Bootstrap Methods in Analyzing CD4 Data

Nana Boateng[*]Qirui Huang[†]Akwasi Kusi-Appiah[‡]

December 12, 2012

1 Introduction

1.1 Background of The CD4 Data

- The $CD4$ data frame has 20 rows and 2 columns. $CD4$ cells are carried in the blood as part of the human immune system. One of the effects of the HIV virus is that these cells die.

- The count of $CD4$ cells is used in determining the onset of full-blown AIDS in a patient. In this study of the effectiveness of a new anti-viral drug on HIV, 20 HIV-positive patients had their $CD4$ counts recorded and then were put on a course of treatment with this drug.

- After using the drug for one year, their $CD4$ counts were again recorded.The aim of the experiment was to show that patients taking the drug had increased $CD4$ counts which is not generally seen in HIV-positive patients.

- Here we are interested in whether our new drug works effectively on our HIV-positive patients after one-year treatment based on the paired data. The percentage change of CD4 counts will be calculated as our test statistics which is given by $T = \frac{\bar{Y}-\bar{X}}{\bar{X}}$. Since we dont know its distribution which is hard to find mathematically too, bootstrap method will be conducted.

2 Regression Analysis

```
> library(boot)
> cd4
  baseline oneyear
1     2.12    2.47
2     4.35    4.61
3     3.39    5.26
4     2.51    3.02
5     4.04    6.36
```

[*] Department of Mathematics, University of Memphis

```
6      5.10    5.93
7      3.77    3.93
8      3.35    4.09
9      4.10    4.88
10     3.35    3.81
11     4.15    4.74
12     3.56    3.29
13     3.39    5.55
14     1.88    2.82
15     2.56    4.23
16     2.96    3.23
17     2.49    2.56
18     3.03    4.31
19     2.66    4.37
20     3.00    2.40
>
> plot(cd4$baseline,cd4$oneyear)
>   cd4.lm <- glm(cd4$oneyear ~ cd4$baseline, data=cd4)
> summary(cd4.lm)

Call:
glm(formula = cd4$oneyear ~ cd4$baseline, data = cd4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3950  -0.5382  -0.1561   0.5856   1.4888

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.6904     0.7878   0.876 0.392377
cd4$baseline   1.0349     0.2330   4.442 0.000315 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for gaussian family taken to be 0.6775183)

    Null deviance: 25.565  on 19  degrees of freedom
Residual deviance: 12.195  on 18  degrees of freedom
AIC: 52.864

Number of Fisher Scoring iterations: 2

> cd4.diag=glm.diag.plots(cd4.lm,ret=T)
> # diagnostic plots
> layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
> plot(cd4.lm)
```

The plot of the the data after oneyear versus baseline year shows there is some association between the two.We further derive a regression model which shows the data from the baseline year is not significant with a p-value of 0.000315.This means the response variable ,which is the data after oneyear does not comeabout wholly depend on the baseline year but some other extra factor which is the treatment the patients were put on.The diagnostic plots of the residuals versus the linear predictor appears not to have any form of association which indicates that the error variance is contant hence the linear model accurate.The observation 5,from the Cook distance plot appears to have a rather higher cooks distance which implies that observation contributes significantly to the model outcome and hence contributes highly to type1 errors. The normal QQ-plot versus the Theoretical quantiles indicates that a normality assumption for the residuals would be inaccurate.

3 Bootstrap Applications

3.1 Confidence Intervals

The Nonparamtric Boot.ci Method

```
 #using diff of means
> library(boot)
> cd4

> diff=cd4$oneyear-cd4$baseline
>  mean.fun <- function(d, i)
+ {    m <- mean(d[i])
+      n <- length(i)
+      v <- (n-1)*var(d[i])/n^2
+      c(m, v)
+ }
> diff.boot <- boot(diff, mean.fun, R = 999)
> boot.ci(diff.boot, type = c("norm", "basic", "perc", "stud"))
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = diff.boot, type = c("norm", "basic", "perc",
    "stud"))

Intervals :
Level      Normal              Basic
95%   ( 0.4690,  1.1463 )   ( 0.4460,  1.1370 )

Level    Studentized          Percentile
95%   ( 0.4444,  1.2157 )   ( 0.4730,  1.1640 )
```

```
Calculations and Intervals on Original Scale
>
```

The Percentile Method

```
 i <- order(diff.boot1$t)
> yo <- diff.boot1$t[i]
> y1 <- (yo[25])
> y2 <- (yo[975])
> (2*(mean(diff.boot1$t0))-(y1))
[1] 0.3662303
> (2*(mean(diff.boot1$t0))-(y2))
[1] 0.1436618
> CIB=c((2*(mean(diff.boot1$t0))-(y2)),(2*(mean(diff.boot1$t0))-(y1)))
> CIB
[1] 0.1436618 0.3662303
```

normal approximation

```
 yb <- mean(diff)

> n <- nrow(cd4)
> vhart=(yb^2)/n
>
> L1=yb-qnorm(.975)*sqrt(vhart)
>
>
>
> U1=yb + qnorm(.975)*sqrt(vhart)
>
>
>
>
>
>    limit= c(L1,U1)
>
>     limit
[1] 0.4521997 1.1578003
>
```

