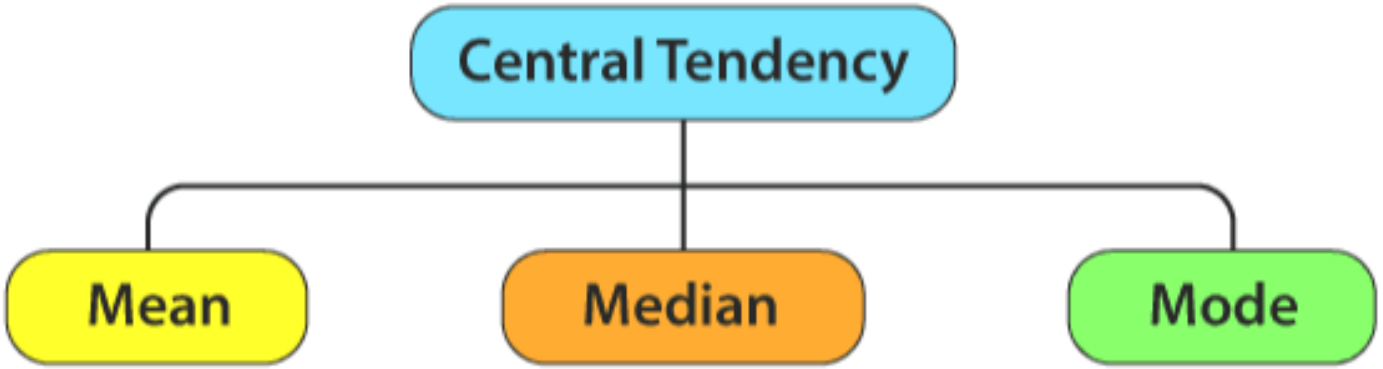
CSE303

LECTURE 2: MEASURES OF CENTRAL TENDENCY AND DISPERSION

CENTRAL tendency

- The central tendency is stated as the statistical measure that represents the single value of the entire distribution or a dataset.
- These three values summarize the dataset using a single value.



Mean

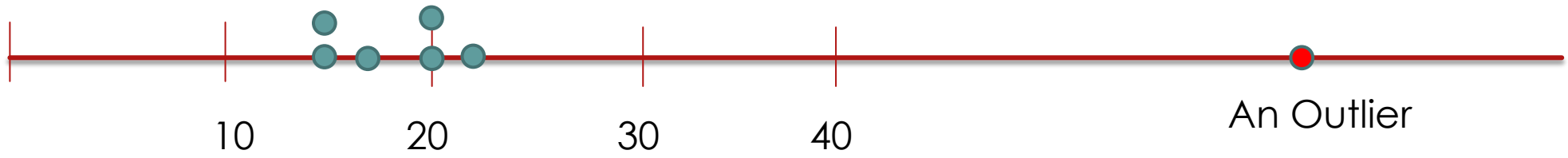
- The sum of all values divided by the number of values.

$$\text{Mean} = \bar{x} = \frac{\sum_i^n x_i}{n}$$

Some issues to consider:

- Mean is influenced by extreme values (Outliers).
- An outlier is a value or an element of a dataset that shows higher deviation from the rest of the values.

Example – AN OutlIER



Trimmed mean

- The average of all values after dropping a fixed number of extreme values from both ends.

$$\text{Trimmed mean} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}$$

- Preferable to use instead of ordinary mean as it can negate the effect of extreme values (Outliers).

median

- The median is the middle number on a sorted list of the data.
- If there is an even number of data values, the middle value is one that is not actually in the data set, but rather the average of the two values that divide the sorted data into upper and lower halves.

- If n is odd, $\frac{n+1}{2}$ th value → **10, 15, 18, 22, 25, 28, 33, 43, 50**
- If n is even, $\frac{\frac{n}{2}\text{ th value} + (\frac{n}{2}+1)\text{ th value}}{2}$ → **10, 15, 18, 22, 25, 28, 33, 43, 50, 55 → (25+28)/2 = 26.5**
- Not influenced by the extreme values (Outliers).

Practice Problem - 1

- A group of 10 students appeared in a test. Their obtained marks are given below.

