STAT 341 Assignment 1

Mushi Wang 20732874

Due Friday September 27 at 9:00am

Note

- Replace "Student Name and ID" with your name and waterloo ID.
- Using RMarkdown or LaTeX is required and no hand-written and/or imported screenshots will be accepted in the assignments. A mark of 0% will be assigned to the questions which were not complied in RMarkdown or LaTeX, and/or those which include hand-written solutions and/or screenshots.
- Organization is part of a full solution. Full marks will be awarded to organized complete solutions and marks will be deducted for unorganized solutions.

Wayne Gretzky Goals

- Wayne Gretzky "The Great One" is a Canadian former professional ice hockey player. He played 20 seasons in the National Hockey League (NHL) and he is considered to be the greatest hockey player ever. The dataset "GretzkyGoals.csv" contains all of Gretzky's goals during his time in the NHL. Here, we will examine the times at which the goals occurred during a sixty-minute game.
- **Note**:
  - For each part below, any plots should be side-by-side in the same figure and they all should be properly labelled.

a) **[5 Marks]** Read-in the data and convert the times into seconds. Remove the overtime goals, which are any goals that occur beyond sixty minute mark of regular play. Then calculate average, median and range for the times Wayne scored goals during a game.

```
goals.results <- read.csv('GretzkyGoals.csv', header = TRUE)
N = nrow(goals.results)
time = numeric(1)
j = 1
for(i in 1 : N) {
  period.char = goals.results$Per.[i]
  if(period.char == "OT"){
    next
  }
  time.char = as.character(goals.results$Time[i])
  if(nchar(time.char) == 5) {
    time.char = c(period.char, substr(time.char,1,2), substr(time.char,4,6))
  } else {
    time.char = c(period.char, substr(time.char,1,1), substr(time.char,3,5))
  }

  time.num = as.numeric(time.char)
  time.num[1] = time.num[1] - 1
  time[j] = sum(time.num*c(1200,60,1))
  j = j + 1
}
mean(time)
```

```
## [1] 1996.215
```

```
median(time)
```

```
## [1] 2028
```

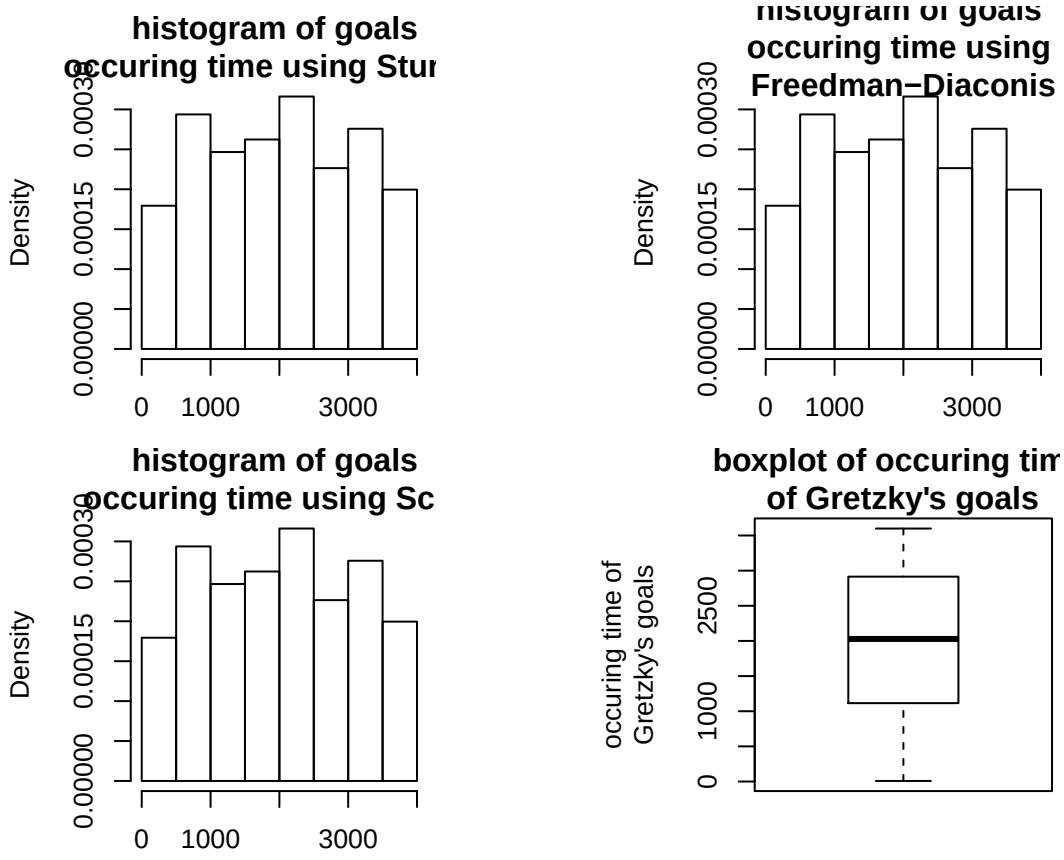
```
range(time)
```

```
## [1]    8 3599
```

The mean is 1996.215 The median is 2028 The range is 8 to 3599

b) **[5 Marks]** Plot three histograms using Sturges, Scott and Freedman-Diaconis rules for the number of bins along with a boxplot. All four plots should be side-by-side in the same figure and they all should be properly labelled. From these three histograms and boxplot does Wayne tend to score at any particular time during the game?