Fisher Transformation of The Correlation Method

```
corr.fun<-function(d,w=rep(1,nrow(d))/nrow(d))
+ {w=w/sum(w)
+ n=nrow(d)
+ m1=sum(d[,1]*w)
+ m2=sum(d[,2]*w)
+ v1=sum(d[,1]^2*w)-m1^2
+
+ v2=sum(d[,2]^2*w)-m2^2
+ rho=(sum(d[,1]*d[,2]*w)-m1*m2)/sqrt(v1*v2)
+ i=rep(1:n,round(n*w))
+ us=(d[i,1]-m1)/sqrt(v1)
+
+ us
+
+ xs=(d[i,2]-m1)/sqrt(v2)
+ L=us*xs-0.5*rho*(us^2+xs^2)
+ c(rho,sum(L^2)/n^2)
+
+ }
>
> cd4.boot=boot(cd4,corr.fun,R=999,stype="w")
>
> boot.ci(cd4.boot, type = c("norm", "basic", "perc", "stud"))
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = cd4.boot, type = c("norm", "basic", "perc",
    "stud"))

Intervals :
Level      Normal              Basic
95%   ( 0.5453,  0.9090 )   ( 0.5779,  0.9448 )

Level    Studentized          Percentile
95%   ( 0.4884,  0.8879 )   ( 0.5015,  0.8684 )
Calculations and Intervals on Original Scale
>
> abc.ci(cd4,corr,conf=0.95)
[1] 0.9500000 0.5208270 0.8510486
>
>
> #fisher transformation
> fisher=function(r)0.5*log((1+r)/(1-r))
> fisher.dot=function(r) 1/(1-r^2)
> fisher.inv=function(z) (exp(2*z)-1)/exp((2*z)+1)
> boot.ci(cd4.boot,h=fisher,hdot=fisher.dot,hinv=fisher.inv,conf=0.95)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = cd4.boot, conf = 0.95, h = fisher, hdot = fisher.dot,
    hinv = fisher.inv)

Intervals :
Level      Normal              Basic             Studentized
95%   ( 0.2357,  0.3394 )   ( 0.2331,  0.3393 )   ( 0.2480,  0.3421 )

Level     Percentile            BCa
95%   ( 0.2458,  0.3420 )   ( 0.2405,  0.3407 )
Calculations on Transformed Scale;  Intervals on Original Scale
>
```

Paired T- Test

```
 a=cd4$baseline
> b=cd4$oneyear
>
>  t.test(b,a, paired=TRUE)

        Paired t-test

data:  b and a
t = 4.4908, df = 19, p-value = 0.0002504
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4298118 1.1801882
sample estimates:
mean of the differences
                  0.805
```

Histogram plots of Difference in Means with Corresponding Normal Density

```
> par(mfrow=c(1,2))
>
> x1=   diff.boot1$t[1:99]
> x2=   diff.boot1$t[1:999]
>
> histnorm=function(x, ...)
```

```
+
+ {
+    if (length(x)>0)
+    {
+       hist(x,freq=FALSE, ...)
+       rug(x)
+       curve(dnorm(x, mean=mean(x), sd=sd(x)), add=TRUE, col="red",
 lty="dotted", xaxt="n")
+       abline(v=mean(x),col="blue")
+       mtext(paste("mean ", round(mean(x),1), "; sd ", round(sd(x),1), "; N
 ", length(x),sep=""), side=1, cex=.75)
+    } # fi
+ } # histnorm
>  histnorm(x1)
>   title(sub="t*=99")
>
>  histnorm(x2)
>  title(sub="t*=999")
```

The Bootstrap Average

```
#bootstrap average
library(boot)
diff

diff.f <- function(data,i)
{
   d<-data[i]
   c(mean(d),var(d))
}




diff.bootnp<-boot( diff, diff.f, R=999)




par(mfrow=c(1,2))


plot(diff.bootnp$t[1:99,1],sqrt(diff.bootnp$t[1:99,2]),axes=F,main=
"",xlim=c(0.2,1.5),ylim=c(0.4,1.2),xlab="Bootstrap average(t*=99)"
,ylab="Bootstrap SD",pch=20,cex=1.5)
axis(1,at=c(0.2,0.4,0.8,1.4,1.5),labels=paste(c(0.2,0.4,0.8,1.4,1.5)))
axis(2,at=c(0,0.2,0.4,0.6,0.8,1.2),labels=paste(c(0,0.2,0.4,0.6,0.8,1.2)))
box()
```

```
plot(diff.bootnp$t[,1],sqrt(diff.bootnp$t[,2]),axes=F,main=""
,xlim=c(0.2,1.5),ylim=c(0.4,1.2),xlab="Bootstrap average(t*=999)",ylab=
"Bootstrap SD",pch=20,cex=1.5)
axis(1,at=c(0.2,0.4,0.8,1.4,1.5),labels=paste(c(0.2,0.4,0.8,1.4,1.5)))
axis(2,at=c(0,0.2,0.4,0.6,0.8,1.2),labels=paste(c(0,0.2,0.4,0.6,0.8,1.2)))
box()
```

The qualitative feature to be read from these plots is that data standard deviation is proportional to data average.The plots indicate that the bootstrap standard deviation is significantly proportinal to the bootstrap average.

