Measure of Variability - Chapter 5 (5.2)

Aldo Mema

2025-06-20

```r
#Measure of Variability - chapter 5  (5.2)

# What is variability?
# How spread out are the data
# How far from the median or mean are the observed data

#load the data

load ("C:/Users/aldom/Msc Data Science/R_Project_Folder/data/aflsmall.Rdata")

# What is the variability of the data?
library(lsr)

who() # see variables
```

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

```r
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
```

```r
#Range
# The range is the difference between the maximum and minimum values in a
dataset.

max(afl.margins)
```

```
## [1] 116
```

```r
min(afl.margins)
```

```
## [1] 0
```

```r
# for this dataset the max value is:116 and the min value is: 0

range(afl.margins)
```

```
## [1]   0 116
```

```r
# Be careful when using range to quantify variability, it is sensitive to
outliers
# Examples with extreme outlier (-100, 2,3,4,5,6,7,8,9,10)
# the range with the data above is not a robust measure of variability as it
is affected by the outlier


# INTERQUARTILE RANGE

# It calculates the difference between the 25Th quantile and the 75th
quantile.
# What is a quartile?
# Quartiles are the values that divide a dataset into four equal parts.
# The first quartile is the value at 25%
# The second quartile is the value at 50% (median)
# The third queritle is the value at 75%

quantile(x = afl.margins, probs = .5) # median caluculation with quantile
function as the book example
```

```
##  50%
## 30.5
```

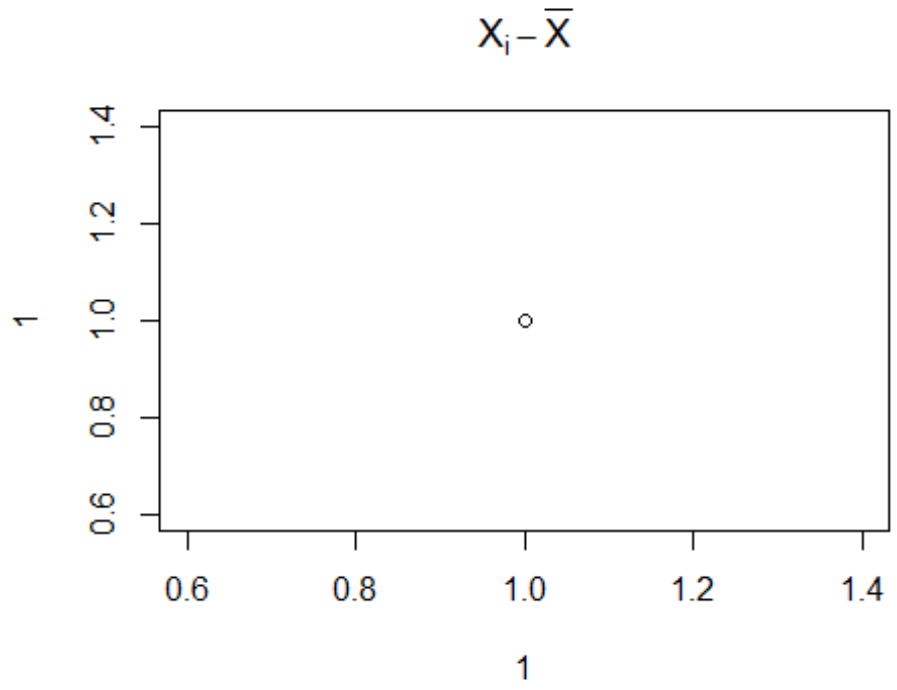
```r
quantile ( x = afl.margins, probs = c(.25,.75)) # 25th and 75th percentiles
```

```
##   25%   75%
## 12.75 50.50
```

```r
# Now note that 50.5 - 12.75 = 37.75 (IQR)

IQR (x = afl.margins)
```

