EPHE 357

Descriptive Statistics

Measures of Central Tendency and Variability

Population:
the set of things (people, rats, brains, neurons, devices, etc.) to which you wish your findings to apply.

Sample:
the hopefully representative members of the population from which you actually collect data.

Variable:
a thing (a concept, property, etc.) that we can name and either qualify (by assigning sub-names or adjectives) or quantify (by counting or measuring).

Descriptive Statistics:

mean, median, mode, variance, standard deviation…

Inferential Statistics:

t-tests, ANOVA, ANCOVA, regression, ICA…

Describing and Visualizing Data

Plot your data!!!

Descriptors: Where is it?

Mean

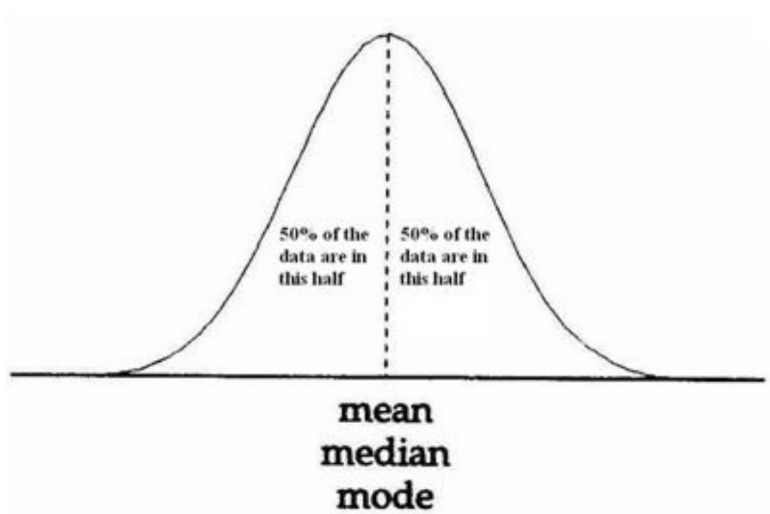
Median

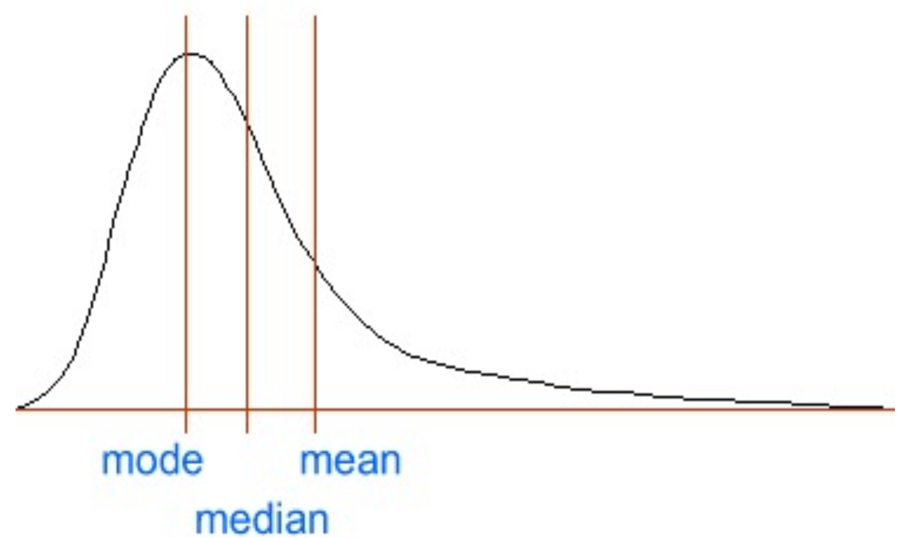
Mode

Mode

The most commonly occurring score.