Graphs Empirical Biases and Variance

```
library(boot)
bias<-matrix(nrow=8,ncol=4)

variance<-matrix(nrow=8,ncol=4)




i<-c(10,20,50,100,200,300,400,500)
bias<-matrix(nrow=8,ncol=4)

variance<-matrix(nrow=8,ncol=4)


for(j in 1:4)
{
diff.boot1<-boot(diff,diff.f,R=1000)
for(k in 1:8)
  { bias[k,j]<-mean(diff.boot1$t[1:i[k]])-mean(diff)
    variance[k,j]<-var(diff.boot1$t[1:i[k]])
  }
}

par(mfrow=c(1,2))
matplot(log(i),bias,axes=F,xlim=c(2.2,6.3),ylim=c(-0.00004,0.08),
xlab="R",ylab="Bias",type="c",pch=20,col=1,lty=1)
```

```
#matplot(i,bias,axes=F,xlim=c(10,500),ylim=c(-50,10),xlab="R",ylab=
"Bias",type="c",pch=20,col=1,lty=1)
matpoints(log(i),bias,lty=1,pch=".",col=1,cex=1.5)
axis(1,at=c(log(10),log(50),log(100),log(500)),labels=paste(c(10,50,100,500)))
axis(2,at=c(-0.00004,-0.004,-0.04,0.0004,0.004,0.04,0.08),labels=
paste(c(-0.00004,-0.004,-0.04,0.0004,0.004,0.04,0.08)))
abline(h=0,lty=3)
#abline(h=mean(diff)^2/NROW(diff),lty=3)
box()

matplot(log(i),variance,axes=F,xlim=c(2.2,6.3),ylim=c(-0.0002,0.08),xlab="R",
ylab="Variance",type="c",pch=20,col=1,lty=1)
matpoints(log(i),variance,lty=1,pch=".",col=1,cex=1.5)
axis(1,at=c(log(10),log(50),log(100),log(500)),labels=paste(c(10,50,100,500)))
axis(2,at=c(-0.0004,-0.004,-0.04,0.0004,0.004,0.04,0.08),labels=
paste(c(-0.0004,-0.004,-0.04,0.0004,0.004,0.04,0.08)))
abline(h=mean(diff)^2/NROW(diff),lty=3)
box()
#title(sub=" ")
```

- The empirical approximations are justified by the law of large numbers
- The bias of the bootstrap sample converges to the exact value under the fitted model as R,increases much more quickly done the variance does

Theoritical Quantile Plots

```
x<-qgamma((1:999)/1000,NROW(diff),rate=NROW(diff)/mean(diff))
y<-qgamma((1:99)/100,NROW(diff),rate=NROW(diff)/mean(diff))

par(mfrow=c(2,2))

qqnorm(diff.boot1$t[1:99],axes=F,pch=20,main="",xlim=c(-1,2.5),ylim
=c(-1,2),xlab="Quantiles of standard normal",ylab="t*",cex=0.9)
axis(1,at=c(-1,-.5,.5,1,1.5,2,2.5),labels=paste(c(-1,-.5,.5,1,1.5,2,2.5)))
axis(2,at=c(-1,-.5,.5,1,1.5,2),labels=paste(c(-1,-.5,.5,1,1.5,2)))
qqline(diff.boot1$t[1:99],lty=3)

box()

qqnorm(diff.boot1$t[1:999],axes=F,pch=20,main="",xlim=c(0,2.5),ylim=c(-1,1.5)
,xlab="Quantiles of standard normal",ylab="t*",cex=0.5)
axis(1,at=c(-1,1,2),labels=paste(c(-1,1,2)))
axis(2,at=c(-1,2,4,6),labels=paste(c(-1,2,4,6)))
```

```
qqline(diff.boot1$t[1:999],lty=3)
box()

qqplot(y,diff.boot1$t[1:99],axes=F,pch=20,main="",xlim=c(0.4,1.2),ylim=c(-1,2),
xlab="Theoritical  quantile",ylab="t*",cex=0.9)
axis(1,at=c(-1,-.5,.5,1,1.5,2),labels=paste(c(-1,-.5,.5,1,1.5,2)))
axis(2,at=c(-1,-.5,.5,1,2),labels=paste(c(-1,-.5,.5,1,2)))
abline(a=0,b=1,lty=3)
box()

qqplot(x,cd4.boot1$t[1:999],axes=F,pch=20,main="",xlim=c(0.3,1.5),ylim
=c(3,4),xlab="Theoritical gamma quantile",ylab="t*",cex=0.5)
axis(1,at=c(0.3,0.6,0.9,1.2,1.5),labels=paste(c(0.3,0.6,0.9,1.2,1.5)))
axis(2,at=c(3,3.5,4),labels=paste(c(3,3.5,4)))
abline(a=0,b=1,lty=3)
box()

title(sub="Theoritical quantiles,qgamma")
```

The plot of the theoritical quantiles indicate that the non-normality of the bootstrap distribution.