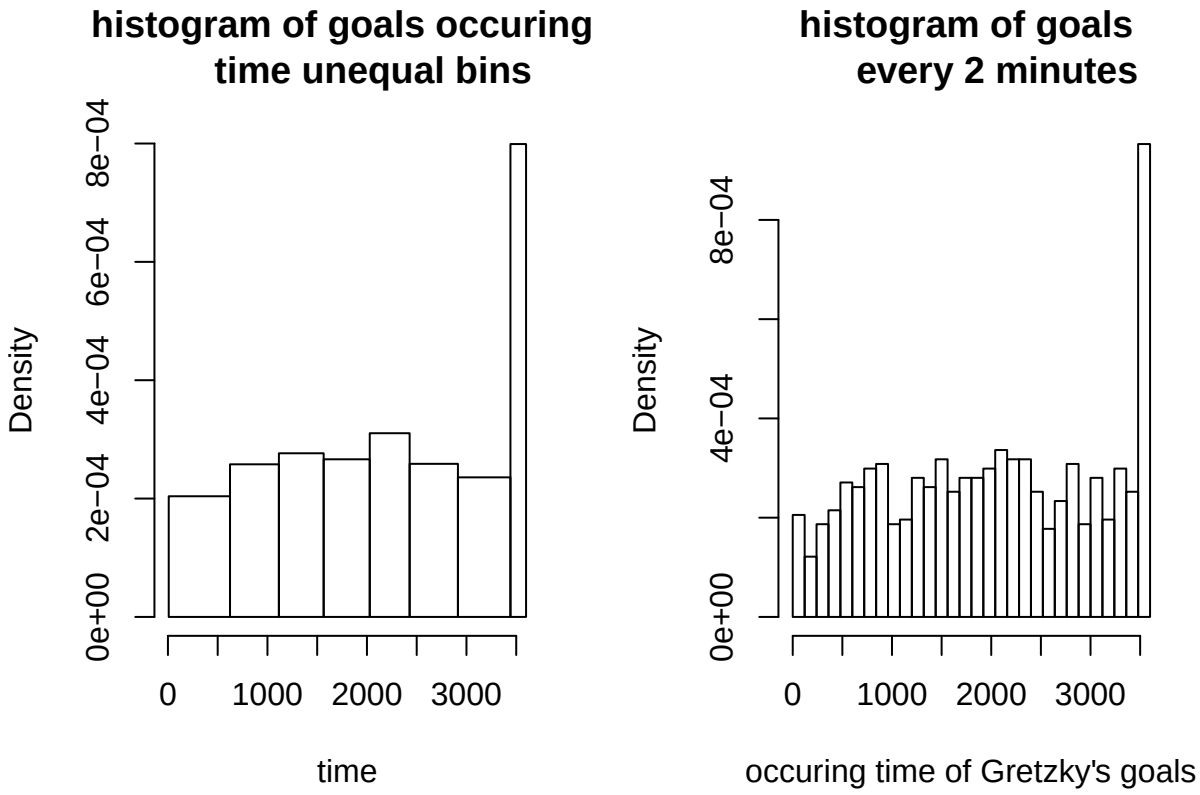
```
par(mfrow = c(2, 2), mar=2.5*c(1,4,1,0.1))
hist(time, prob=TRUE, xlab="occuring time of Gretzky's goals", main="histogram of goals
occuring time using Sturges")
hist(time, breaks="FD", prob=TRUE, xlab="occuring time of Gretzky's goals", main="histogram of goals
occuring time using
Freedman-Diaconis")
hist(time, breaks="scott", prob=TRUE, xlab="occuring time of Gretzky's goals", main="histogram of goals
occuring time using Scott")
boxplot(time, main="boxplot of occuring time
of Gretzky's goals", ylab="occuring time of
Gretzky's goals")
```



Wayne has no tendency to score at at any particular time during the game

c) **[5 Marks]** Construct two histograms, one using unequal bins (using the same number of bins used by part b) and another that the breaks the 60 minute game in two minutes interval. From these two histograms does Wayne tend to score at any particular time during the game?

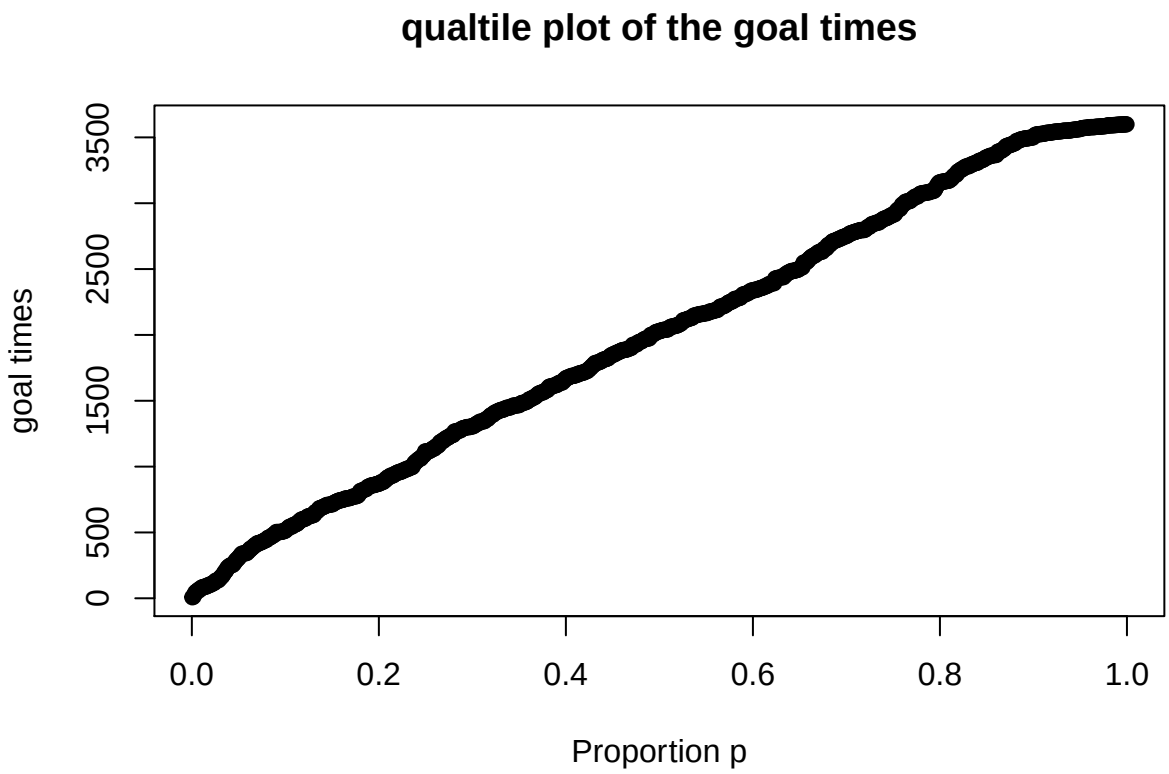
```
par(mfrow = c(1, 2))
hist(time, breaks=quantile(time, p=seq(0, 1, length.out=9)), prob=TRUE, main="histogram of goals occuri
     time unequal bins" )
hist(time, breaks=seq(0, 3600, length.out=31), prob=TRUE, xlab="occuring time of Gretzky's goals", main=
     every 2 minutes")
```