i.e., the score obtained from the largest number of subjects



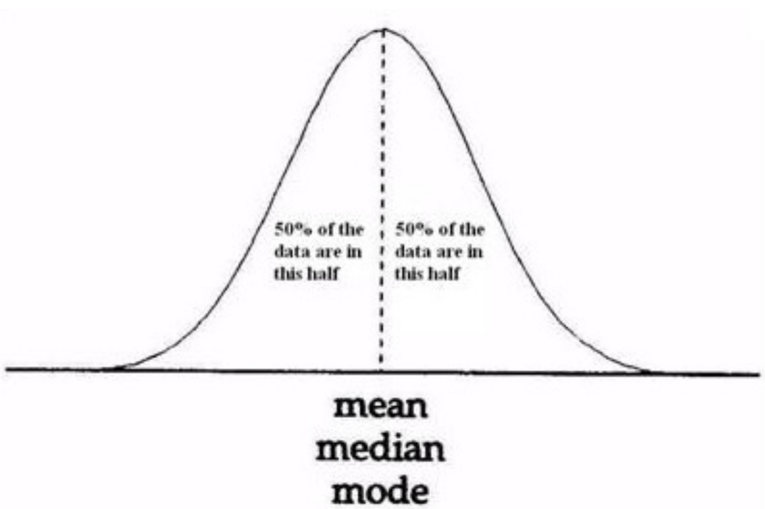


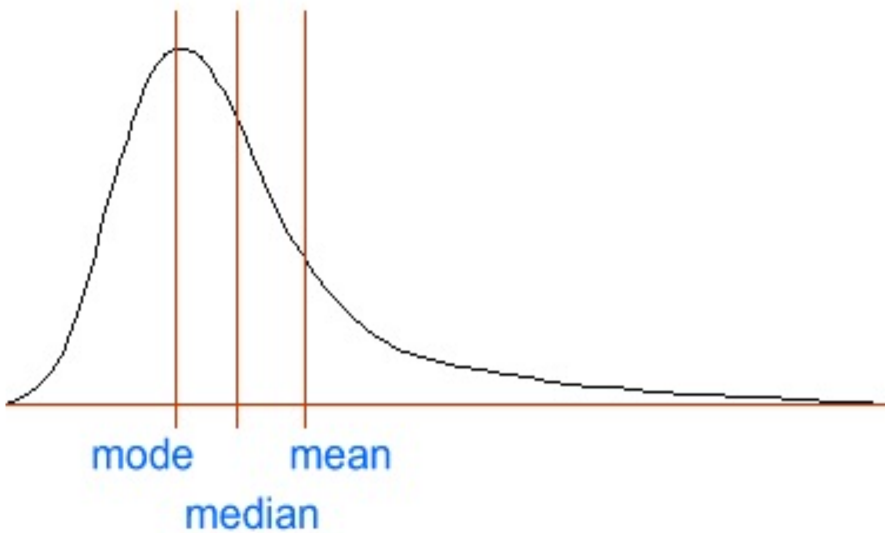
Mode: Problems

Consider the following. Take a group of undergraduate students and ask them how many cigarettes they smoke in one day. It is quite possible, that the mode will be zero, but what does this tell us about the behaviour of the group?

Median

The score that corresponds to the point at or below which 50% of the scores fall. (i.e., the 50th percentile).





Median: Problems

1. Not easily enterable into equations.

2. Not as stable from sample to sample as the mean.

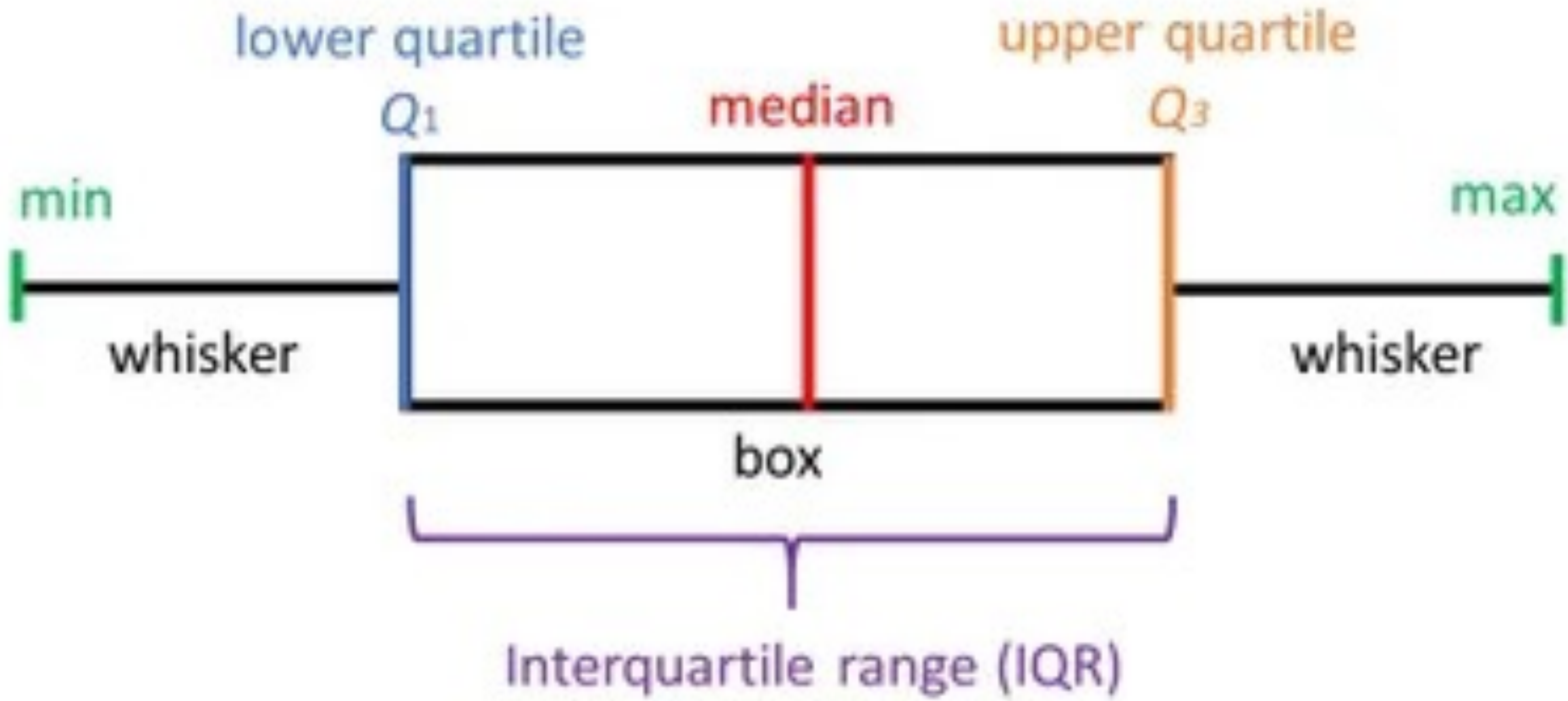
Median: Advantages

Mean Wage = $33.25

Median Wage = $8.25

Not sensitive to extreme scores

Box Plots


lower quartile
upper quartile
Q_1
median
Q_3
min
max
whisker
whisker
box
Interquartile range (IQR)

Mean

$$\bar{x} = \frac{\sum x}{n}$$

The most commonly reported measure of central tendency.

