南京邮电大学

实 验 报 告

（ 2022 / 2023 学年 第 一 学期）

课程名称	非参数统计
实验名称	非参数统计上机实验

学生姓名	王畅	班级学号	B20070412
学院(系)	理学院	专　　业	应用统计学

目录

1.研究不同种族之间政党支持问题

有代码如下：

```
c1<-c(341,405,105);
c2<-c(103,11,15);
data1<-rbind(c1,c2)
#不妨假设H0：种族和政党间是无关的
#列联表检验
chisq.test(data1)
#p-value <2e-16，远小于0.05，认为不能接受H0

```

不妨假设 H0：种族和政党间是无关的

通过列联表检验有：

p-value <2e-16，远小于 0.05，认为不能接受 H0

2.研究肺炎和疾病之间的继承关系

本实验通过 Fisher 精确性检验检验二者的关系，应有代码如下：

```
#Fisher 精确性检验
compare<-matrix(c(6,4,1,9),nrow=2, ncol=2 )

fisher.test(compare, alternative = "greater")
```

得到的结果有：

不妨假设 H0：二者之间没有关系

```
p-value = 0.02864
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.258605      Inf
```

此时应有 P 值为 0.02864<0.05，认为接受原假设

既认为二者之间没有关系

3.对不同年龄段的人群进行电视节目的调查

对于将样本按照年龄分层的数据，我选择用 Mantelhaen 检验进行检验，应有代码如下：

```
#Mantel-Haenszel 检验
c1<-c(87,70,45);
c2<-c(91,86,15);
c3<-c(41,38,10);
data1<-rbind(c1,c2,c3)
#假设 H0：三种年龄的关注度一样
x=array(data1,c(3,3,3))
```

```
8. mantelhaen.test(x)
```

不妨假设：H0：三种年龄的关注度一样

应有结果：

```
1.  mantelhaen.test(x)
2.
3.  Cochran-Mantel-Haenszel test
4.
5. data:  x
6. Cochran-Mantel-Haenszel M^2 = 54.252, df = 4, p-value = 4.6
   6e-11
```