Wayne tended to score at last mintues of games.

d) **[3 Marks]** Construct a quantile plot of the goal times. What feature does the quantile plot exhibit?

```
qvals <- sort(time)
pvals <- ppoints(length(qvals))
plot(pvals, qvals, pch = 19,
     xlim=c(0,1),
     xlab = "Proportion p",
     ylab = "goal times",
     main = "qualtile plot of the goal times")
```

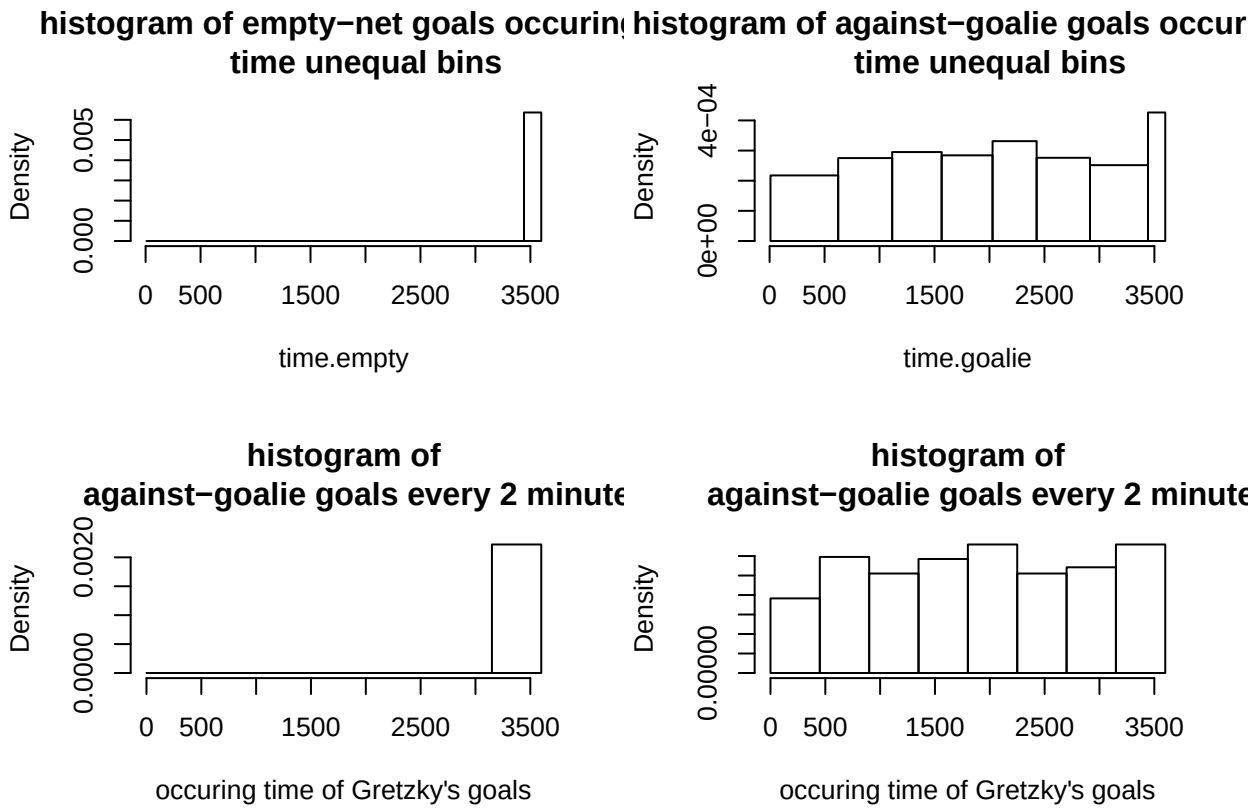


The graph indicates a linear relationship. However the tail shows that Wayne scored a little more at last minutes of games. In conclusion Wayne has a little tendency to score more goals at the end of games.

e) **[5 Marks]** Partition the goal times into empty-net goals and against-goalie goals. Construct a histogram using the same number of bins used by part b) and using varying bins widths for each group. Comment on the differences among the groups?

```
time.empty = numeric(1)
time.goalie = numeric(1)
j = 1;
k = 1;
m = 1;
for(i in 1:N) {
  period.char = goals.results$Per.[i]
  if(period.char == "OT"){
    next
  }
  if(goals.results$Goalie[i] == "") {
    time.empty[j] = time[m]
    j = j + 1
  } else {
    time.goalie[k] = time[m]
    k = k + 1
  }
  m = m + 1
}

par(mfrow = c(2, 2))
```

```r
hist(time.empty, breaks=quantile(time, p=seq(0, 1, length.out=9)), prob=TRUE, main="histogram of empty-n
     time unequal bins" )
hist(time.goalie, breaks=quantile(time, p=seq(0, 1, length.out=9)), prob=TRUE, main="histogram of agains
     time unequal bins" )
hist(time.empty, breaks=seq(0, 3600, length.out=9), prob=TRUE, xlab="occuring time of Gretzky's goals",
     against-goalie goals every 2 minutes")
hist(time.goalie, breaks=seq(0, 3600, length.out=9), prob=TRUE, xlab="occuring time of Gretzky's goals"
     against-goalie goals every 2 minutes")
```



The empty-net goals only occured at the end of certain games, since empty net only occurs in last mintues. However, against-goalie goal occured reasonablely evenly through out the games.