Properties of the Mean

1. The mean is sensitive to the exact value of all the scores in the distribution

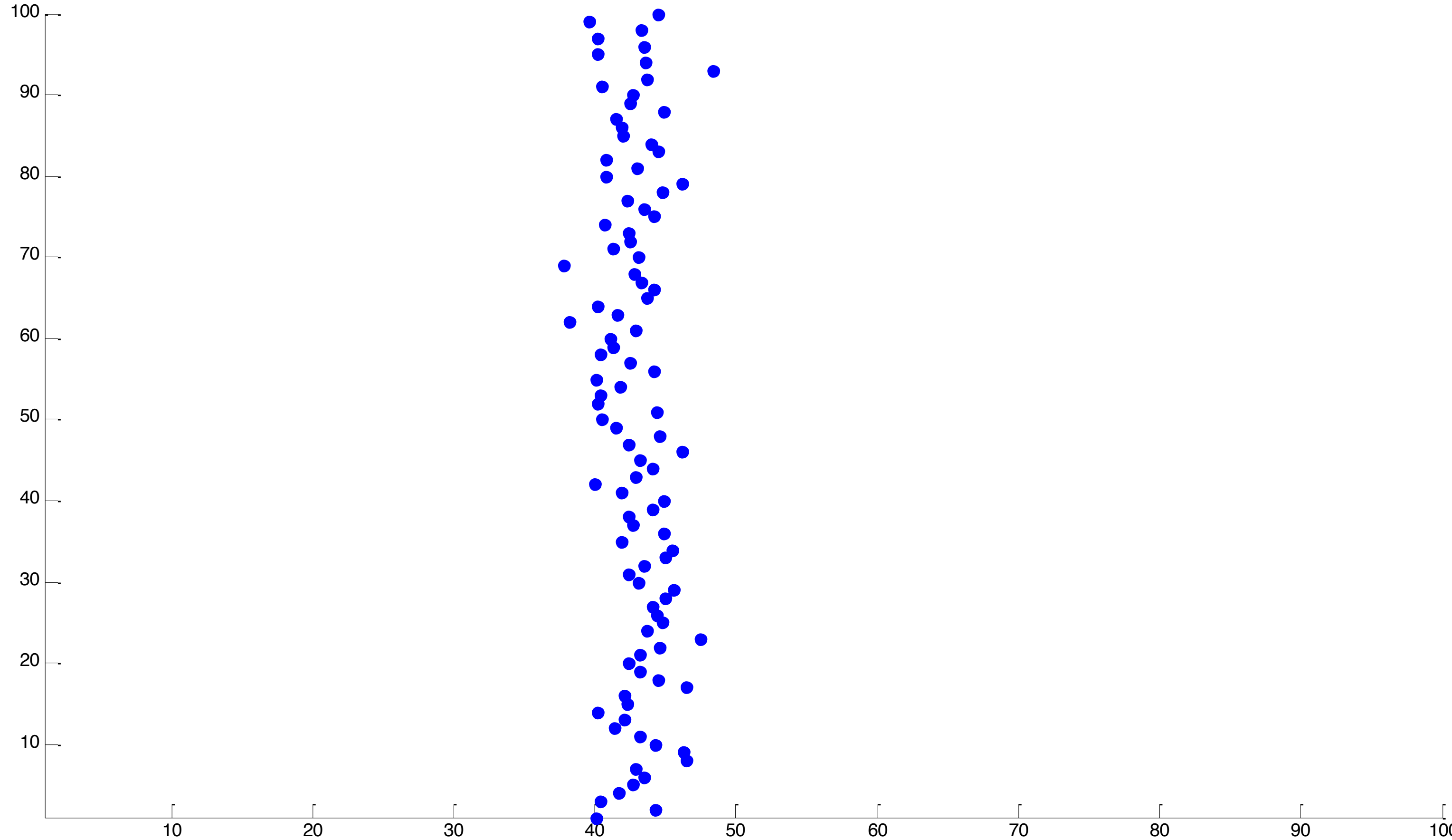
Properties of the Mean

2. The sum of the deviations about the mean is zero

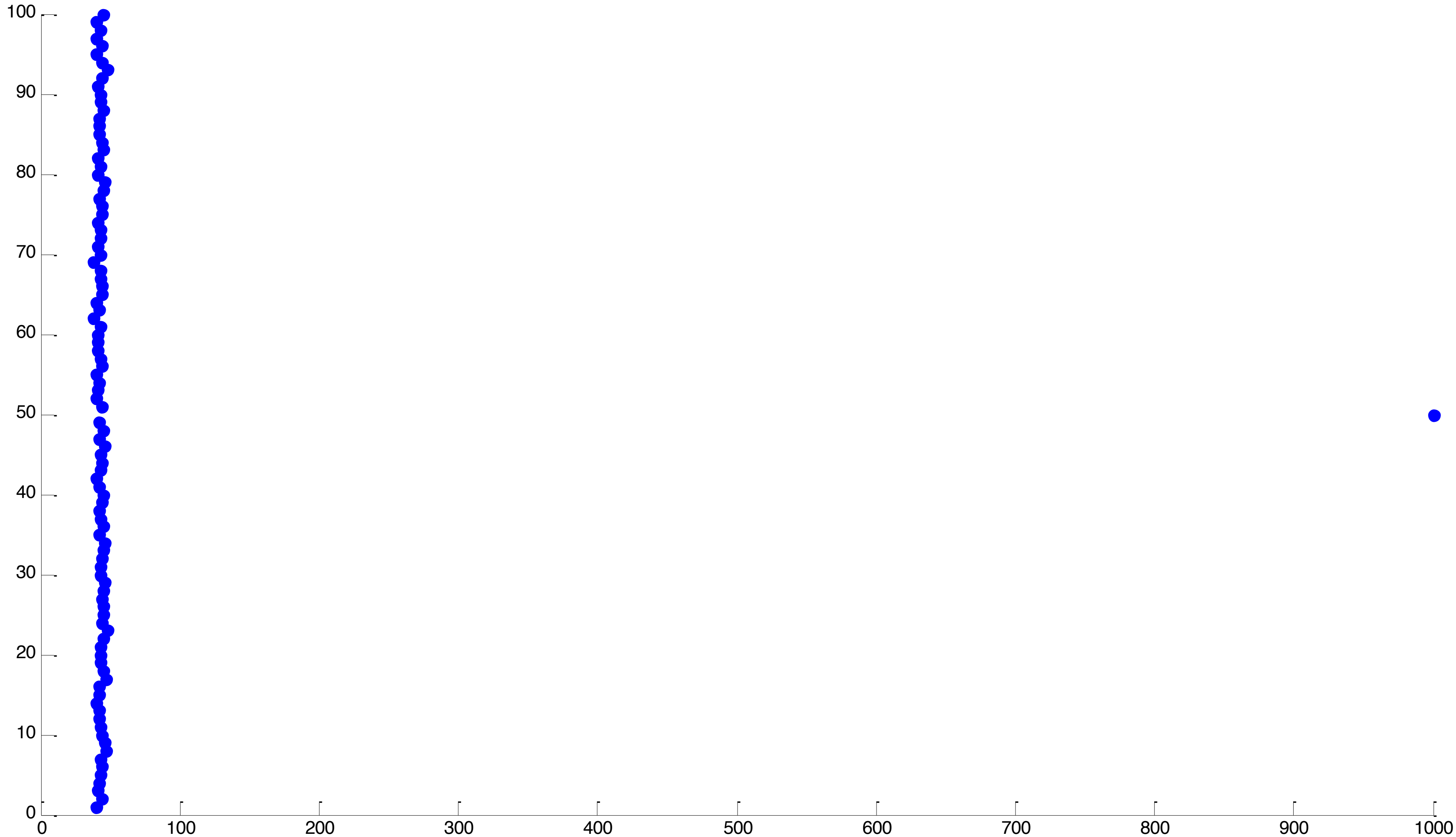
X	$X-\overline{X}$
334	-76.2
387	-23.2
431	20.8
521	110.8
378	-32.2
$\overline{X} =$ 410.2	$\sum$ 0

