Math 183
Statistical Methods

Eddie Aamari
S.E.W. Assistant Professor

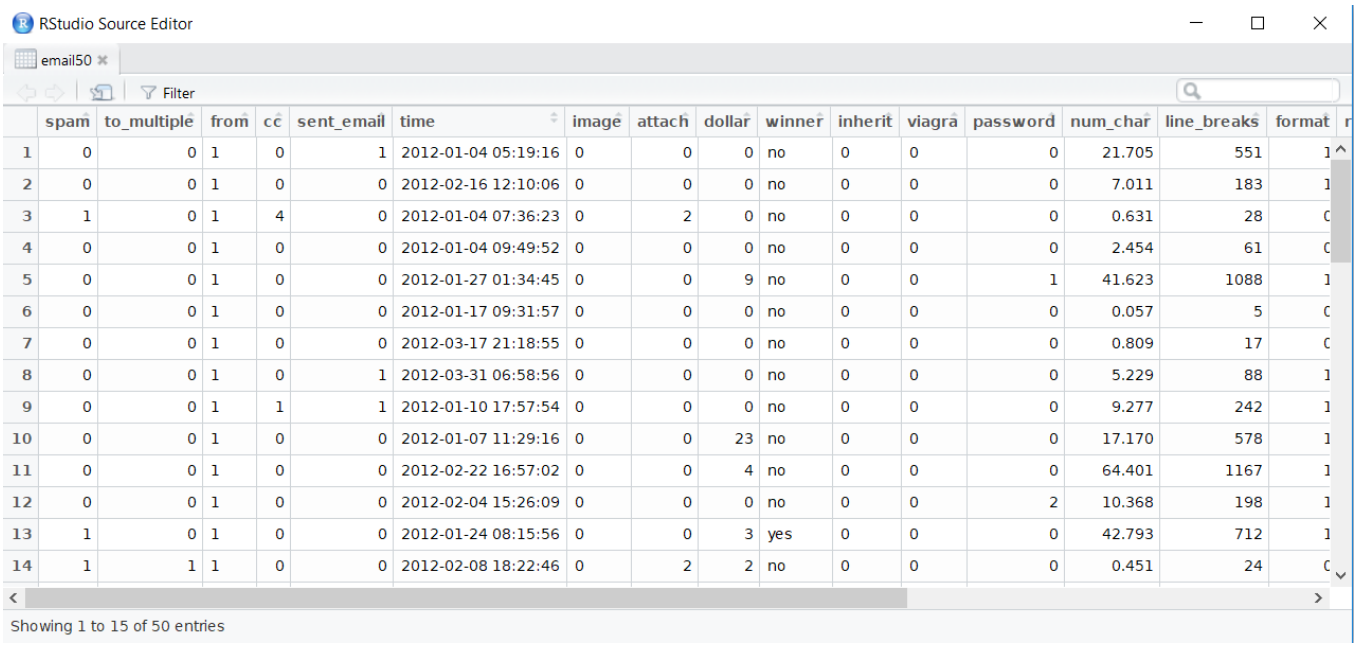
eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today: Chapter 2 (end)

- Mean, median, quantiles
- Box plots
- Measures of spread
- Data transformations

Summarizing Data

Real life data are complex and hardly understandable at first sight.



RStudio Source Editor

email50

Filter

	spam	to_multiple	from	cc	sent_email	time	image	attach	dollar	winner	inherit	viagra	password	num_char	line_breaks	format	r
1	0	0	1	0	1	2012-01-04 05:19:16	0	0	0	no	0	0	0	21.705	551	1	
2	0	0	1	0	0	2012-02-16 12:10:06	0	0	0	no	0	0	0	7.011	183	1	
3	1	0	1	4	0	2012-01-04 07:36:23	0	2	0	no	0	0	0	0.631	28	0	
4	0	0	1	0	0	2012-01-04 09:49:52	0	0	0	no	0	0	0	2.454	61	0	
5	0	0	1	0	0	2012-01-27 01:34:45	0	0	9	no	0	0	1	41.623	1088	1	
6	0	0	1	0	0	2012-01-17 09:31:57	0	0	0	no	0	0	0	0.057	5	0	
7	0	0	1	0	0	2012-03-17 21:18:55	0	0	0	no	0	0	0	0.809	17	0	
8	0	0	1	0	1	2012-03-31 06:58:56	0	0	0	no	0	0	0	5.229	88	1	
9	0	0	1	1	1	2012-01-10 17:57:54	0	0	0	no	0	0	0	9.277	242	1	
10	0	0	1	0	0	2012-01-07 11:29:16	0	0	23	no	0	0	0	17.170	578	1	
11	0	0	1	0	0	2012-02-22 16:57:02	0	0	4	no	0	0	0	64.401	1167	1	
12	0	0	1	0	0	2012-02-04 15:26:09	0	0	0	no	0	0	2	10.368	198	1	
13	1	0	1	0	0	2012-01-24 08:15:56	0	0	3	yes	0	0	0	42.793	712	1	
14	1	1	1	0	0	2012-02-08 18:22:46	0	2	2	no	0	0	0	0.451	24	0	

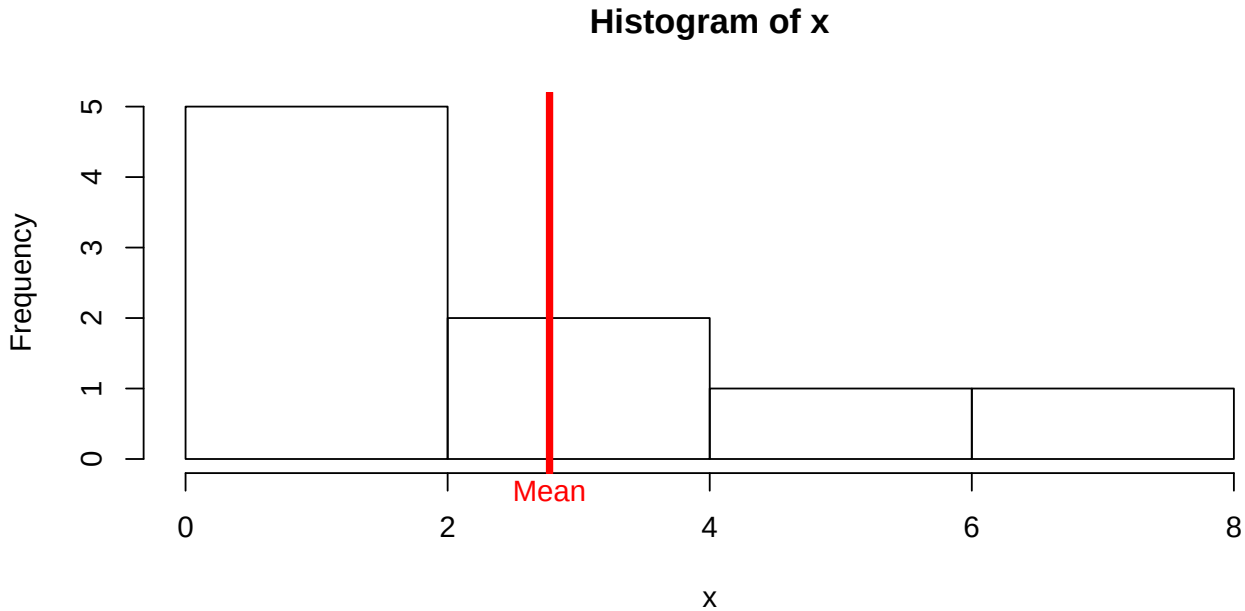
Showing 1 to 15 of 50 entries

Need to summarize them!

