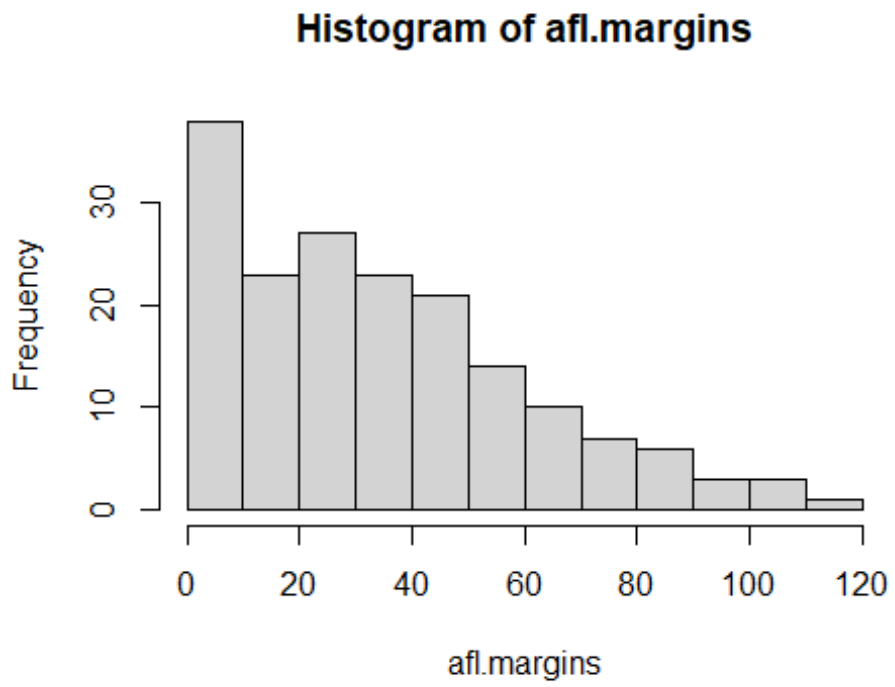
Chapter 5 Measures of Central tendency

Aldo M

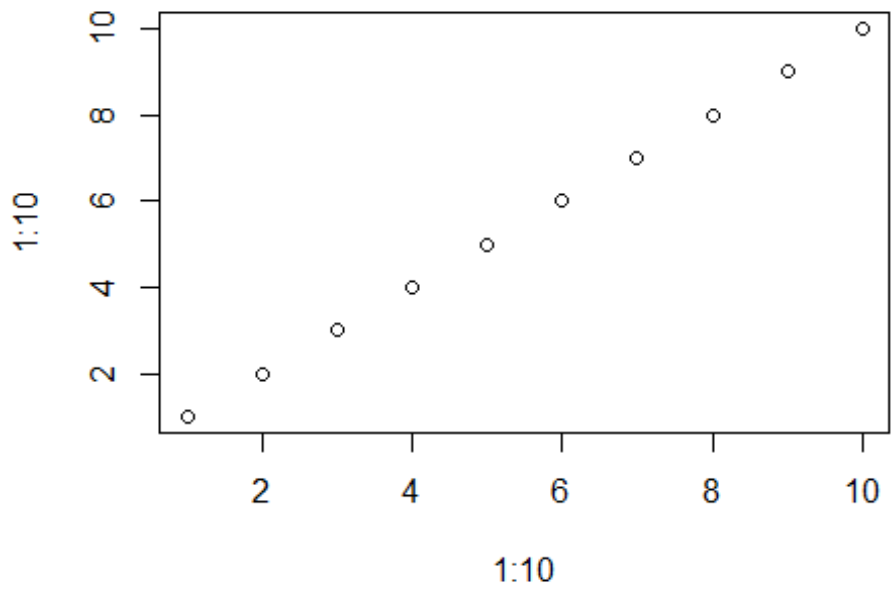
2025-06-18

```
# descriptive statistics chapter 5

load( "C:/Users/aldom/Msc Data Science/R_Master_Folder/data/aflsmall.Rdata" )
library(lsr)
who()

##    -- Name --      -- Class --   -- Size --
##    afl.finalists   factor        400
##    afl.margins     numeric       176

# this file consist in two variables of the Australian Football League (AFL)

# afl.margins = winning (number of points)
# afl.finalists (names of the 400 teams)

print(afl.margins)

##   [1]  56  31  56   8  32  14  36  56  19   1   3 104  43  44  72   9  28
25
##  [19]  27  55  20  16  16   7  23  40  48  64  22  55  95  15  49  52  50
10
##  [37]  65  12  39  36   3  26  23  20  43 108  53  38   4   8   3  13  66
67
##  [55]  50  61  36  38  29   9  81   3  26  12  36  37  70   1  35  12  50
35
##  [73]   9  54  47   8  47   2  29  61  38  41  23  24   1   9  11  10  29
47
##  [91]  71  38  49  65  18   0  16   9  19  36  60  24  25  44  55   3  57
83
## [109]  84  35   4  35  26  22   2  14  19  30  19  68  11  75  48  32  36
39
## [127]  50  11   0  63  82  26   3  82  73  19  33  48   8  10  53  20  71
75
## [145]  76  54  44   5  22  94  29   8  98   9  89   1 101   7  21  52  42
21
## [163] 116   3  44  29  27  16   6  44   3  28  38  29  10  10

#starting descriptive statistics with an histogram (helps to get sense of
what the data look like)

hist(afl.margins)
```