Density Function Plots

```
library(boot)
cd4
cd42<-cd4[1:10,]
cd42<-cd4[1:20,]

t1<-mean(cd42$baseline)/mean(cd42$oneyear)
t2=mean(cd4$baseline)/mean(cd4$oneyear)
t2

v1L=sum((cd42$baseline-t1*cd42$oneyear)^2)/(nrow(cd42)*mean(cd42$baseline))^2


v2L<-sum((cd4$baseline-t2*cd4$oneyear)^2)/(nrow(cd4)*mean(cd4$baseline))^2

ratio<-function(data,i)
{ tstar<-mean(data$baseline[i])/mean(data$oneyear[i])
  vstar<-sum((data$baseline[i]-tstar*data$oneyear[i])^2)/(mean(data$baseline[i])*nrow(data)
  c(tstar,vstar)
}

#bootstrapping,geting t* and the corresponding densities
```

```
cd42.boot<-boot(cd42,ratio,R=999)


mb1<-mean(cd42.boot$t[,1])-t1

v1<-var(cd42.boot$t[,1])

x1<-sort(cd42.boot$t[,1])

d.a1<-dnorm(x1-t1,mean=mb1,sd=sqrt(v1))

d.b1<-dnorm(x1-t1,mean=0,sd=sqrt(v1L))
d.kde1<-rep(0,999)

bw1<-density(x1-t1)$bw
for(j in 1:999)
{ for(k in 1:999)
     { d.kde1[j]<-d.kde1[j]+dnorm((x1[j]-x1[k])/bw1,mean=0,sd=1)/(999*bw1) }
}
density1<-cbind(d.a1,d.b1,d.kde1)

cd4.boot<-boot(cd4,ratio,R=999)
mb2<-mean(cd4.boot$t[,1])-t2

v2<-var(cd4.boot$t[,1])

x2<-sort(cd4.boot$t[,1])
d.a2<-dnorm(x2-t2,mean=mb2,sd=sqrt(v2))
d.b2<-dnorm(x2-t2,mean=0,sd=sqrt(v2L))
d.kde2<-rep(0,999)
bw2<-density(x2-t2)$bw
for(j in 1:999)
{ for(k in 1:999)
     { d.kde2[j]<-d.kde2[j]+dnorm((x2[j]-x2[k])/bw2,mean=0,sd=1)/(999*bw2) }
}
density2<-cbind(d.a2,d.b2,d.kde2)

#now finally plotting

# Define colors to be used for cars, trucks, suvs
#plot_colors <- c(rgb(r=0.0,g=0.0,b=0.9), "red", "forestgreen")

# Start PDF device driver to save output to figure.pdf
#pdf(file="C:/Desktop/graph1", height=3.5, width=5)
```

```
par(mfrow=c(1,2))
matplot(x1-t1,density1,axes=F,main="",xlim=c(-0.2,0.2),xlab="t*-t"
,ylim=c(0,11),ylab="PDF",pch="",col=1:3,cex=0.1)
matlines(x1-t1,density1,lty=c(2,3,1))
axis(1,at=c(-0.2,-0.1,0,0.2),labels=paste(c(-0.2,-0.1,0,0.2)))
axis(2,at=c(0,2,4,6,8,10),labels=paste(c(0,2,4,6,8,10)))
box()

matplot(x2-t2,density2,axes=F,main="",xlim=c(-0.2,0.2),xlab="t*-t"
,ylim=c(0,11.5),ylab="PDF",pch="",col=1:3,cex=0.1)
matlines(x2-t2,density2,lty=c(2,3,1))
axis(1,at=c(-0.2,-0.1,0,0.1,0.2),labels=paste(c(-0.2,-0.1,0,0.1,0.2)))
axis(2,at=c(0,2,4,6,8,10,12),labels=paste(c(0,2,4,6,8,10,12)))
box()

title(" Density Estimates plots ")
```

- The approximation of the probability distribution function of t^* can be obtained using simulation results
- The density estimates plot shows the effect of small sample size on the accuracy of the normal distributions
- The density estimates for $t^* - t$ based on 999 non-parametric simulations for the entire CD4 data set.