For numerical data, we use their **center** and **spread**.

These are called “statistics”.

Data Centrality: Mean



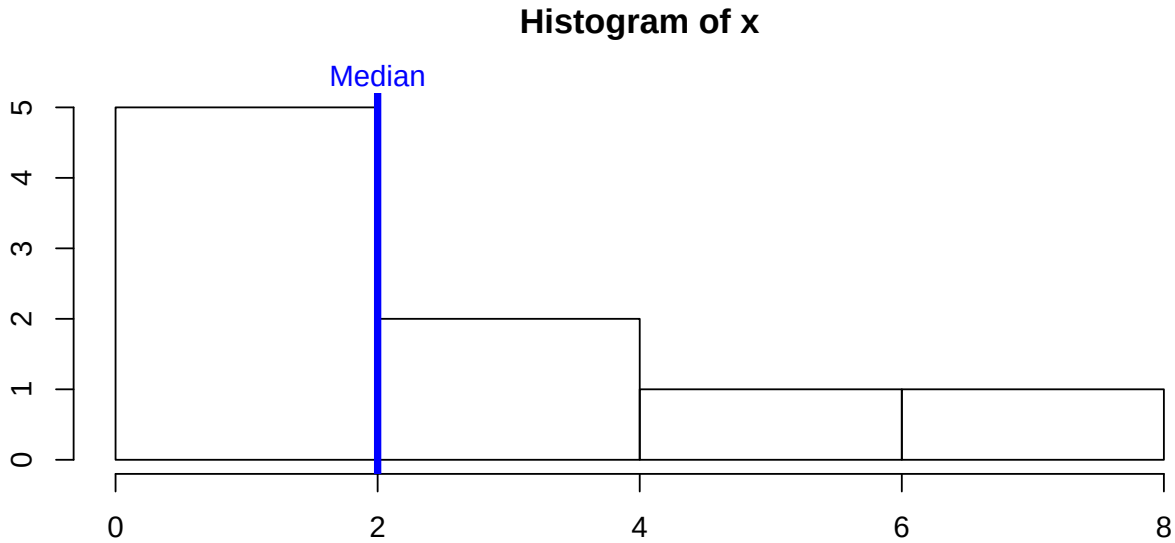
The **mean** of $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

```
> x = c(0,2,1,3,8,6,4,0,1)
> mean(x)
[1] 2.777778
```

mean(x) is the balance point on the histogram,
the value that sets the torque to zero.

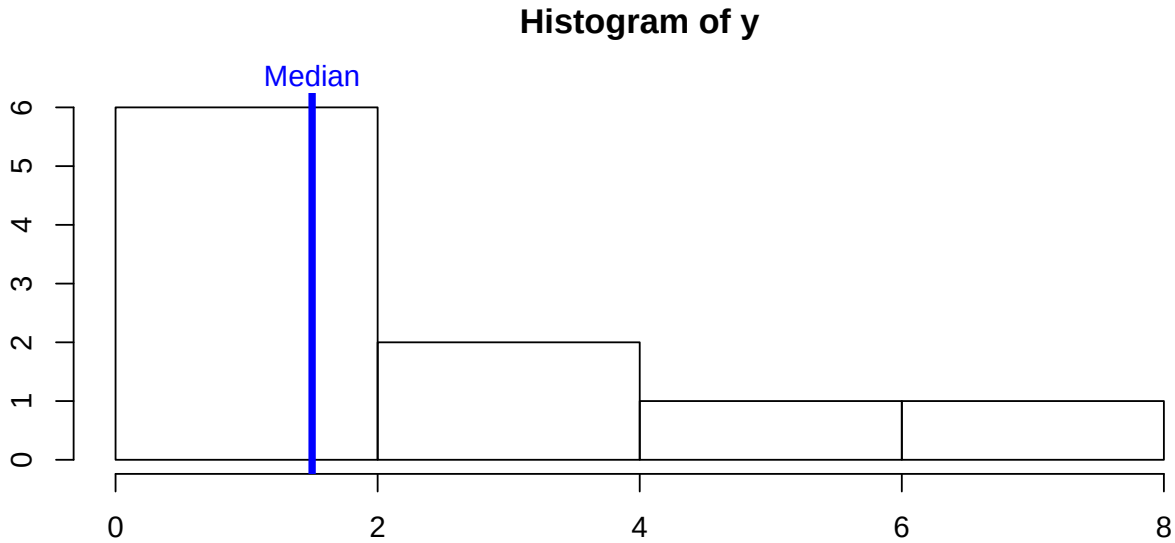
Data Centrality: Median



The **median** of $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the value that splits the sample into two halves.

```
> x = c(0,2,1,3,8,6,4,0,1)
> sort(x)
 [1] 0 0 1 1 2 3 4 6 8
> median(x)
 [1] 2
```

Data Centrality: Median

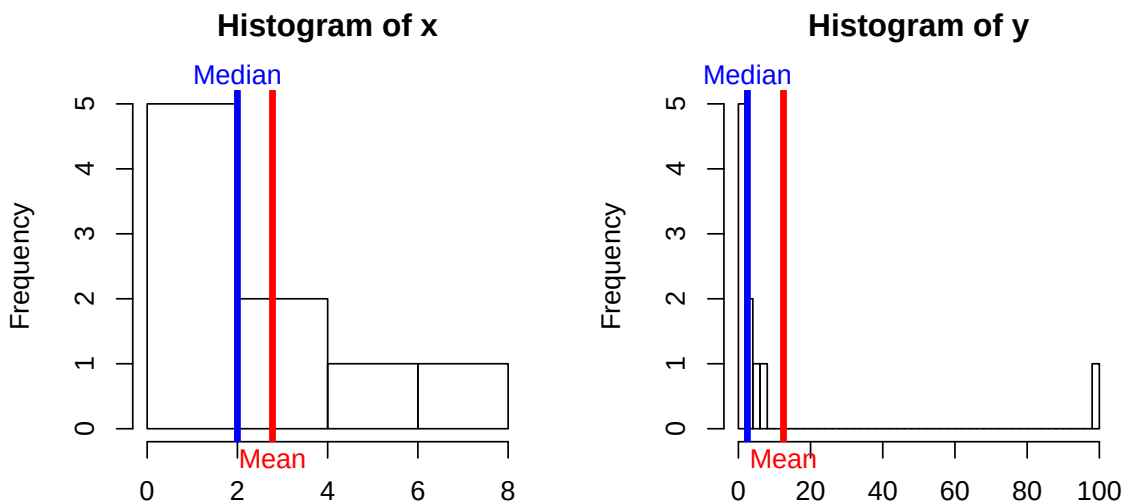


If sample size is even, take the average of the two "middle" points.

```
> y = c(0,2,1,3,8,6,4,0,1,1)
> sort(y)
 [1] 0 0 1 1 1 2 3 4 6 8
> median(y)
[1] 1.5
```