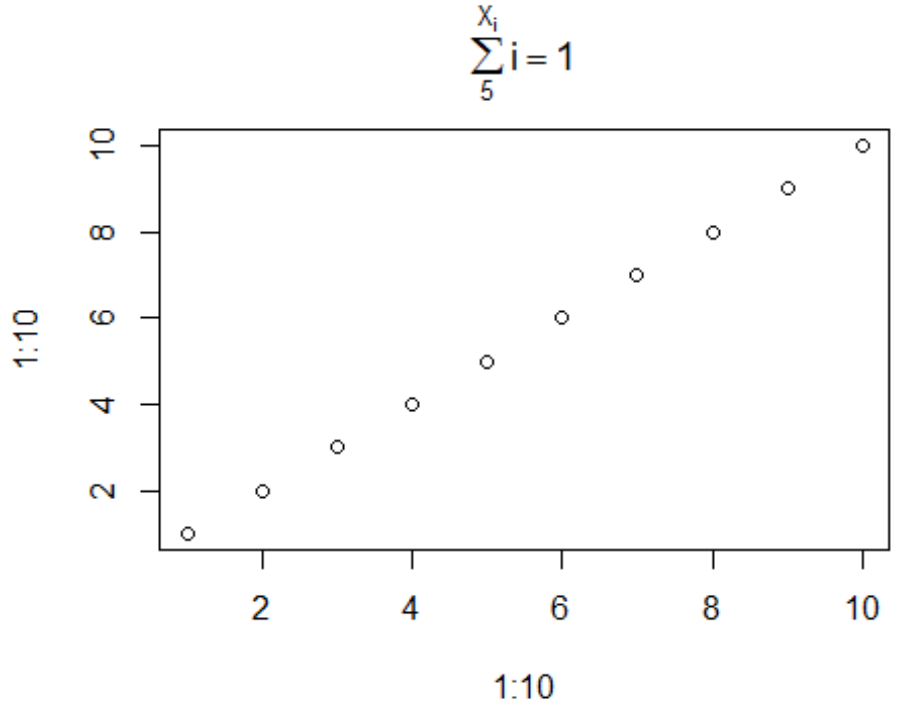



```r
# Mesure of central tendency (understand the average, middle of where the
data is)

# Mean = average | add all the values and divide by the total number of
values

# First piece of notation is N (number of observations)

# X is the traditional label of the observations X1 (first observation),
X2(second observation),
# Xn (last observation)

# Xi (i-th observations)

# Sample mean formula:
# X̄ = (X₁ + X₂ + ... + X_N) / N

plot(1:10, 1:10, main=expression(bar(X) == (X[1] + X[2] + "..." + X[N])/N))
```

