# DATA 500
# Lecture 5

Statistics by Numbers

# What is Statistics (singular)?

**Statistics is the science of uncertainty**
**(Estimation, Approximation, Prediction)**

# What are (basic) statistics? (plural)

Sample Size (total): ***N or n*** (sometimes ***n*** for a group):

Individual measurement / Observation: $\boldsymbol{x}_{\mathbf{i}}$

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \ldots, \boldsymbol{x}_n$$

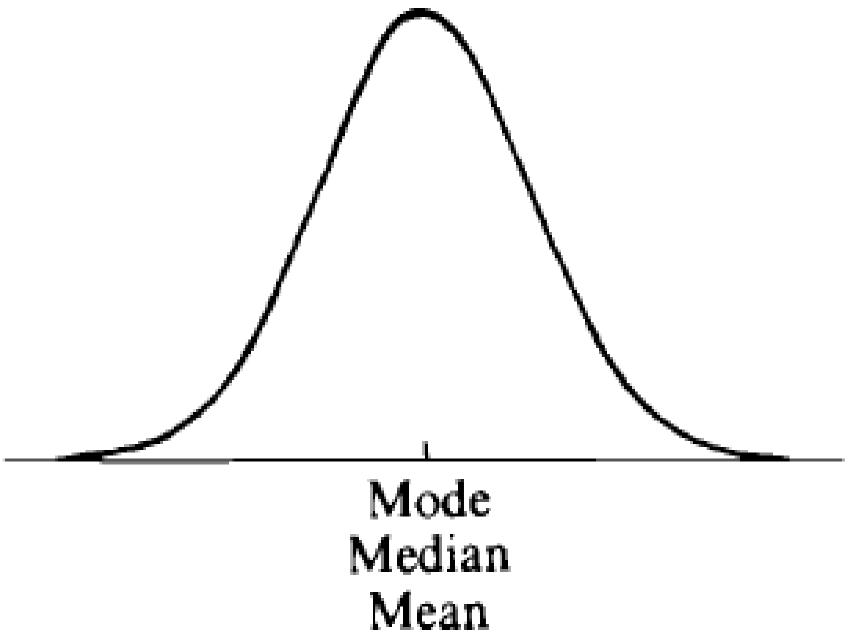
Sum of all observations $= \sum_{i=1}^{n} x_i = \sum \boldsymbol{x}_{\mathbf{i}}$

**Sample Mean** $= \sum x_i / n = \bar{x}$ ("x bar")

$$sample\ mean = \bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_i x_i$$

# Basic Statistics (plural): Central tendency

**Mean** = $\bar{x}$
**Median:** (midpoint; 50$^{th}$ percentile)
**Mode:** most common value

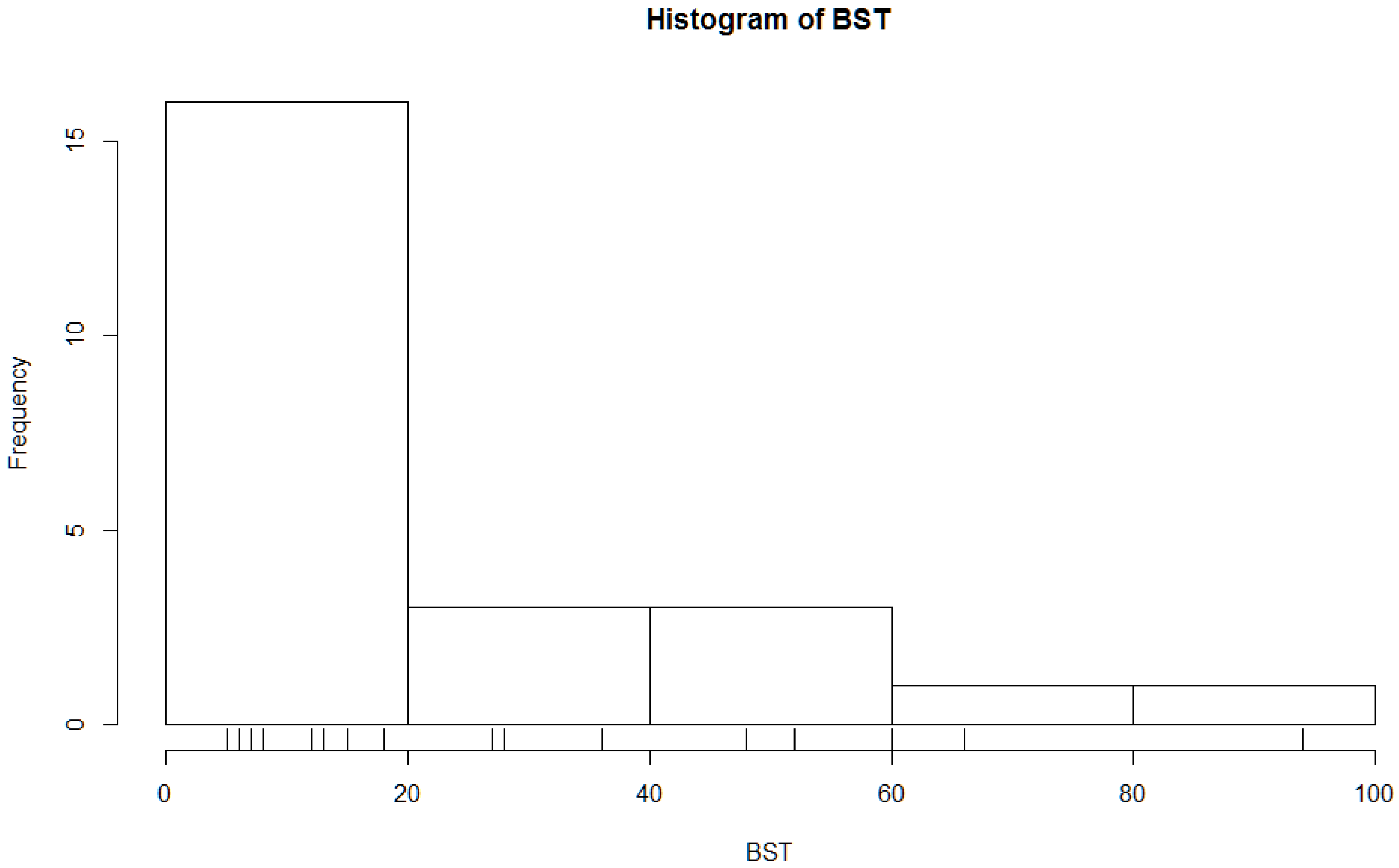



The Normal Distribution

# A brief look at data: BST

```
# Business startup times vector
BST <- scan("http://math.mercyhurst.edu/~sousley/STAT_139/data/BST.vec");

# Executive pay vector
exec.pay <- scan("http://math.mercyhurst.edu/~sousley/STAT_139/data/exec.pay.vec");

# Parent and child statures dataframe
galton <- read.csv("http://math.mercyhurst.edu/~sousley/STAT_139/data/galton.csv");
```

```
# Let's make a histogram
hist(BST)
rug(BST)
mean(BST)
[1] 23.625

median(BST)
[1] 13
```



# A brief summary of data: Quartiles

## Quartiles

Quartiles are cutoffs for dividing the sample into four parts:

1st : 25th percentile

2nd : 50th percentile - median

3rd : 75th percentile

min 1st, 25th median 3rd, 75th max

**Quint**iles divide the sample into 5 parts (20,40, 60,80)

**Deciles** divide the sample into tenths

```
# Get the quartiles  Qu. = quartile
summary(BST)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00    7.00   13.00   23.62   30.00   94.00
```

# Calculating the median (the 50$^{th}$ percentile)

**First, sort the data from smallest to largest**

**In odd-numbered datasets**,

it is the value of the $i = n/2 + 0.5$ observation

**Example:**

51 observations, 51/2 = 25.5 + 0.5 = 26$^{th}$ obs

1st…25th (25 obs) **26th** 27th…51st (25 obs)

**In even-numbered datasets**,

it is the **mean** of the values of $i = n/2$ and $(n/2)+1$

example:

24 observations, $i_1$ = 24/2 = 12 , $i_2$ = (24/2)+ 1 = 13

1st…12th (12 obs) **12th 13th** 13th…24th (12 obs)