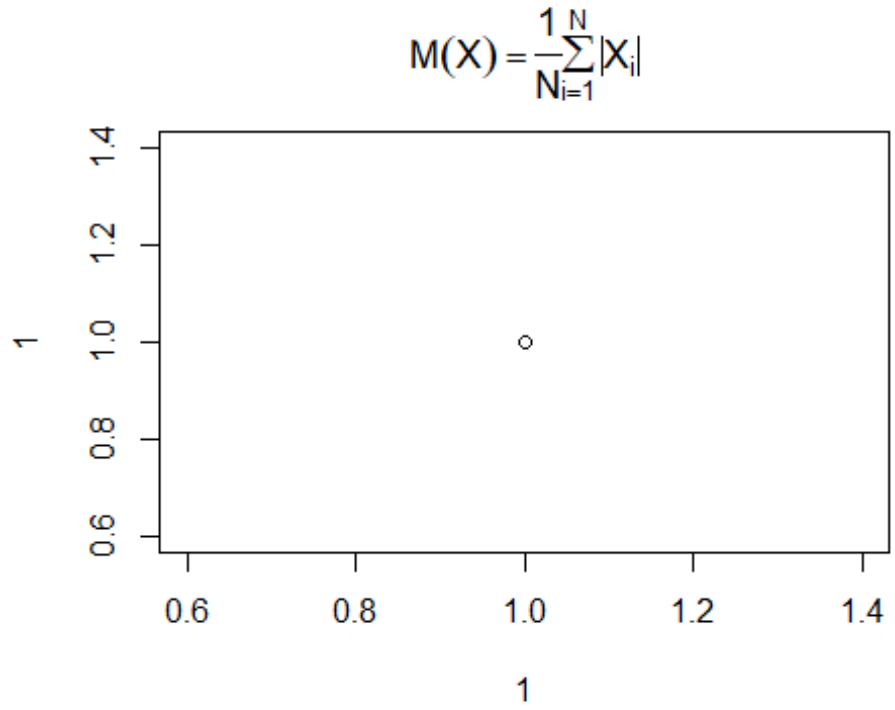
```
## [1] 37.75
```

```r
# The IQR is a range SPANNED by the "middle half" of the data
# One quarter of the data falls below 25% another quarter is above 75%
leaving the MIDDLE HALF OF THE DATA LYING IN BETWEEN THE TWO! WHICH IS THE
IQR (RANGE COVERED BY THE MIDDLE)

# BOTH OF THE ABOVE MESURES RELY ON THE IDEA OF THE MESURE OF THE SPREAD OF
THE DATA

# Next way of thinking --->

# A diffrent approach to resovle the problem is to firstly select the
meaninful reference point (usually mean or median) and then report a typical
DEVIATION from the reference point.

# What is typical deviation? usually the mean or median absolute deviations
from the actual mean or median

#MEAN Absolute Deviation (from the mean)

# Let's pretend that ((afl.margins) in this dataset) there are only 5 games
played with winning margins of 56,31,56,8,32

# first we calculate the mean of this 5 games margins

X = (56 + 31 + 56 + 8 + 32) / 5

# the next step is to convert each of our observations Xi into a deviation
score (deviation from the mean)

# this is done by calculating the difference between each observation and the
mean (

# Display the formula as text in the console
plot(1, 1, main=expression(X[i] - bar(X)))
```




```
# for the first observation (56) the dviation is equal to 56 - 36.6 = 19.4
# is important to convert this deviations to absolute deviations so we
convert any negative value to a positive value

# for the second observation (31) the deviation is equal to 31 - 36.6 = -5.6
(absolute value is 5.6)
# for the third observation (56) the deviation is equal to 56 - 36.6 = 19.4
# for the fourth observation (8) the deviation is equal to 8 - 36.6 = -28.6
(absolute value is 28.6)
# for the fifth observation (32) the deviation is equal to 32 - 36.6 = -4.6
(absolute value is 4.6)

# Now we can calculate the mean absolute deviation by taking the average of
these absolute deviations
aboslute_deviation = (19.4 + 5.6 + 19.4 + 28.6 + 4.6) / 5

# mean_absolute_deviation = 15.52

X <- c(56, 31,56,8,32)   # enter the data
X.bar <- mean( X )       # step 1. the mean of the data
AD <- abs( X - X.bar )   # step 2. the absolute deviations from the mean
AAD <- mean( AD )        # step 3. the mean absolute deviations
print( AAD )             # print the results

## [1] 15.52
```