此时 P 值为：4.66e-11，P 值过小，接受原假设，认为不同年龄段对于不同电视节目的关注相同

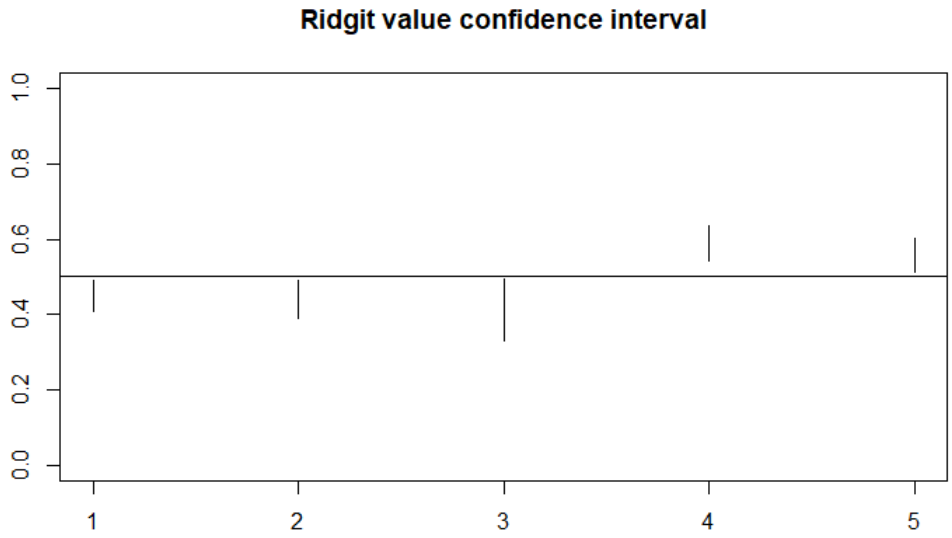
4.对电信公司手机满意度进行调查

对于本实验，我们选择 Ridit 检验进行检验，应有代码如下：

```
1. data<-data.frame(read.table("result.txt"))
2. c1<-c(data[1:5,1])
3. c2<-c(data[7:11,1])
4. c3<-c(data[13:17,1])
5. c4<-c(data[19:23,1])
6. c5<-c(data[25:29,1])
7. manyi<-rbind(c1,c2,c3,c4,c5);
8. ridi.test<-function(x)
9. {
10.  order.num=ncol(x)
11.  treat.num=nrow(x)
12.  rowsum=rowSums(x)#O_i.
13.  colsum=colSums(x)#O_.i
14.  total=sum(rowsum)
15.  N=(colsum/2)[1:order.num]+c(0,(cumsum(colsum))[1:order.nu
   m-1])
16.  ri=N/total#每个顺序类的得分
```

```
  p_coni=x/outer(rowsum,rep(1,order.num),"*")##概率阵—i 水平下属于第 j 顺序类的概率、
  pi.=rowsum/total
  score=p_coni%*%ri
  confi_inter=matrix(c(score-1/sqrt(3*rowsum),score+1/sqrt(3*rowsum)))

  if(length(rle(sort(ri))$lengths)==length(ri))#不打结
  {
    w=(12*total/(total+1)*sum(rowsum*(score-0.5)^2))
  }
  if(length(rle(sort(ri))$lengths)<=length(ri))#打结
  {
    tao<-rle(sort(ri))$lengths
    T=1-sum(tao^3-tao)/(order.num^3-order.num)
    w=(12*total/((total+1)*T))*sum(rowsum*(score-0.5)^2)}
  pvalue=pchisq(w,treat.num-1,lower.tail = FALSE)
  list(score,confi_inter=confi_inter,W=w,pvalue=pvalue)
}
options(digits = 4)##设结果为4 位有效数字

res_data=ridi.test(manyi)
graph_data<-res_data$confi_inter
x11()
plot(0,0,ylim = c(0,1),xlim = c(1,5),xlab = "function",ylab = "",main="Ridgit value confidence interval",col="gray7")

abline(h=0.5)
for(i in 1:(nrow(graph_data)/2))lines(c(i,i),c(graph_data[i],graph_data[i+5]))
```

假设 H0：各个满意度之间没有差异



得到的检验结果有：

Riditscore	Confidence Interval		W	P-value
0.4492	0.409	0.4893	37.85	1.20E-07
0.439	0.3888	0.4893		
0.411	0.3293	0.4926		
0.5877	0.5412	0.6342		
0.558	0.5129	0.6031		