# A brief look at data: BST

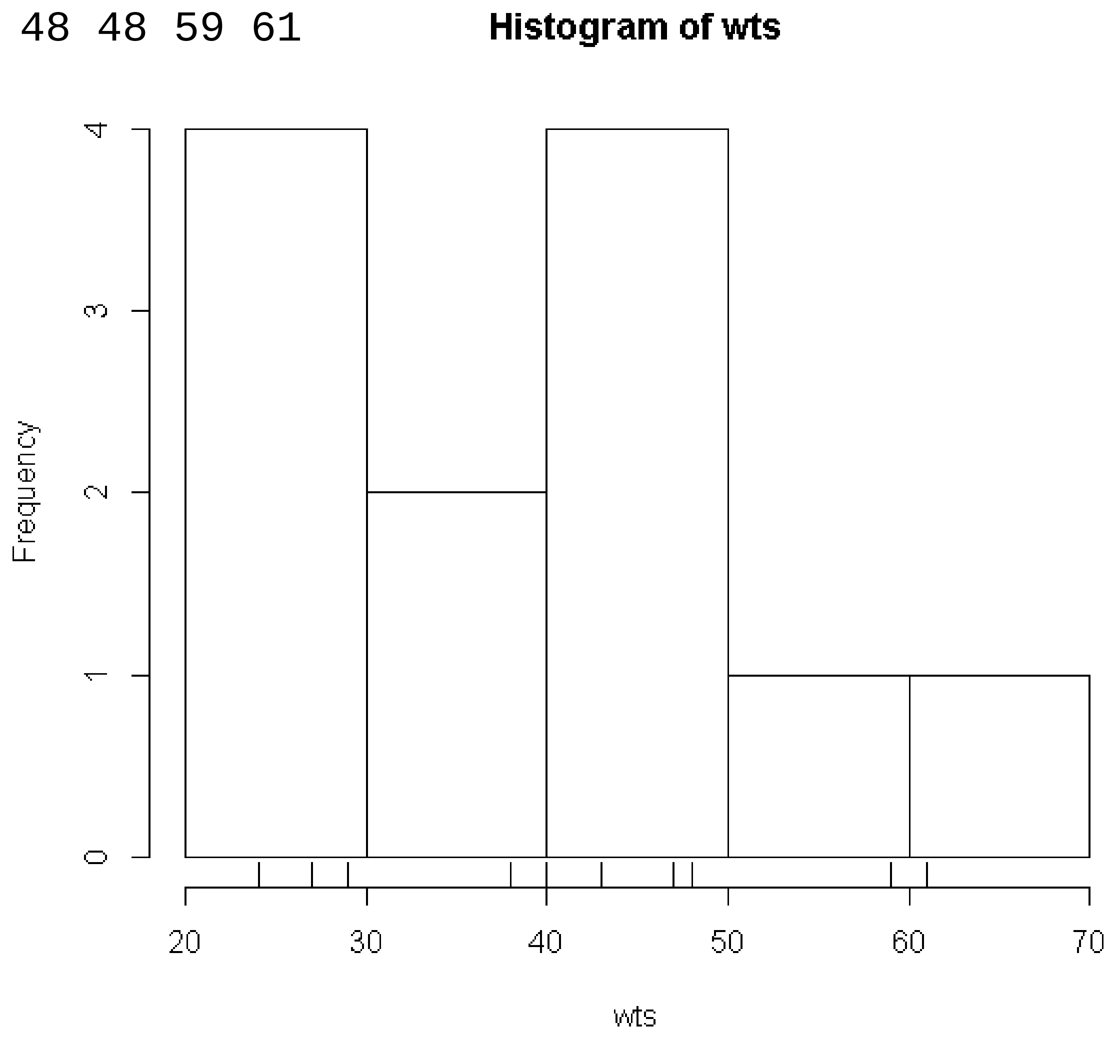
```
sort(BST)
[1] 5 5 5 5 6 7 7 7 8 12 12 13 13 15 18 18 27 28 36 48 52 60 66 94
length(BST)
[1] 24

# Get the quartiles  Qu. = quartile
summary(BST)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00    7.00   13.00   23.62   30.00   94.00
```

**Five number summary: min, Q1, median, Q3, MAX**

**- tells us a bit about central tendency and spread**

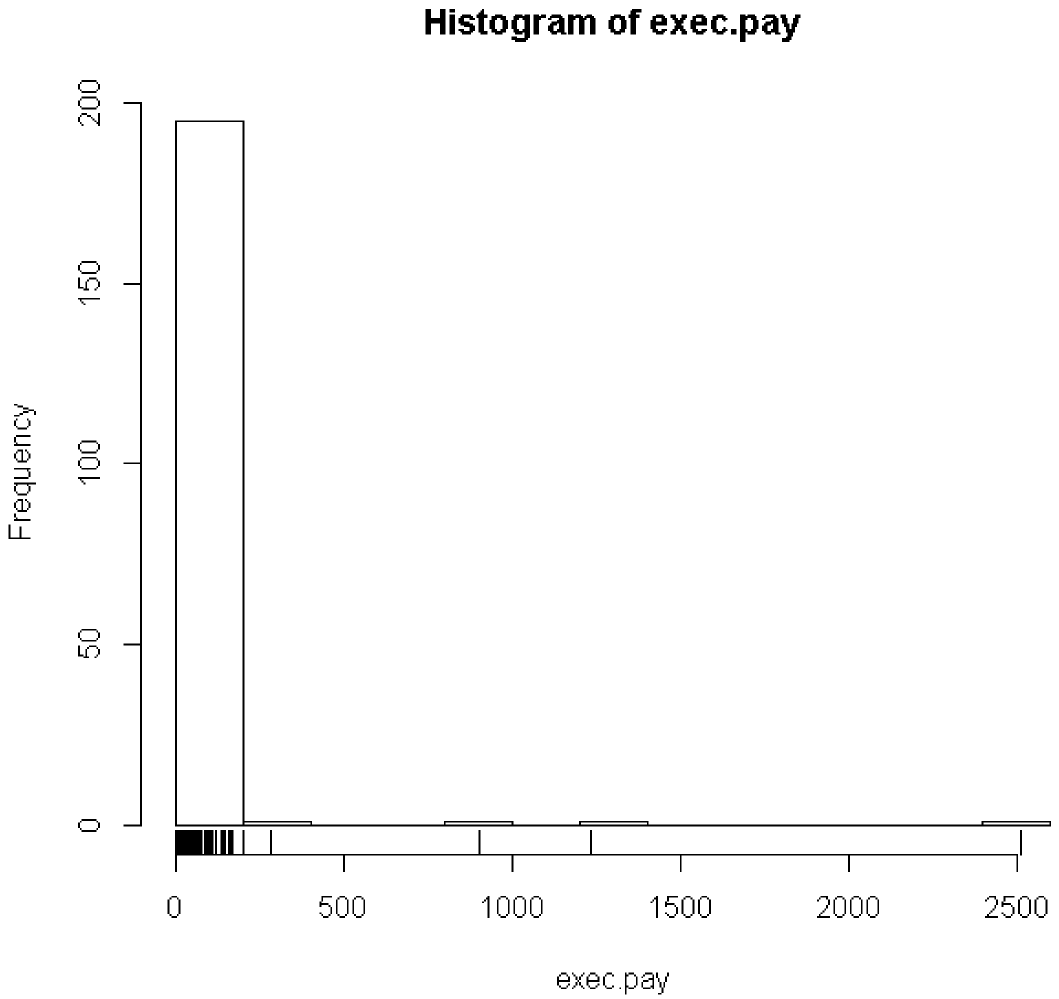
# A brief look at data: wts

```
# read 12 weights of 4-year-old children into a vector
wts <- c(38, 43, 48, 61, 47, 24, 29, 48, 59, 24, 40, 27)
# sort the vector
sort(wts)
 [1] 24 24 27 29 38 40 43 47 48 48 59 61

#Let's make a histogram
hist(wts)
rug(wts)
mean(wts)
[1] 40.66667
median(wts)
[1] 41.5
```