4 Conclusion

A paired t test of the CD4 data proved the alternative hypothesis to be true that there is significant difference between the mean of the CD4 counts in the baseline year and the CD4 counts after oneyear. That confidence intervals dont contain zero along with our p value indicates that the new drug is effective for our HIV positive patients. This shows the anti-viral drug must have contributed significantly to theincrease in CD4 counts

References

[1] Bootstarp Methods and Their Application.A. C. Davidson and D.V. Hinkley.Cambridge University Press,1997

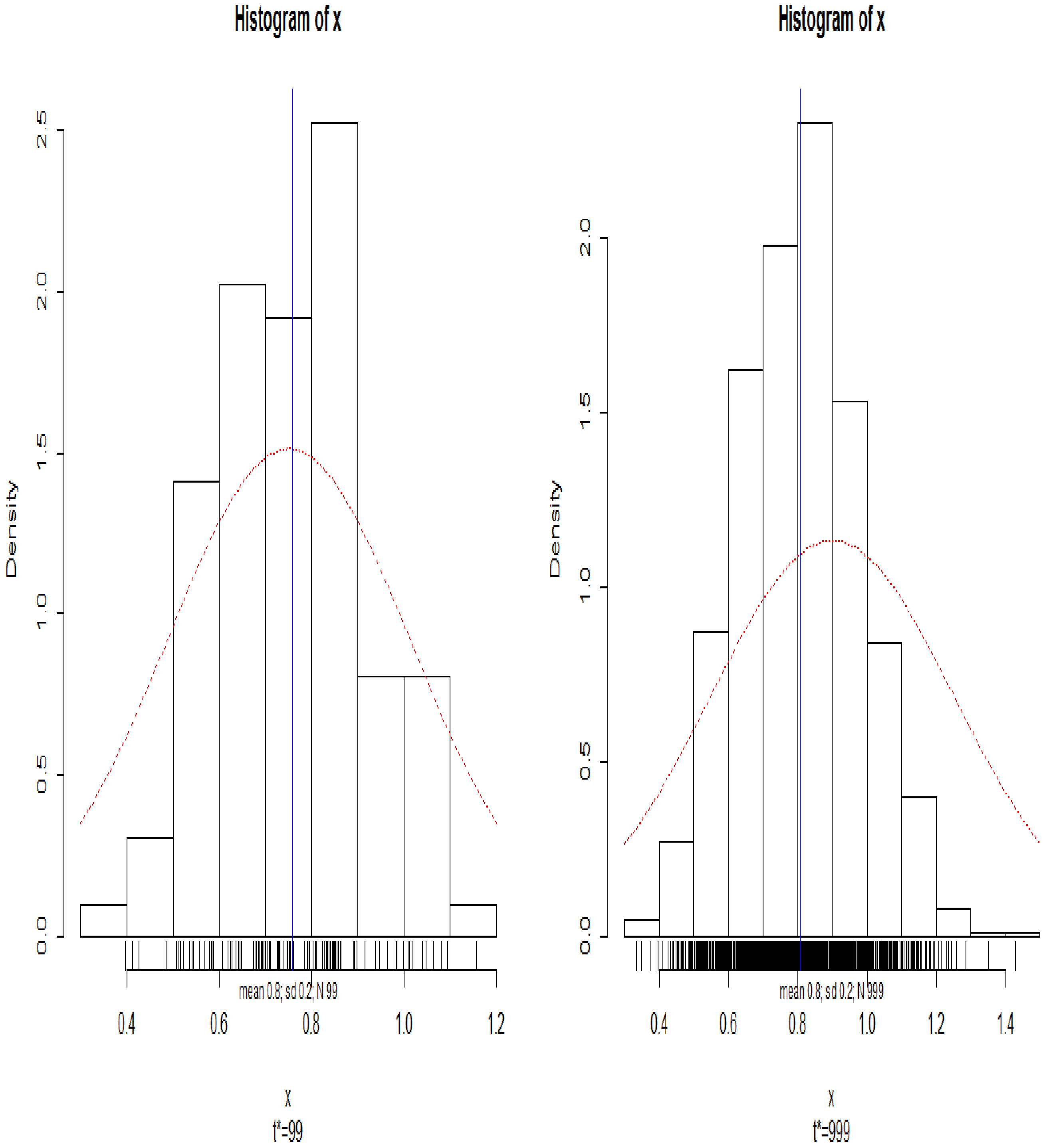

Figure 1: Histogram plots of Difference in Means with Corresponding Normal Density