World Health Organization (WHO) on life expectancy

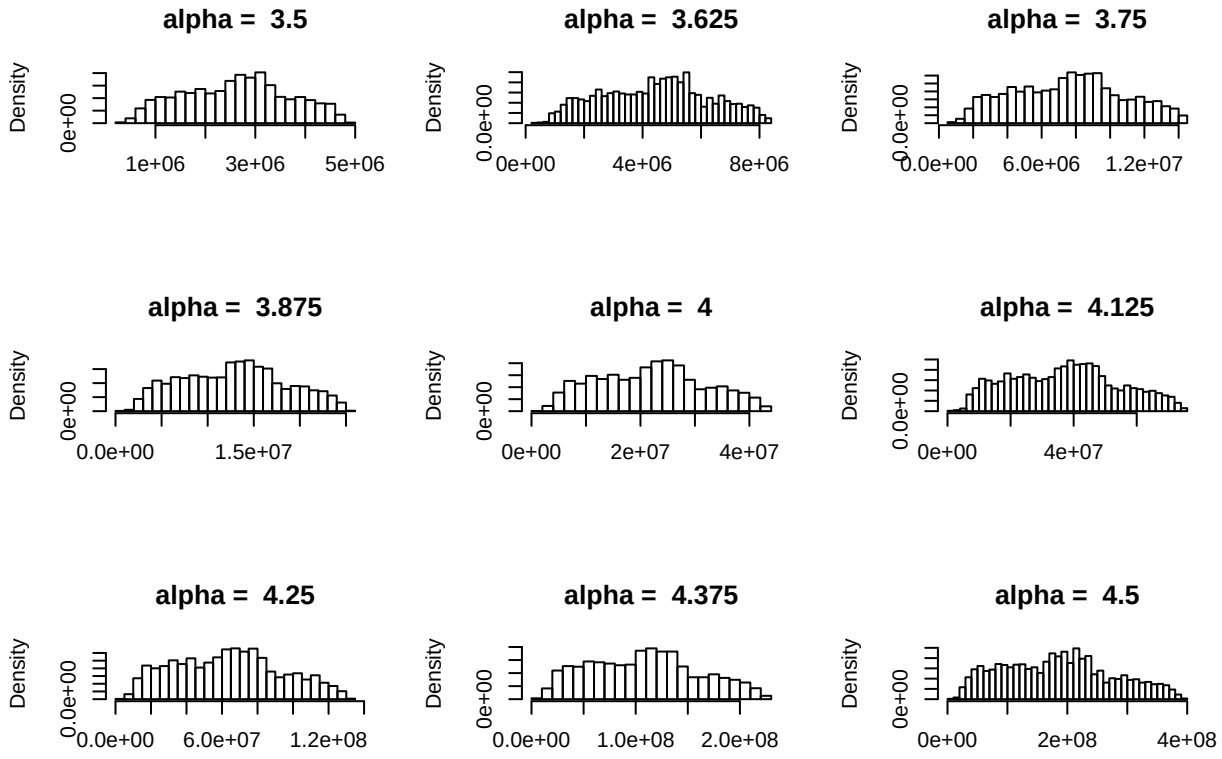
In this question you will be analyzing data for WHO on life expectancy. The data is in the file "WHO_life.csv" posted on LEARN. Below is the `powerfun` from the course notes for your convenience.

```r
powerfun <- function(x, alpha) {
  if(sum(x <= 0, na.rm=TRUE) > 0) stop("x must be positive")
  if (alpha == 0)
    log(x)
  else if (alpha > 0) {
    x^alpha
  } else -x^alpha
}
```

- The variables are `Country`, `Year`
  - `LB.XXXX` the life expectancy at birth (years) for Males, Females & Both, and
  - `L60.XXXX` Life expectancy at age 60 (years) for Males, Females & Both.

a) **[3 Marks]** What range of powers (the values of α) make the distribution of the life expectancy at birth (years) for males symmetric?

