CHAPTER 16
16.1, 16.2, 16.3

Exploratory Data Analysis

&

Summary Statistics

I understand that absorbing information at the beginning of the semester is like getting a drink from a fireplug.

Overcoming the friction and fighting through the firehose of information will be complete by the end of this week!

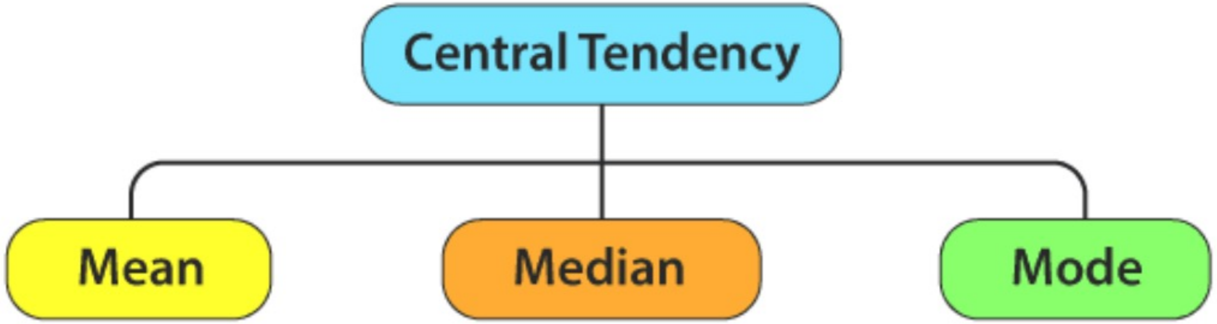
To date:

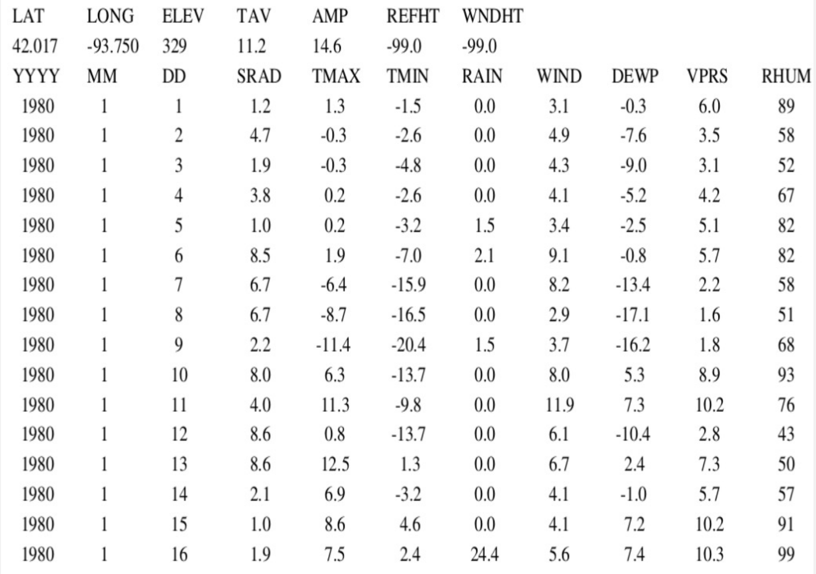
- You have access to Canvas and you have explored the links
- You have access to Piazza
- You've read the syllabus for tonight's quiz
- You've installed Anaconda (for access to Jupyter Notebooks)
- You've completed the NumPy/Pandas tutorial
- You have worked thru NB00 ad NB01

Today we cover the ideas that you read about in sections:

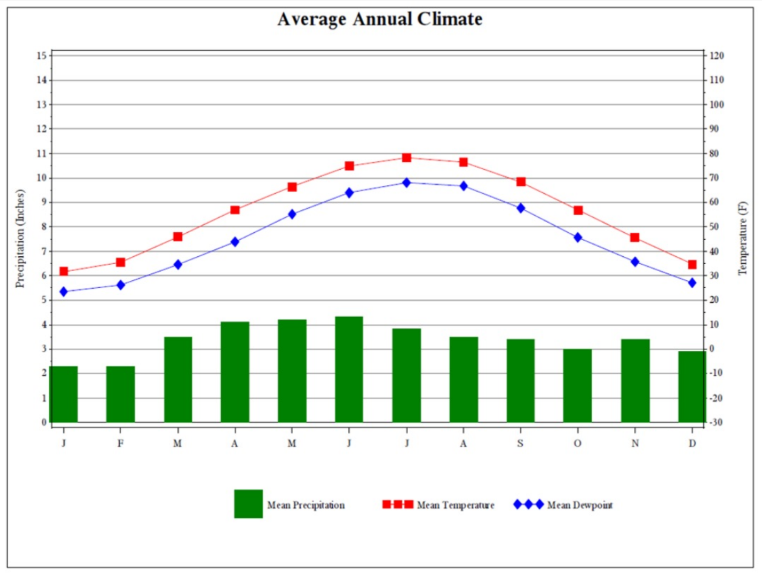
16.1, 16.2, & 16.3

- Mean
- Median
- Mode
- Quartiles
- Sample Variance
- Sample Standard Deviation
- Inner Quartile Range (IQR)
- 5-number summary