# A brief look at data: exec.pay

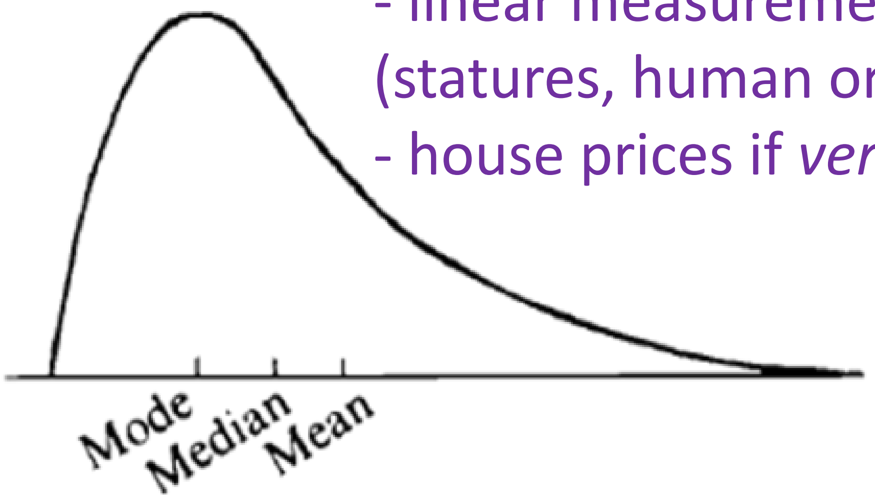
```
# executive pay
mean(exec.pay)
[1] 59.88945
median(exec.pay)
[1] 27
hist(exec.pay)
rug(exec.pay)


# compare to wts from before:
mean(wts)
[1] 40.66667
median(wts)
[1] 41.5
```

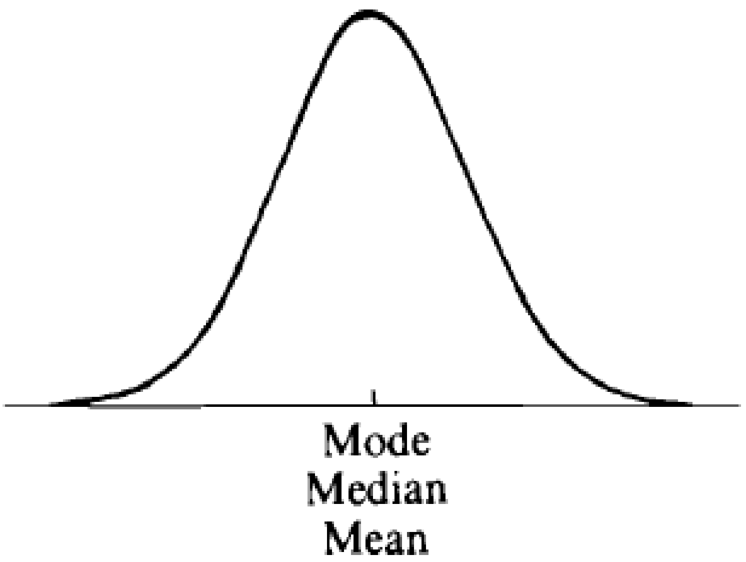


# Medians and Means

**Medians are better** when data are not normally distributed
- salaries (always some very large, some very small)
- household wealth
- house prices (if large area)
- waiting times for ...

**Means and Medians** are fine when data are normally distributed
- linear measurement of anything within groups
(statures, human or aphid femurs, etc.)
- house prices if *very small neighborhood* (depends)




The Normal Distribution

# Spread: Percentiles, Quantiles, Quartiles, Quintiles

25th Percentile example:

the value in the sample that satisfies:

25% of sample < value and (100-25%) of sample ≥ value

```
x <- 1:100 # 0,1,2,3,4,5
length(x)
[1] 100
# quantile is the function to get percentiles
quantile(x, 0.25)
  25%
25.75
# the default quantiles are four divisions (quartiles)
quantile(x)
    0%    25%    50%    75%   100%
  1.00  25.75  50.50  75.25 100.00

# quintiles: 5 divisions
quantile(x, seq(0, 1, by=0.20)) # divide by fives: quintiles
   0%   20%   40%   60%   80%  100%
  1.0  20.8  40.6  60.4  80.2 100.0
```

# Spread: Typical Statistics

## Maximum, Minimum, Range (Max-Min)

```
#R gives you the min and max in one function
range(wts)
[1] 24 61

# to get the range, use the diff function
diff(range(wts))
[1] 37
```

# Spread: Typical Statistics

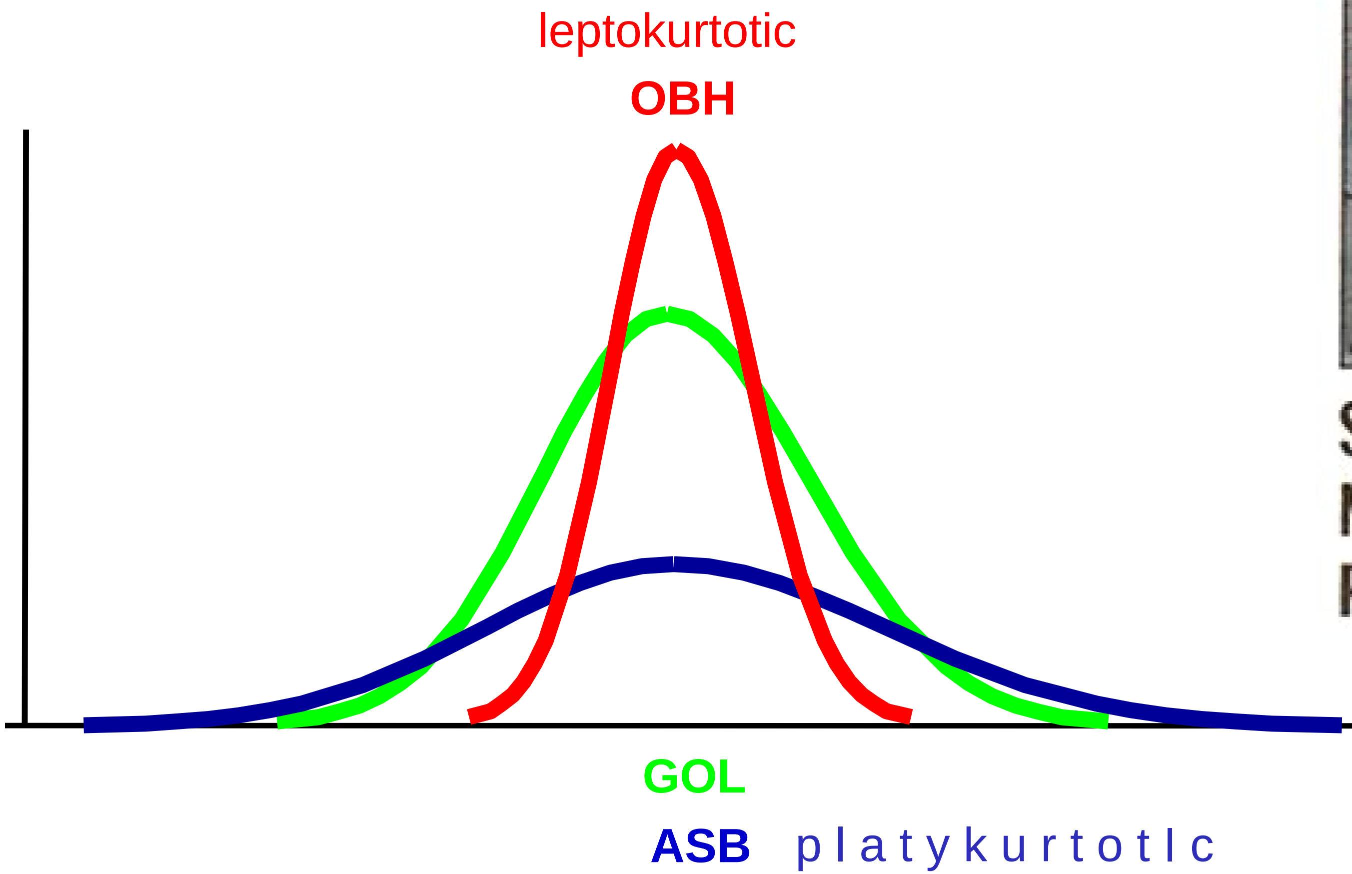
Standard deviation $s$, SD (sample)

$$\textit{sample standard deviation} = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$$