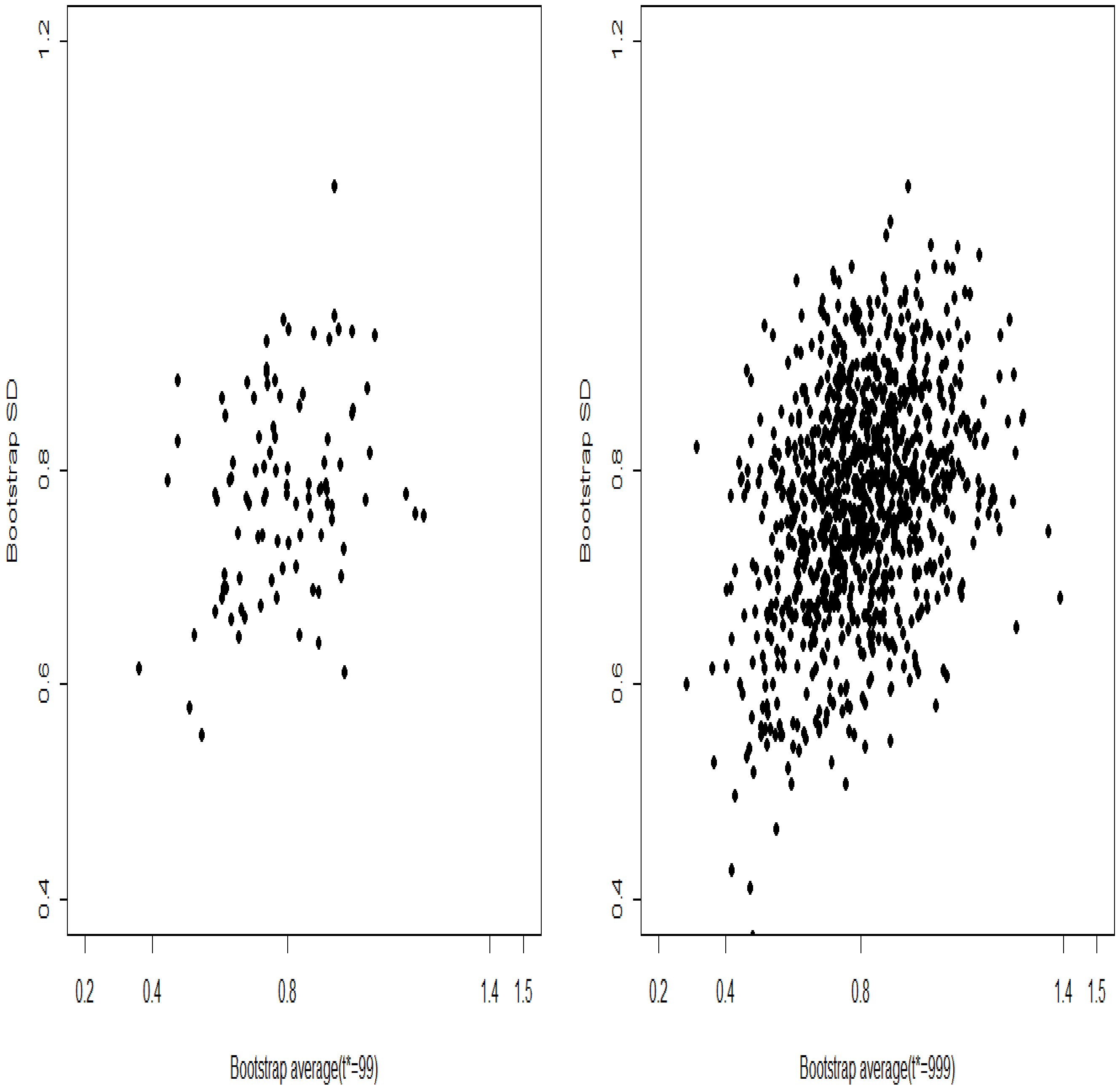

Figure 2: Scatter plots of Sample Standard Deviation versus Sample Average generated byNon-Parametric Bootstrap