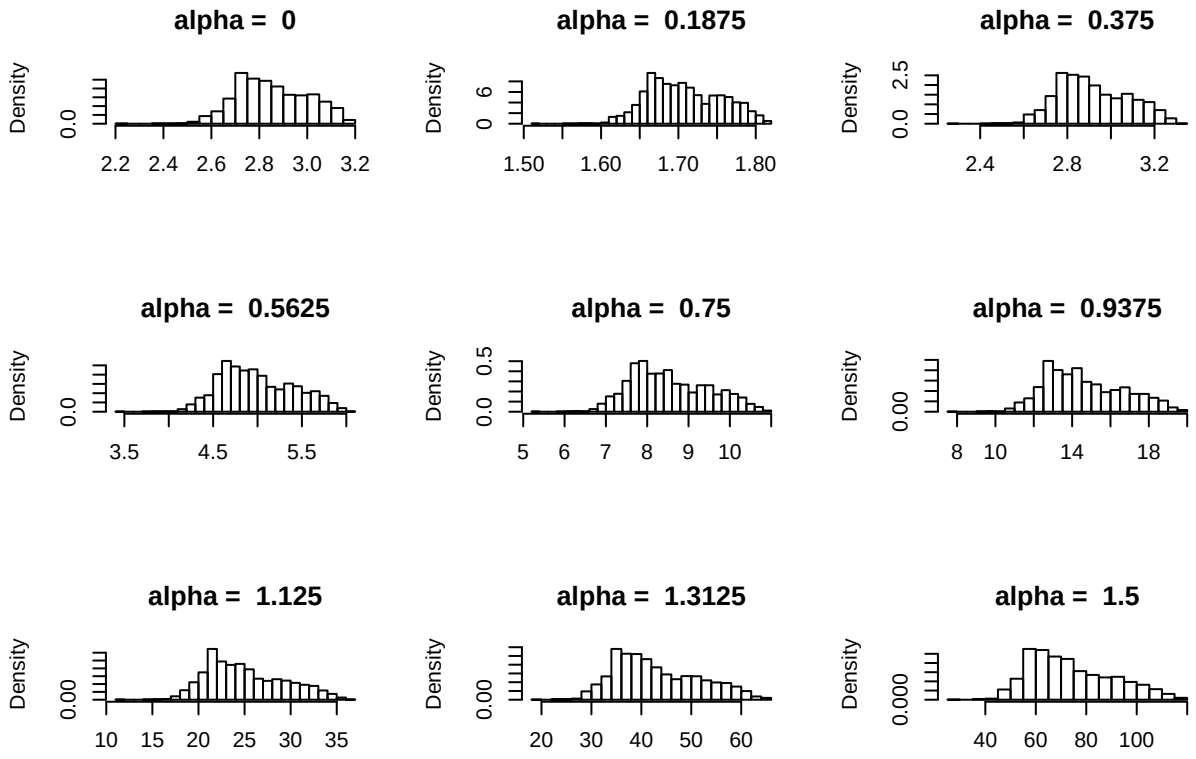
```
par(mfrow=c(3,3))
who.results <- read.csv('WHO_life.csv', header = TRUE)
a = seq(3.5, 4.5, length.out=9)

for(i in 1:9) {
  hist(powerfun(who.results$LB.Male, a[i]), prob=TRUE, main=paste("alpha = ", a[i]), xlab = "", breaks =
}
```



The range of α is [3.5, 4]

b) **[3 Marks]** What range of powers (the values of α) make the distribution of the life expectancy at age 60 (years) for males symmetric?

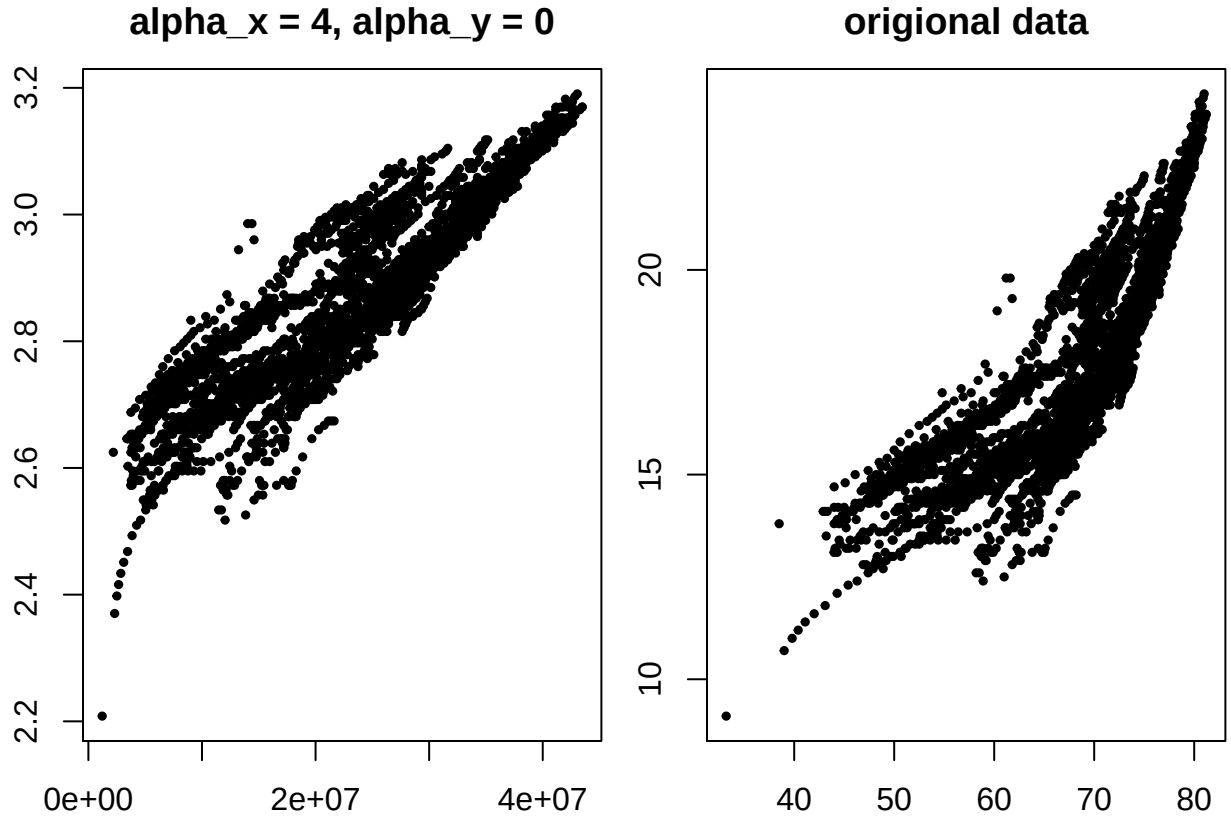
```
par(mfrow=c(3,3))
who.results <- read.csv('WHO_life.csv', header = TRUE)
a = seq(0, 1.5, length.out=9)

for(i in 1:9) {
  hist(powerfun(who.results$L60.Male, a[i]), prob=TRUE, main=paste("alpha = ", a[i]), xlab = "", breaks
}
```