LAT	LONG	ELEV	TAV	AMP	REFHT	WNDHT				
42.017	-93.750	329	11.2	14.6	-99.0	-99.0				
YYYY	MM	DD	SRAD	TMAX	TMIN	RAIN	WIND	DEWP	VPRS	RHUM
1980	1	1	1.2	1.3	-1.5	0.0	3.1	-0.3	6.0	89
1980	1	2	4.7	-0.3	-2.6	0.0	4.9	-7.6	3.5	58
1980	1	3	1.9	-0.3	-4.8	0.0	4.3	-9.0	3.1	52
1980	1	4	3.8	0.2	-2.6	0.0	4.1	-5.2	4.2	67
1980	1	5	1.0	0.2	-3.2	1.5	3.4	-2.5	5.1	82
1980	1	6	8.5	1.9	-7.0	2.1	9.1	-0.8	5.7	82
1980	1	7	6.7	-6.4	-15.9	0.0	8.2	-13.4	2.2	58
1980	1	8	6.7	-8.7	-16.5	0.0	2.9	-17.1	1.6	51
1980	1	9	2.2	-11.4	-20.4	1.5	3.7	-16.2	1.8	68
1980	1	10	8.0	6.3	-13.7	0.0	8.0	5.3	8.9	93
1980	1	11	4.0	11.3	-9.8	0.0	11.9	7.3	10.2	76
1980	1	12	8.6	0.8	-13.7	0.0	6.1	-10.4	2.8	43
1980	1	13	8.6	12.5	1.3	0.0	6.7	2.4	7.3	50
1980	1	14	2.1	6.9	-3.2	0.0	4.1	-1.0	5.7	57
1980	1	15	1.0	8.6	4.6	0.0	4.1	7.2	10.2	91
1980	1	16	1.9	7.5	2.4	24.4	5.6	7.4	10.3	99



GIVEN A SET OF DATA, HOW DO YOU SUMMARIZE IT AND ACCURATELY DESCRIBE THE RESULTS?

QUANTITATIVELY, YOU CAN DESCRIBE THE CENTER OF THE DATA.

PICTURES ARE WORTH A THOUSAND WORDS.

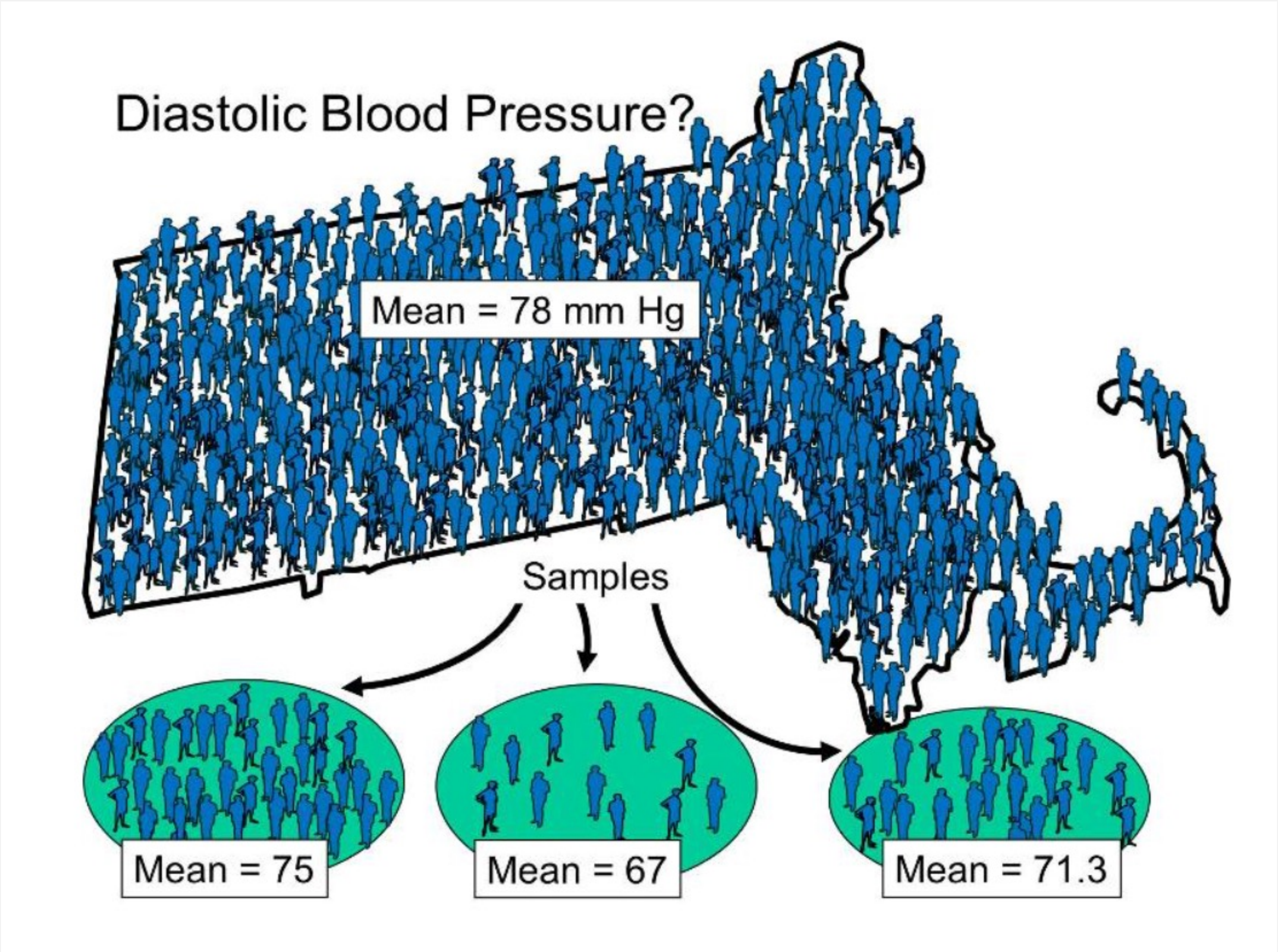
PICTURES ARE WORTH A FEW NUMBERS.

- When we describe a set of data, we are most often describing a sample.

- A sample is a (hopefully) representative set of data that comes from a population.

- We want information about a population, and we get that information by looking at and describing a sample of that population.

- In particular we are often looking at a variable of interest (VOI)



More vocabulary (about samples and data collection)

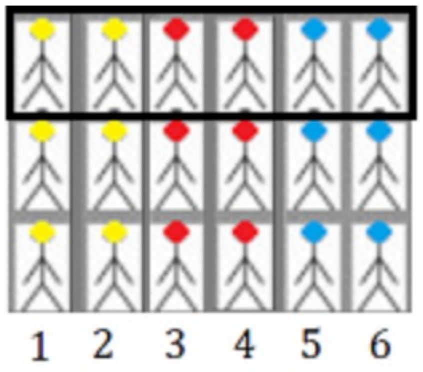
The **sample frame** is the source material or device from which the sample is drawn.