当 alpha=0.5 的时候，拒绝 H0，认为满意度之间是存在差异的

问项 4、5 满意度比较高，问项 1、2、3 满意度较低

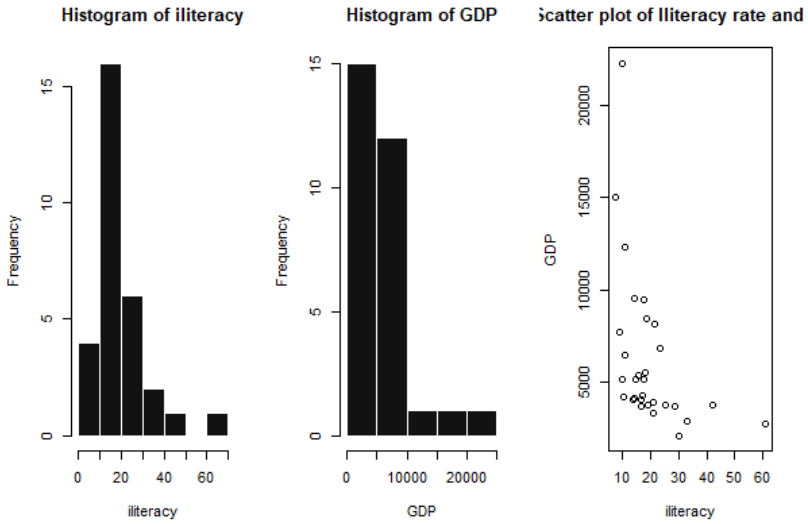
5.检验文盲率和人均 GDP 之间的关系

本实验运用 Pearon,Spearman 和 Kendall 检验法进行检验，代码如下：

```
iliteracy<-c(7.33,10.80,15.60,8.86,9.70,18.52,17.71,21.24,23.20,14.24,13.82,17.97,10.00,10.15,17.05,10.94,20.97,16.40,16.59,17.40,14.12,18.99,30.18,28.48,61.13,21.00,32.88,42.14,25.02,14.65)
GDP<-c(15044,12270,5345,7730,22275,8447,9455,8136,6834,9513,4081,5500,5163,4220,4259,6468,3881,3715,4032,5122,4130,3763,2093,3715,2732,3313,2901,3748,3731,5167)
par(mfrow = c(1, 3))
hist(iliteracy,border = F,col = "gray7")
hist(GDP,border = F,col = "gray7")
plot(iliteracy,GDP,main="Scatter plot of Iliteracy rate and GDP")
#编写 cor.pearson、cor.spearman,cor.kendall 函数
cor.pearson<-function(x,y)
{
  x1<-x-mean(x)
  y1<-y-mean(y)
  numerator<-sum(x1*y1)
  denominator<-sqrt(sum(x1^2)*sum(y1^2))
  cor<-numerator/denominator
  Z<-cor*sqrt(length(x)-2)/sqrt(1-cor^2)
  p_value<-2*pt(Z,length(x)-2)
  list(p_value=p_value,cor=cor)

}


cor.spearman<-function(x,y)
{
  x.rank<-rank(x)
  y.rank<-rank(y)
  x1<-x.rank-mean(x.rank)
  y1<-y.rank-mean(y.rank)
  numerator<-sum(x1*y1)
  denominator<-sqrt(sum(x1^2)*sum(y1^2))
  cor<-numerator/denominator
  Z<-cor*sqrt(length(x)-2)/sqrt(1-cor.test^2)
  p_value<-2*pt(Z,length(x)-2)
  list(p_value=p_value,cor=cor)
```

```
}

cor.kendall<-function(x,y)
{
  options(digits = 4)
  n<-length(x)
  s=0
  c=0
  for(i in 1:(n-1)){
    for(j in (i+1):n){
      s=s+sign((iliteracy[i]-iliteracy[j])*(GDP[i]-GDP[j]))
      c=c+1
    }
  }
  tau <- 2/(n*(n-1))*s
  Z <- tau*sqrt((9*n*(n-1))/(2*(2*n+5)))
  p_value<-2*pnorm(Z)
  list(p_value=p_value,cor=cor)
}
#则有以下的结果
cor.test(iliteracy,GDP)
cor.pearson(iliteracy,GDP)
cor.test(iliteracy,GDP,method="spearman")
cor.spearman(iliteracy,GDP)
cor.kendall(iliteracy,GDP)
cor.test(iliteracy,GDP,method="kendall")
```

得到三种检查方法的图像有：



通过观察图像，提出假设：

H0：GDP 和文盲率之间呈现正相关关系

应有三种检验的结果如下：

Pearson's

```
1.  cor.test(iliteracy,GDP)
2.
3.  Pearson's product-moment correlation
4.
5. data:  iliteracy and GDP
6. t = -2.657, df = 28, p-value = 0.01287
7. alternative hypothesis: true correlation is not equal to 0
8. 95 percent confidence interval:
9.  -0.6964200 -0.1055303
10.sample estimates:
11.       cor
12.-0.4487388
1.  > cor.pearson(iliteracy,GDP)
2. $p_value
3. [1] 0.01287223
```