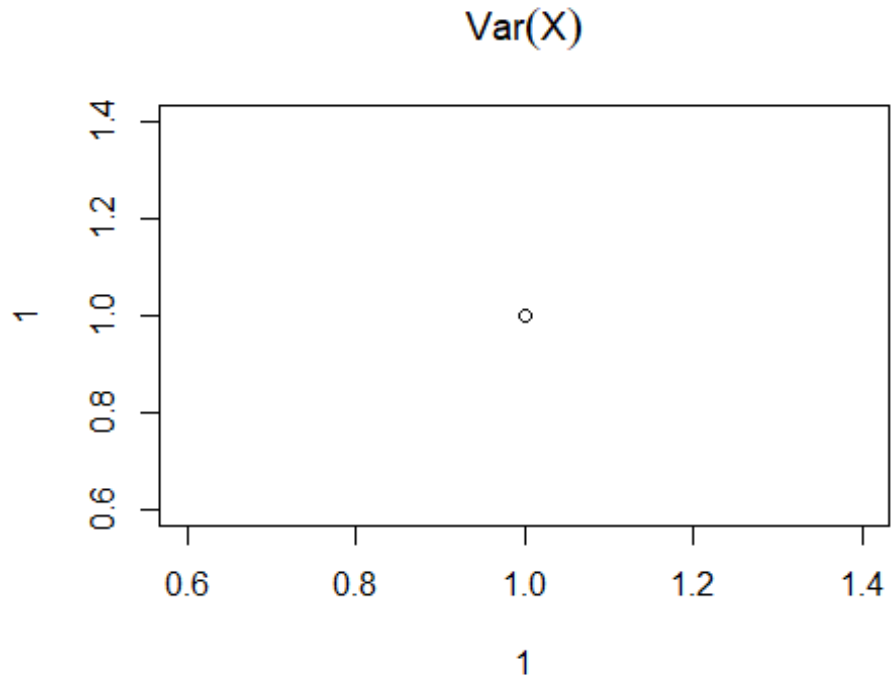
```r
# in Mathematical formula this is:
# Display the formula in a plot title
plot(1, 1, main=expression(M(X) == frac(1, N) * sum(abs(X[i]), i==1, N)))
```



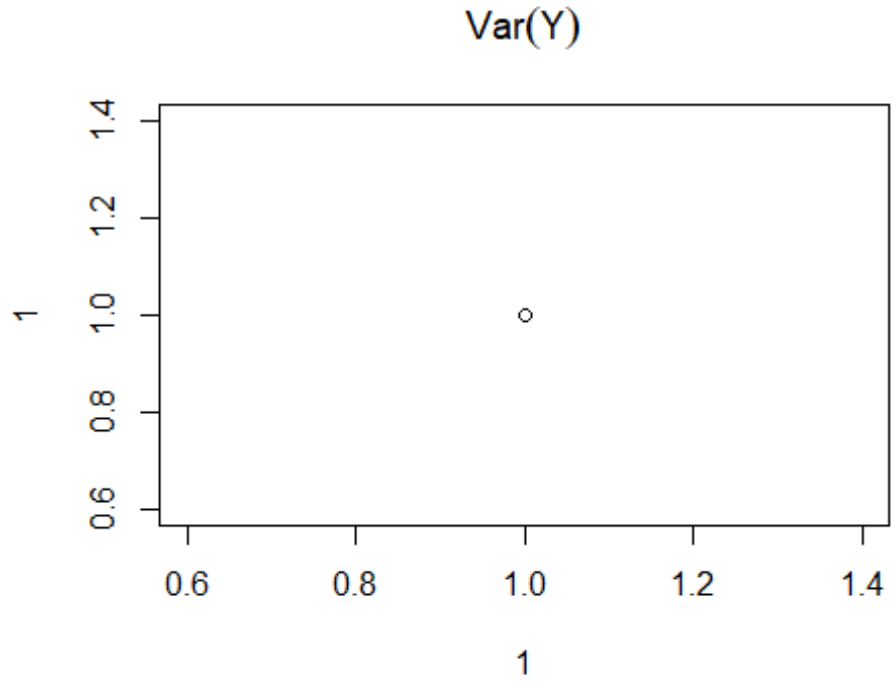

```r
# in the lsr package there is a simple command to do all of this and is
called AAD()

library(lsr)
aad( X )
```

```
## [1] 15.52
```