Properties of the Mean

3. The mean is sensitive to extreme scores



Mean = 42.86



100
90
80
70
60
50
40
30
20
10
0
0
100
200
300
400
500
600
700
800
900
1000

Mean = 52.45

Properties of the Mean

4. The sum of the squared deviations of all the scores about the mean is a minimum

X	$X-\overline{X}$	$(X-\overline{X})^2$
334	-76.2	5806.44
387	-23.2	538.24
431	20.8	432.64
521	110.8	12276.64
378	-32.2	1036.84

$\overline{X} =$ 410.2

$\sum$ 20090.8

Properties of the Mean

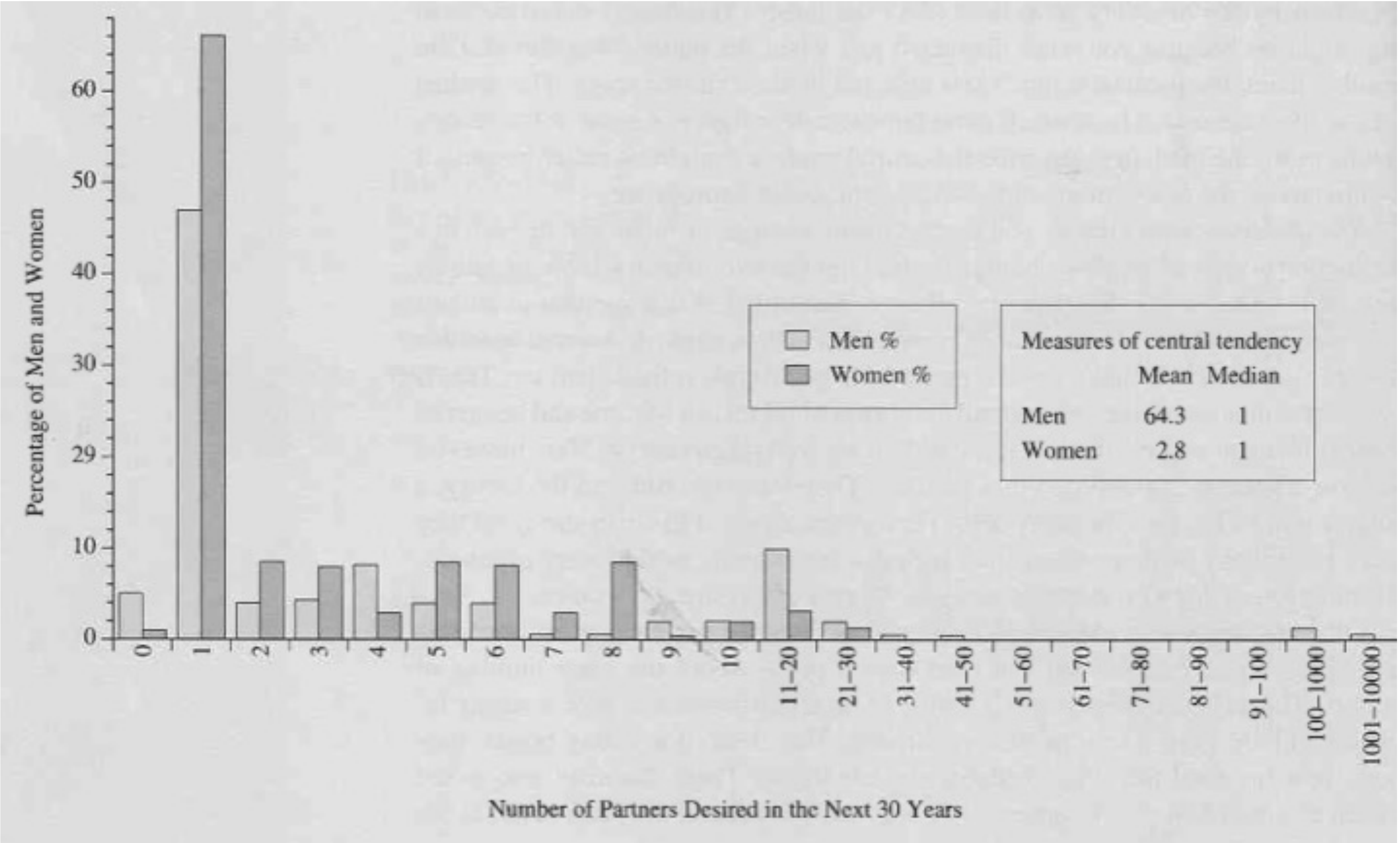
5. Under most circumstances, the mean is least subject to sampling variation

Overall Mean

$$\overline{X} = \frac{\sum \text{All scores}}{N}$$

$$\overline{X} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2 + \ldots\ldots + n_n \overline{X} n}{n_1 + n_2 + \ldots . n_n}$$

Mean: Problems



Sensitivity to outlying values

Mean: Problems

Value might not exist in the actual data

Mean: Advantages

1. It can be algebraically manipulated

2. If you draw several sample means from a population, they would reflect a more stable estimate of the population mean than sample medians and modes

Caution: Note that the statement says “If you draw several”

Descriptors: How fat is it?