spearman

```
1. > cor.test(iliteracy,GDP,method="spearman")
2.
3.  Spearman's rank correlation rho
4.
5. data:  iliteracy and GDP
```

```
6. S = 7331.3, p-value = 0.0001851
7. alternative hypothesis: true rho is not equal to 0
8. sample estimates:
9.        rho
10.-0.6309934
```

kendall

```
1.  cor.kendall(iliteracy,GDP)
2. $p_value
3. [1] 0.0002731
```

得到三种结果的 P 值都相当小，拒绝原假设 H0，认为 GDP 和文盲率之间呈现负相关关系

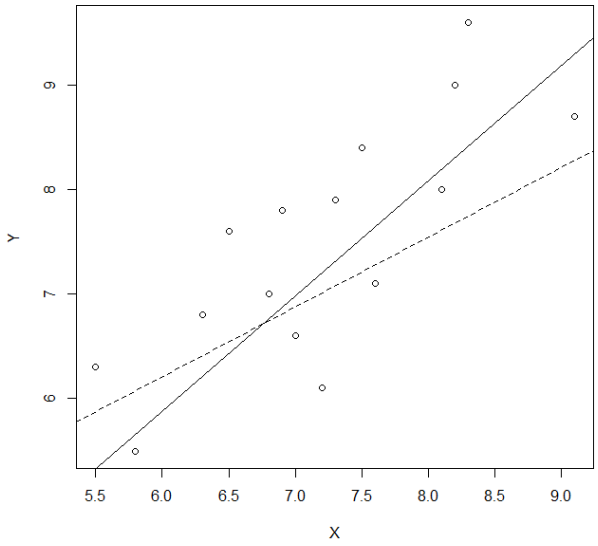
6.计算销售的中位数回归直线

通过 Brown-Mood 法计算回归直线，代码如下：

```
1. #中位数回归直线
2. X<-c(9.1,8.3,7.2,7.5,6.3,5.8,7.6,8.1,7.0,7.3,6.5,6.9,8.2,6.
   8,5.5);
3. Y<-c(8.7,9.6,6.1,8.4,6.8,5.5,7.1,8.0,6.6,7.9,7.6,7.8,9.0,7.
   0,6.3);
4. cyx=coef(lm(X~Y))
5. md=median(X)
6. x1<-X[X<=md]
7. x2<-X[X>md]
8. y1<-Y[Y<=md]
9. y2<-Y[Y>md]
10.md1=median(x1)
11.md2=median(x2)
12.mw1=median(y1)
13.mw2=median(y2)
14.beta=(mw2-mw1)/(md2-md1)
15.alpha<-median(X-beta*X)
16.x11()
```

```
plot(X,Y)
abline(alpha,beta)
abline(c(cyx),lty=2)
```

得到的直线有：



7.对歌手打分进行一致性检验

本实验通过多元 Kendall 检验法检验，代码如下：

```
#Kendall 检验法
#载入数据
data1<-read.table("result2.txt")
data2<-as.numeric(data1$V1)
data<-matrix(data2,nrow=10,ncol=12)
data=t(data)
#不妨假设H0：裁判打分是一致的
zhibiao<-rep(0,120)
rzhibiao<-matrix(zhibiao,12,10)
#求每组的秩和
for(i in 1:10)
{

  br1<-rank(data[,i])
  print(br1)
```

```
  for(t in 1:12){
        rzhibiao[t,i]=br1[t]
      }
}
#发现原秩表是有结表
#计算R_i
R.i<-rep(0,12)
for(i in 1:12){
  br2=0
  for(t in 1:10){
    br2=br2+rzhibiao[i,t]
  }
  R.i[i]<-br2
}
print(R.i)
qiuheR.j=0
for(i in 1:12){
  qiuheR.j=qiuheR.j+(R.i[i])^2;
}
#求有结系数T
T<-11*(2^3-2)+(4^3-4)
#计算总系数W_c
W=(12*qiuheR.j-(3*10^2*12*(12+1)^2))/(10^2*(12^3-12)-10*T)
print(W)
#得到计算的W值为0.1135611
#近似卡方值应有
Xc<-10*(12-1)*W
print(Xc)
#得到卡方近似值为12.49172，查表有X_0.95,11为19.675
#12.49172<19.675
#拒绝原假设 认为评委打分是不一致的
```