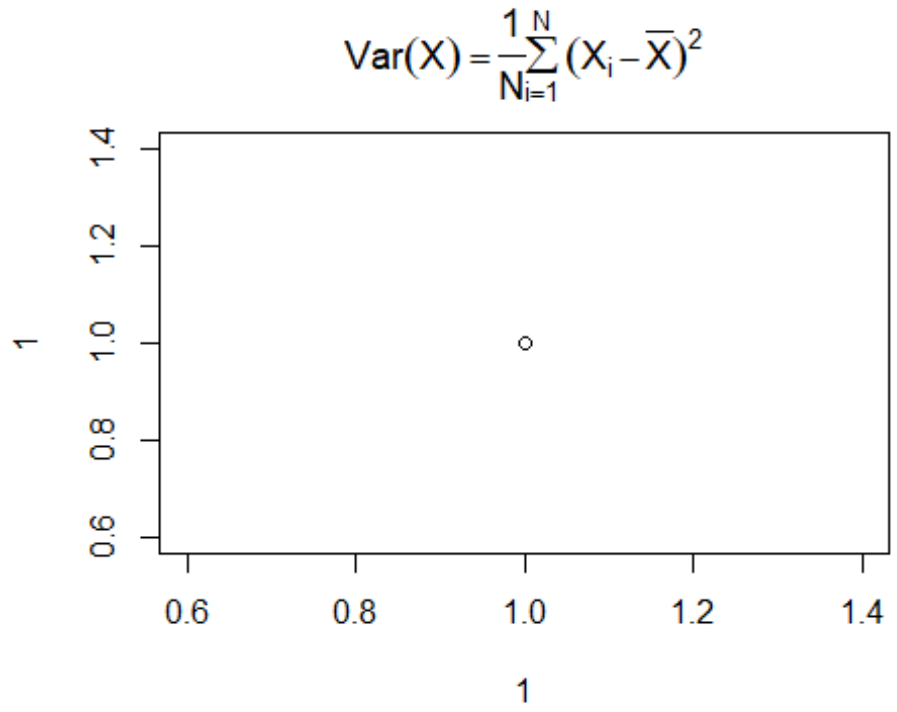
```r
#VARIANCE

# from a mathematical perspective there is a solid reason to prefer SQUARED
DEVIATIONS rater then absolute deviations!

# When we do that we obtain the measure VARIANCE
# Display Var(X)
plot(1, 1, main=expression(Var(X)))
```




```
# Display Var(Y)
plot(1, 1, main=expression(Var(Y)))
```



```r
# Now we suppose we want to define a NEW variable Z (which is the sum of the
two Z = X+Y)

# It turns out that Z = Var(X) + Var(Y) in this case - but is not always true

# the Variance of a data X is sometimes written as Var (X) or s^2 (s squared)

# the forulas for calculatin the variance of a set of obsrvations are plotted
below

# Display variance formula in a plot title
plot(1, 1, main=expression(Var(X) == frac(1, N) * sum((X[i] - bar(X))^2,
i==1, N)))
```




```r
# This is the same formula to calculate the mean absolute deviation EXCEPT
that that is "SQUARED DEVIATION"

# Variance sometimes is referred ad "mean squared deviation" (MSD)

# Table 5.1 in the book shows the calculations for the variance of a set of
observations

# Game numbers
i <- 1:5
```

```r
# Values of X
X <- c(56, 31, 56, 8, 32)

# Mean of X
mean_X <- mean(X)

# Deviation from the mean
deviation <- X - mean_X

# Squared deviation
squared_deviation <- (X - mean_X)^2

# Create data frame
table5.1 <- data.frame(
  i = i,
  X_i = X,
  Deviation = deviation,
  Squared_Deviation = squared_deviation
)

# View the table
print(table5.1)
```

```
##   i X_i Deviation Squared_Deviation
## 1 1  56      19.4            376.36
## 2 2  31      -5.6             31.36
## 3 3  56      19.4            376.36
## 4 4   8     -28.6            817.96
## 5 5  32      -4.6             21.16
```

```r
# The last collumn contains the squared deviation
# So we average them

( 376.36 + 31.36 + 376.36 + 817.96 + 21.16 ) / 5
```

```
## [1] 324.64
```

```r
# to do all of this in r without typing the numbers:

mean( (X - mean(X) )^2)
```

```
## [1] 324.64
```

```r
# same answer but to much typing - in short

var(X) # wait this is not the same as 324.64, what is going on?
```

```
## [1] 405.8
```

```r
# The reason is that the formula above is the population variance, but we are
working with a sample of the population with only 5 data points
```

```r
# let's switch to the full dataset

# calculate the variance

mean( (afl.margins - mean(afl.margins))^2)
```