Simple random sample: randomly select people from a sample frame.

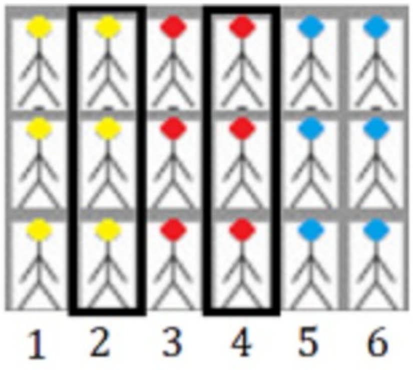
Systematic sample: Order the sample frame. Choose an integer *k*. Sample every k^{th} unit in the sample frame.

Census sample: sample everyone/everything in the population.

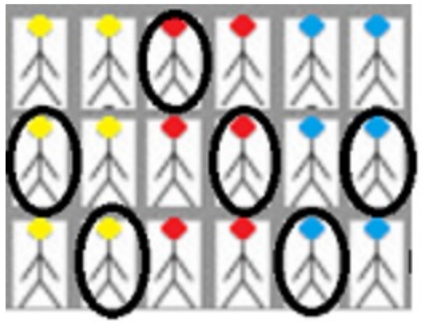
Stratified sample: if you have a heterogeneous population that can be broken up into homogeneous groups, randomly sample from each group proportionate to their prevalence in the population.



Quota Sampling



Cluster Sampling



Stratified Sampling

Example: Suppose that the city of Boulder wants to estimate its per-household income via a phone survey. They call every 50th number on a list of Boulder phone numbers between 6 PM and 8 PM.

What is the **population**?

What is the **sample**?

What is the **Variable of Interest**?

One way to describe a set of data is to explain what the ‘center’ of the data looks like. To do this, we use *measures of central tendency*: Mean, Median, Mode.

consider these exam scores:

{50, 55, 70, 75, 80, 80, 95}

Question: What is a ‘typical’ score?

To answer this question, we tend to describe *center* scores (not the 50 or the 95).

But what does *center* mean?

The literal middle score is 75 (the median)

The most common score is an 80 (the mode)

The scores have an average of 71.7 (the mean)

The Mean

$$\text{Arithmetic Mean} = \frac{a_1 + a_2 + a_3 + \ldots + a_n}{n} = \sum_{i=1}^{n} \frac{a_i}{n}$$

a_1 is the first value
a_2 is the second value
a_n is the last value
n is the number of values
$\bar{x}$ is the arithmetic mean

For our exam score example: {50, 55, 70, 75, 80, 80, 95}

$$\bar{x} = \sum_{i=1}^{7} \frac{a_i}{7} = \frac{50 + 55 + 70 + 75 + 80 + 80 + 95}{7} = 71.714$$

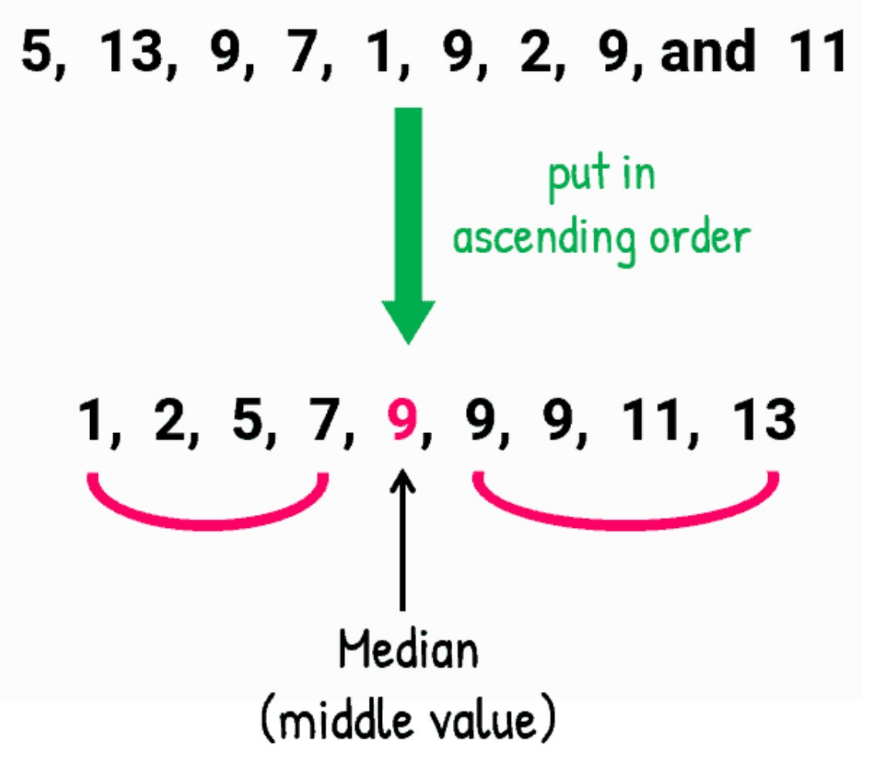
The median

If n is odd, then median is a_i

for $i = \left\lceil \frac{n}{2} \right\rceil$ for odd n.