The range of α is [0.25, 1.125]

c) **[3 Marks]** Using $\alpha = 4$ as the power for `x = LB.Male` and $\alpha = 0$ as the power for `y = L60.Male`, plot the transformed variables. Between the transformed and original data, which one is better-suited for linear modeling?

```
par(mfrow=c(1, 2), mar=2.5*c(1,1,1,0.1))

plot( powerfun(who.results$LB.Male, 4), powerfun(who.results$L60.Male, 0), pch = 19, cex=0.5, xlab = ""
     main = "alpha_x = 4, alpha_y = 0")
plot( who.results$LB.Male, who.results$L60.Male, pch = 19, cex=0.5, xlab = "", ylab = "",
     main = "origional data")
```



The transformed data is better-suited for linear modeling

Investigating influence and sensitivity of the geometric mean

a) The geometric mean for the population $\mathcal{P}) = \{y_1, \ldots, y_N\}$ is

$$a(\mathcal{P}) = a(y_1, \ldots, y_N) = \left(\prod_{i=1}^{N} y_i\right)^{1/N}$$

i) **[3 Marks]** Derive the sensitivity curve for the geometric mean and write it as a function y and $a(\mathcal{P})$

$$\begin{aligned} SC(y : a(\surd)) &= N(\left(y \prod_{i=1}^{N-1} y_i\right)^{1/N} - \left(\prod_{i=1}^{N-1} y_i\right)^{1/(N-1)}) \\ &= N(y^{\frac{1}{N}} a(p)^{\frac{N-1}{N}} - a(p)) \end{aligned}$$

ii) **[3 Marks]** Write the infuence of the geometric mean as function of y_u and $a(\mathcal{P})$.

$$\begin{aligned}\Delta(a,u) &= (\prod_{i=1}^{n} y_i)^{\frac{1}{N}} - (\prod_{i=1 \text{ and } i\neq u}^{n} y_i)^{\frac{1}{N-1}} \\ &= a(\mathcal{P}) - (\frac{a(\mathcal{P})^N}{y_u})^{\frac{1}{N-1}} \\ &= a(\mathcal{P}) - \frac{a(\mathcal{P})^{\frac{N}{N-1}}}{y_u{}^{\frac{1}{N-1}}} \\ &= a(\mathcal{P})(1 - \frac{a(\mathcal{P})}{y_u{}^{\frac{1}{N-1}}})\end{aligned}$$

b) The measure of sensitivity and influence does not have to depend on the difference between the attribute values. Instead we might define the **sensitivity-ratio** (SR) for non-negative attributes as

$$SR\,(y;\alpha(\mathcal{P})) = \left[\frac{\alpha(y_1,...,y_{N-1},y)}{\alpha(y_1,...,y_{N-1})}\right]^N$$

i) **[3 Marks]** Derive the SR for the geometric mean as function of y and $a(\mathcal{P})$

$$\begin{aligned}SR\,(y;\alpha(\mathcal{P})) &= \left(\frac{(y\prod_{i=1}^{N-1} y_i)^{1/N}}{(\prod_{i=1}^{N-1} y_i)^{1/(N-1)}}\right)^N \\ &= \frac{y\prod_{i=1}^{N} y_i}{(\prod_{i=1}^{N-1} y_i)^{N/(N-1)}} \\ &= \frac{\prod_{i=1}^{N} y_i}{(\prod_{i=1}^{N-1} y_i)(\prod_{i=1}^{N-1} y_i)^{1/(N-1)}} \\ &= \frac{y}{(\prod_{i=1}^{N-1} y_i)^{1/(N-1)}} \\ &= \frac{y}{a(\mathcal{P})}\end{aligned}$$

ii) **[3 Marks]** A measure of influence can be constructed with the ratio as well. Here we define the **influence-ratio** (IR) for non-negative attribute as

$$IR\,(a,u) = \left[\frac{a(\mathcal{P})}{a(y_1,\ldots,y_{u-1},,y_{u+1},y_N)}\right]^N$$

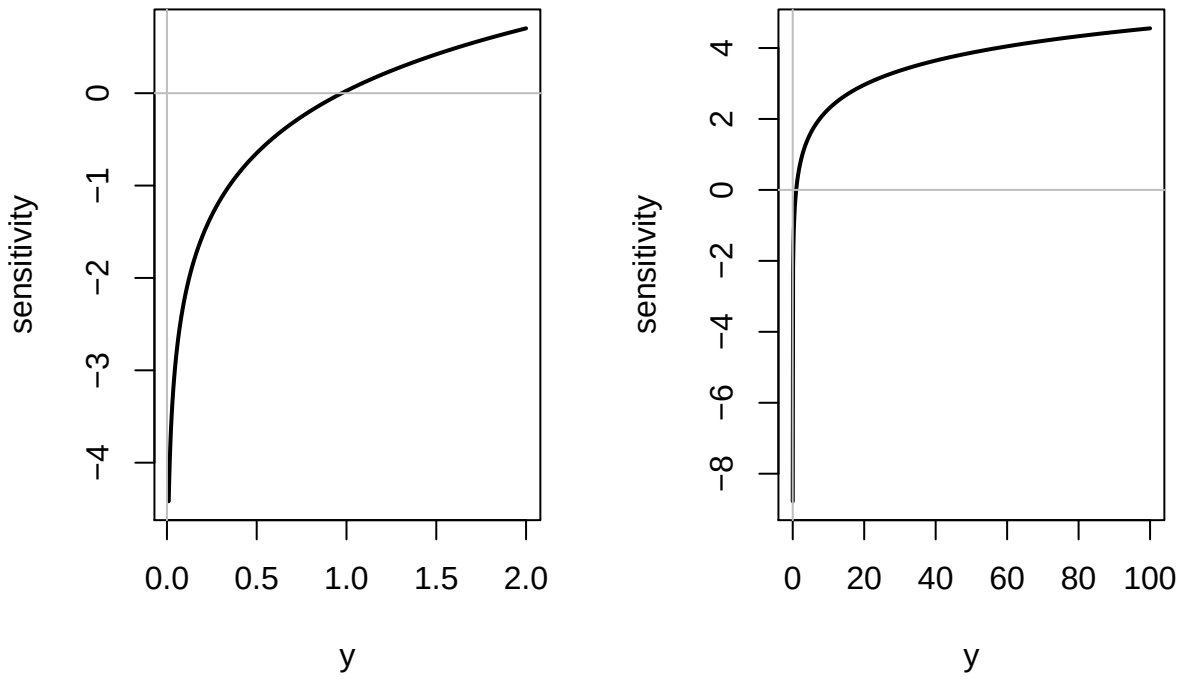
- Derive the influence-ratio for the geometric mean as function of y_u and $a(\mathcal{P})$

