.3、提取数据表

```
data<-Table(gds4794)

head(data)

#查看数据表的行、列数

ncol(data)

nrow(data)

#肿瘤数据

tumor<-data[,3:25]

#正常组织数据

normal<-data[,26:67]

#data第1列设定为tumor和normal行标题

rownames(tumor)<-data[,1]

rownames(normal)<-data[,1]
```

2.4、随机抽样

```
#随机抽取1行数据

n=1

#按行随机抽样

row.name = rownames(tumor)

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

sam.row.name #查看抽中的数据行【某个基因在不同样本中的表达水平】

tumor_expression_level <- tumor[sam.row.name,]

normal_expression_level <- normal[sam.row.name,]
```

2.5、数据类型转换

```
a<-unlist(tumor_expression_level)

b<-unlist(normal_expression_level)
```

2.6、计算均值和方差

```
#计算均值和方差

a_average<-mean(a,na.rm=TRUE)

a_sd<-sd(a,na.rm=TRUE)

b_average<-mean(b,na.rm=TRUE)

b_sd<-sd(b,na.rm=TRUE)

#查看结果

a_average

a_sd

b_average

b_sd
```

2.7、等方差和异方差t检验的比较

```
t.test(a,b,var.equal=TRUE) #等方差t检验

t.test(a,b) #异方差t检验
```

2.8、统计绘图

```
#概率密度分布图【结果纪录】

x1=min(a,b)

x2=max(a,b)

y1=0

y2=0.0002
```

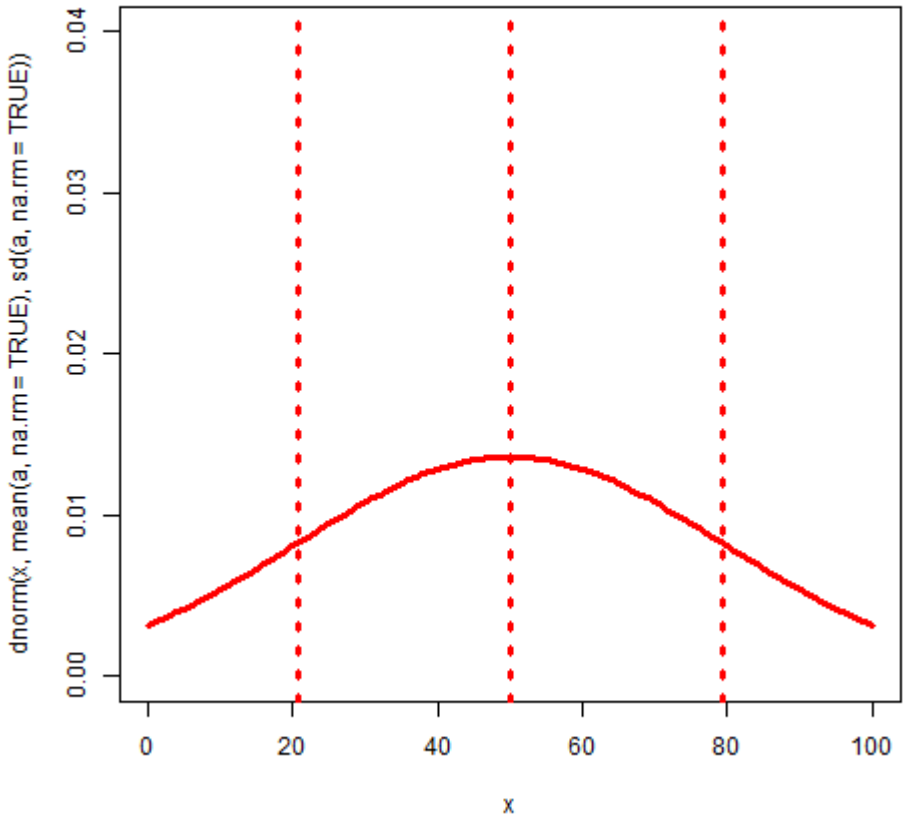
```
#a
png(file = "t_test_1DEG.png")
curve(dnorm(x,mean(a,na.rm=TRUE), sd(a,na.rm=TRUE)),
xlim=c(x1,x2),ylim=c(y1,y2), col="blue", lwd=3)
abline(v=a_average,lty=3, lwd=3, col="blue") #  增加均值线
abline(v=a_average+a_sd, lty=3, lwd=3, col="blue") #  增加标准差线
abline(v=a_average-a_sd, lty=3, lwd=3, col="blue") #  增加标准差线
#b
curve(dnorm(x,mean(b,na.rm=TRUE), sd(b,na.rm=TRUE)), add=TRUE,
xlim=c(0,100),ylim=c(0,0.0002), col="red", lwd=3)
abline(v=b_average, lty=3, lwd=3, col="red") #  增加均值线
abline(v=b_average+b_sd, lty=3, lwd=3, col="red") #  增加标准差线
abline(v=b_average-b_sd, lty=3, lwd=3, col="red") #  增加标准差线
dev.off()
```

## 三、实验结果：

1.