If n is even, then median is $\frac{a_i + a_j}{2}$

for $i = \frac{n}{2}$ and $j = \frac{n+2}{2}$



For our exam score data: {50, 55, 70, 75, 80, 80, 95}

median is $a_4 = 75$

For the data: {50, 55, 70, 75, 80, 80, 95, 100}

median is $\frac{a_4 + a_5}{2} = \frac{75 + 80}{2} = 77.5$

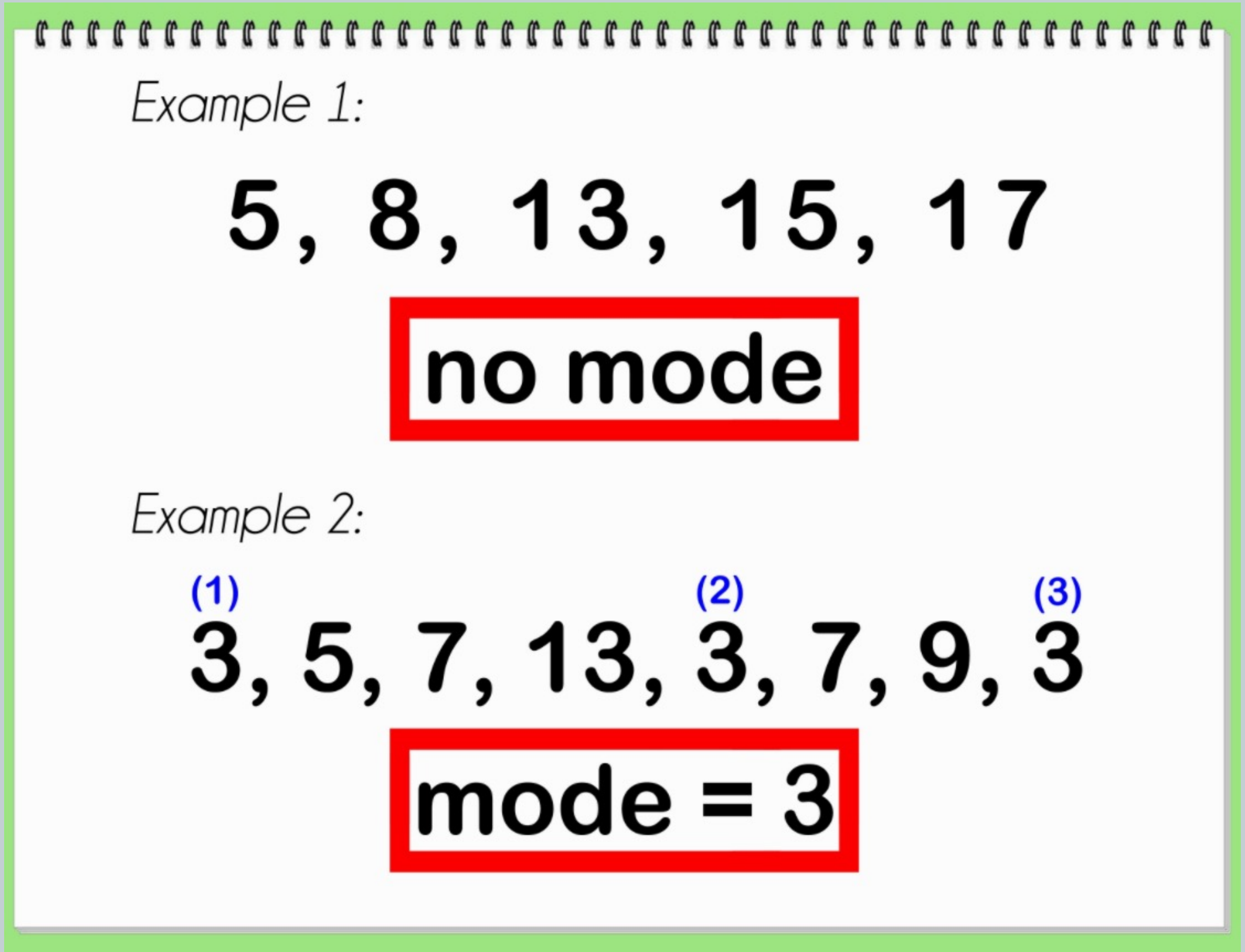
Notice the median for the second set of data wasn't even an actual data point in the set!

The mode

The mode is the most common data point to occur.

For our exam score data:
{50, 55, 70, 75, 80, 80, 95}
The mode is 80.

For the data:
{50, 55, 70, 75, 80, 80, 95, 95}
The data is bimodal: 80 & 95



Example 1:

5, 8, 13, 15, 17

no mode

Example 2:

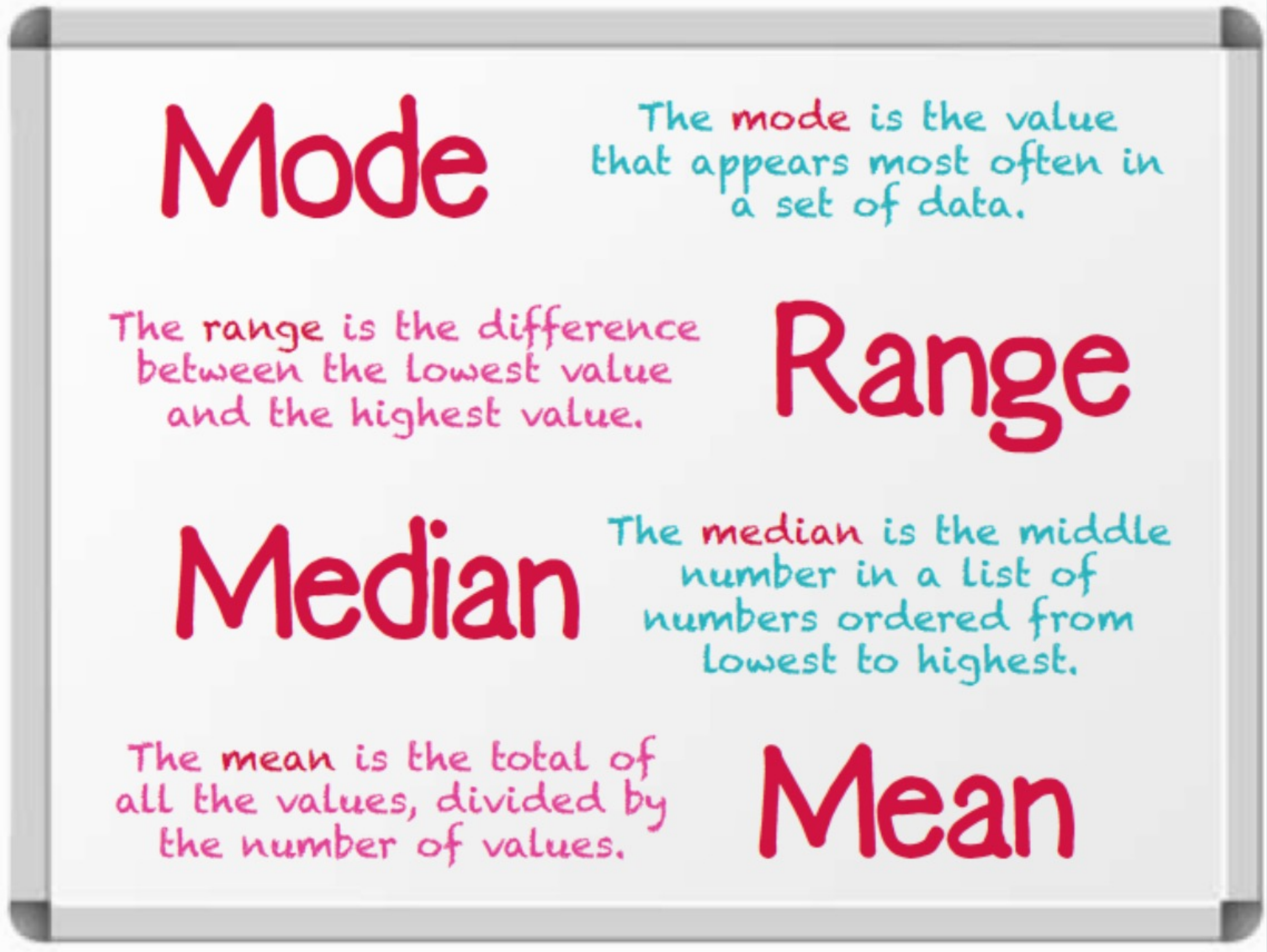
(1) 3, 5, 7, 13, (2) 3, 7, 9, (3) 3

mode = 3

The Range

The range is the difference between the highest and lowest data points.

For our exam score data: {50, 55, 70, 75, 80, 80, 95}
the range is $95 - 50 = \mathbf{45}$



Mode

The mode is the value that appears most often in a set of data.

The range is the difference between the lowest value and the highest value.

Range

Median

The median is the middle number in a list of numbers ordered from lowest to highest.

The mean is the total of all the values, divided by the number of values.

Mean

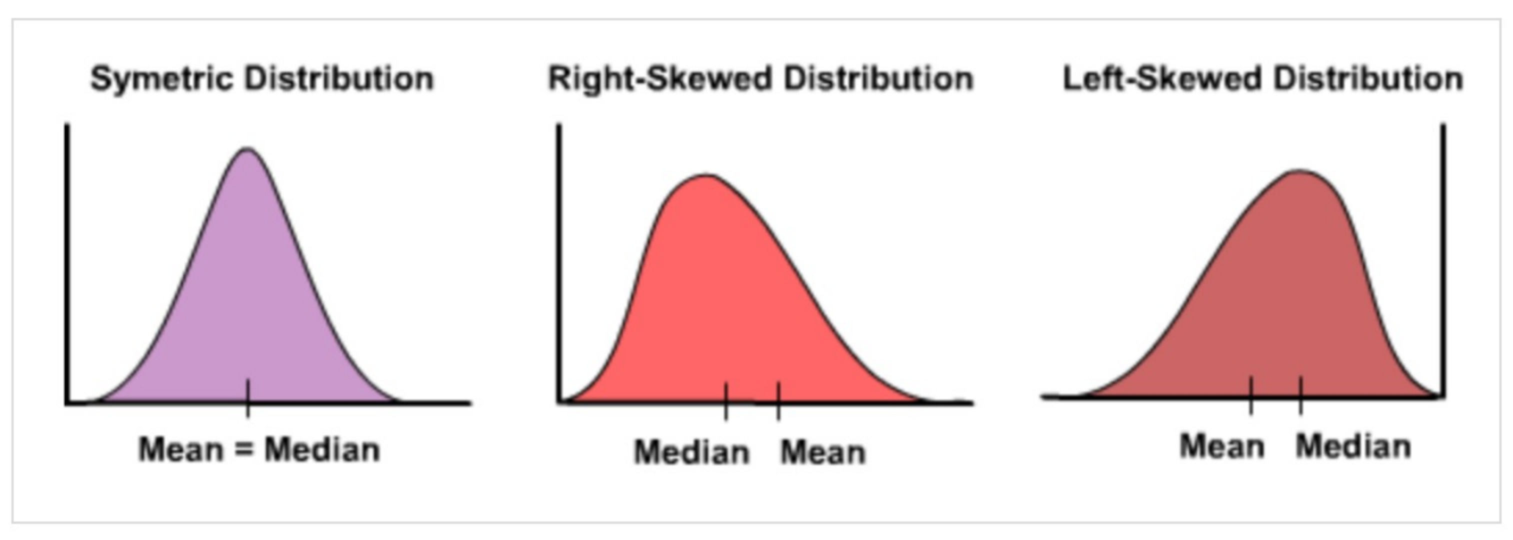
If a factory was advertising for a new worker, then would they use the mean, the median, or the mode to best describe the position in an interview?

In this particular factory, there are 23 employees.

1 Owner: makes $4800 per week
1 Boss makes $2000 per week
6 managers make $500 per week
5 foremen make $400 per week
10 workers make $200 per week

mean: $600 per week
median: $400 per week
mode: $200 per week

‘Symmetric’ data sets are more ‘honest’



We often graph data to get an overall picture,

i.e. to avoid being misled by a single number.

- Some data comes out symmetric (peoples heights, mean = median)
- Some data comes out right-skewed (a difficult exam, mean > median)
- Some data comes out left-skewed (an easy exam, mean < median)

It has been advised that a more genuine data picture is provided with a five-number summary: The minimum, the maximum, the sample median, and the 25th and 75th empirical percentiles, aka the lower quartile and the upper quartile.