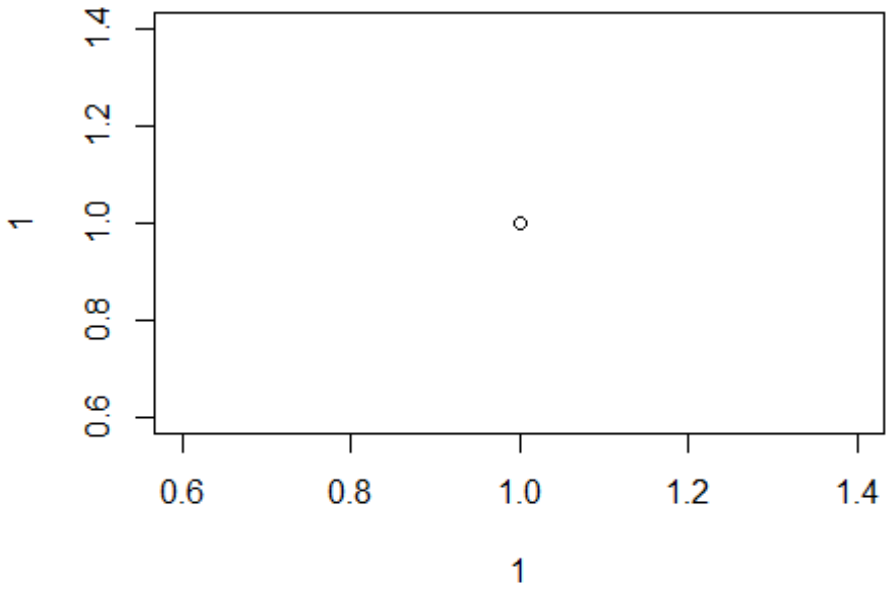
```
## [1] 675.9718
```

```r
# now let's use var()

var(afl.margins)
```

```
## [1] 679.8345
```

```r
# still a difference in the result but with a larger dataset the difference is smaller

# This is because R is using a slightly different formula to calculate the variance, which is called the sample variance. (N-1)

# Display sample variance formula in a plot title
plot(1, 1, main=expression(s^2 == frac(1, N-1) * sum((X[i] - bar(X))^2, i==1, N)))
```

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$$



```r
# verify the number to be the same as var()

sum( (X-mean(X))^2 ) / 4
```

```
## [1] 405.8
```

```r
# the division for N-1 instead of N is explain in chapter 10 and is called
Bessel's correction, and it is used to correct the bias in the estimation of
the population variance from a sample.

# The sample variance is an unbiased estimator of the population variance,
which means that on average it will give you the correct value for the
population variance.

# At the end of to we interpret this for descrptive statistics ?

# Descrprive statistc is used to describe things and the interpretation of
variance is not human friendly but a matematical approach to resolve the
problem

# The variance is a measure of how much the data points deviate from the
mean, and it is expressed in squared units of the original data.

# The larger the variance, the more spread out the data points are from the
mean.

# The smaller the variance, the more clustered the data points are around the
mean.

# The variance is not a very intuitive measure of variability, as it is
expressed in squared units of the original data, which makes it hard to
interpret.

# A more intuitive measure of variability is the standard deviation, which is
the square root of the variance.

# STANDARD DEVIATION
#What is standard deviation?
# The standard deviation is a measure of how much the data points deviate
from the mean, and it is expressed in the same units as the original data.

# because we are humans and not robots we need a mesure expressed in the same
units as the data itself

# the solution is to take the squre root of the variance known as the
standard deviation
#also called root mean square deviation (RMSD)

# no humans can understan of what a viariance of 324.68 means - but we can
```