假设 H0：裁判打分是一致的

通过计算得到的结果有：

得到计算的 W 值为 0.1135611，近似卡方值应有

Xc<-10*(12-1)*W

```
> print(Xc)
[1] 12.49172
```

得到卡方近似值为 12.49172，查表有 X_0.95,11 为 19.675

12.49172<19.675

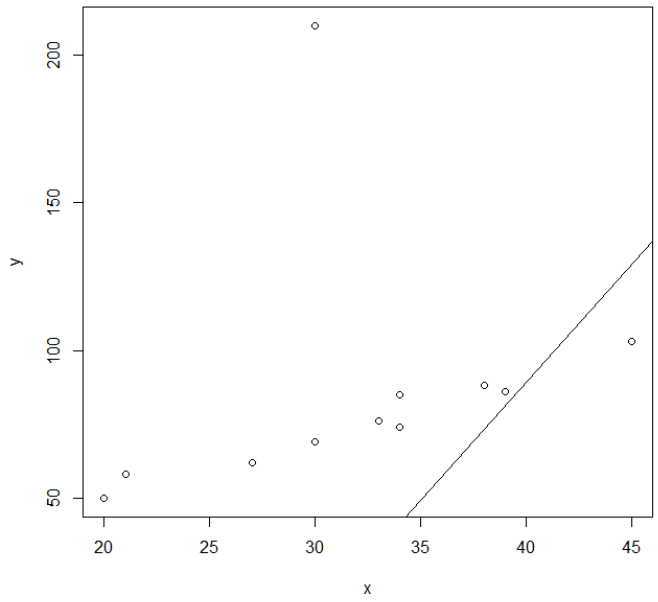
拒绝原假设 认为评委打分是不一致的

8.对某种材料进行离群数据分析

对于原数据有：

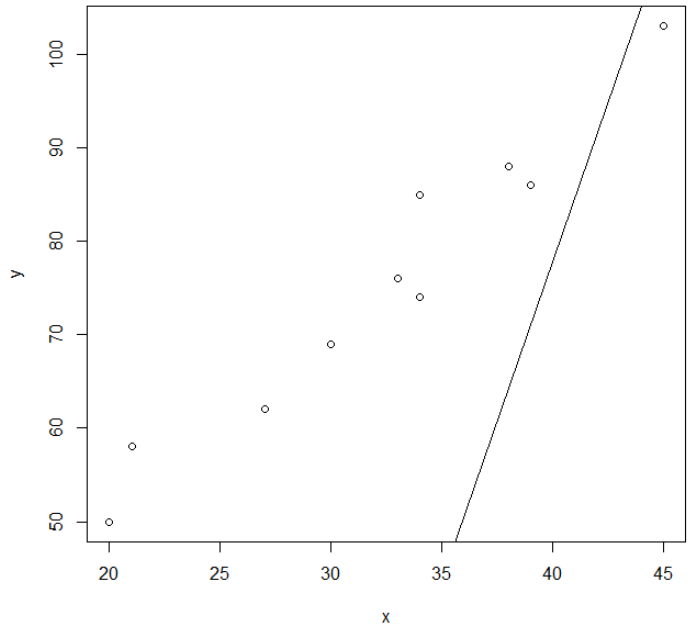
```
x<-c(33,45,30,20,39,34,34,21,27,38,30)
y<-c(76,103,69,50,86,85,74,58,62,88,210)

cyx=coef(lm(x~y))
md=median(x)
x1<-x[x<=md]
x2<-x[x>md]
y1<-y[y<=md]
y2<-y[y>md]
md1=median(x1)
md2=median(x2)
mw1=median(y1)
mw2=median(y2)
beta=(mw2-0)/(md2-md1)
alpha<-median(x-beta*x)
x11()
plot(x,y)
abline(alpha,beta)
abline(c(cyx),lty=2)
#两线重合 是一致的
#认为出现了离群点：(30,210)
#删去后作图如下
```