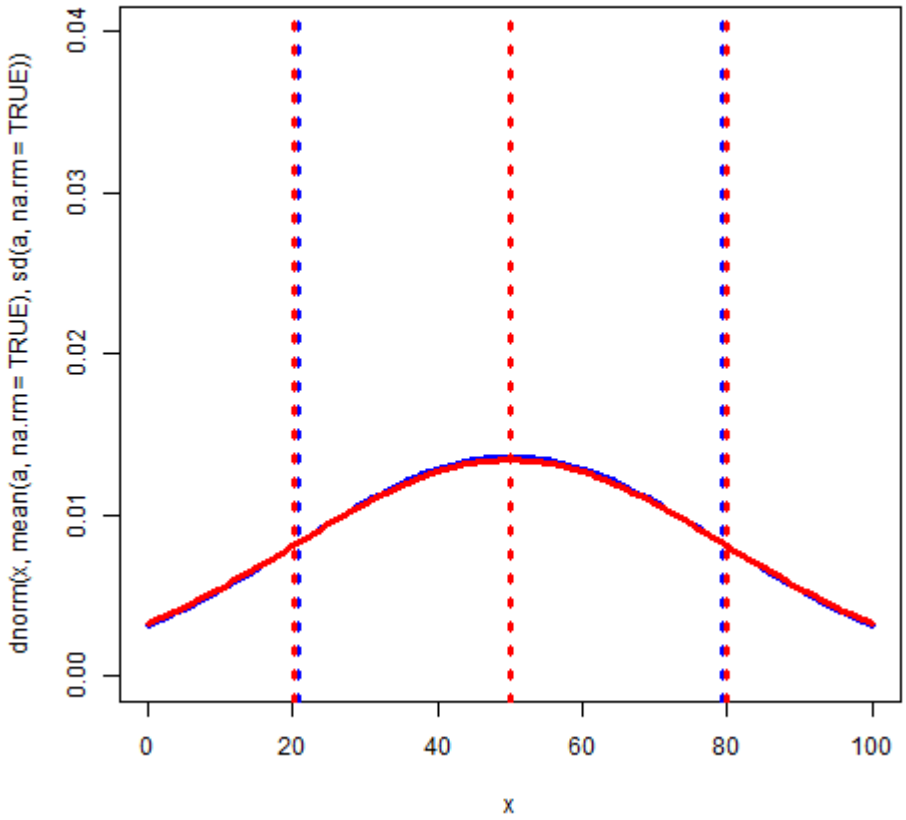
1.1

```
> t.test(a,b,var.equal=TRUE) #等方差t检验

        Two Sample t-test

data:  a and b
t = 0, df = 198, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.172607  8.172607
sample estimates:
mean of x mean of y
       50        50

> t.test(a,b) #异方差t检验

        Welch Two Sample t-test

data:  a and b
t = 0, df = 198, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.172607  8.172607
sample estimates:
mean of x mean of y
       50        50
```


dnorm(x, mean(a, na.rm = TRUE), sd(a, na.rm = TRUE))
0.00
0.01
0.02
0.03
0.04
0
20
40
60
80
100
x