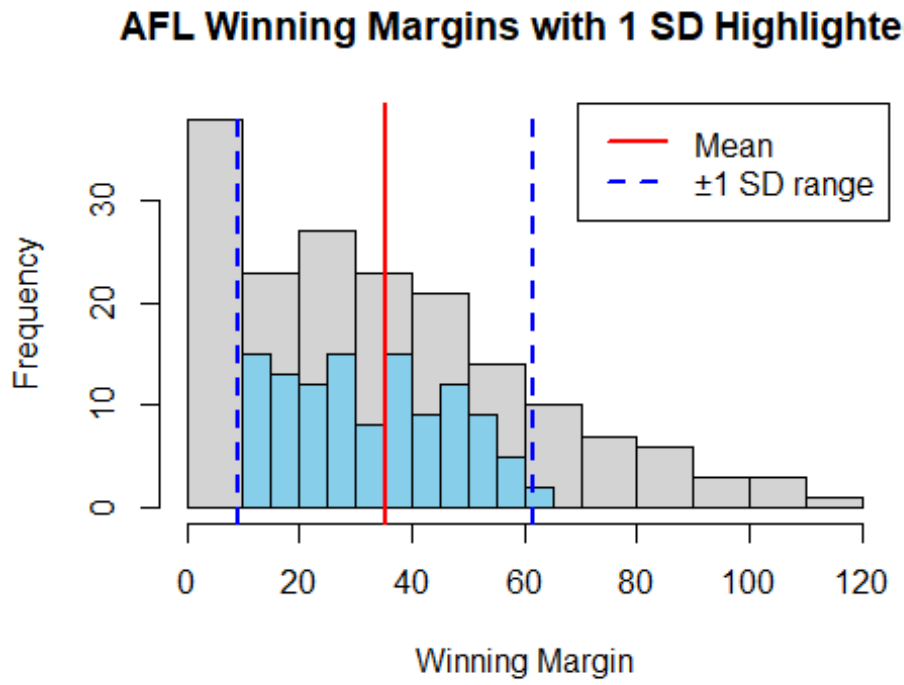
```
easier understan a standard deviation of 18.01 points!

# Display population standard deviation formula in a plot title

# R function is sd()

# but what R calculates is slight difference called sample standard deviation
N-1 rather then N

# this is referred as  σ^ (sigma hat) and is an unbiased estimator of the
population standard deviation

plot(1, 1, type = "n", axes = FALSE, xlab = "", ylab = "", main = "")
text(1, 1, expression(hat(sigma) == sqrt(frac(1, N - 1) * sum((X[i] -
bar(X))^2, i == 1, N))), cex = 1.5)
```

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

```
# in our exercise

sd( afl.margins )

## [1] 26.07364

# Interpreting standard deviations is slightly more complex. Because the
standard deviation is derived from the variance, and the variance is a
quantity that has little to no meaning that makes sense to us humans, the
standard deviation doesn't have a simple interpretation. As a consequence,
```