Range

Deviation

Variance

Standard Deviation

Coefficient of Variation

Range

Range: Maximum Score – Minimum Score

Deviation

Deviation = Score – Mean

Average Deviation = 0

Variance

Variance: the average of the squared deviations

$$s^2 = \frac{\sum (X - \overline{X})^2}{N}$$

Variance

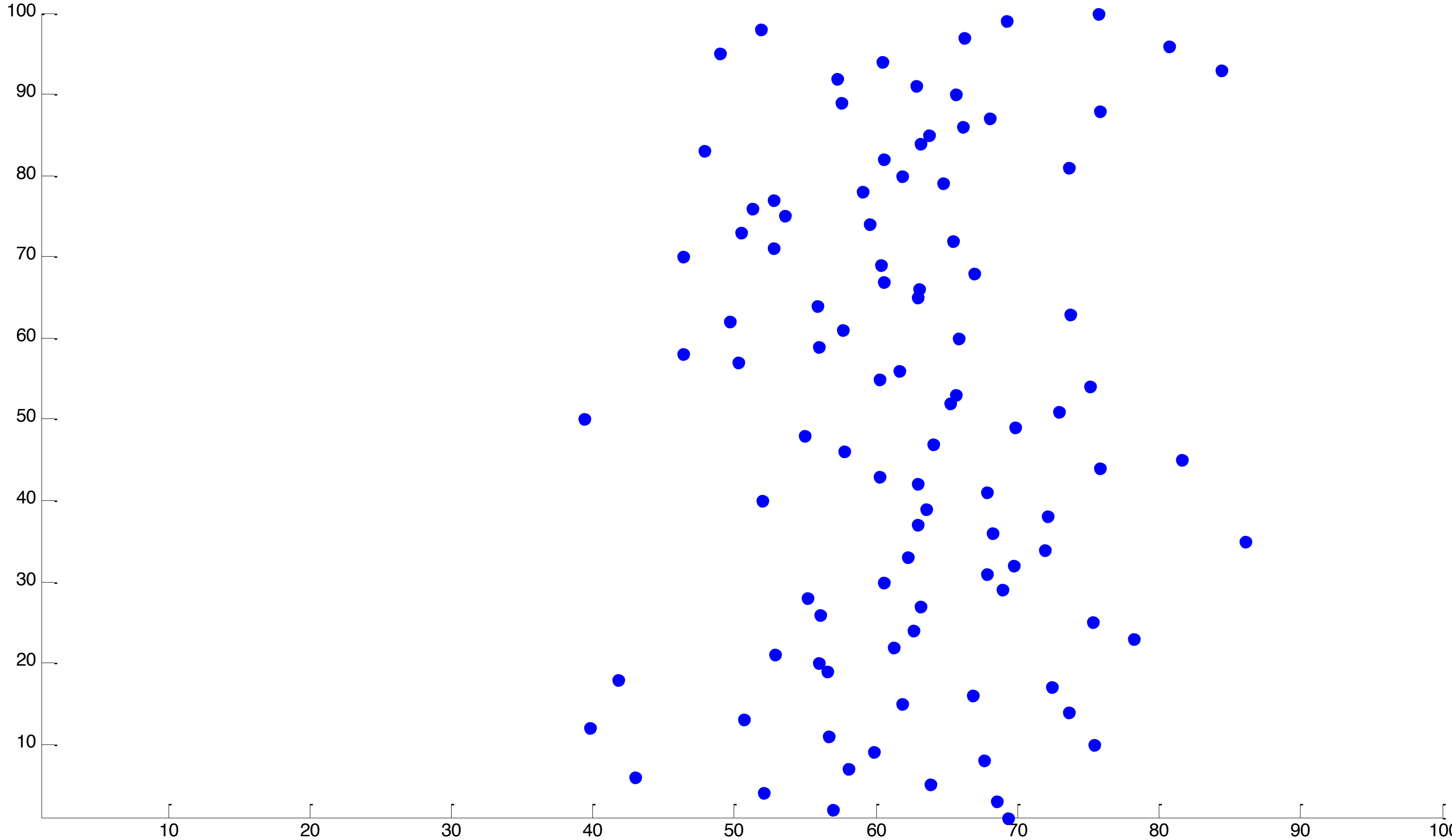
or is it…

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

Why do we use N-1 for samples?