$$\overline{X} = (X_1 + X_2 + \ldots + X_N) / N$$

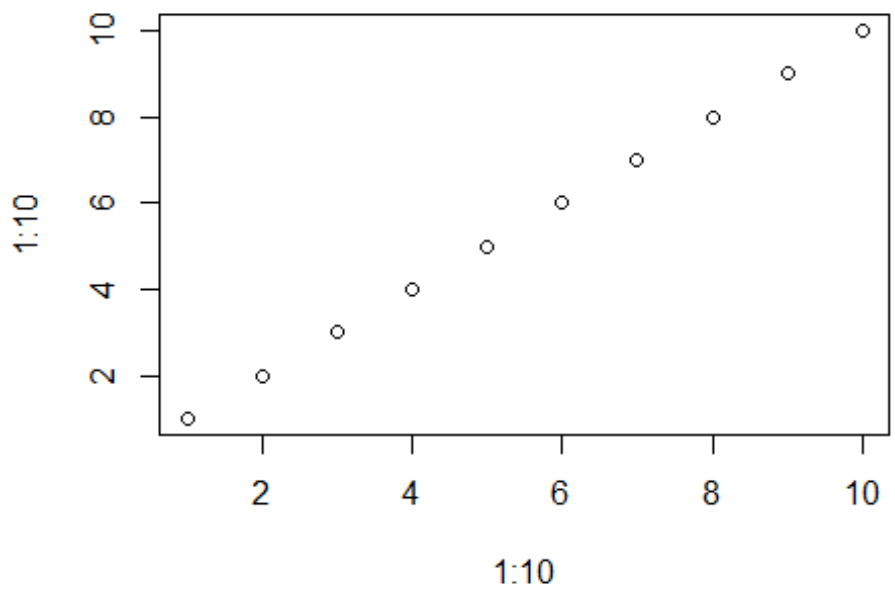



```r
# this formula in short will be with the use of the summation symbol ∑
# Summation notation:
# sum from i=1 to 5 of X_i
# LaTeX: \sum_{i=1}^{5} X_i

plot(1:10, 1:10, main=expression(sum(i==1, 5, X[i])))
```




```r
# in this case i=5 as the book exercise

# the sum add up all the observations (X1+X2+X3+X4+X5)


# Formula for the sample mean:
# X̄ = (1/N) * sum from i=1 to N of X_i
# LaTeX version:
# \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i

plot(1:10, 1:10, main=expression(bar(X) == (1/N) * sum(i==1, N, X[i])))
```

$$\overline{X} = (1/N)\sum_{N}^{X_i} i = 1$$




```
# Now I will be calculating the mean in R

(56 + 31 + 56 + 8 + 32) / 5 # simple form

## [1] 36.6

sum( afl.margins ) #large obervations real word scenarios

## [1] 6213

sum( afl.margins[1:5] ) #sum of the first five observations

## [1] 183

sum( afl.margins[1:5] ) / 5 #mean calculation

## [1] 36.6

mean( x = afl.margins ) # mean calculation

## [1] 35.30114

# The Median calculation (second mesure of central tendency) - the middle
value

# 8,31,32,56,56 --- 32 is the middle value

# 8,14,31,32,56,56 --- 31 and 32 are the middle values (avergae is 31.5)
```

```r
#what happens within R?

sort( x = afl.margins )
```