$$\begin{aligned}IR\,(a,u) &= \left(\frac{(\prod_{i=1}^{N} y_i)^{1/N}}{(\prod_{i=1 \text{ and } i\neq u}^{N-1} y_i)^{1/(N-1)}}\right)^N \\ &= \frac{a(\mathcal{P})^N}{(\frac{a(\mathcal{P})^N}{y_u})^{\frac{N}{N-1}}} \\ &= \frac{1}{(\frac{a(\mathcal{P})}{y_u})^{\frac{N}{N-1}}} \\ &= \frac{y_u^{\frac{N}{N-1}}}{a(\mathcal{P})^{\frac{N}{N-1}}}\end{aligned}$$

c) The population provided in *returns2.txt* is the monthly returns of an investment over a period of 20 years.

i) **[2 Marks]** Plot the sensitivity curve (SC) of the geometric mean for this population over the ranges $[0.01, 2]$ & $[0.0001, 100]$. No comments required.

```
returns2 = read.csv("returns2.csv")

N = length(returns2)
data = returns2$returns

geo.mean = function(x){
  if(any(x < 0)){
    return('All variate values must be positive!')
  }
  return( (prod(x)) ^ (1/length(x)) )
}

sc = function(y.pop, y, attr, ...) {
  N <- length(y.pop) + 1
  Map(function(y) {  N*(attr(c(y,y.pop),...) - attr(y.pop,...))} ,y   )
}

y1 <- seq(0.01, 2, length.out=1000)
y2 <- seq(0.0001,100, length.out=1000)

par(mfrow=c(1,2))
plot(y1, sc(data, y1, geo.mean), type="l", lwd = 2,
     main="Sensitivity curve for Geometric Mean in [0.01,2]",
     xlab='y' , ylab="sensitivity")
abline(h=0, v=0, col="grey")

plot(y2, sc(data, y2, geo.mean), type="l", lwd = 2,
     main="Sensitivity curve for Geometric Mean in [0.0001, 100]",
     xlab='y' , ylab="sensitivity")
abline(h=0, v=0, col="grey")
```