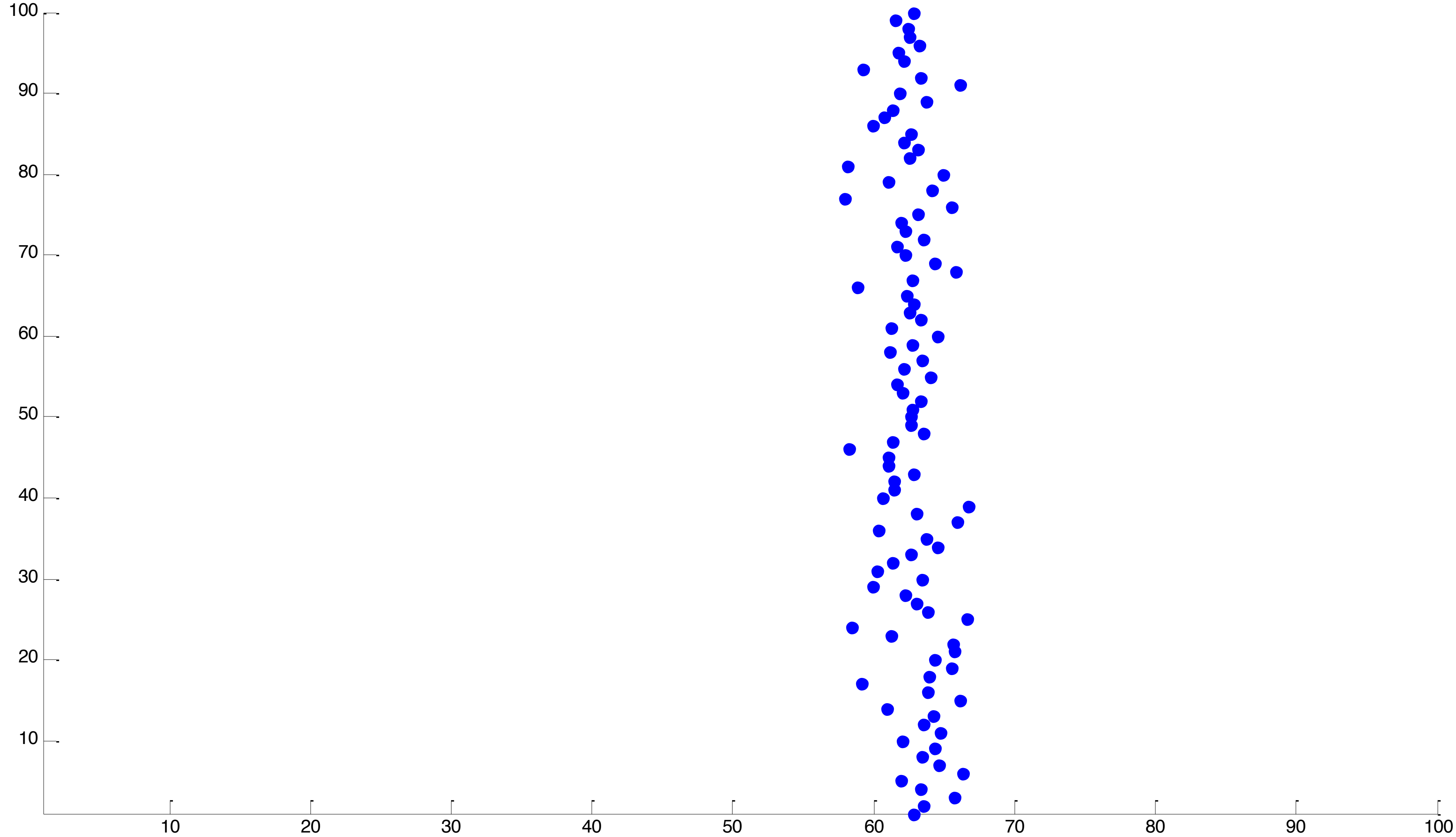
Standard Deviation

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}}$$



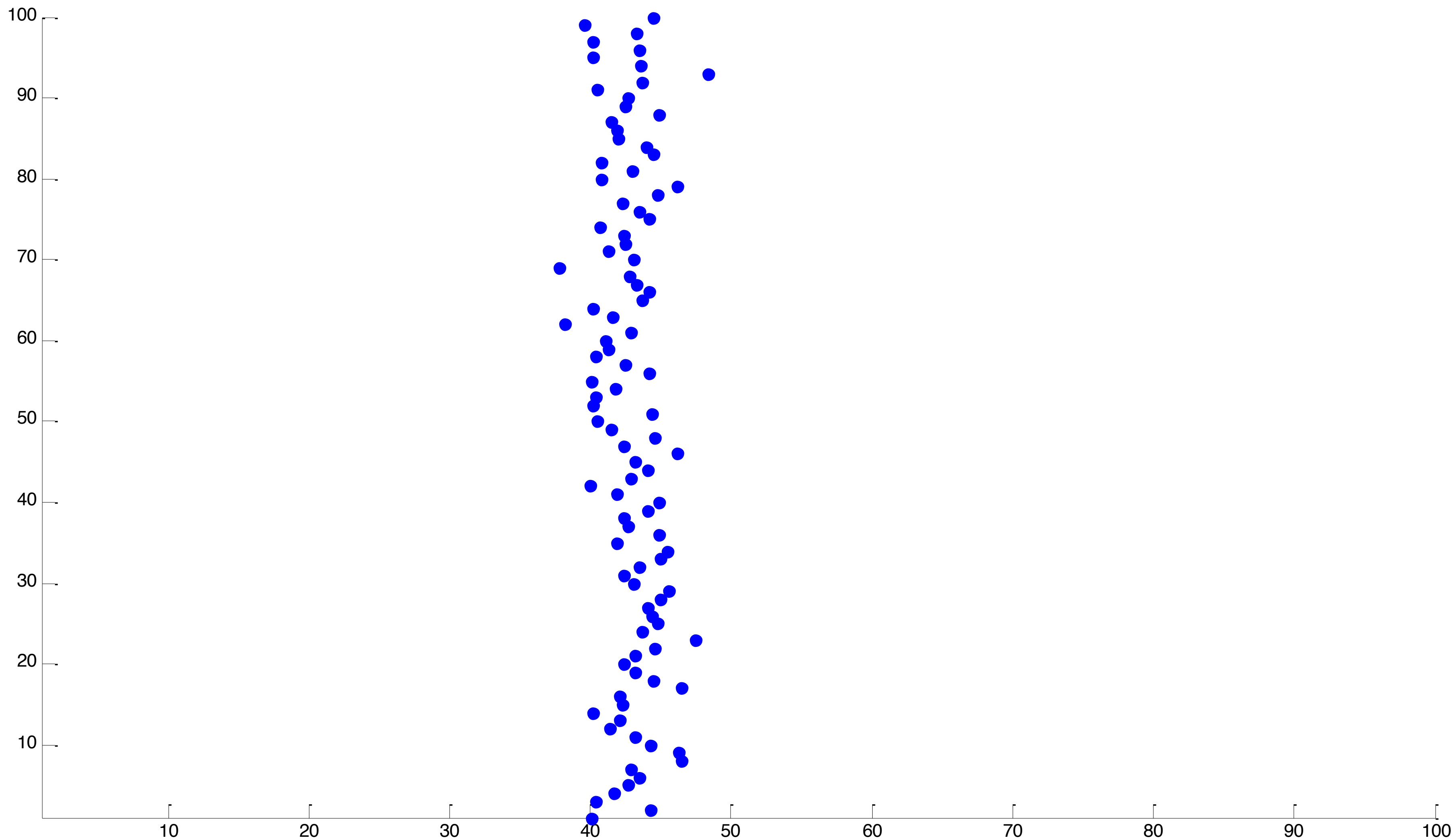
Mean = 63

Standard Deviation = 10



Mean = 63

Standard Deviation = 2



Mean = 43

Standard Deviation = 2

Properties of the Standard Deviation

1. Provides a measure of dispersion
2. Sensitive to each score in the distribution
3. Stable with regard to sampling fluctuations