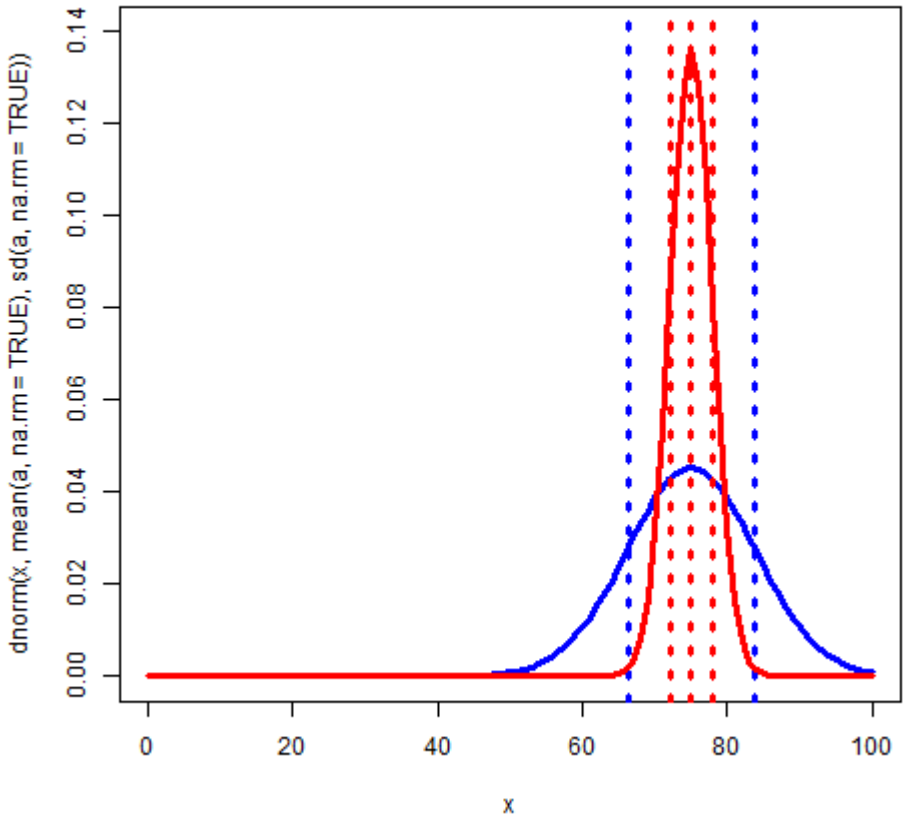
1.2

```
> t.test(a,b,var.equal=TRUE) #等方差t检验

        Two Sample t-test

data:  a and b
t = 1.3929e-15, df = 148, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.0809  10.0809
sample estimates:
mean of x mean of y
       50        50

> t.test(a,b) #异方差t检验

        Welch Two Sample t-test

data:  a and b
t = 1.3858e-15, df = 96.801, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.1764  10.1764
sample estimates:
mean of x mean of y
       50        50
```


dnorm(x, mean(a, na.rm = TRUE), sd(a, na.rm = TRUE))
0.00
0.01
0.02
0.03
0.04
0
20
40
60
80
100
x

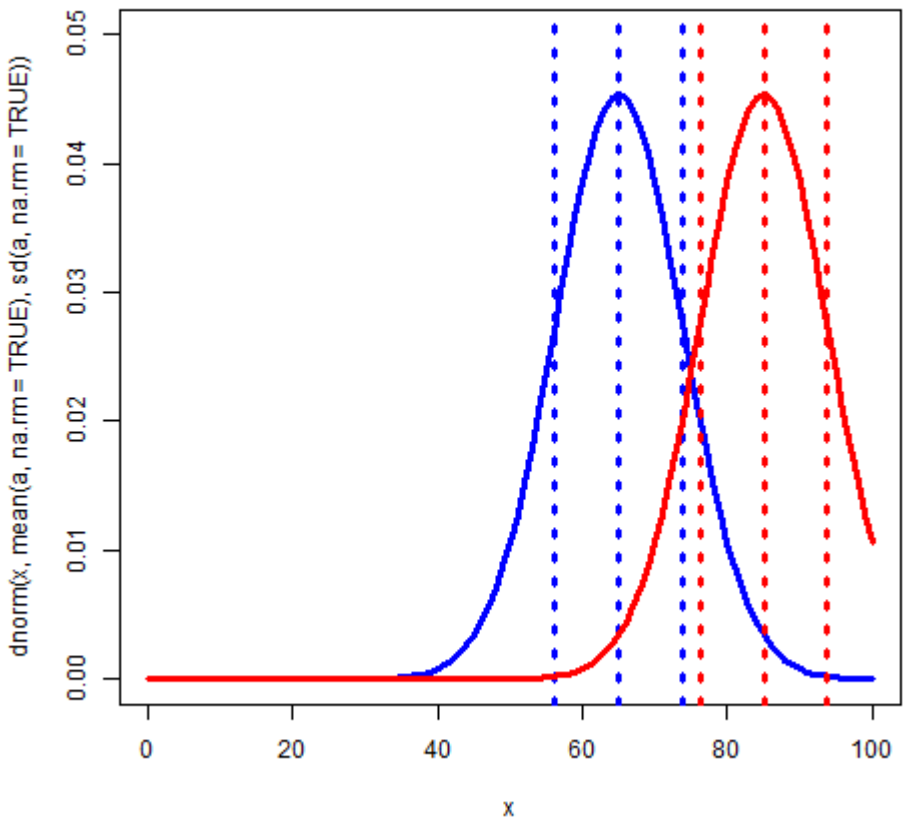
1.3

```
> t.test(a,b,var.equal=TRUE) #等方差t检验

        Two Sample t-test

data:  a and b
t = 0, df = 198, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.827451  1.827451
sample estimates:
mean of x mean of y
       75        75

> t.test(a,b) #异方差t检验

        Welch Two Sample t-test

data:  a and b
t = 0, df = 120.73, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.83467  1.83467
sample estimates:
mean of x mean of y
       75        75
```


dnorm(x, mean(a, na.rm = TRUE), sd(a, na.rm = TRUE))
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0
20
40
60
80
100
x