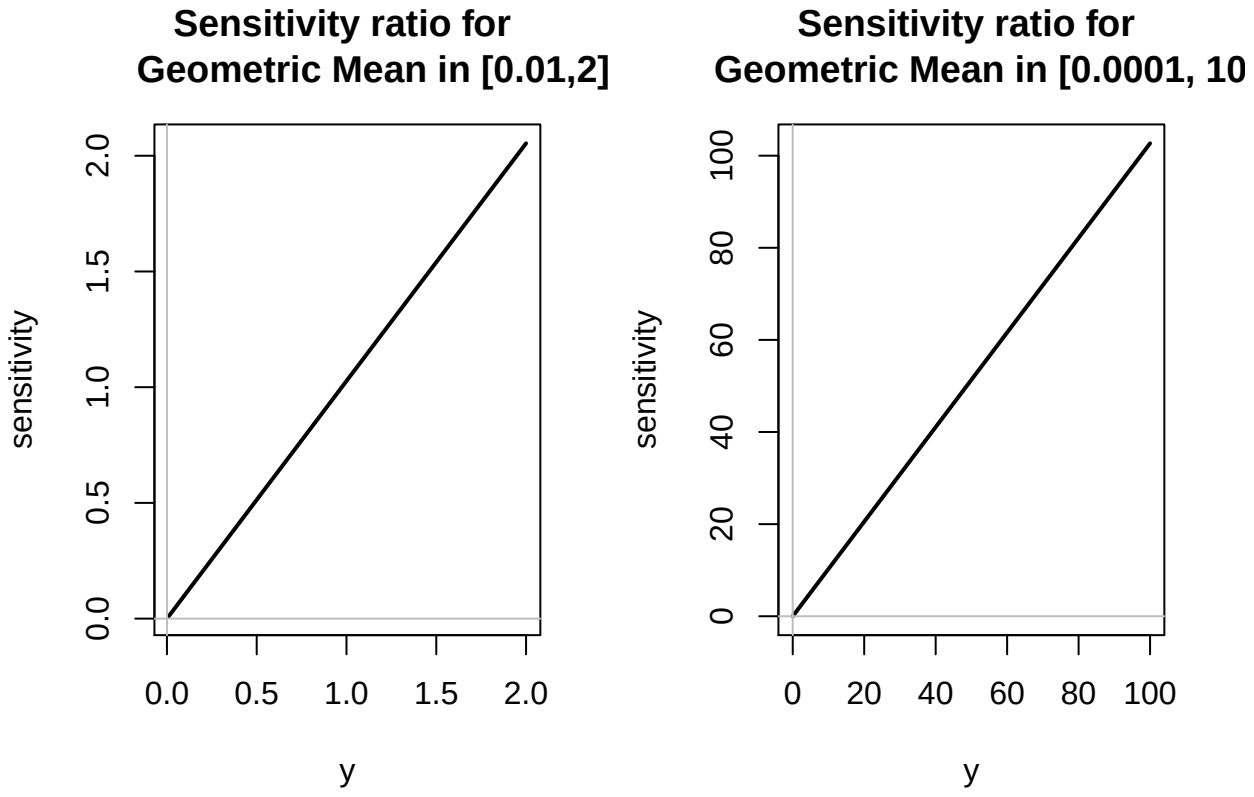


ii) **[2 Marks]** Write a function similar to `sc` called `sr` such that the
- inputs are a population `y.pop`, a sequence or vector of y values `y` and an attribute function `a
- ouput is the **sentivity-ratio** for each y value.

```
sr = function(y.pop, y, attr, ...) {
  N <- length(y.pop) + 1
  Map(function(y) {  y / attr(y.pop,...)} ,y)
}
```

iii) **[3 Marks]** Plot the sensitivity-ratio (SR) of the geometric mean for this population over the ra

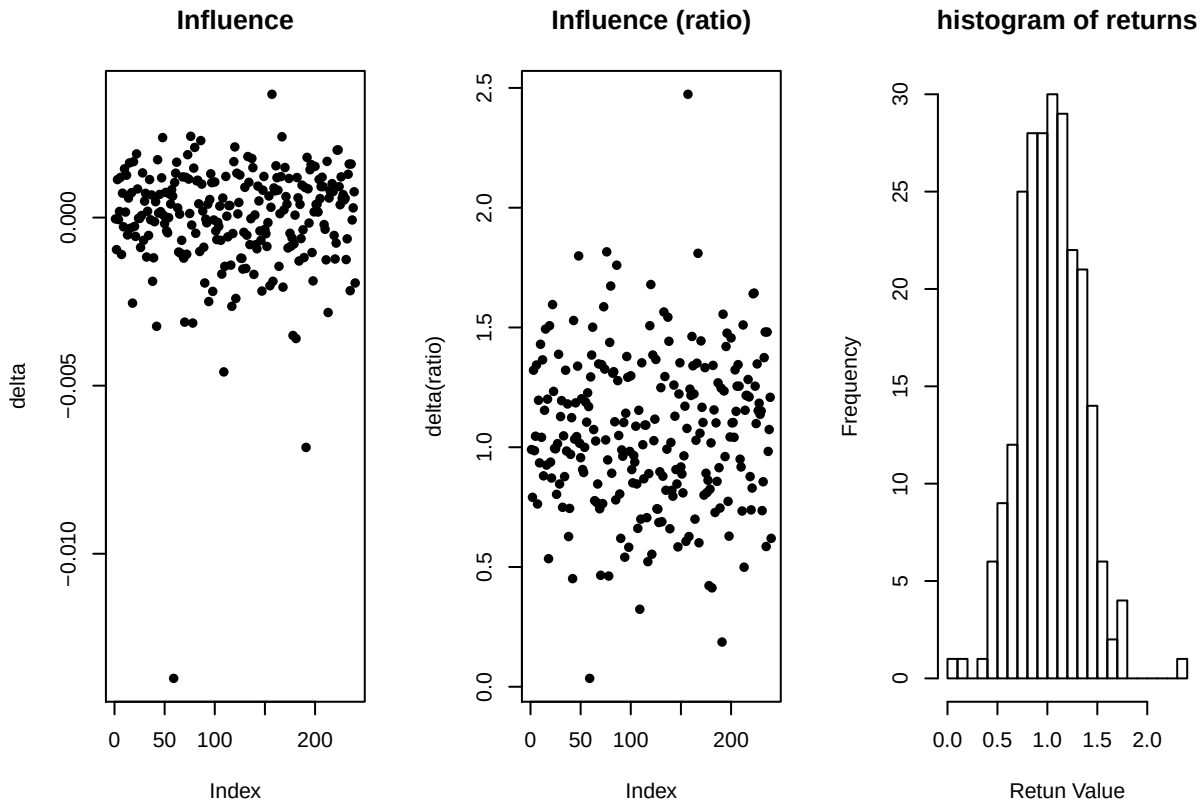
```
par(mfrow=c(1,2))
plot(y1, sr(data, y1, geo.mean), type="l", lwd = 2,
     main="Sensitivity ratio for
     Geometric Mean in [0.01,2]",
     xlab='y' , ylab="sensitivity")
abline(h=0, v=0, col="grey")

plot(y2, sr(data, y2, geo.mean), type="l", lwd = 2,
     main="Sensitivity ratio for
     Geometric Mean in [0.0001, 100]",
     xlab='y' , ylab="sensitivity")
abline(h=0, v=0, col="grey")
```



The sensitivity curves get higher without bound as $y- > \infty$. Thus a single obersivation can change geometric mean by huge amount. Geometric mean may not be a good attribute to represent the location of a population when extreme value exists.

d) **[4 Marks]** Using the same population, plot the influence values from a) ii) and b) ii) for the geometric mean and a histogram of the data, and comment on the plots. Use Freedman-Diaconis rule for the number of bins. Comment on the influential observations based on each measure.

```
N = length(data)
delta = prod(data)^(1/N) - (prod(data)/data)^(1/(N-1))
ratio = (prod(data)^(1/N) / (prod(data)/data)^(1/(N-1)))^N

par(mfrow=c(1,3))
plot(delta, main="Influence",pch=16 )
plot(ratio, ylab  = "delta(ratio)", main="Influence (ratio)",pch=16 )
hist(data, main="histogram of returns", xlab="Retun Value", breaks='FD')
```



There are 3 or 4 values that has relatively high influence than the other. Such as the points with delta value less than -0.005 and the point with delta greater than 2 in the second graph.