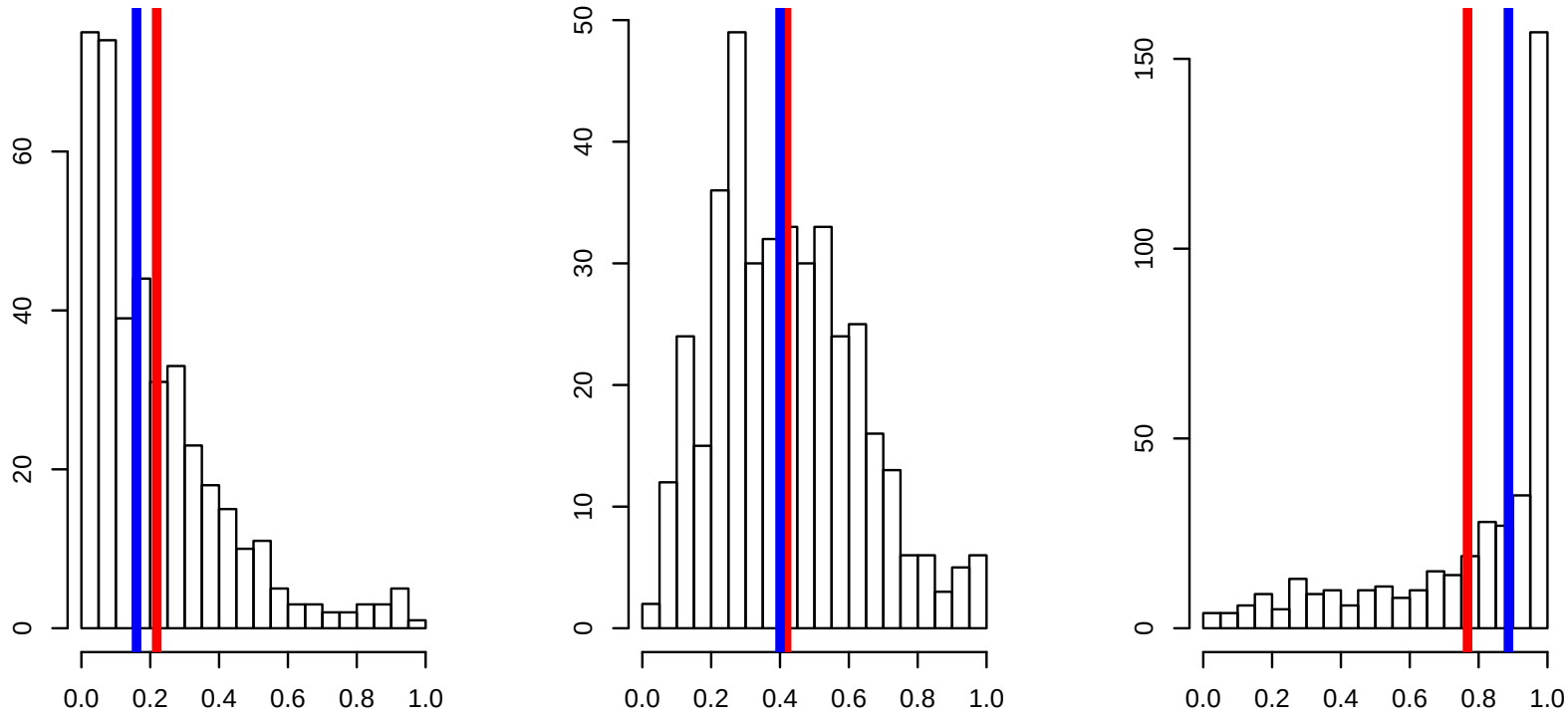
Median VS Mean: Outliers

```
> x = c(0,2,1,3,8,6,4,0,1)
> mean(x);median(y)
[1] 2.777778
[1] 1.5
> y = c(0,2,1,3,8,6,4,0,1,100)
> mean(y);median(y)
[1] 12.5
[1] 2.5
```



Median VS Mean: Skew

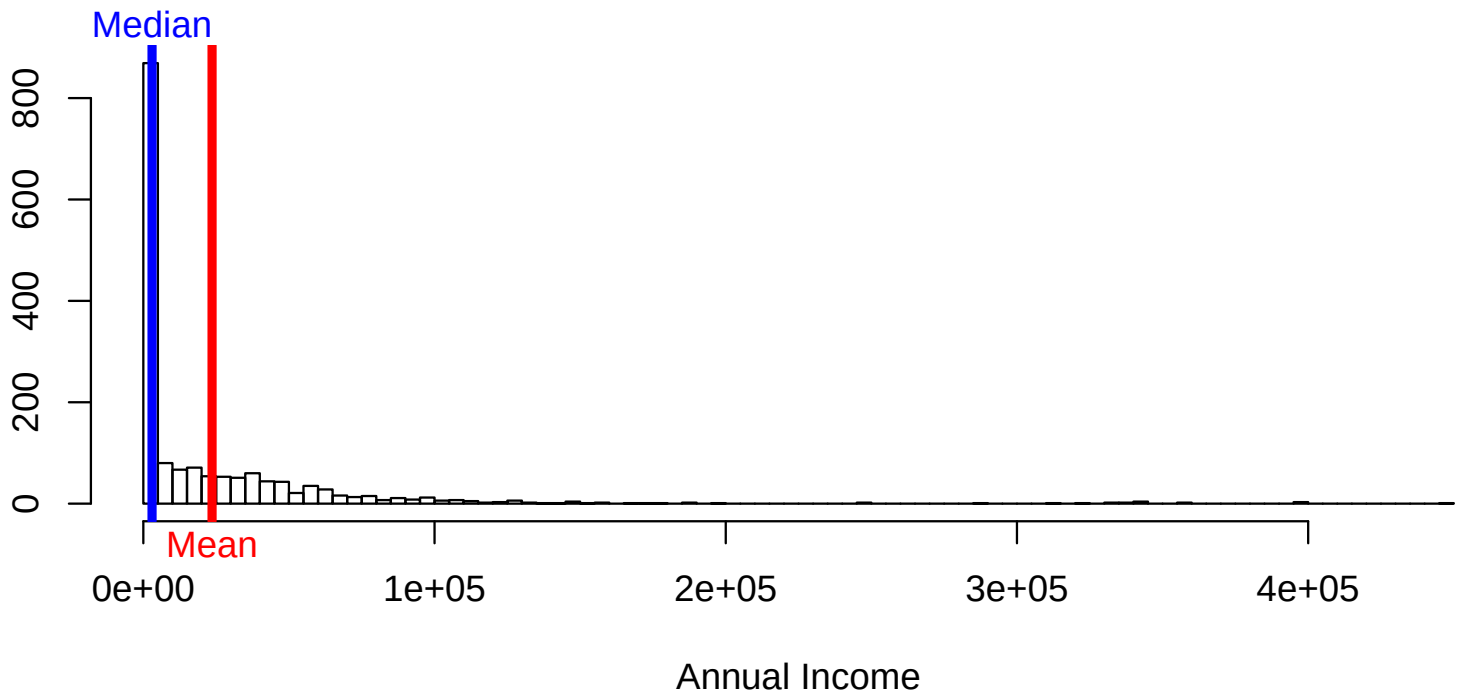
Where are the mean and the median located?