Standard Deviation

$$s = \sqrt{\frac{SS}{N-1}}$$

$$SS = \sum (x - \bar{x})^2$$

$$SS = \sum X^2 - \frac{(\sum X)^2}{N}$$

Coefficient of Variation

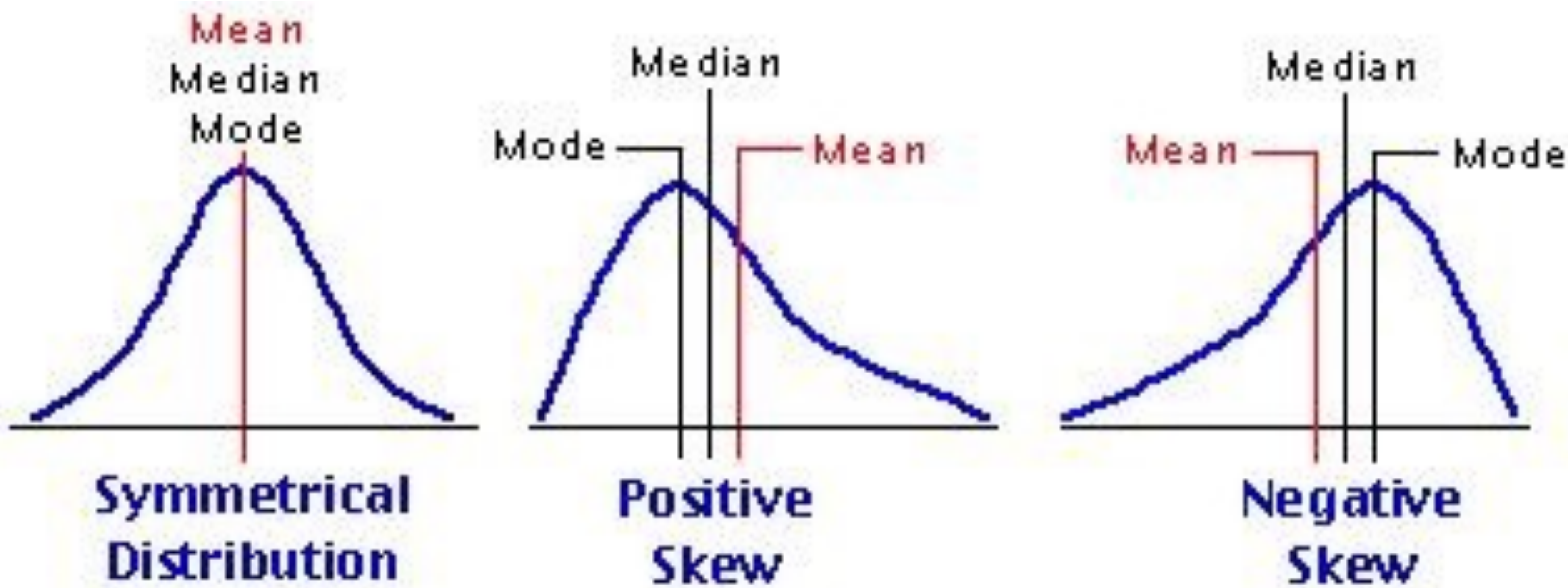
$$c = \frac{\sigma}{\mu}$$

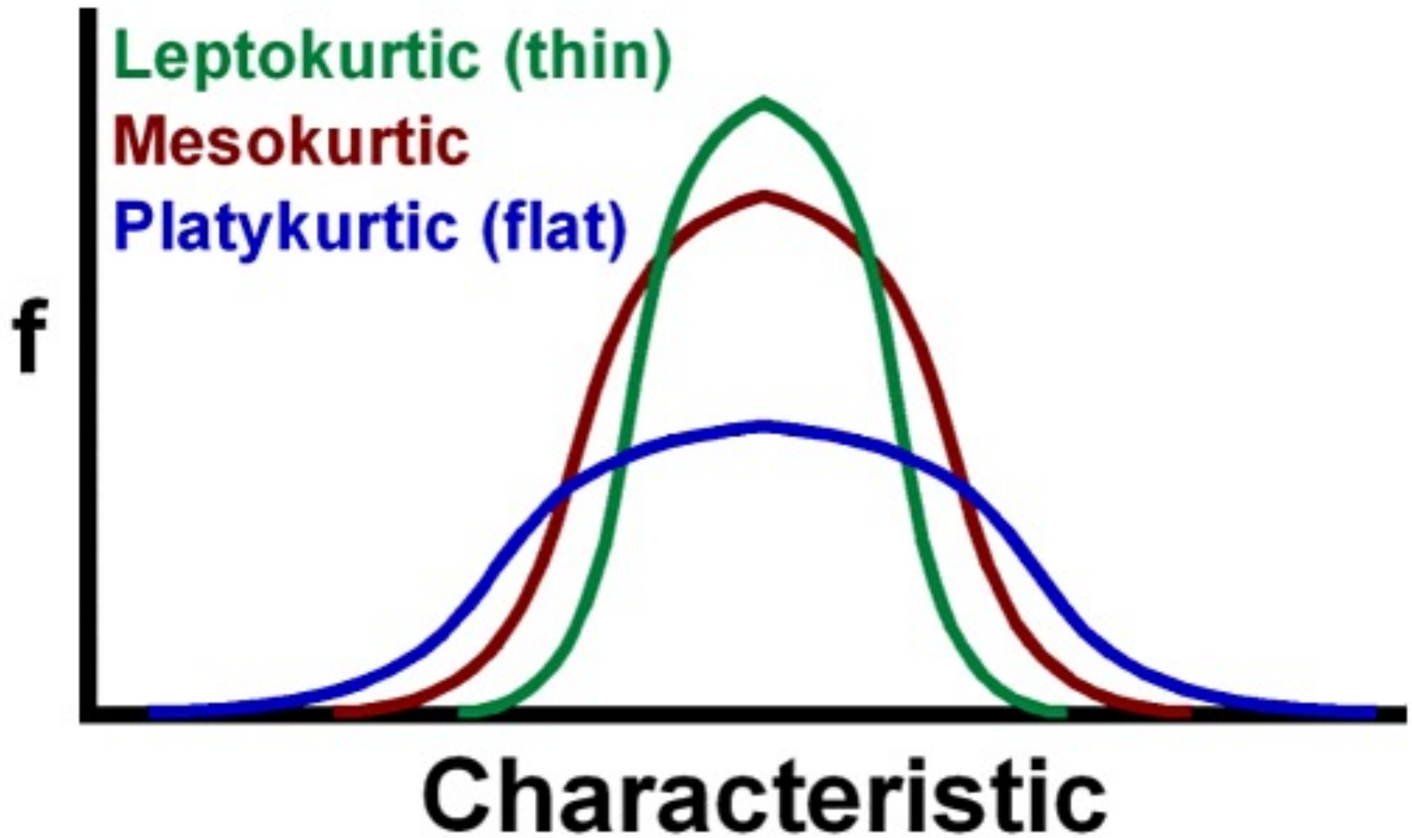
fyi: the inverse of this is typically referred to as the signal to noise ratio

Descriptors: What shape is it?

Skew

Kurtosis


Mean
Median
Mode
Symmetrical
Distribution
Median
Mode
Mean
Positive
Skew
Median
Mean
Mode
Negative
Skew



Leptokurtic (thin)
Mesokurtic
Platykurtic (flat)
f
Characteristic

$$\textbf{Skew} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_j - \bar{x}}{s} \right)^3$$