$s$ is the square root of the variance ($s^2$)

s.d. is provided in original units

# The *Standard Deviation* is a measure of spread

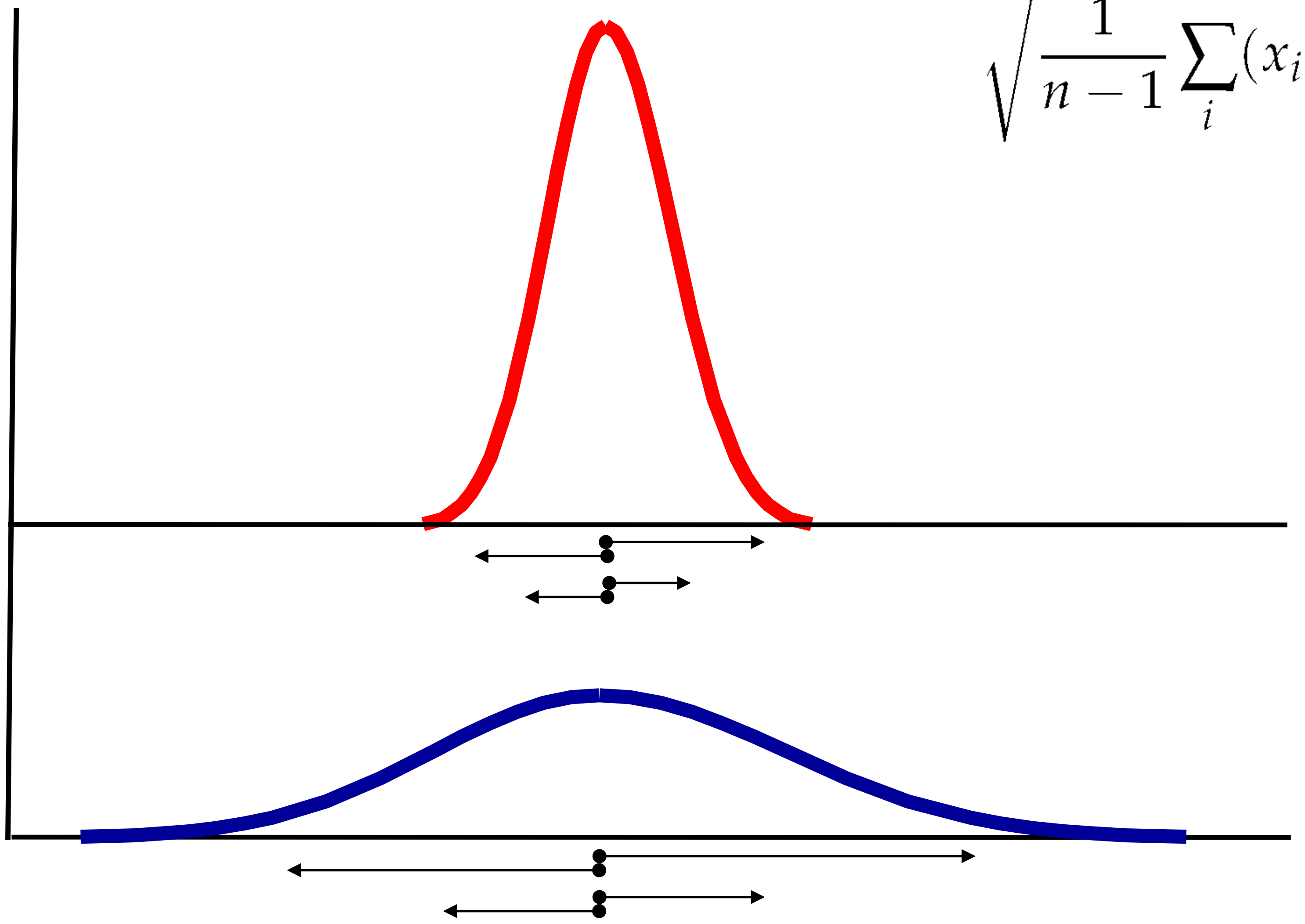



$$\sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$$

Standard Deviation Not Enough For Perverted Statistician

# The *Standard Deviation* is a measure of spread

$$\sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$$

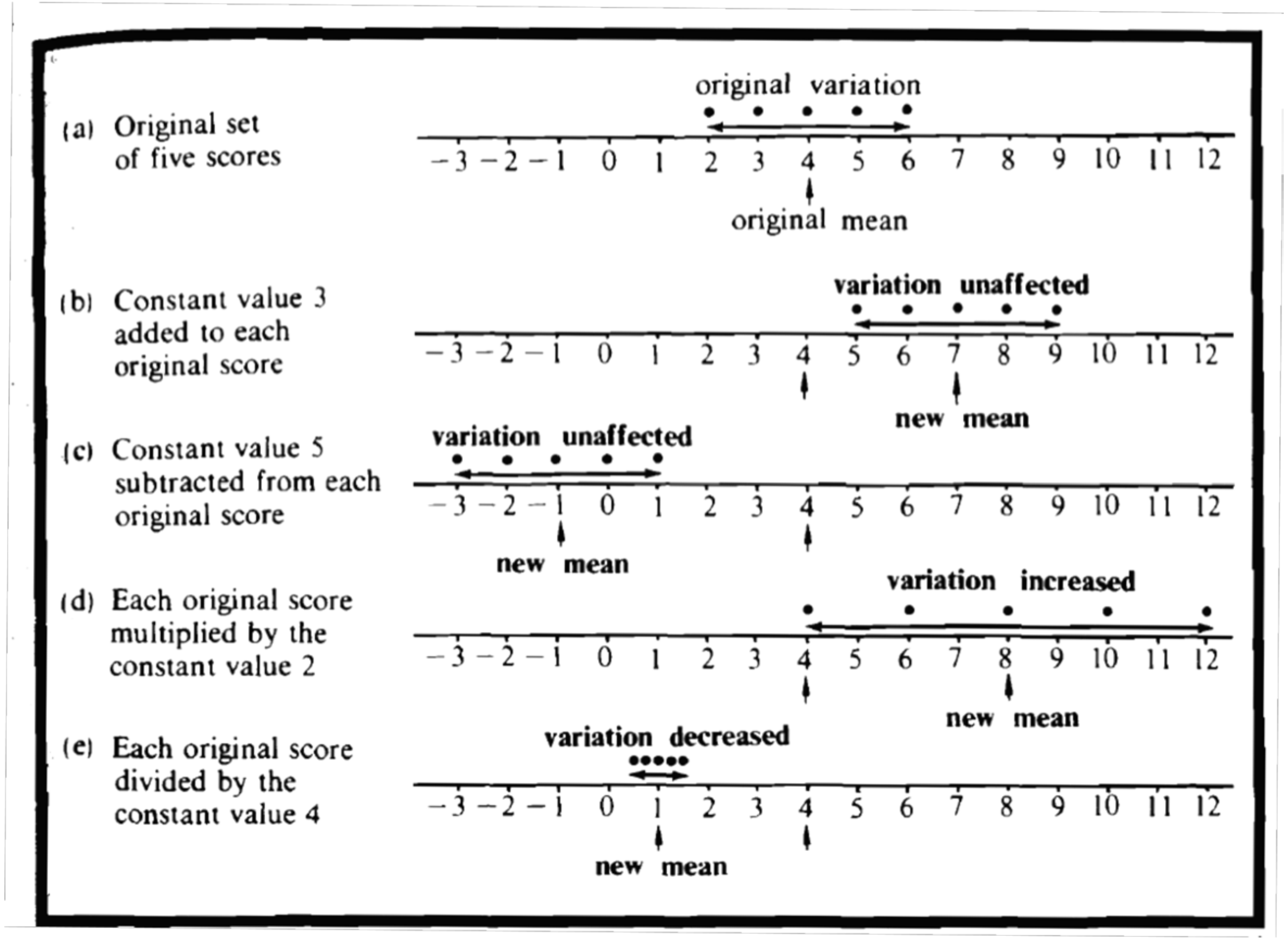


# Standard Deviation

```
# in R, use the sd function
sd(wts)
[1] 12.75171
sd(exec.pay)
[1] 207.0435
sd(BST)
[1] 23.82876
```

Variance is used indirectly in other calculations because the units are squared.

