**source**("http://www.openintro.org/stat/data/cdc.R")

**head**(cdc)

A screenshot of a cell phone

Description automatically generated

**Investigate the file**

The file contains information of 20000 individuals related to their general health which is divided into 5 categories (good, very goood, excellent, fair, poor). The second column “exerany” contains binary/categorical values (0/1). Same is the case for “hlthplan” and “smoke100”. The height, weight, desired weight and age are the continious data.The column gender is the nominal data as it represents the individual as male(m) or female(f)

**Describe the object**

1 How many parameters *(*The columns are parameters)

**ncol**(cdc)

9 parameters are there

2 How many cases? *(The rows are the cases)*

**nrow**(cdc)

20000 cases

*We can also check dimension to get the answer for above two questions*

**dim**(cdc)

## [1] 20000 9

3. How many Males?

**sum**(cdc**$**gender**==** m)

9569 males

3. How many Females?

**sum**(cdc**$**gender**==**f)

10431 females

4. How many older than 25? 35?

**sum**(cdc**$**age**>**25)

17272 older than 25

**sum**(cdc**$**age**>**35)

13216 older than 35

5. Show the distribution of age as a histogram

hist(cdc$age, xlab = "age", main ="Histogram" )

A screenshot of a cell phone

Description automatically generated

***Describe one continuous variable:***

We are choosing “Height”

summary((cdc$height))

Min. 1st Qu. Median Mean 3rd Qu. Max.

48.00 64.00 67.00 67.18 70.00 93.00

*Mean*

*mean(cdc$height)*

67.1829

Median

*median(cdc$height)*

67

Mode

getmode <- function(cdc$height) {

uniqv <- unique(cdc$height)

uniqv[which.max(tabulate(match(cdc$height, uniqv)))]

}

result <- getmode(cdc$height)

print(result)

Mode = 66

Variance

var((cdc$height))

17.0235

Standard Deviation

sd(cdc$height)

4.125954

Standard Error

sd(cdc$height)/sqrt(length(cdc$height))

0.0291749

IQR

IQR(cdc$height)

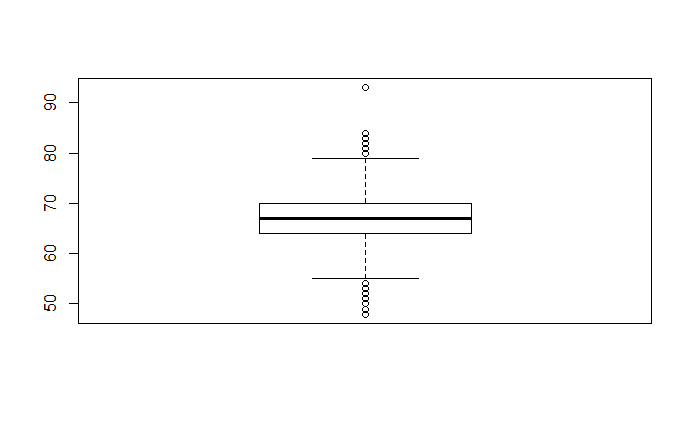
6

Show the data as Histogram

A screenshot of a cell phone

Description automatically generated

Show the data as Boxplot



*Show the data as Histogram (for women only)*

newdata <- cdc[which(cdc$gender=="f" ),]

head(newdata)

h<-hist(newdata$height)

A screenshot of a cell phone

Description automatically generated

*Show the data as Boxplot(for women)*

boxplot(newdata$height)

A screenshot of a cell phone

Description automatically generated

Describe one categorical variable

We will be describing “Gender”

Frequency table

table(cdc$gender)

m f

9569 10431

Depict the frequency table using a barplot

*barplot(table(cdc$gender)*

*,main = "Gender counts", xlab = "Gender", ylab = "Frequency")*

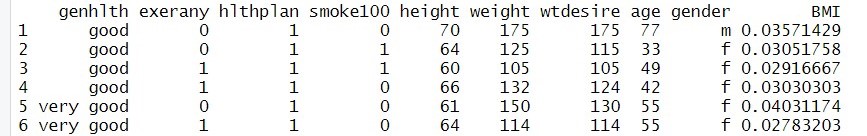
*A screenshot of a cell phone

Description automatically generated*

Computing BMI (Weight/height\*height)

cdc$BMI<-cdc$weight/(cdc$height\*cdc$height)

head(cdc)



Subsetting data into Men and Women

men<-cdc[which(cdc$gender=='m'), ]

head(men)

MEN

A screenshot of a cell phone

Description automatically generated

women<-cdc[which(cdc$gender=='f'), ]

head(women)

WOMEN

A screenshot of a cell phone

Description automatically generated

*Create two histograms for both datasets*

*men<-cdc[which(cdc$gender=='m'), ]*

*head(men)*

*hist(men$BMI,xlab ="BMI", main = " Men's Histogram") A screenshot of a cell phone

Description automatically generated*

WOMEN’s histogram

women<-cdc[which(cdc$gender=='f'), ]

head(women)

hist(women$BMI,xlab ="BMI", main = "Women's Histogram")

A screenshot of a cell phone

Description automatically generated

Compare the means between men and women form these datasets

Compare Means

summary(men)A close up of text on a white background

Description automatically generated

summary(women)

A close up of text on a white background

Description automatically generated

Mean BMI in men is 0.03830

Mean BMI in women is 0.03662