— Median — Mean

Right/Left skewness inform on the respective positions of the median and the mean?

Median VS Mean: Practice



Median VS Mean

Moral:

- The median is robust to outliers and skew.
- The mean is not.

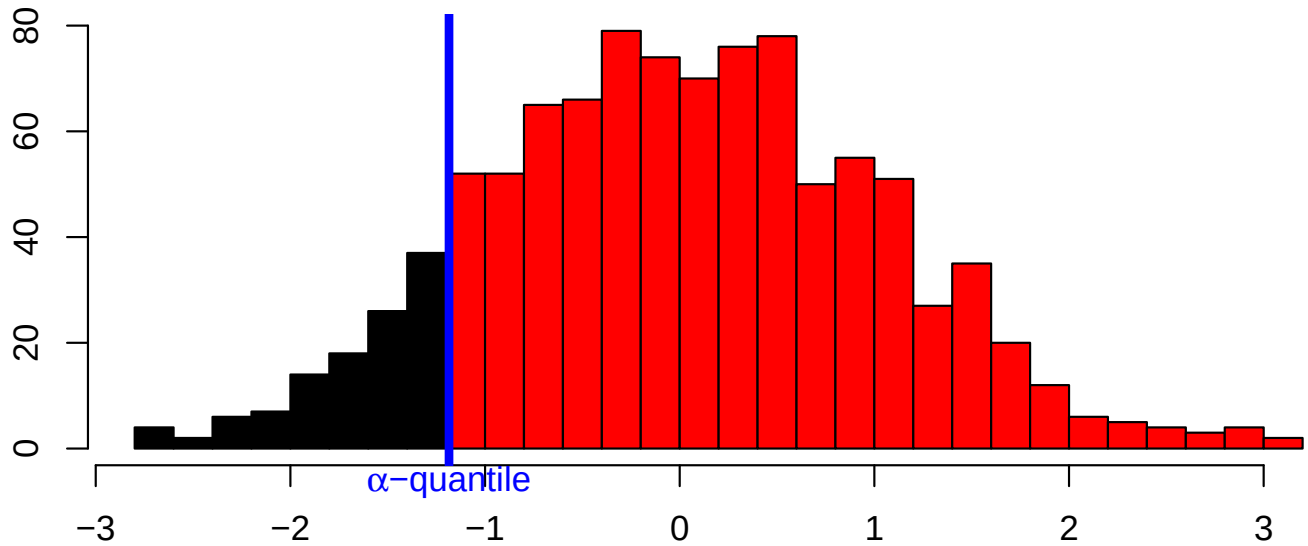
But:

- The mean easier to handle and much more popular.

Data Position: Quantiles

For $0 \leq \alpha \leq 1$, the α-**quantile** is the value that splits data into:

- Proportion α on the left
- Proportion $(1 - \alpha)$ on the right

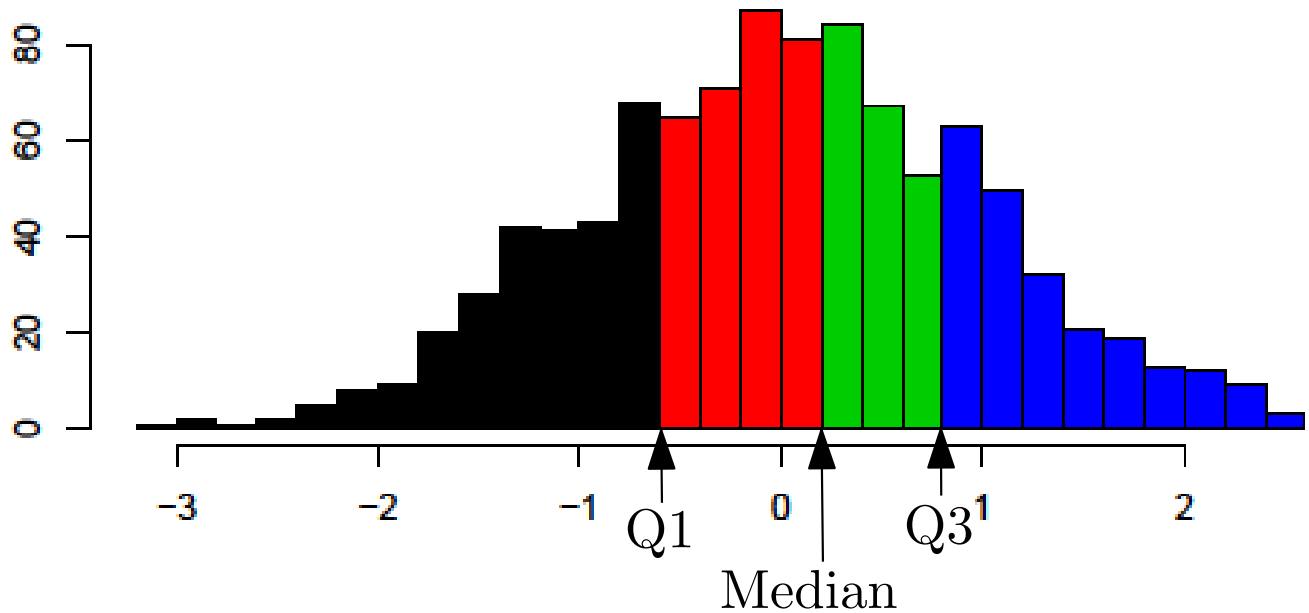


R function: `quantile(x,alpha)`

Data Position: Quantiles

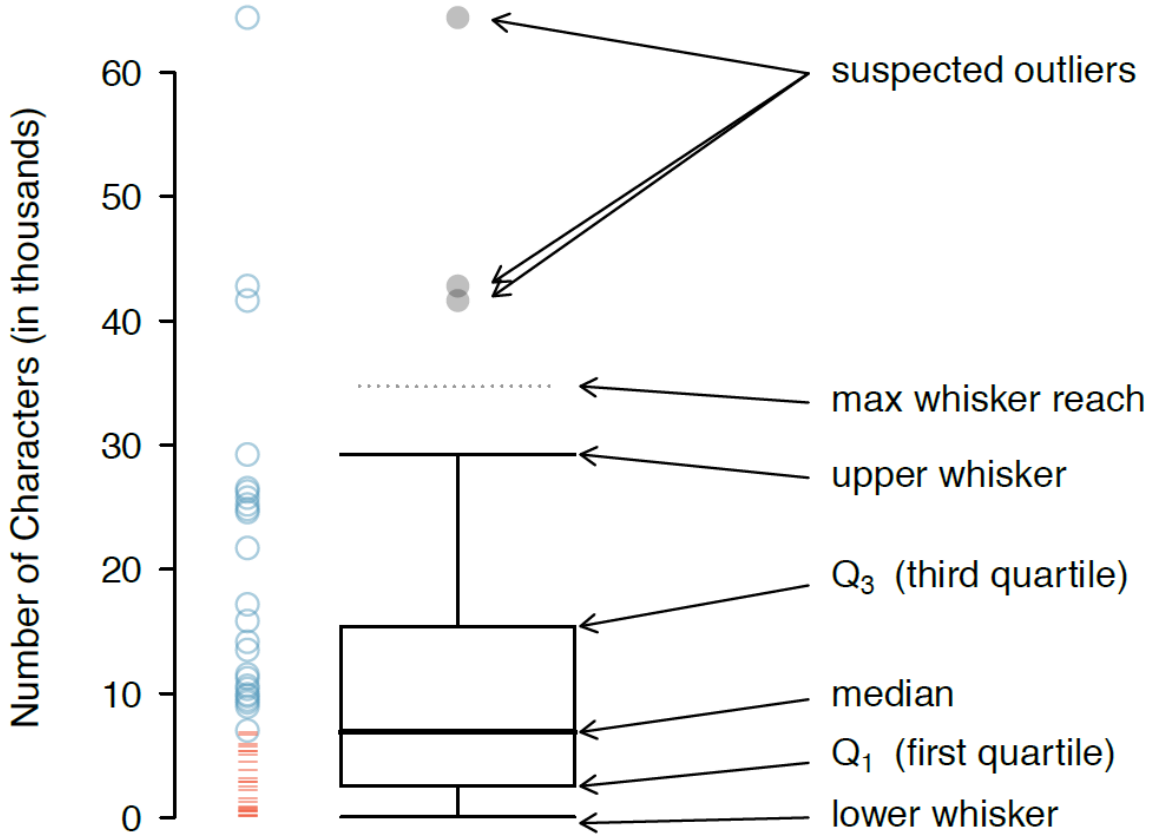
Some quantiles have special names:

- 25%-quantile: 1st quartile (Q2)
- 50%-quantile: median
- 75%-quantile: 3rd quartile (Q3)



Box Plot

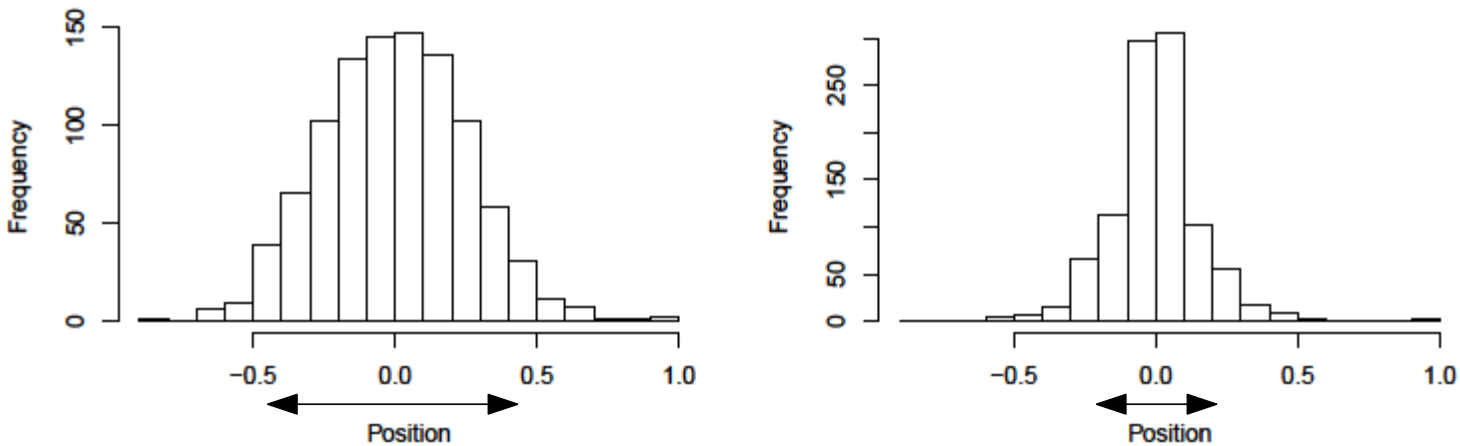
A **box plot** (or whisker plot) is a visualization of these quantiles.



R function: `boxplot(x)`.

Spread of a Distribution

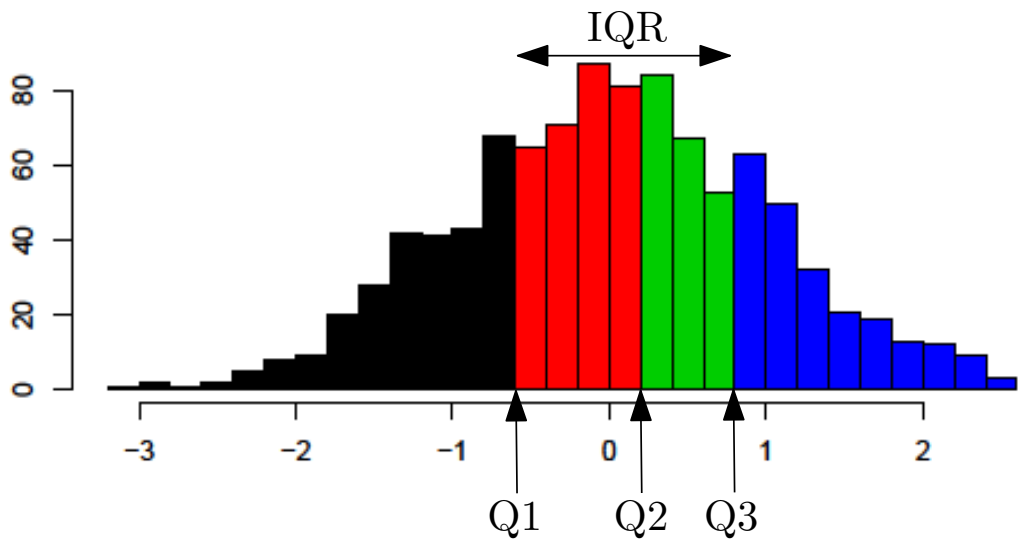
A centrality statistics is not sufficient to describe data fully.



We would like an indicator of where “most of the data” lie.