# Units matter!
# The Effect of Conversions / Changes




$\bar{x}_n = \bar{x}_o + 3$

$\bar{x}_n = \bar{x}_o - 5$

New $s$: $s * 2$

New $s$: $\frac{s}{4}$

# Scale-free measures of variation: *Z* scores

The standard deviation (s) depends on units:

convert inches to cm: New $s$ = old $s$ * 2.54

With $s$, we can calculate "Z" scores.

*Z* scores are simply scaled differently

- they are scaled by the standard deviation
- they are unitless!

# "$Z$ " scores (standardized data)

Observations converted to standard deviation scores

$\bar{x}$ = 50 mm, $s$ = 4.5 mm;

$$z\text{-score} = \frac{x_i - \bar{x}}{s}$$

$x_8$ = 59 mm, $Z(x_8)$ = 2.0
$x_9$ = 41 mm, $Z(x_9)$ = -2.0
$x_5$ = 49 mm, $Z(x_5)$ = -0.22
$x_6$ = 50 mm, $Z(x_6)$ = 0.0

**Nearly all *Z* scores are actually *t* scores:**

$$t = \frac{(x_i - \bar{x})}{s}$$

$Z$ scores are unitless

**We can summarize the relationship of any value to the sample using ONE number!**

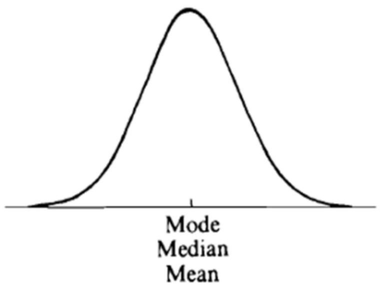
# Typical Statistics: shape

## Skewness (symmetry, taper)

### Normal distribution

skewness = 0

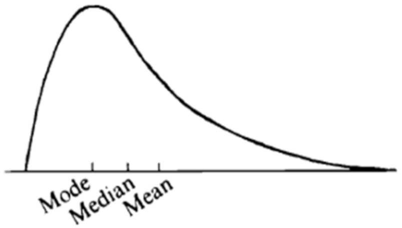
median = mean




### Positive skew

(skewed to the right)

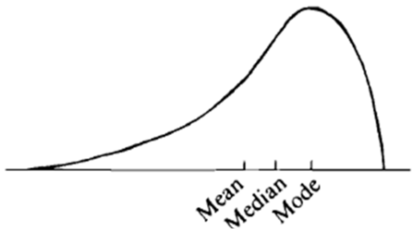
median < mean




### Negative skew

(skewed to the left)

mean < median




$$z\text{-score} = \frac{x_i - \bar{x}}{s}$$

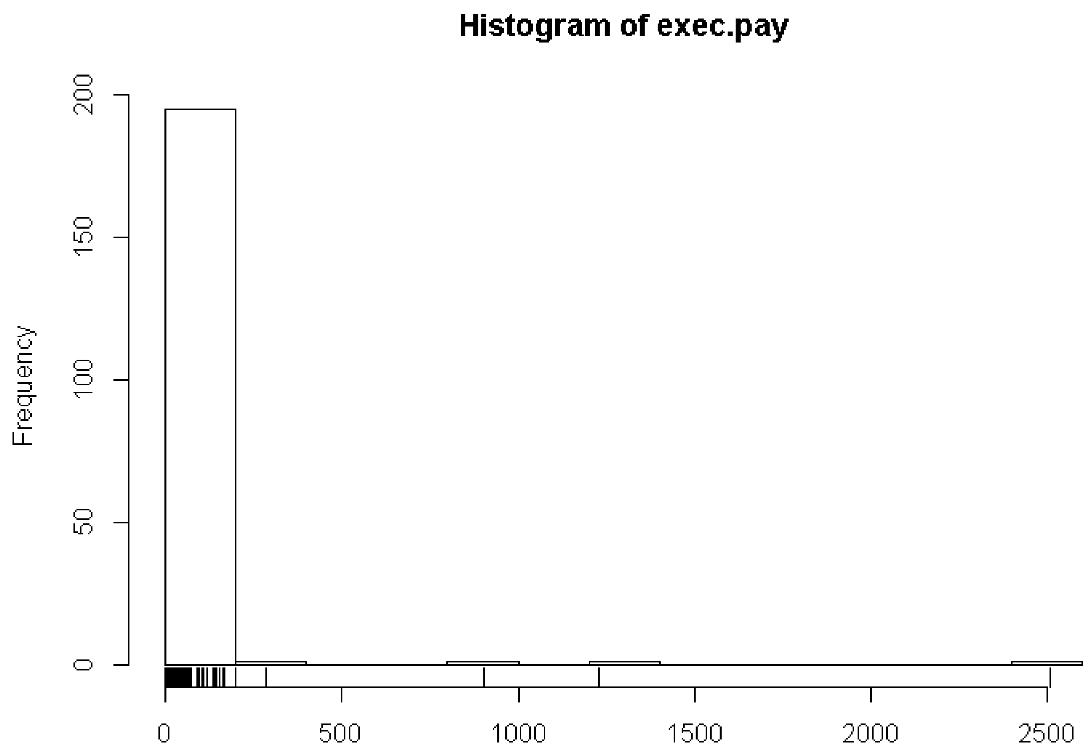
$$skewness = \sqrt{n}\frac{\sum(x_i - \bar{x})^3}{\left(\sum(x_i - \bar{x})^2\right)^{3/2}} = \frac{1}{n}\sum z_i^3$$

# Numerical Skewness

```
# certain things require R packages;
# get the e1071 package for skewness function;
install.packages("e1071");
# load into memory;
library(e1071);
```

## Look at executive pay

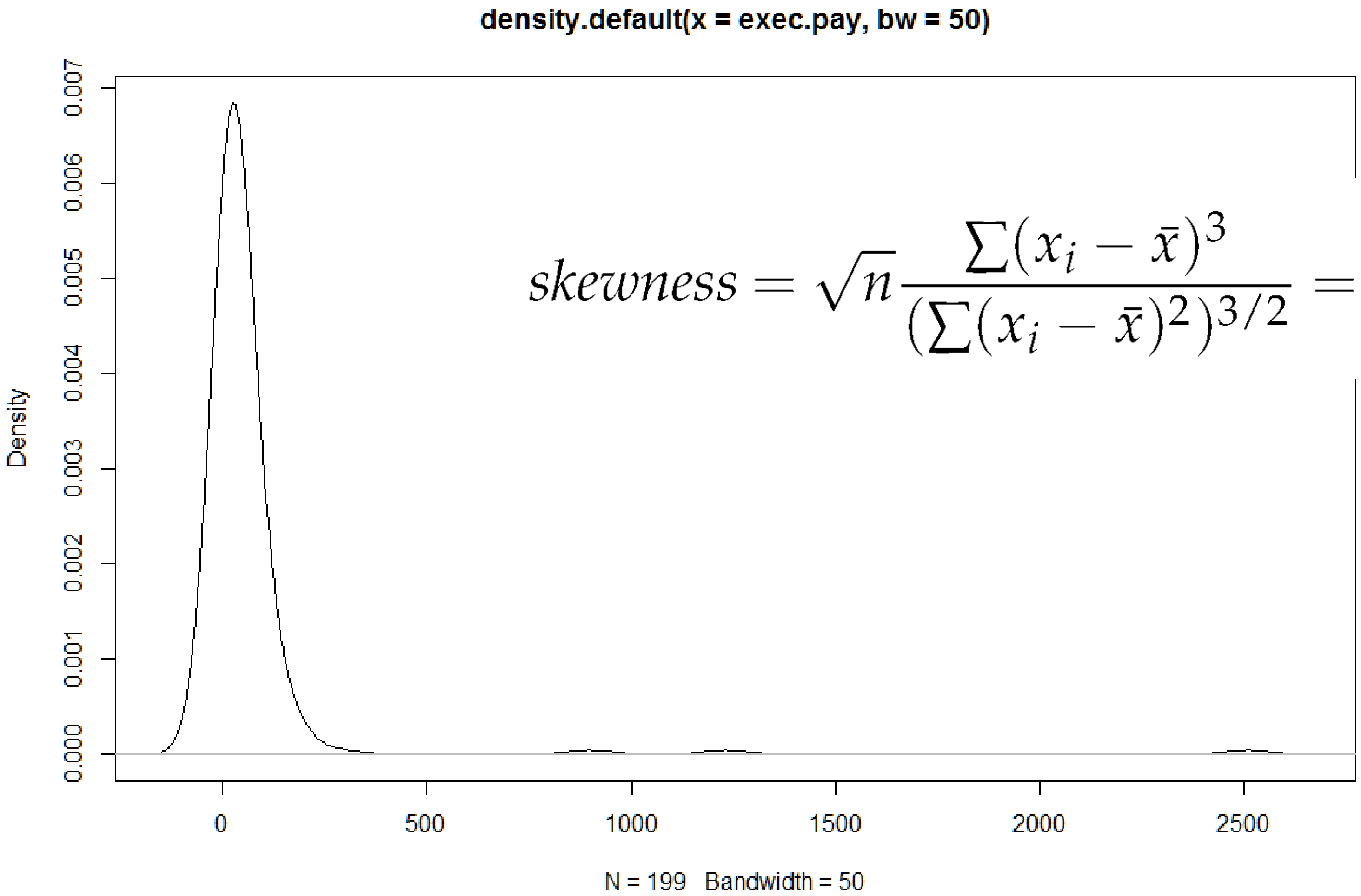
```
mean(exec.pay)
[1] 59.88945
median(exec.pay)
[1] 27
hist(exec.pay)
rug(exec.pay)

skewness(exec.pay)
[1] 9.578542
```