拟合的曲线为：



通过观察认为出现离群点(30,210)

删去离群点后拟合图像如下：



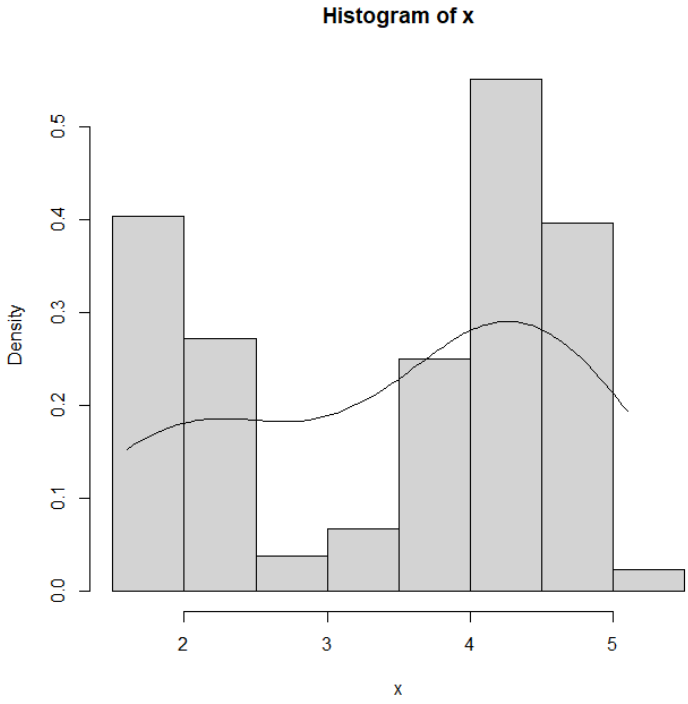
并且通过观察认为 Theil 和 Brown-mood 方法作线性回归的方程是一致的，两个线性方程曲线在图像上很好重合。