1.4

```
> t.test(a,b,var.equal=TRUE) #等方差t检验

        Two Sample t-test

data:  a and b
t = -16.086, df = 198, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.45178 -17.54822
sample estimates:
mean of x mean of y
       65        85

> t.test(a,b) #异方差t检验

        Welch Two Sample t-test

data:  a and b
t = -16.086, df = 198, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.45178 -17.54822
sample estimates:
mean of x mean of y
       65        85
```



2：

2.4："200660_at"

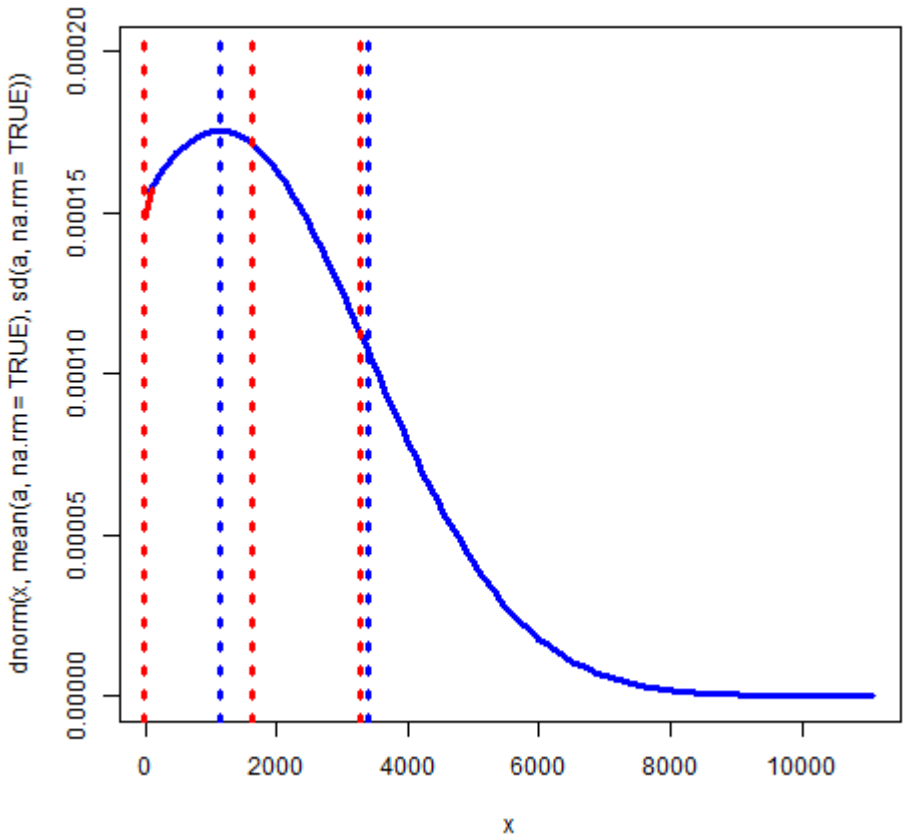
2.6

```
> a_average
[1] 1131.996
> a_sd
[1] 2273.411
> b_average
[1] 1629.21
> b_sd
[1] 1639.022
```

2.7

```
> t.test(a,b,var.equal=TRUE) #等方差t检验

        Two Sample t-test

data:  a and b
t = -1.0169, df = 63, p-value = 0.3131
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1474.3223   479.8946
sample estimates:
mean of x mean of y
 1131.996  1629.210

> t.test(a,b) #异方差t检验

        Welch Two Sample t-test

data:  a and b
t = -0.92542, df = 34.794, p-value = 0.3611
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1588.1904   593.7626
sample estimates:
mean of x mean of y
 1131.996  1629.210
```

2.8




四、讨论：

1：

t检验是用t分布理论来推论差异发生的概率，从而比较两个平均数的差异是否显著。两独立样本T检验又分为等方差T检验和异方差T检验。等方差T检验主要用于同一样本的两组数据或明确知道两组数据的方差一致的情况，异方差用于两组数据方差不一致的情况。而实验一由于都是随机生成一组数据，所以只有1.2中数据量不同而导致其结果有略微差别，其余结果均相同。其结果亦没有显著性差异。

2：

实验中抽出的基因探针在患者和正常人中间的表达水平相差及其微

小，可能是由于个体差别，也没有显著性差别，所以可以断定此基因与癌症无关。