$$skewness = \sqrt{n}\frac{\sum(x_i - \bar{x})^3}{(\sum(x_i - \bar{x})^2)^{3/2}} = \frac{1}{n}\sum z_i^3$$

# Numerical Skewness

```
skewness(exec.pay)
[1] 9.578542
```




$$skewness = \sqrt{n}\frac{\sum(x_i - \bar{x})^3}{(\sum(x_i - \bar{x})^2)^{3/2}} = \frac{1}{n}\sum z_i^3$$

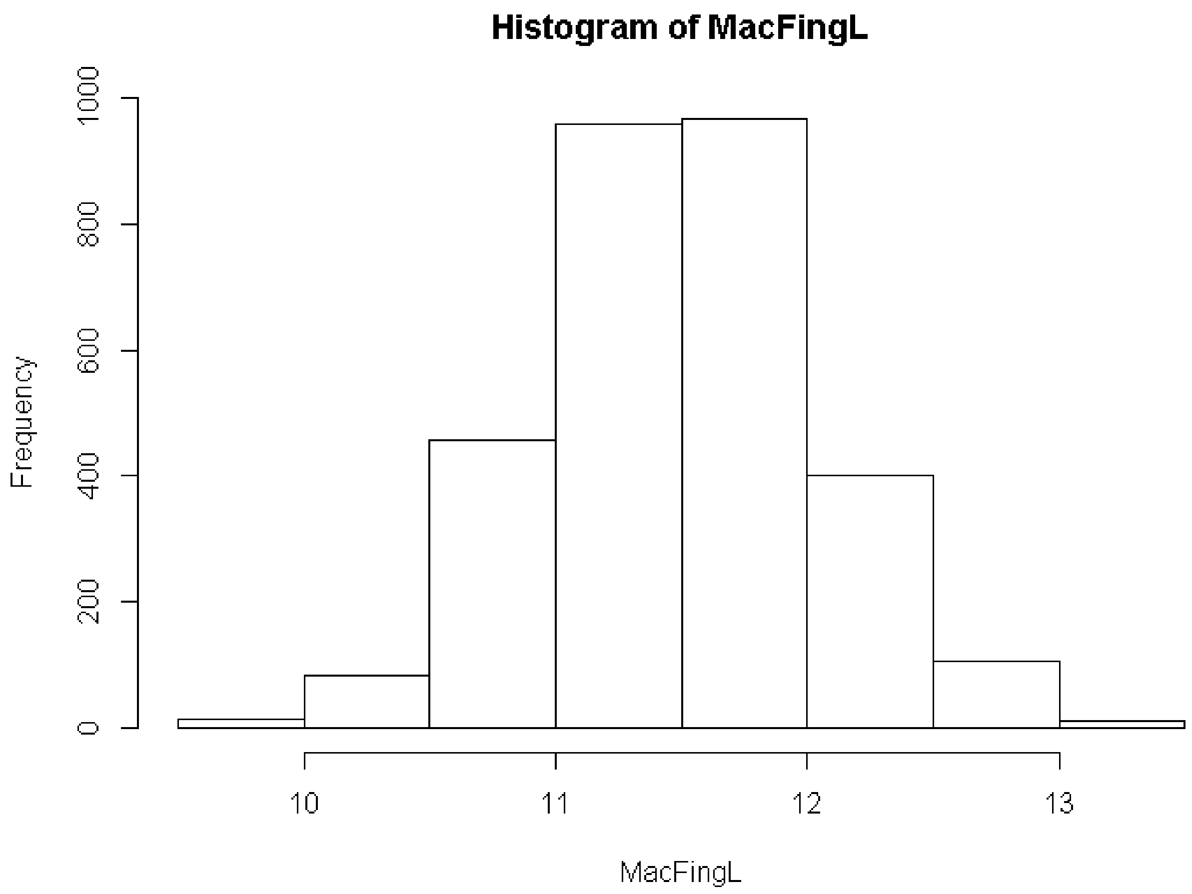
```
plot(density(exec.pay, bw = 50))
```

# Skewness

```
# look at the Macdonell finger length data
MacFingL <- scan("http://math.mercyhurst.edu/~sousley/STAT_139/data/MacFingL.vec");
```

## Look at finger length

```
mean(MacFingL)
[1] 11.45

median(MacFingL)
[1] 11.45

skewness(MacFingL)
[1] 0.05155381
```

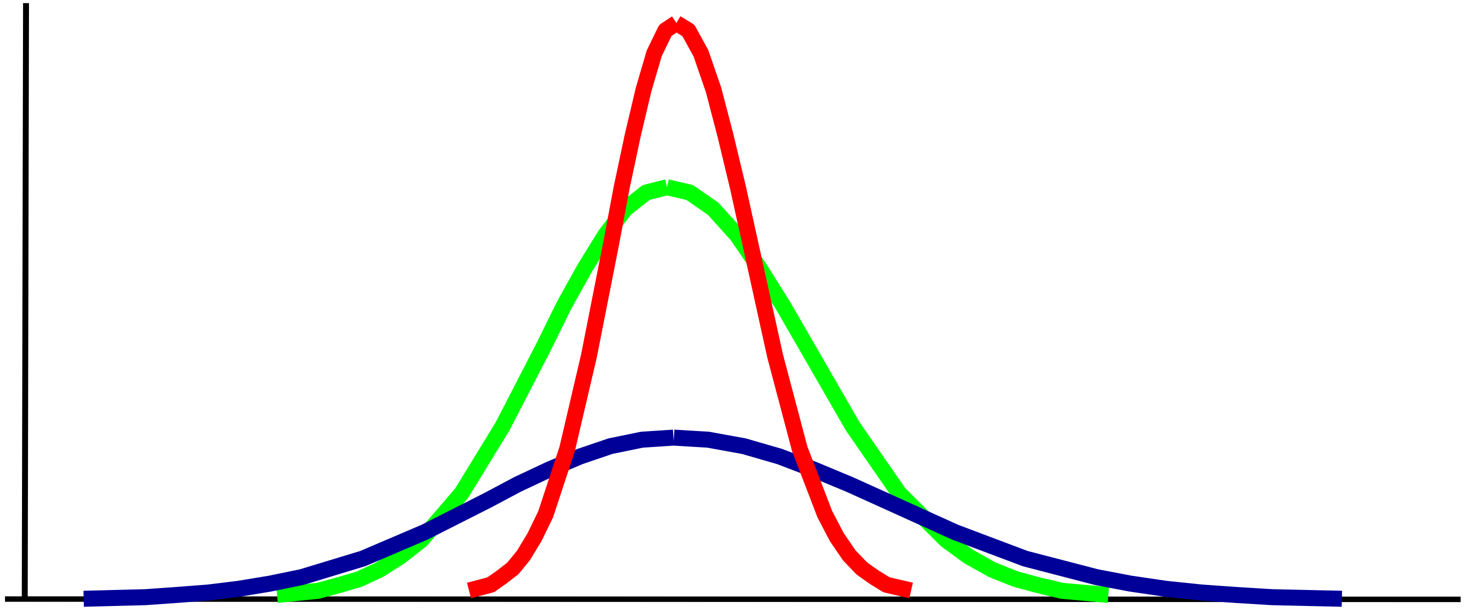



$$skewness = \sqrt{n}\frac{\sum(x_i - \bar{x})^3}{\left(\sum(x_i - \bar{x})^2\right)^{3/2}} = \frac{1}{n}\sum z_i^3$$

