library(faraway)

data(pima)

d=pima

names(d)

d$diastolic[d$diastolic==0]=NA

d$glucose[d$glucose==0]=NA

d$triceps[d$triceps==0]=NA

d$bmi[d$bmi==0]=NA

d$insulin[d$insulin==0]=NA

Question 1#

par(mfrow=c(1,2))

x=is.na(d$triceps)

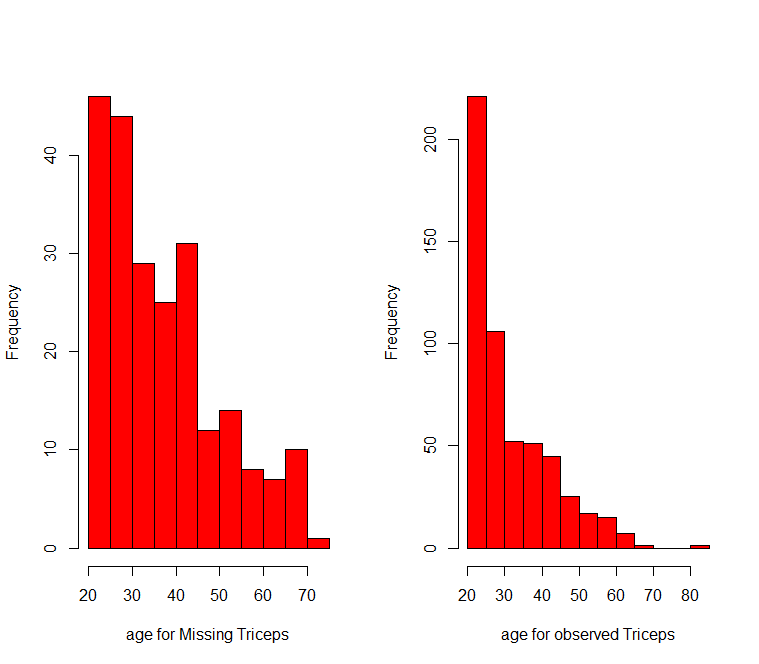
hist(d$age[x==TRUE],xlab="age for Missing Triceps",main="",col="red",border="black")

hist(d$age[x==FALSE],xlab="age for observed Triceps",main="",col="red",border="black")

# the other way to plot hist of age for observed triceps is:

y=is.finite(d$triceps)

#hist(d$age[y],xlab="age for observed Triceps",main="Histogram of age for observed triceps",col="red",border="black")



اطلاعاتی که از نمودار هیستوگرام میتوان دریافت کرد عبارت اند از

توزیع داده ها که ایا نرمال هستند یا چولگی دارد.

داده های پرت

داده با فراوانی بیشتر

پراکندگی داده ها

در نمودار سمت راست، با محاسبه میانگین و انحراف معیار احتمالا میتوان گفت افرادی که سن آن ها بالای 80 سال است داده پرت محسوب میشوند. معمولا 3 برابر انحراف معیار را داده پرت حساب میکنند. پراکندگی داده ها برای افرادی که چین پوستی مشاهده شده دارند بیشتر است و بیشتر افراد در رده سنی 20 تا 25 سال هستند. میانگین داده ها حدود 40 است.

بیشترین فراوانی مربوط به افراد در بازه سنی 20 تا 25 سال است.

در نمودار سمت چپ، بیشترین فراوانی مربوط به افراد در بازه سنی 20 تا 30 سال است و کمترین فراوانی برای افراد با سن حداقل 70 است.

Question 2#

mean\_insulin=mean(d$insulin,na.rm=TRUE)

larger=which(d$insulin>mean\_insulin & is.na(d$bmi)==FALSE)

smaller=which(d$insulin=<mean\_insulin & is.na(d$bmi)==FALSE)

mean(d$bmi[larger])

mean(d$bmi[smaller])

The other way:

Larger= which(d$insulin>mean\_insulin)

smaller=which(d$insulin=<mean\_insulin)

mean(d$bmi[larger],na.rm=TRUE)

mean(d$bmi[smaller],na.rm=TRUE)

#Question 3

# average and standard deviation for insulin attribute for both levels of test attribute.

mean\_insulin\_test\_1=mean(d$insulin[d$test==1],na.rm=TRUE)

mean\_insulin\_test\_1

sd\_insulin\_test\_1=sd(d$insulin[d$test==1],na.rm=TRUE)

sd\_insulin\_test\_1

mean\_insulin\_test\_0=mean(d$insulin[d$test==0],na.rm=T)

mean\_insulin\_test\_0

sd\_insulin\_test\_0=sd(d$insulin[d$test==0],na.rm=T)

sd\_insulin\_test\_0

The other way:

tapply(d$insulin,d$test,mean,na.rm=TRUE)

tapply(d$insulin,d$test,sd,na.rm=TRUE)

CIs\_positive\_sign=function(mean,sd,n){return(mean+1.96\*(sd/sqrt(n)))}

CIs\_negative\_sign=function(mean,sd,n){return(mean-1.96\*(sd/sqrt(n)))}

n\_1=length(which(d$test==1 & is.na(d$insulin)==FALSE))

n\_0=length(which(d$test==0 & is.na(d$insulin)==FALSE))

CIs\_test\_1=c(CIs\_positive\_sign (mean\_insulin\_test\_1,sd\_insulin\_test\_1,n\_1), CIs\_negative\_sign (mean\_insulin\_test\_1,sd\_insulin\_test\_1,n\_1))

CIs\_test\_0=c(CIs\_positive\_sign (mean\_insulin\_test\_0,sd\_insulin\_test\_0,n\_0), CIs\_negative\_sign (mean\_insulin\_test\_0,sd\_insulin\_test\_0,n\_0))

c("test1",CIspositive\_test\_1,"test2",CIspositive\_test\_0)

بازه ها اشتراک ندارند و به این معناست که اگر انسولین فردی در بازه مربوط به CIs\_test\_1 باشد با اطمینان 95 درصد میتوان گفت دارای دیابت است و در بازه درگیر میتوان گفت فاقد دیابت است.