Spread of a Distribution

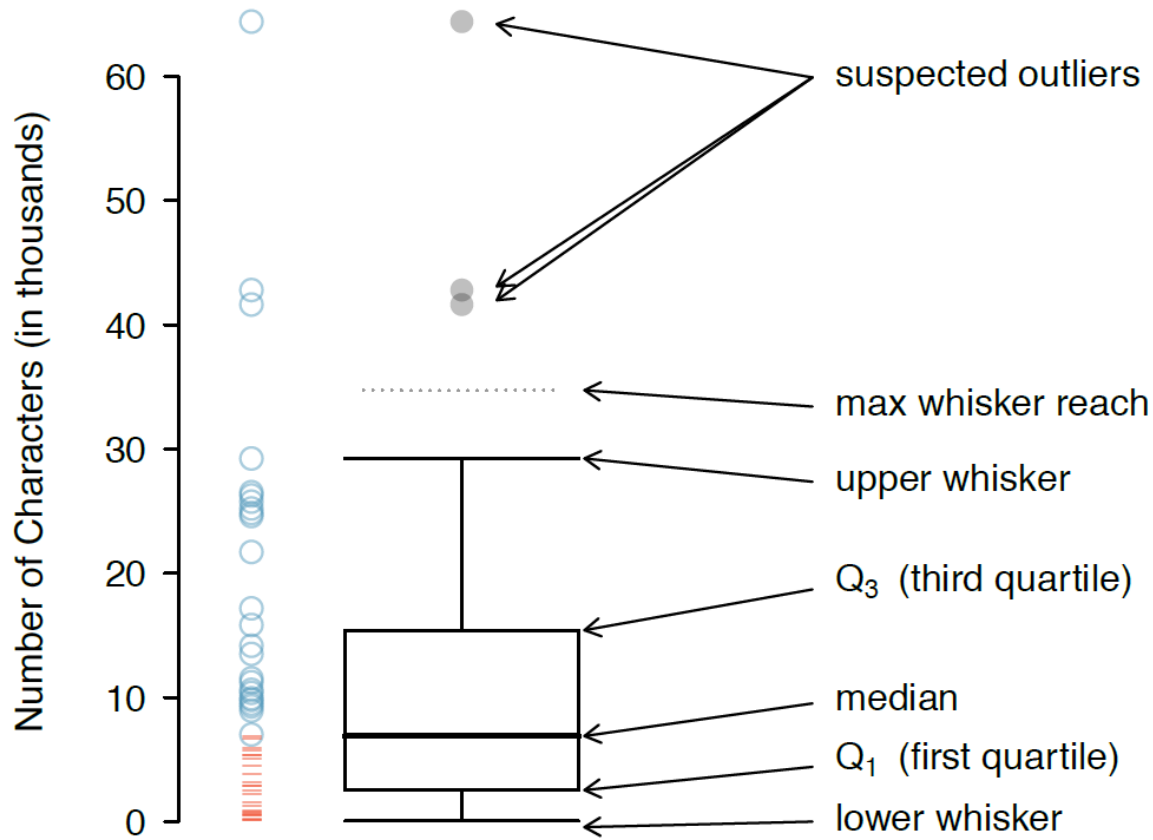
- **Range** $= \max(data) - \min(data)$
  - Easy to compute
  - Not resistant to outliers; only takes into account two points (extremes)
- **Interquartile Range (IQR)** $= Q3 - Q1$.



- Resistant to outliers and skew.
- Not very popular; Hard to handle and to compute.

Back to Box Plots

The range and the IQR can be vizualized on a box plot.



Standard Deviation

The sample **standard deviation** of **x** is defined as:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

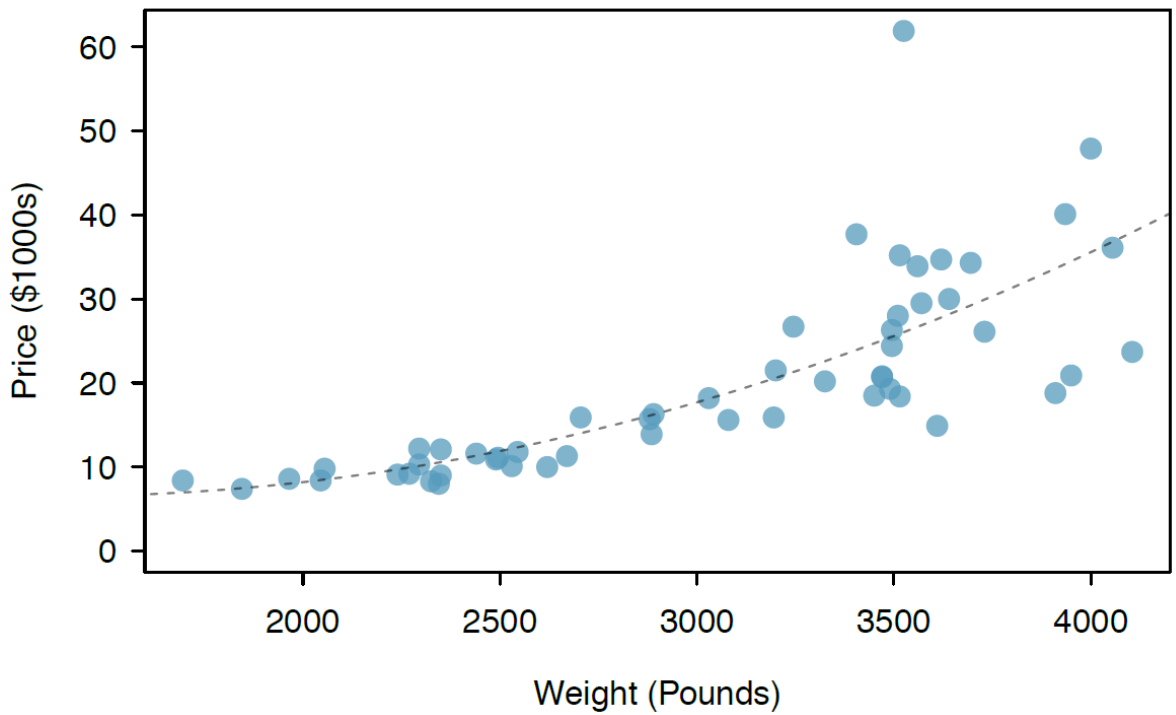
- Sum of squares: to get positive values
- Square root: to undo squaring action and have s with same units as the x_i's.
- $(n-1)$ instead of n: explained later in the course.

R function: `sd(x)`.

- Very popular notion of spread; easy to handle mathematically
- Sensitive to outliers and skew; hard to explain to non-statisticians.