# Kurtosis (concentration)

**leptokurtotic  = POSITIVE kurtosis**

**OBH**



**GOL = standard normal kurtosis = 0**

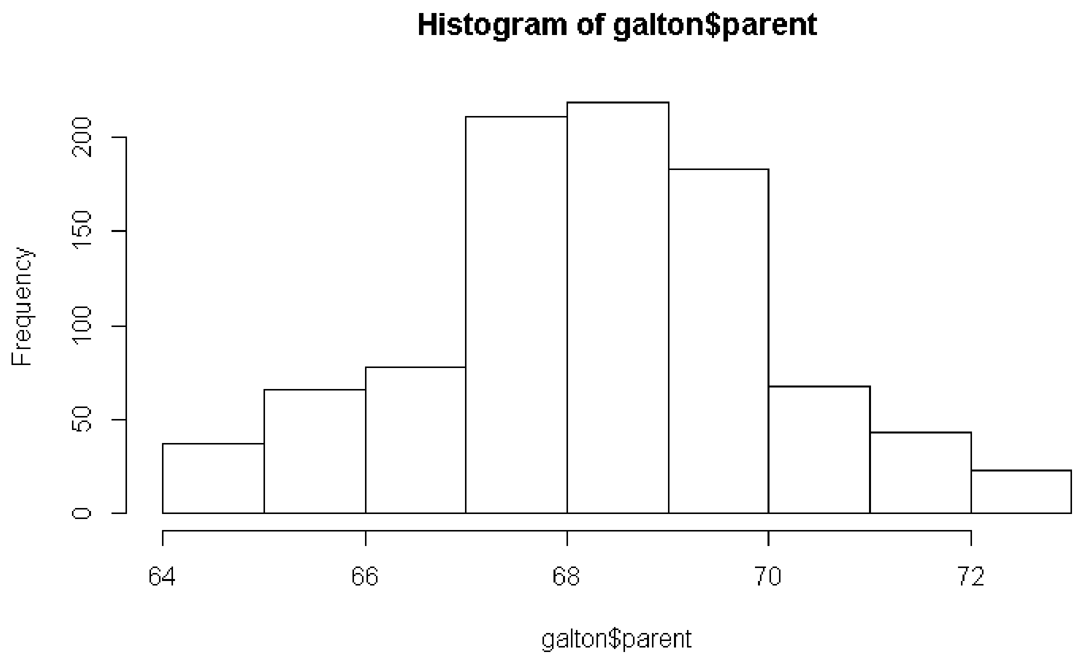
**ASB**

**p l a t y k u r t o t I c = negative kurtosis**

$$kurtosis = n\frac{\sum(x_i-\bar{x})^4}{\left(\sum(x_i-\bar{x})^2\right)^2} - 3 = \frac{1}{n}\sum z_i^4 - 3$$

# Numerical Kurtosis

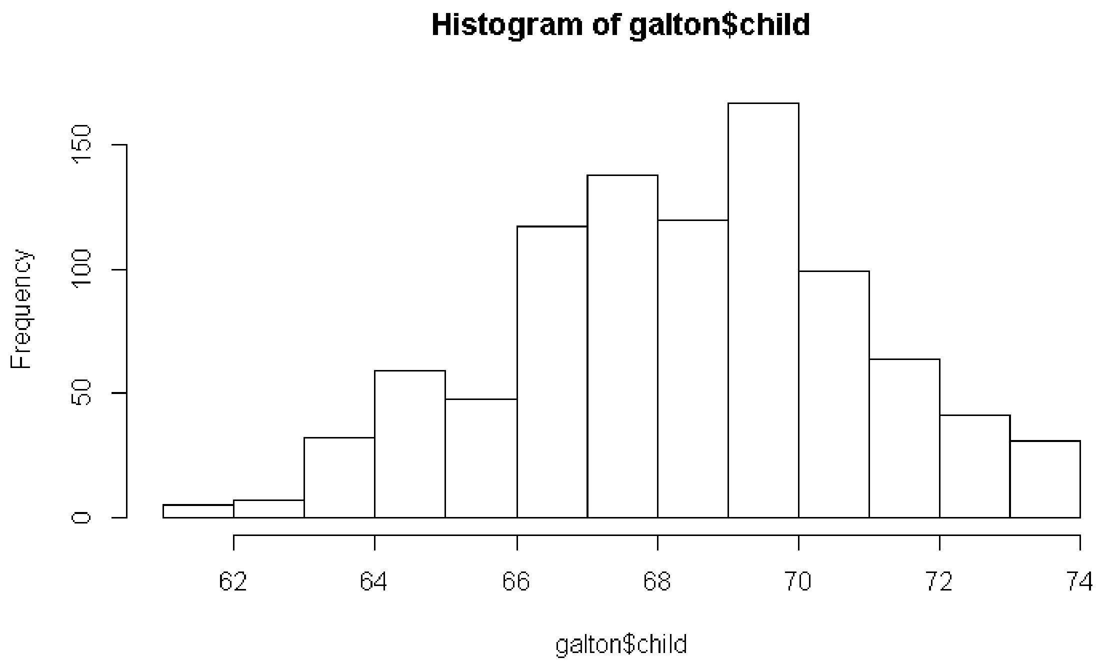
```
# look at statures of PARENTS in Galton's data
galton <- read.csv("http://math.mercyhurst.edu/~sousley/STAT_139/data/galton.csv");

mean(galton$parent)
[1] 68.30819
median(galton$parent)
[1] 68.5
skewness(galton$parent)
[1] -0.03503614

kurtosis(galton$parent)
[1] 0.05104267

# get the range
diff(range(galton$parent))
[1] 9
sd(galton$parent)
[1] 1.787333

hist(galton$parent)
```




$$kurtosis = n\frac{\sum(x_i - \bar{x})^4}{(\sum(x_i - \bar{x})^2)^2} - 3 = \frac{1}{n}\sum z_i^4 - 3$$