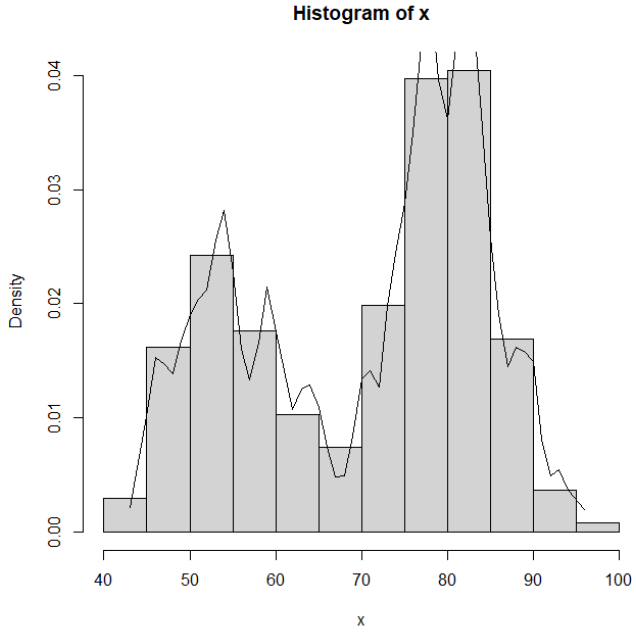
9.对老忠实泉的三个分数据进行核估计

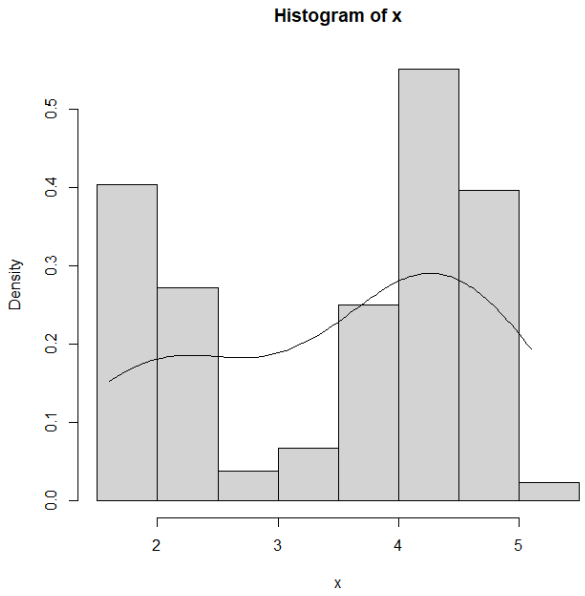
核估计代码如下：

```
require(stats); require(graphics)
f.tit <-  "faithful data: Eruptions of Old Faithful"

ne60 <- round(e60 <- 60 * faithful$eruptions)
all.equal(e60, ne60)              # 相对差异 ~ 1/10000
table(zapsmall(abs(e60 - ne60))) # 0、0.02 或 0.04
faithful$better.eruptions <- ne60 / 60
te <- table(ne60)
te[te >= 4]                       # (太多) 5 的许多倍数!
plot(names(te), te, type = "h", main = f.tit, xlab = "Eruption time (sec)")

plot(faithful[, -3], main = f.tit,
     xlab = "Eruption time (min)",
     ylab = "Waiting time to next eruption (min)")
lines(lowess(faithful$eruptions, faithful$waiting, f = 2/3, iter = 3),
      col = "red")

ker.density=function(x,h){
  x=sort(x)
  n=length(x);s=0;t=0;y=0
  for(i in 2:n)
    s[i]=0
  for(i in 1:n){
    for(j in 1:n)
      s[i]=s[i]+exp(-((x[i]-x[j])^2)/(2*h*h))
    t[i]=s[i]
  }
  for(i in 1:n)
    y[i]=t[i]/(n*h*sqrt(2*pi))
  z=complex(re=x,im=y)
  hist(x,freq=FALSE)
  lines(z)
}
x11()
#分别作三个数据的核估计如下:
ker.density(faithful$better.eruptions,0.8)
```

选定系数为 0.8 时，有核估计图像如下：







10. 阐述非参数方法和参数方法的区别和联系

我认为参数检验是在总体分布形式已知的情况下，对总体分布的参数如均值、方差等进行推断的方法。但是，在数据分析过程中，由于种种原因，我们往往无法对总体分布形态作简单假定，此时参数检验的方法就不再适用了。非参数检验正是一类基于这种考虑，在总体方差未知或知道甚少的情况下，利用样本数据对总体分布形态等进行推断的方法。

在参数检验中，我们一般检验的指标是样本数据的均值，并且在已知样本分布的情况下对于参数进行估计，参数方法的优点是当样本数据符合参数条件时，其估计的精度相当高，并且有很高的检验效率。但是，这无疑对于样本数据的来源是要求苛刻的，并且对样本数据的

独立性，方差，何其分布情况都有了要求，在真正的数据中很难判断和断定他们的真实来源和准确性。

在非参数检验中，我们检验的目标一般是样本的中位数，并且我们对于样本的分布是位置的，这虽然对样本的估计增加了难度，但是同时也给样本估计放宽了条件，我们不再需要知道样本的分布就能估计其参数或者进行检验，其适用的范围更广泛，并且通过求秩检验的方法是相对简单的方法与思路。但是，其缺点是：如果符合参数检验的数据不小心用了非参数检验，其检验结果及其的不好，并且非参数检验的结果的精确性与准确性相对不是那么高。中位数很容易受到样本的离散程度的影响（想起来在描述统计学中学习到的离散系数，其本身就对样本中位数的估计有极大的影响）。