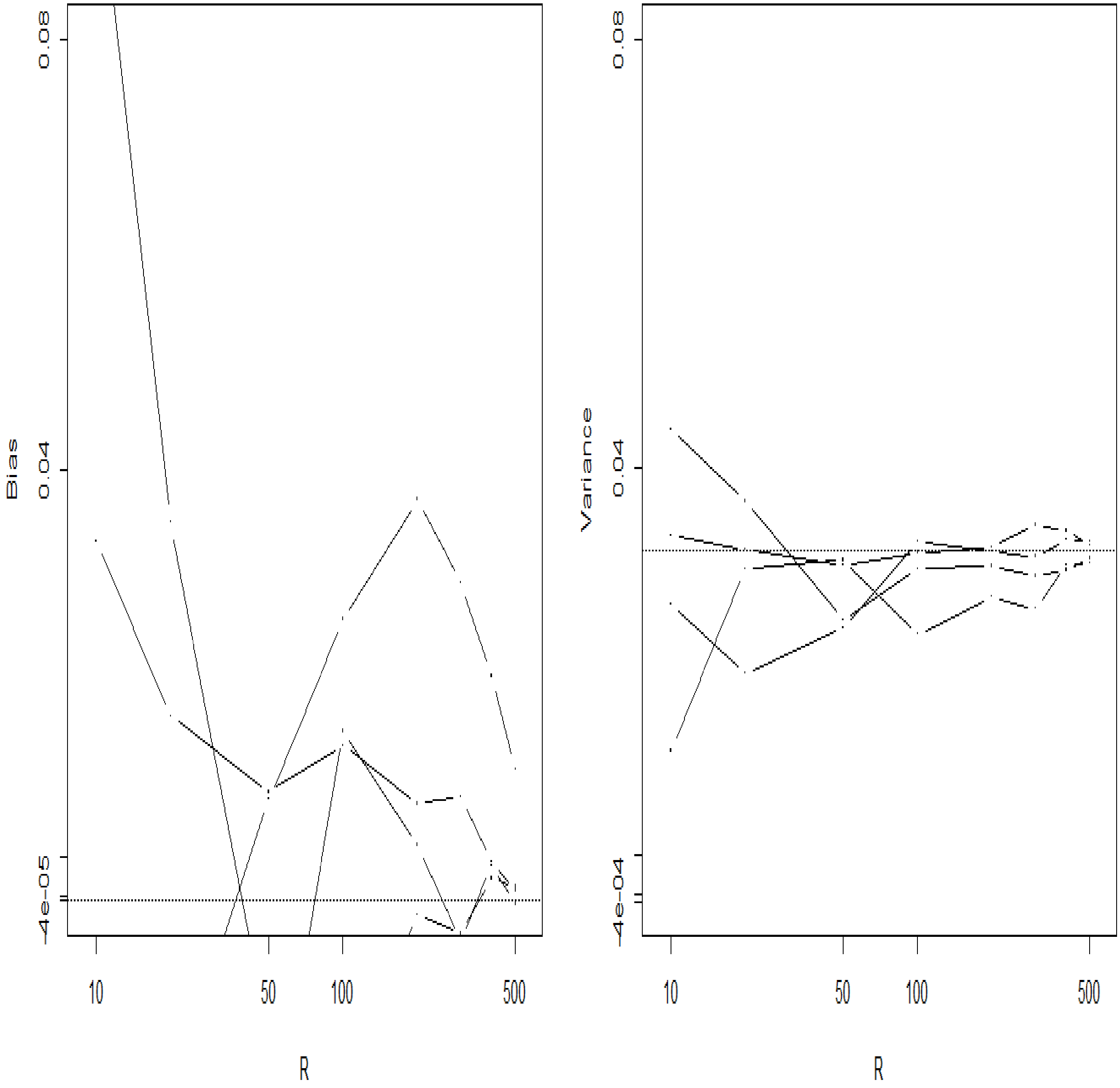

Figure 3: Empirical Biases and Variances of

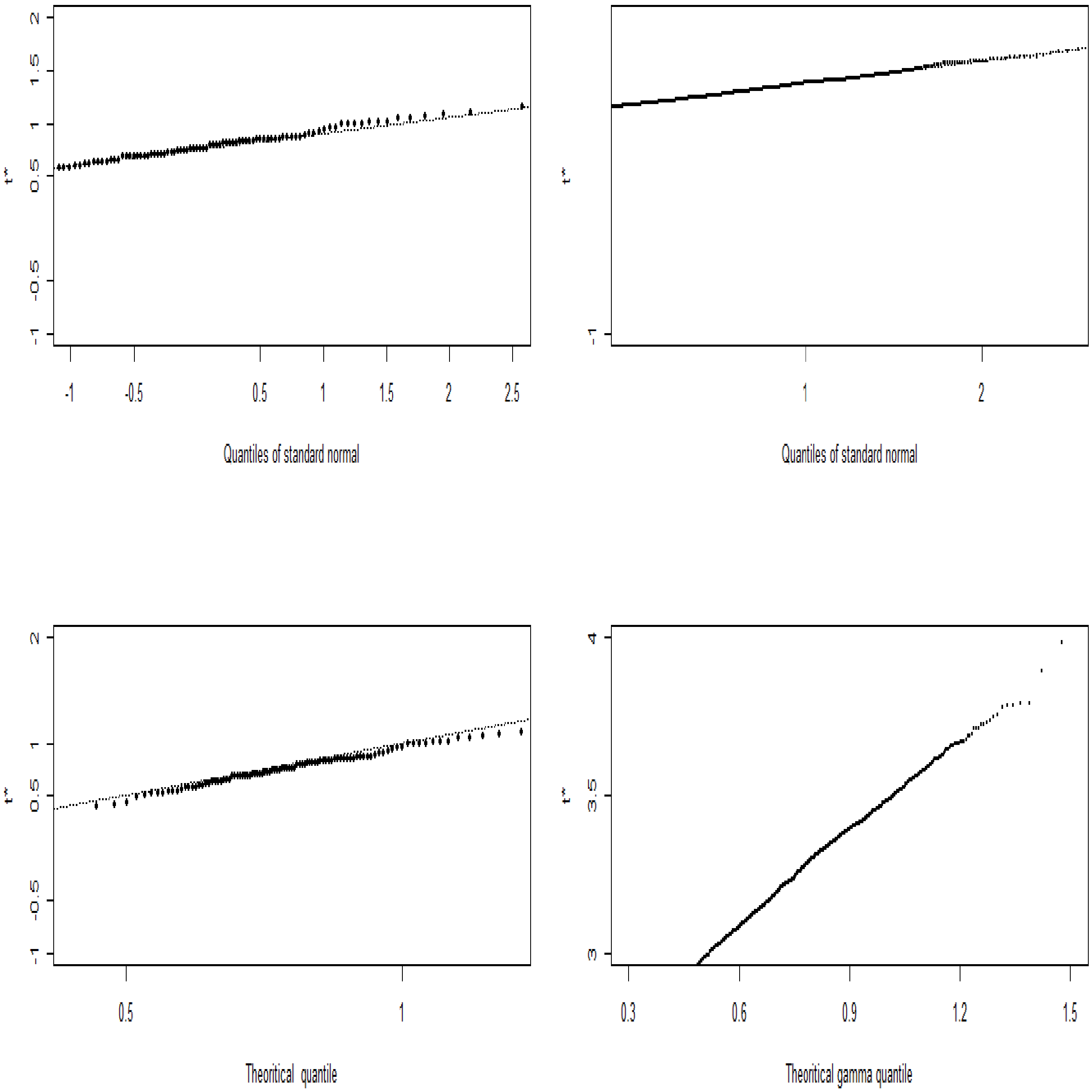