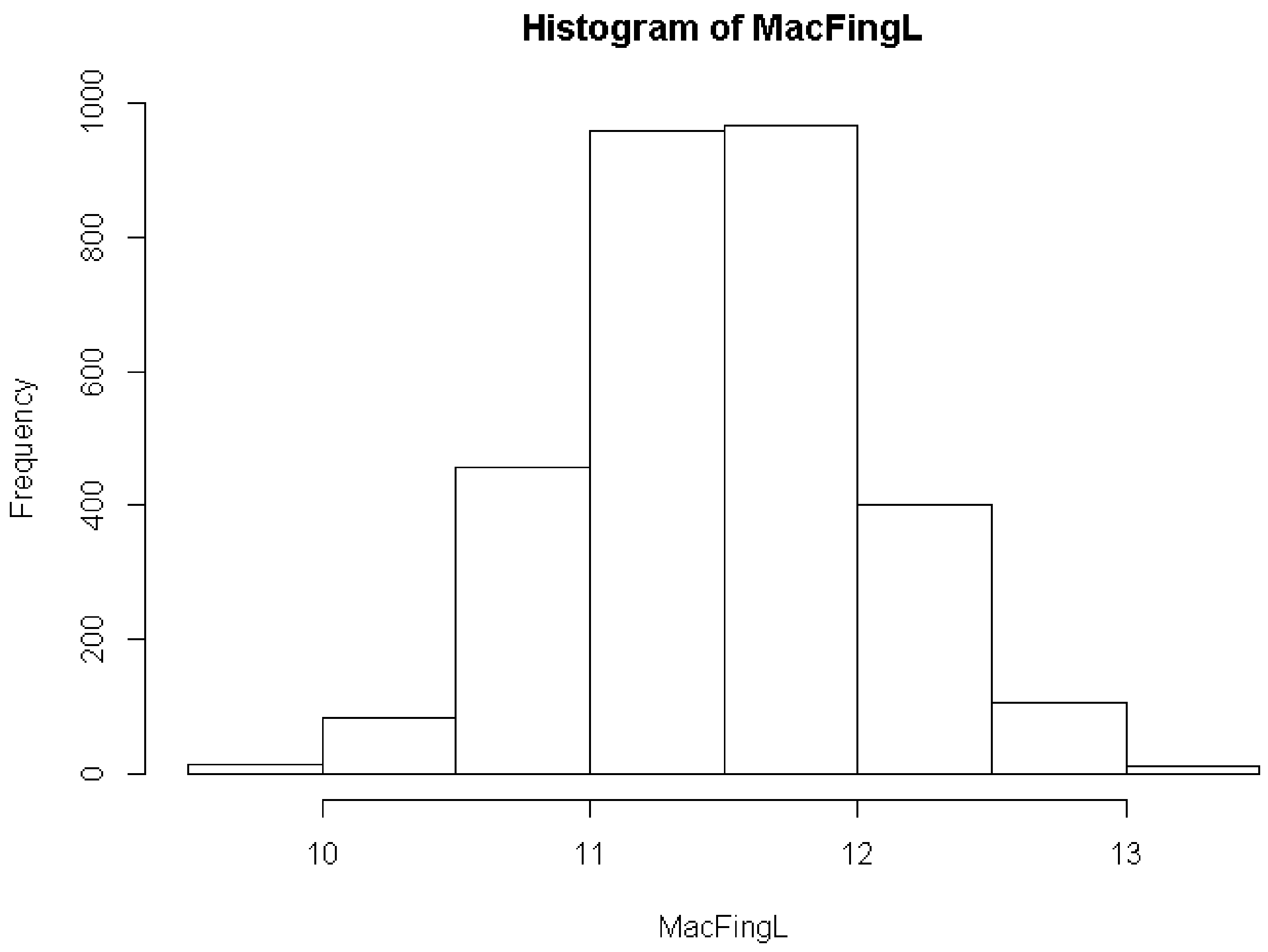
# Numerical Kurtosis

```
# look at statures of children this time in Galton's data

mean(galton$child)
[1] 68.08847
median(galton$child)
[1] 68.2
skewness(galton$child)
[1] -0.08762607

kurtosis(galton$child)
[1] -0.3500438

# get the range
diff(range(galton$child))
[1] 12
sd(galton$child)
[1] 2.517941

hist(galton$child)
```



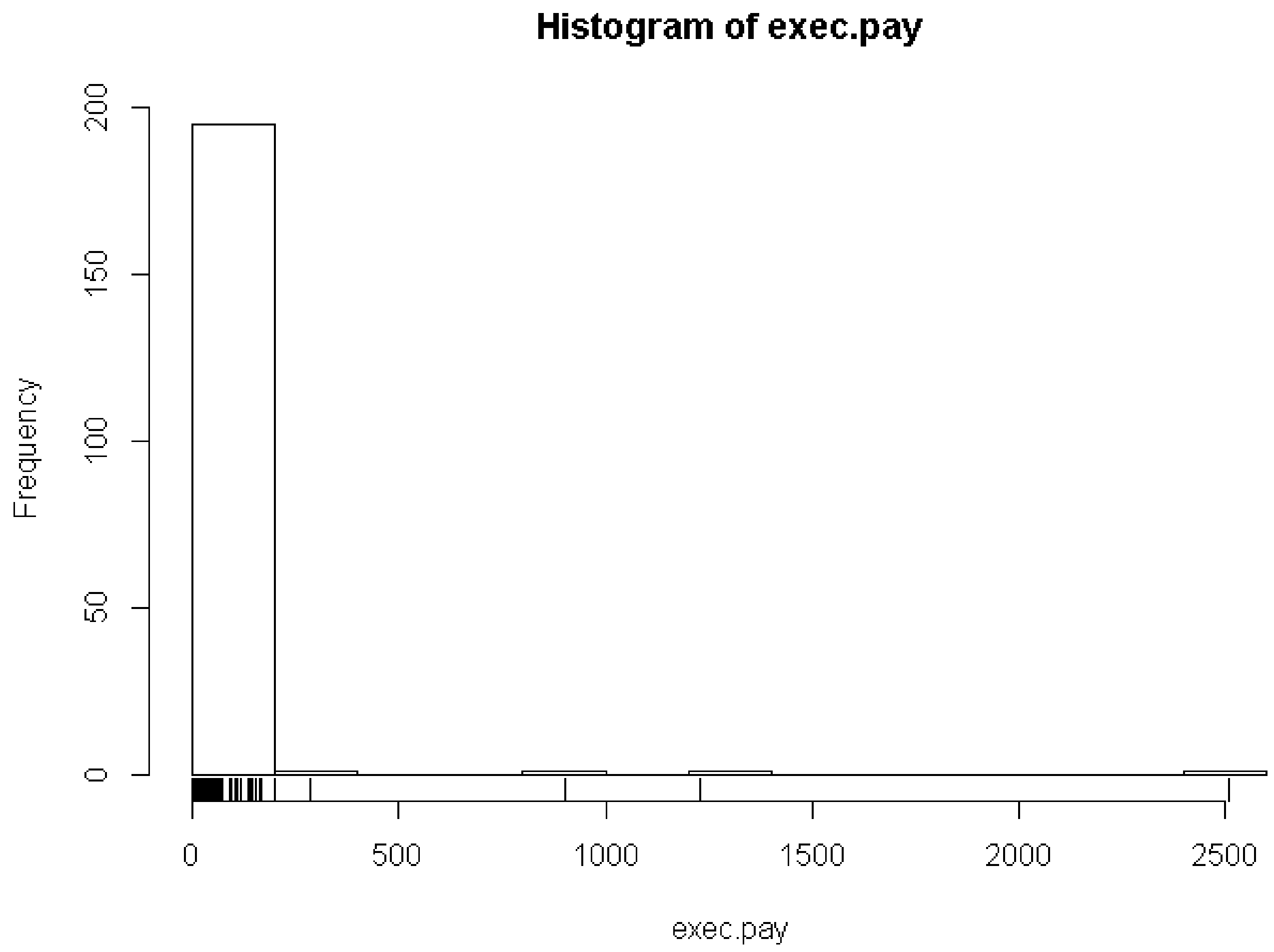

$$kurtosis = n\frac{\sum(x_i-\bar{x})^4}{(\sum(x_i-\bar{x})^2)^2} - 3 = \frac{1}{n}\sum z_i^4 - 3$$

# Kurtosis in Finger data




```
kurtosis(MacFingL)
[1] 0.1068209
```

# Kurtosis in Income data




```
skewness(exec.pay)
[1] 9.578542
kurtosis(exec.pay)
[1] 102.064
```

# Homework 4

**Problems from VCH2-2014**

**2.4 (think!** Use c only when rep, :, or seq will not work.**)**

**2.13**

**2.14 (install package HistData to get Arbuthnot)**

**2.18**

**and...**

# Homework 4 (continued)

## Calculate the Standard Deviation of some measurements in *R*

```
Stdata <- c(10,10,15,20,35,40,40,40,45,45,50)
```

**Problem E1:**
**Calculate the mean and standard deviation using ONLY the sum and length functions as well as exponentiation, multiplication, addition, division, and subtraction.**

**- use as few regular numbers as possible**
**- assigning to variables will make your code more readable**
**(and easier to program)**

| |
|---|
| 10 |
| 10 |
| 15 |
| 20 |
| 35 |
| 40 |
| 40 |
| 40 |
| 45 |
| 45 |
| 50 |

**Due Thursday Feb 8 before class**
**- follow same homework format**