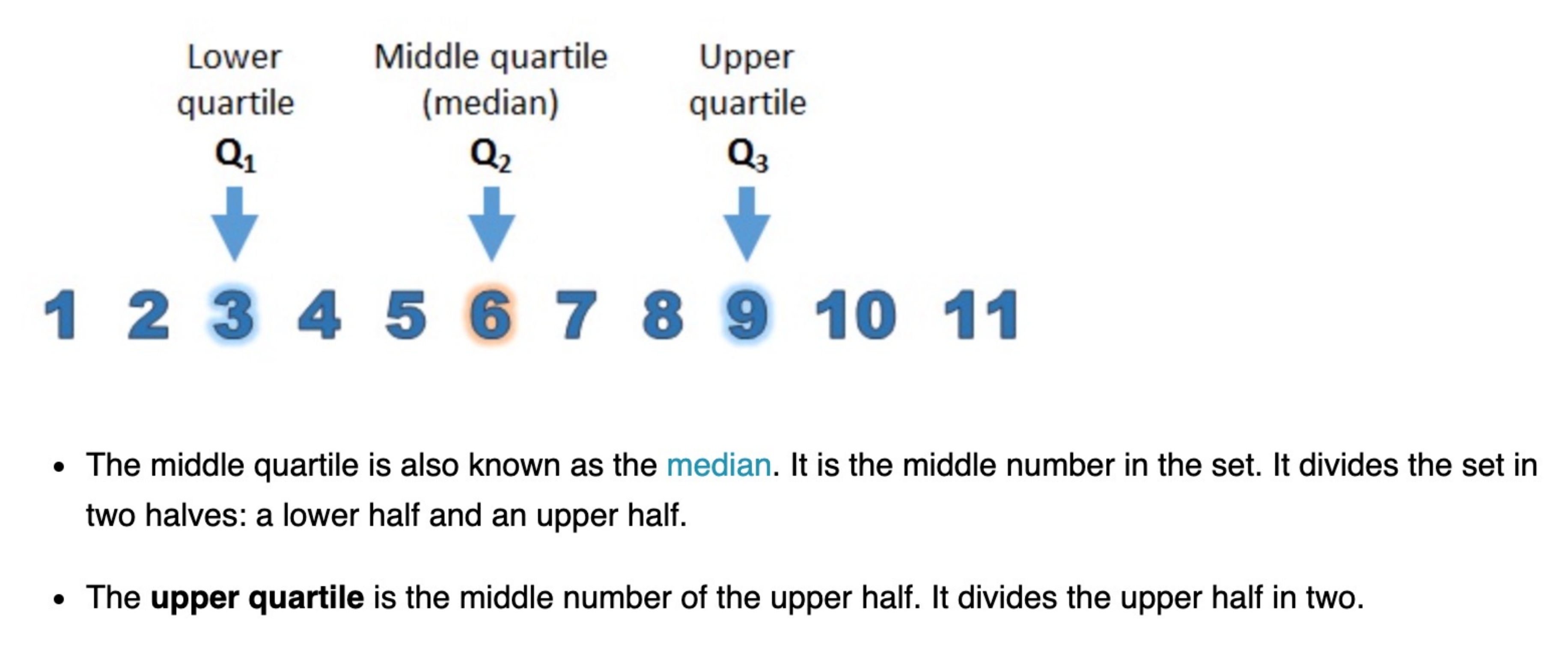
How are these quartiles calculated?

The median divides a dataset in two more or less equal parts: about half of the elements are less than the median and about half of the elements are greater than the median.

More generally we can divide the dataset in two parts in such a way that a proportion p is less than a certain number and a proportion $1-p$ is greater than this number. Such a number is called the p^{th} empirical quantile or the $100p$ empirical percentile and is denoted $q_n(p)$.

Bottom line… when calculating by hand, exclude the median.

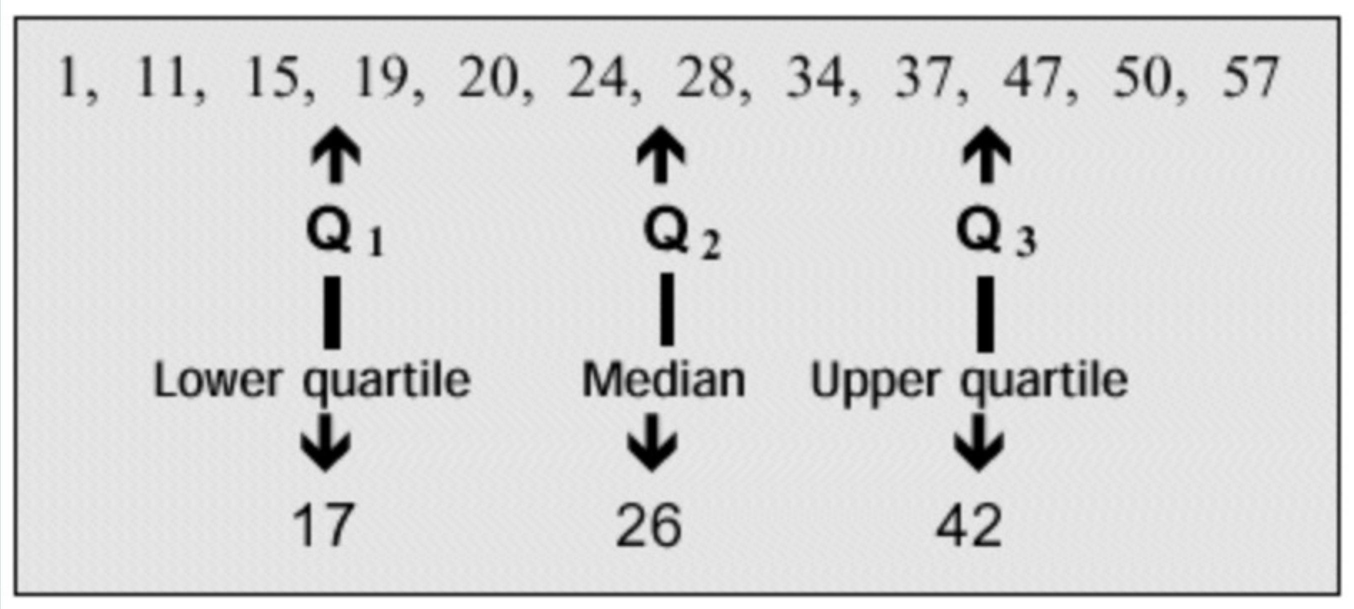
If there is an odd amount of data:



- The middle quartile is also known as the median. It is the middle number in the set. It divides the set in two halves: a lower half and an upper half.
- The **upper quartile** is the middle number of the upper half. It divides the upper half in two.