Two Numerical Variables: Scatter Plots

When we wantsto study jointly two numerical variables, we can use a **scatter plot**. (R function `plot(var1,var2)`)



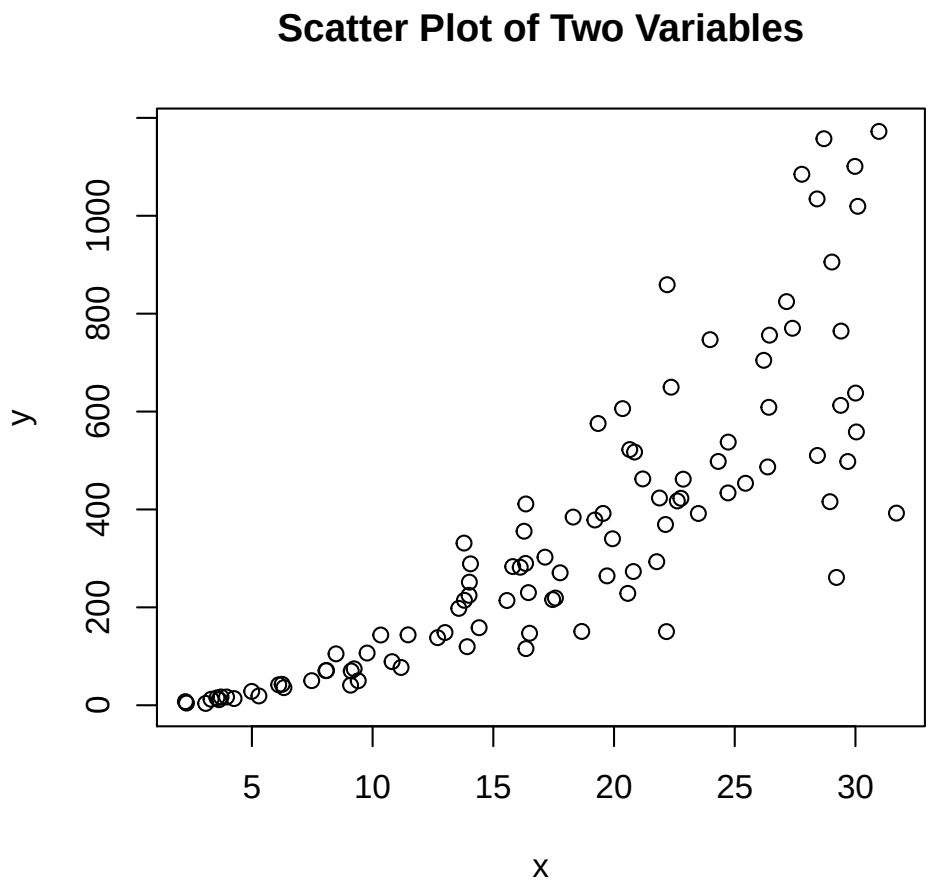
A scatterplot of the price and the weight of 54 cars.

A pair of variables are either related in some way (**associated**) or not (**independent**).

Caution: Association is NOT Causality!

Scatter Plots

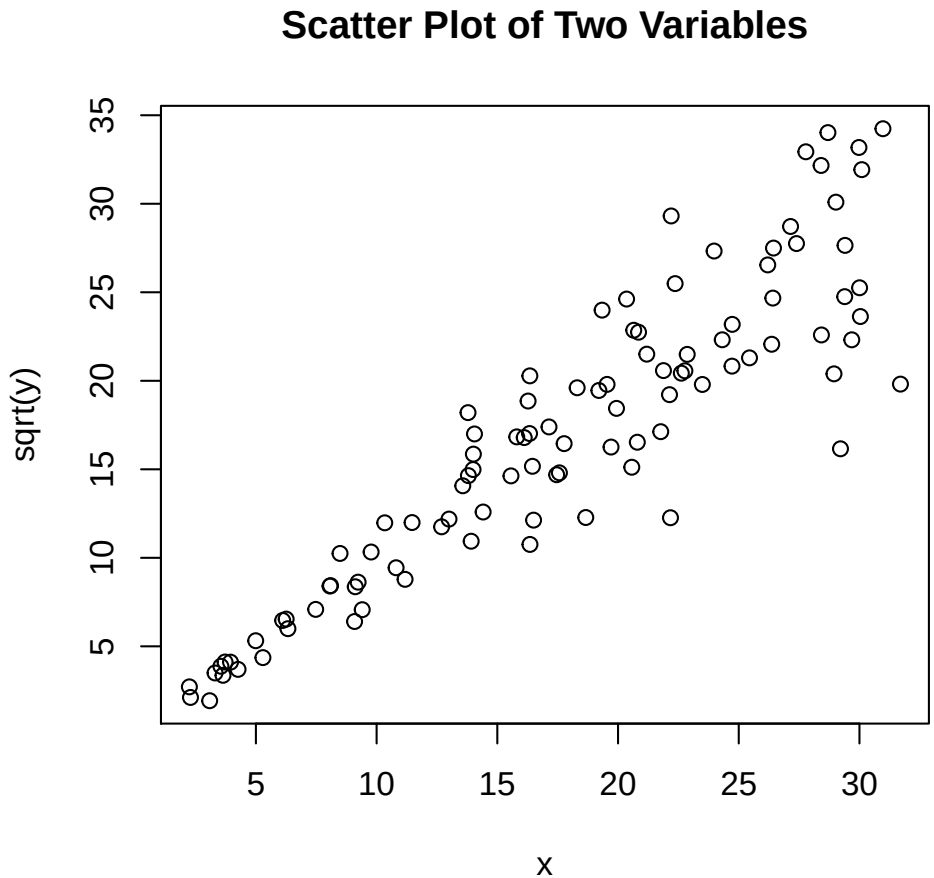
The sense of the association (positive or negative) can usually be visualized...



... But their precise relation may not be so clear!

Data Transformation: Use 1

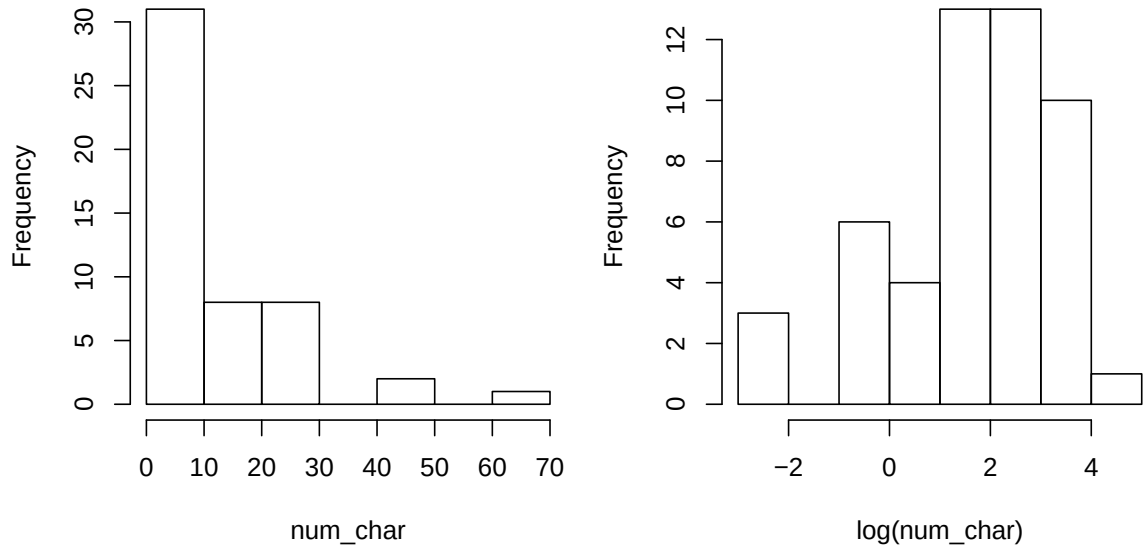
Applying a function to either one or the two variables may help catching their precise relation.