most of us just rely on a simple rule of thumb: in general, you should expect 68% of the data to fall within 1 standard deviation of the mean, 95% of the data to fall within 2 standard deviation of the mean, and 99.7% of the data to fall within 3 standard deviations of the mean. This rule tends to work pretty well most of the time, but it's not exact: it's actually calculated based on an assumption that the histogram is symmetric and "bell shaped."

```
# Calculate mean and standard deviation
mean_val <- mean(afl.margins)
sd_val <- sd(afl.margins)

# Define range within 1 standard deviation
lower_bound <- mean_val - sd_val
upper_bound <- mean_val + sd_val

# Plot histogram
hist(afl.margins, breaks = 10, col = "lightgray", main = "AFL Winning Margins
with 1 SD Highlighted",
     xlab = "Winning Margin")

# Shade bars within one standard deviation
hist_in_range <- hist(afl.margins[afl.margins >= lower_bound & afl.margins <=
upper_bound],
                      breaks = 10, col = "skyblue", add = TRUE)

# Add vertical lines for mean and standard deviations
abline(v = mean_val, col = "red", lwd = 2)
abline(v = c(lower_bound, upper_bound), col = "blue", lty = 2, lwd = 2)

# Add legend
legend("topright", legend = c("Mean", "±1 SD range"), col = c("red", "blue"),
lwd = 2, lty = c(1, 2))
```




```
# The standard deviation is a measure of how much the data points deviate from the mean, and it is expressed in the same units as the original data.

# The larger the standard deviation, the more spread out the data points are from the mean.

# The smaller the standard deviation, the more clustered the data points are around the mean.

# The standard deviation is a more intuitive measure of variability than the variance, as it is expressed in the same units as the original data.

# The standard deviation is also used to calculate the confidence intervals for the mean, which is a range of values that is likely to contain the true population mean.

# The confidence interval is calculated as the mean plus or minus a certain number of standard deviations, depending on the desired level of confidence.

# For example, a 95% confidence interval is calculated as the mean plus or minus 1.96 times the standard deviation.

# This means that we can be 95% confident that the true population mean lies within this range.
```

```r
# In summary, the standard deviation is a measure of how much the data points deviate from the mean, and it is expressed in the same units as the original data. It is a more intuitive measure of variability than the variance, and it is used to calculate the confidence intervals for the mean.

# The standard deviation is a useful measure of variability, as it allows us to understand how much the data points deviate from the mean, and it is expressed in the same units as the original data.

# ------------------------------------------------

# Median Absolute deviation (MAD)

# The median absolute deviation (MAD) is a robust measure of variability that is less affected by outliers than the standard deviation.

# It is calculated as the median of the absolute deviations from the median.

# Using median everywhere

# mean absolute deviation from the mean:
mean( abs(afl.margins - mean(afl.margins)) )
```

```
## [1] 21.10124
```

```r
# *median* absolute deviation from the *median*:
median( abs(afl.margins - median(afl.margins)) )
```

```
## [1] 19.5
```

```r
# MAD attempts to describe a typicall deviation from a typical value in the data set
#If you want to use it to calculate MAD in the exact same way that I have described it above, the command that you need to use specifies two arguments: the data set itself x, and a constant that I'll explain in a moment. For our purposes, the constant is 1, so our command becomes

mad( x = afl.margins, constant = 1 )
```

```
## [1] 19.5
```

```r
# using 1 as constant in real word context is helpful to outlier vulnerability, as it is less affected by outliers than the standard deviation.
# constant = 1 part, this is pretty straightforward.

mad( afl.margins ) # this is the same as above in R
```

```
## [1] 28.9107
```

```
# book tip -- I should point out, though, that if you want to use this
“corrected” MAD value as a robust version of the standard deviation, you
really are relying on the assumption that the data are (or at least, are
“supposed to be” in some sense) symmetric and basically shaped like a bell
curve. That’s really not true for our afl.margins data, so in this case I
wouldn’t try to use the MAD value this way.




#WHICH MEASURE TO USE?

# Range ---<

# GIVES THE SPREAD OF THE DATA, is very VULNERABLE TO OUTLIERS and is it
often usless unless you have a good reason of caring of the extrems data

# IQR (interwuartile range)---<

# TELLS WHERE THE MIDDLE HALF of the data sits. Very useful and used!

# Mean Absolute Deviation ---<

# TELLS HOW FAR ON AVERAGE THE OBSERVATIONS ARE FROM TH EMEAN. Is very
iterpretable (less attractive then standard deviation due to minor issues)

# Variance---<

# Tells the average squared deviation from the mean. it is very
mathematicall, probably the right way to describe variation around the mean.
but is completealy uninterpretable because it doesn't use the same units as
the data.

# STANDARD DEVIATION ---<

# This is the square root of variance. This is expressed in the same untis as
the data so it can be interpreted very well. When the mean is the mesure of
central tendency SD is the defult . This is the most popular mesure of
variation

# Mean absolute deviation ---<

# the typical (median) from the median value. in the raw form its simple and
interpretable and is a robust way to estimate the standard deviation for some
kinds of datasets. Not used a lot not common

## IN SHORT IQR AND SD ARE EASILY THE TWO MOST COMMON MESURES USED TO REPORT
THE VARIABI
```