The lower quartile, Q_1, is the middle number of the lower half.
The upper quartile, Q_3, is the middle number of the upper half.

And if there are an even amount of data:

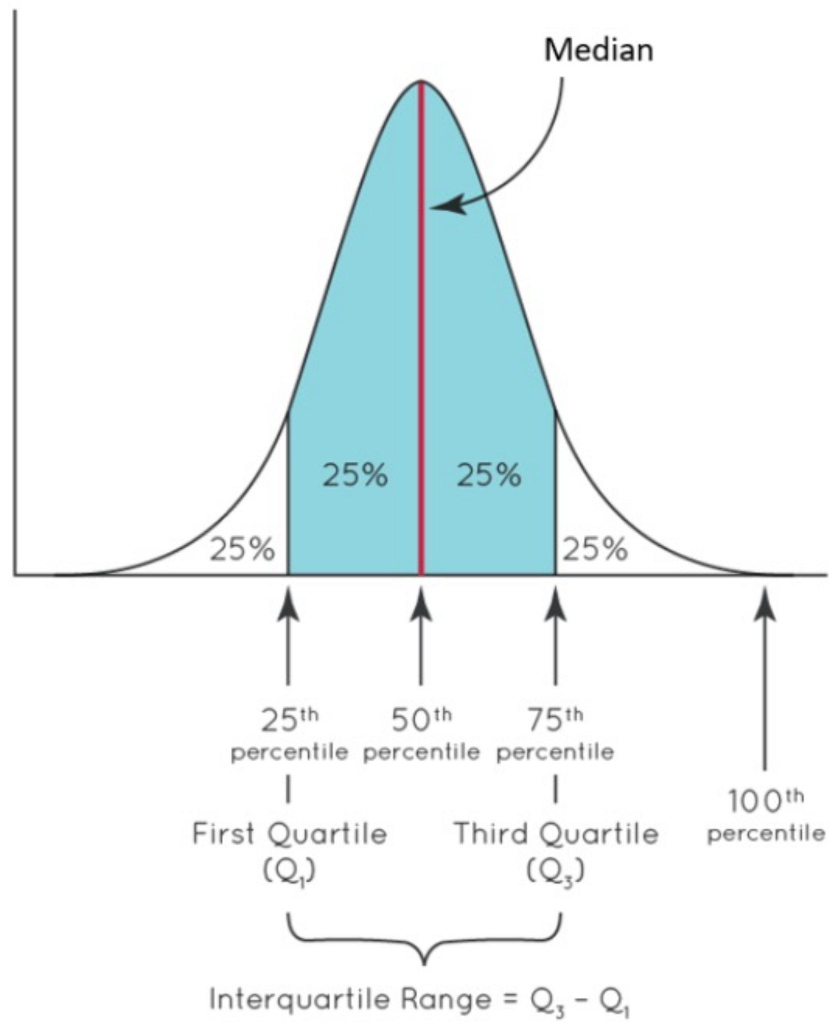


Then Q_1 and Q_3 are calculated as the average of the middle two numbers.

Note that quartiles are calculated differently via different methods:
Tukey method, linear interpolation (Python), Moore and McCabe method, Mendenhall and Sincich method… do you include the median or exclude the median?

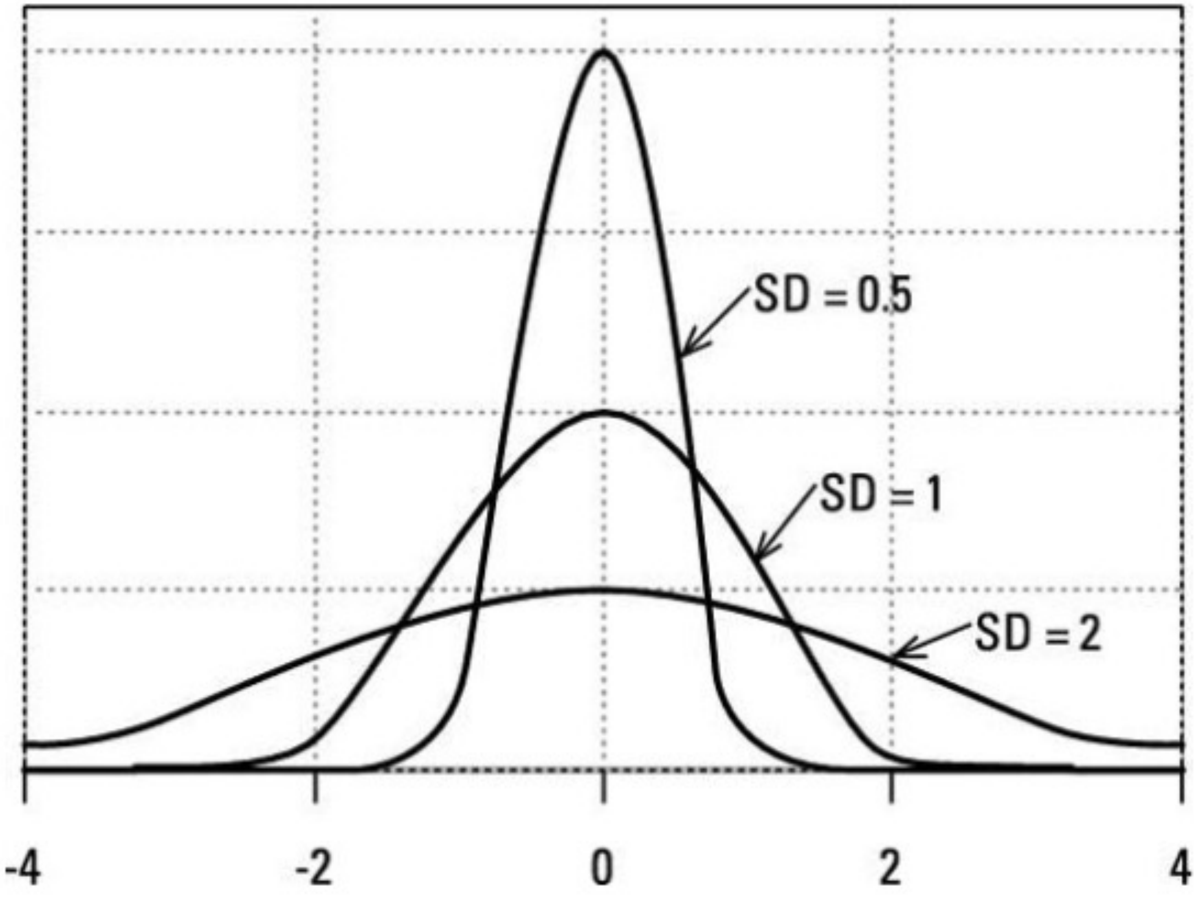
comparison of methods link.