Plotting $(x, \sqrt{y})$ show a linear behavior: $y \simeq (ax + b)^2$ for some a and b to be estimated. (see Regression, Chapter 7).

Data Transformation: Use 2

Transformed data are sometimes easier to work with. Indeed, the transformed data are much less skewed and outliers are usually less extreme.

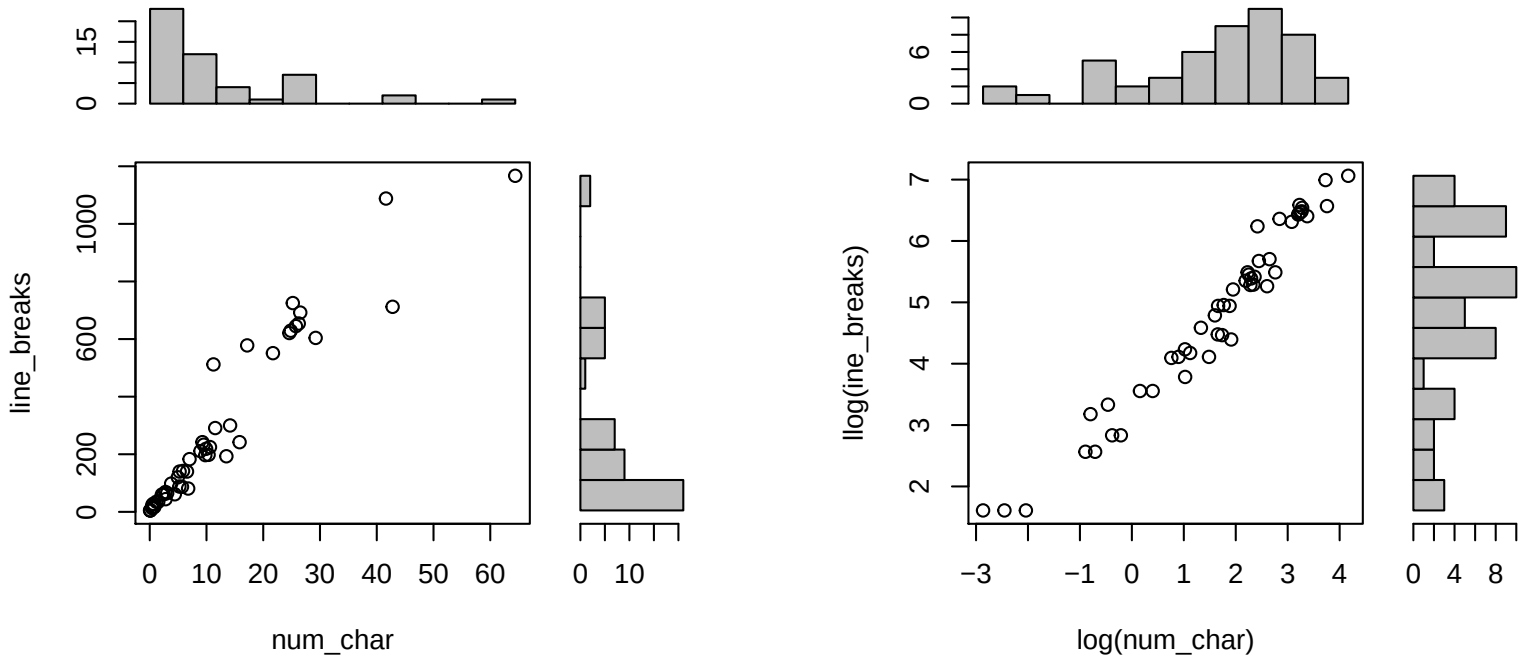


Standard data transformation functions:

- logarithm: when (positive) data skewed right with many values close to 0, and outliers
- arcsine: for data in the interval $[-1, 1]$. Mostly used in biology and chemistry.

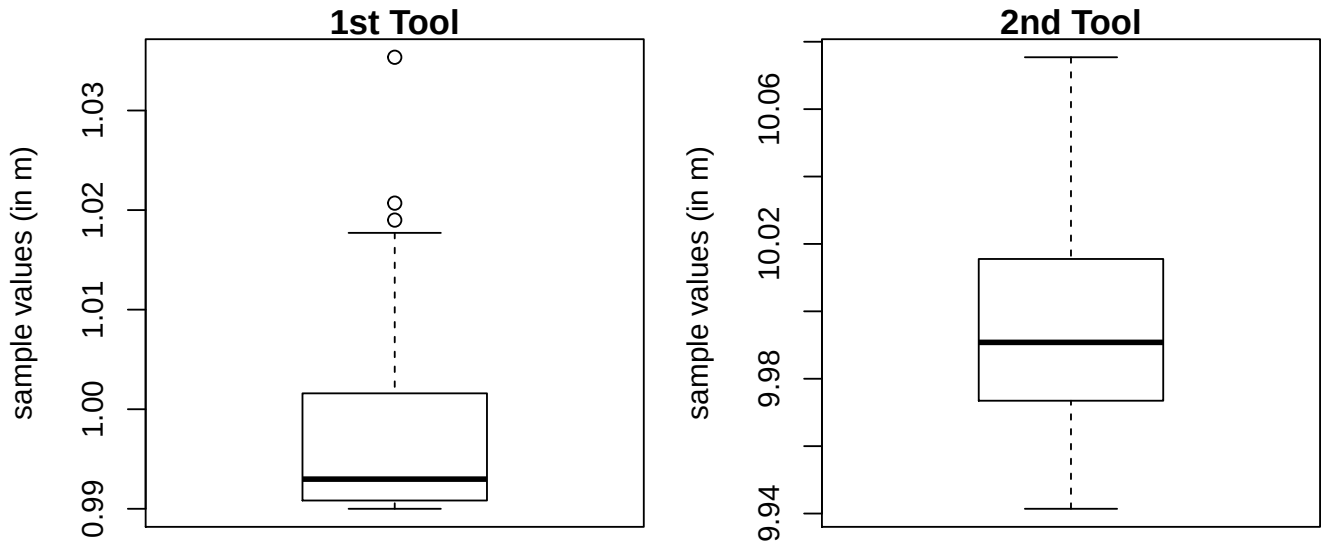
Data Transformation: Use 2

Using the "log" transformation on two variables.



Exercise

Calibrating two measurement tools on a 10 meters benchmark, you collect 50 repeated measures of the same length in vectors **x** (1st tool) and **y** (2nd tool). Which tool is best? Give arguments.




```
> mean(x);sd(x) #First tool
 [1] 0.9981183
 [1] 0.01049968
> mean(y);sd(y) #Second tool
 [1] 9.993498
 [1] 0.02855839
```