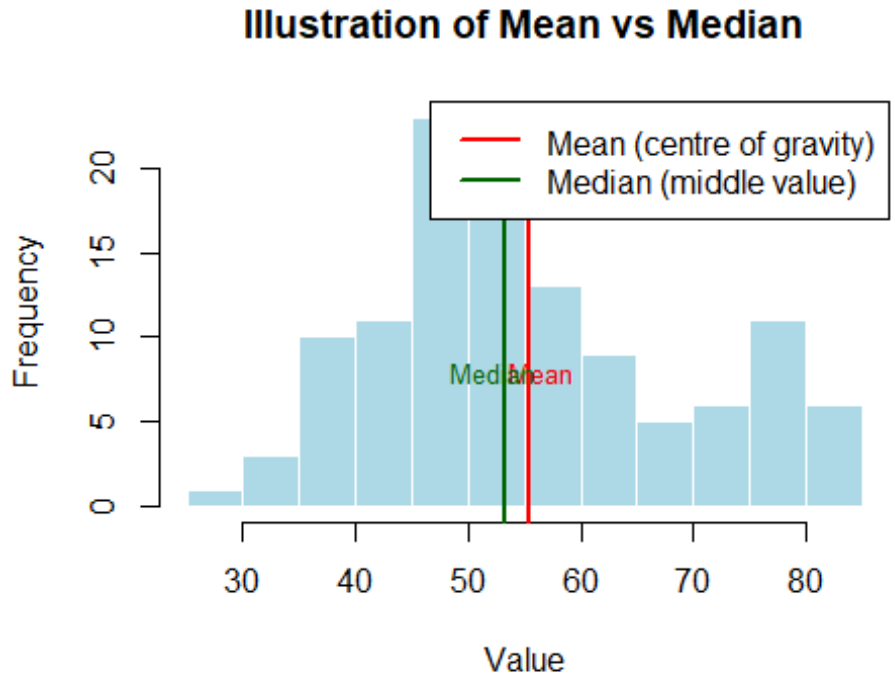
```
##   [1]   0   0   1   1   1   1   2   2   3   3   3   3   3   3   3   3   4
4
##  [19]   5   6   7   7   8   8   8   8   8   9   9   9   9   9   9  10  10
10
##  [37]  10  10  11  11  11  12  12  12  13  14  14  15  16  16  16  16  18
19
##  [55]  19  19  19  19  20  20  20  21  21  22  22  22  23  23  23  24  24
25
##  [73]  25  26  26  26  26  27  27  28  28  29  29  29  29  29  29  30  31
32
##  [91]  32  33  35  35  35  35  36  36  36  36  36  36  37  38  38  38  38
38
## [109]  39  39  40  41  42  43  43  44  44  44  44  44  47  47  47  48  48
48
## [127]  49  49  50  50  50  50  52  52  53  53  54  54  55  55  55  56  56
56
## [145]  57  60  61  61  63  64  65  65  66  67  68  70  71  71  72  73  75
75
## [163]  76  81  82  82  83  84  89  94  95  98 101 104 108 116
```

```r
median( x = afl.margins )
```

```
## [1] 30.5
```

```r
#below there is a visualization of where to locate median and mean in the
dataset
# Sample data: a right-skewed distribution
set.seed(123)
data <- c(rnorm(100, mean=50, sd=10), rnorm(20, mean=80, sd=5))

# Compute mean and median
data_mean <- mean(data)
data_median <- median(data)

# Create histogram
hist(data,
     breaks=20,
     col="lightblue",
     main="Illustration of Mean vs Median",
     xlab="Value",
     border="white")

# Add vertical lines for mean and median
abline(v=data_mean, col="red", lwd=2)
abline(v=data_median, col="darkgreen", lwd=2)
```

```r
# Add legend
legend("topright",
       legend=c("Mean (centre of gravity)", "Median (middle value)"),
       col=c("red", "darkgreen"),
       lwd=2)

# Add text annotations
text(data_mean + 1, 8, "Mean", col="red", cex=0.8)
text(data_median - 1, 8, "Median", col="darkgreen", cex=0.8)
```




```r
# the mean is the centre of gravity and the median is the middle value

# If your data are nominal scale, you probably shouldn't be using either the
mean or the median. Both the mean and the median rely on the idea that the
numbers assigned to values are meaningful. If the numbering scheme is
arbitrary, then it's probably best to use the mode

# If your data are ordinal scale, you're more likely to want to use the
median than the mean, The median only makes use of the order information in
your data (i.e., which numbers are bigger), but doesn't depend on the precise
numbers involved. That's exactly the situation that applies when your data
are ordinal scale. The mean, on the other hand, makes use of the precise
numeric values assigned to the observations, so it's not really appropriate
for ordinal data.
```

```r
# For interval and ratio scale data, either one is generally acceptable.
Which one you pick depends a bit on what you're trying to achieve. The mean
has the advantage that it uses all the information in the data (which is
useful when you don't have a lot of data), but it's very sensitive to extreme
values


# TRIMMED MEAN

# To calculate a trimmed mean, what you do is "discard" the most extreme
examples on both ends (i.e., the largest and the smallest), and then take the
mean of everything else

dataset <- c( -15,2,3,4,5,6,7,8,9,12 )

mean ( x = dataset)

## [1] 4.1

mean ( x= dataset, trim = .1)

## [1] 5.5

# mode is the value that occurs the most frequentely

head(afl.finalists, 25) # who as plays the most finals? showing the 25 rows

##  [1] Hawthorn    Melbourne   Carlton     Melbourne   Hawthorn    Carlton
##  [7] Melbourne   Carlton     Hawthorn    Melbourne   Melbourne   Hawthorn
## [13] Melbourne   Essendon    Hawthorn    Geelong     Geelong     Hawthorn
## [19] Collingwood Melbourne   Collingwood West Coast  Collingwood Essendon
## [25] Collingwood
## 17 Levels: Adelaide Brisbane Carlton Collingwood Essendon Fitzroy ...
Western Bulldogs

# producing a frequency table

table( afl.finalists )

## afl.finalists
##         Adelaide         Brisbane          Carlton      Collingwood
##               26               25               26               28
##         Essendon          Fitzroy        Fremantle          Geelong
##               32                0                6               39
##         Hawthorn        Melbourne  North Melbourne    Port Adelaide
##               27               28               28               17
##         Richmond         St Kilda           Sydney       West Coast
##                6               24               26               38
## Western Bulldogs
##               24
```

```r
# now we can see who's played the most finals

modeOf( x = afl.finalists )
```

```
## [1] "Geelong"
```

```r
# how many number is the modal frequancy? ( number of final games plays at
this occasion)

maxFreq( x = afl.finalists )
```

```
## [1] 39
```

```r
# Taken together, we observe that Geelong (39 finals) played in more finals
than any other team during the 1987-2010 period.

# mode in this case is caluclated on nominal scale data ( median and menas
will be uselss)

# if the scale was ratio scale for example the mesure you need is mean or
median

# guess the exact margin - this is a betting example observig that 8 of 176
games (45%) by picking a random football game

modeOf( x = afl.margins )
```

```
## [1] 3
```

```r
maxFreq( x = afl.margins )
```

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