GRAPHICALLY, THE QUARTILES HELP DESCRIBE THE MIDDLE OF THE DATA.



WE CAN ALSO EXPLAIN DATA BY DESCRIBING ITS SPREAD AND VARIABILITY

WHICH GRAPH SHOWS LOW VARIABILITY AND WHICH SHOWS HIGH VARIABILITY?



Sample Variance: Variance is a measurement of spread or dispersion of observations within a given dataset. Variance measures how far each observations is from mean.

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

But there is a major problem with variance and that is the difficulty of interpreting the units of variance.

How does one interpret squared percents, squared dollars, or squared yen?

This problem is mitigated through the use of the standard deviation.

Sample Standard Deviation (SD)

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

Advantages of SD:

1] Shows how much data is clustered around the mean

2] It gives a more accurate idea of how the data is distributed

3] Not as affected by extreme values

4] Good estimate of variation of a data set if the distribution is normal

Disadvantage of SD: It doesn't give you the full range of the data.

Imagine 5 quiz scores: {6, 3, 8, 5, 3}

Compute the mean, median, mode

Compute the Q_2 (the median)

Compute Q_1 and Q_3(the quartiles)

Compute the variance

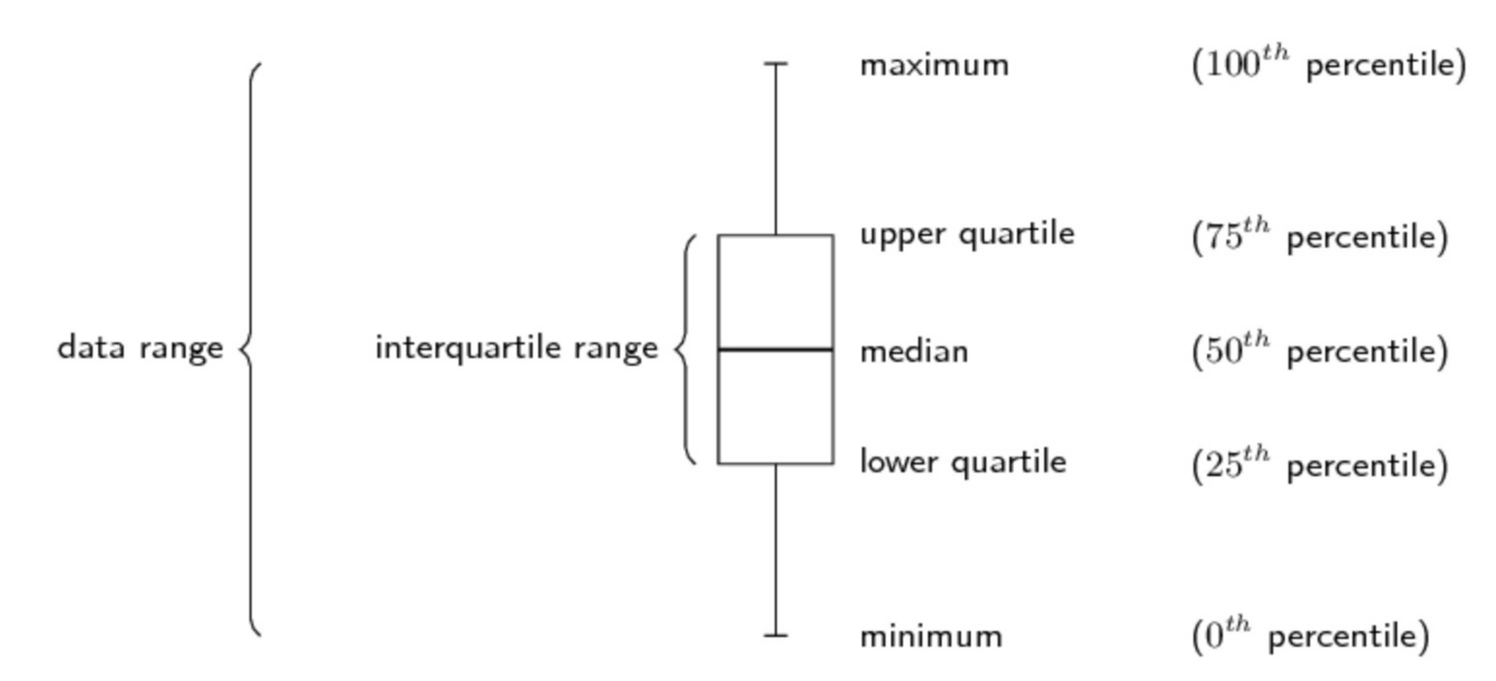
Compute the SD

Compute the range

Compute the IQR

Giving the 5-number summary of a data set is a standard.

1. minimum
2. Q_1
3. Q_2
4. Q_3
5. Maximum


maximum
$(100^{th}$ percentile)
upper quartile
$(75^{th}$ percentile)
data range
interquartile range
median
$(50^{th}$ percentile)
lower quartile
$(25^{th}$ percentile)
minimum
$(0^{th}$ percentile)