$$\textbf{Kurtosis} = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_j - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Confidence Intervals

What is a confidence interval?

It is a range within which we believe the true value of the mean will fall.

Typically, we use 95% or 99% Confidence Intervals

But what does it really mean?

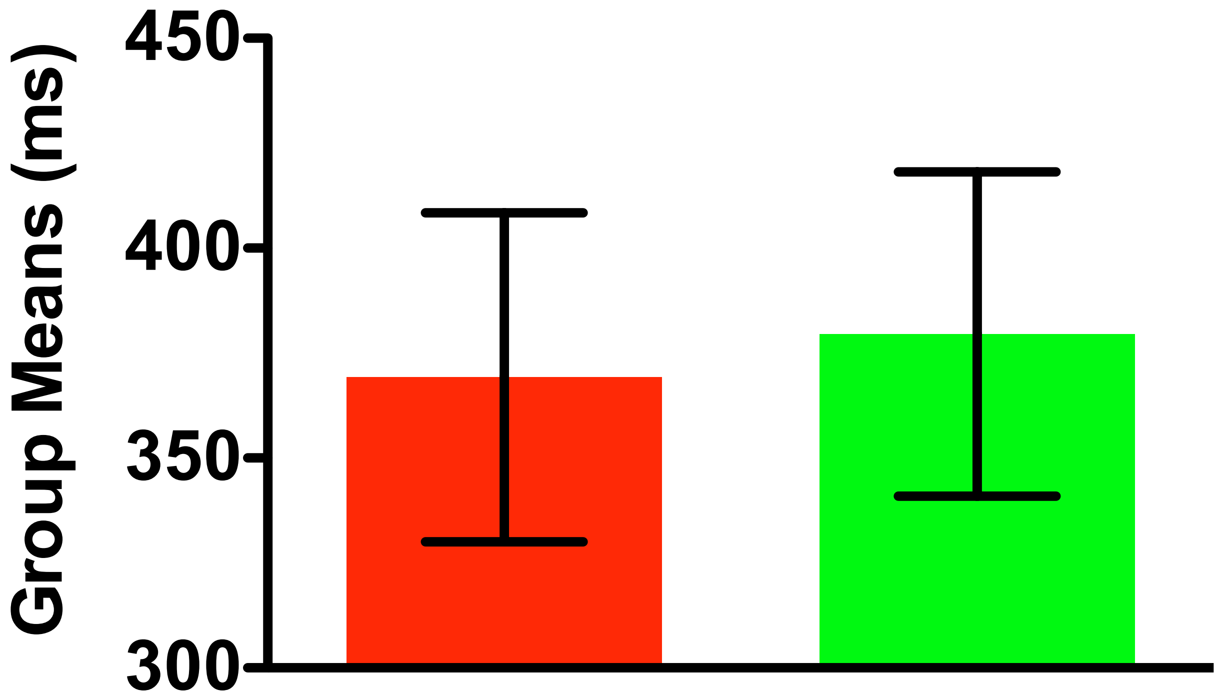
What a confidence interval truly means is that we have a 95% confidence that the true value of the mean is within the prescribed range.

Thus, if we took 100 samples, on average, we would expect 5 samples to have means outside of the prescribed range.

95% Confidence Intervals

Mean +/- CI

$$\bar{x} \pm 1.96 \times \frac{s}{\sqrt{N}}$$


Group Means (ms)
450
400
350
300