Figure 4: Quantile plots of Difference in Mean of CD4 Data

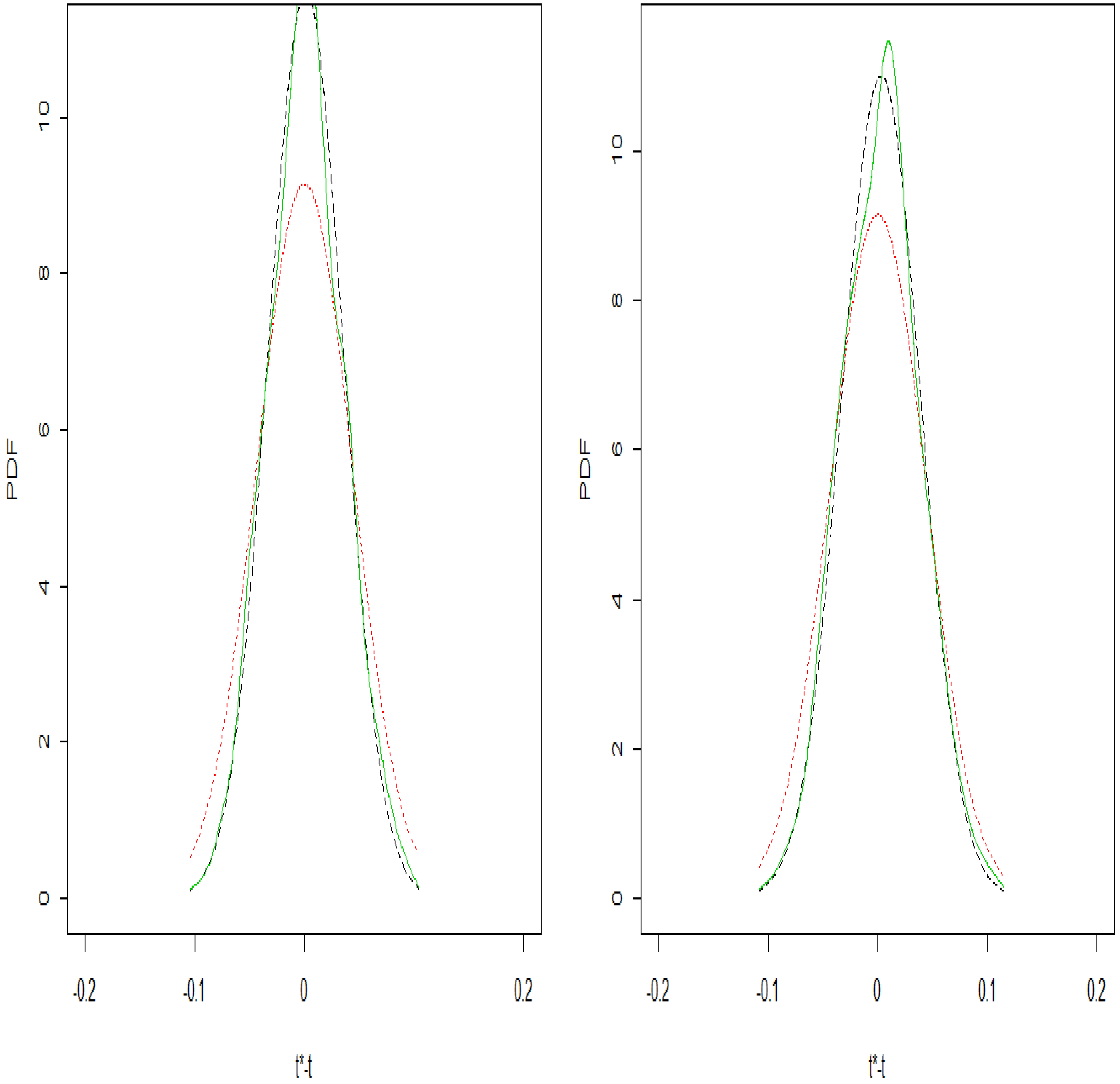

Figure 5: Kernel Density,$N(b, v)$ and $N(0, V_l)$