60, 70, 80, 75, 65, 70, 80, 70, 65, 65

Sorted Order

60 65 65 65 70 70 70 75 80 80

Median = 70

Mean = 70

Practice Problem - 2

- A group of 10 students appeared in a test. Their obtained marks are given below.

60, 70, 80, 75, 65, 70, 100, 70, 65, 65

Mean = 720/10 = 72
Without considering the 100, mean = 620/9 = 68.89

For Trimmed mean,
Sorting the dataset: ~~**60**~~ 65 65 65 70 70 70 75 80 ~~**100**~~
For p = 1 → 60 and 100 will be discarded.
Then trimmed mean = 70

Median = 70

Practice Problem - 3

- A group of 10 students appeared in a test. Their obtained marks are given below.

60, 70, 80, 75, 65, 70, 10, 70, 65, 65

Mean = 630/10 = 63
Without considering the 10, mean = 620/9 = 68.89

For Trimmed mean,
Sorting the dataset: ~~**10**~~ 60 65 65 65 70 70 70 75 ~~**80**~~
For p = 1 → 10 and 80 will be discarded.
Then trimmed mean = 67.5

Median = 67.5

WEIGHTED MEAN

- It is calculated by multiplying each data value $\mathbf{x_i}$ by a weight $\mathbf{w_i}$ and dividing their sum by the sum of the weights $\mathbf{(w_i)}$.

$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i}^{n} w_i}$$

- Some values are intrinsically more variable than others, and highly variable observations are given a lower weight.

MODE

- The value that occurs most frequently.
- Not that much useful.

DISPERSION

- In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed. Common examples of measures of statistical dispersion are the **variance, standard deviation,** and **interquartile range.**

- Dataset 1 → 30 40 50 60 70 → Mean = 50, Median = 50, Variance = lower
- Dataset 2 → 0 25 50 75 100 → Mean = 50, Median = 50, Variance = higher

Mean absolute deviation

- The mean of the absolute value of the deviations from the mean.

$$\text{Mean absolution deviation} = \frac{\sum_{i=1}^{n} | x_i - \bar{x} |}{n}$$

- Dataset 1 → 30 40 50 60 70 → Mean = 50, Median = 50, Mean Absolute Deviation = 12
- Dataset 2 → 0 25 50 75 100 →
  - Mean = 50, Median = 50,
  - Mean Absolute Deviation = 30

Standard Deviation

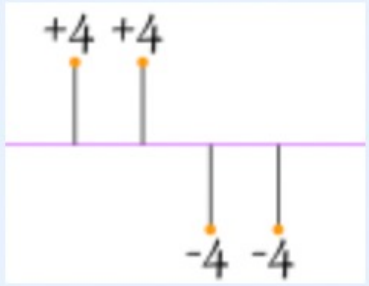
- The Standard Deviation is a measure of how spread out numbers are.
- Its symbol is **σ** (the greek letter sigma)
- The formula is easy: it is the **square root** of the **Variance.**
- So now you ask, "What is the Variance?"

Variance

- The Variance is defined as:
  - The average of the **squared** differences from the Mean.
- To calculate the variance follow these steps:
  - Work out the Mean (the simple average of the numbers).
  - Then for each number: subtract the Mean and square the result (the *squared difference*).
  - Then work out the average of those squared differences. (Why Square?)

Why squared difference?

If we just add up the differences from the mean ... the negatives cancel the positives:



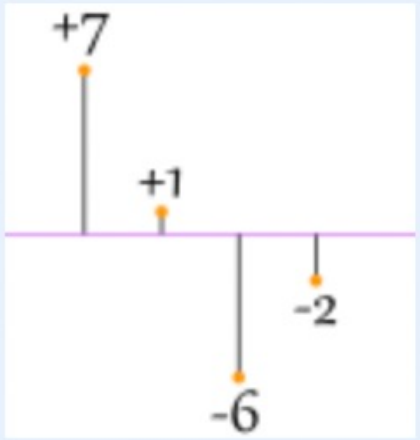
$$\frac{4 + 4 - 4 - 4}{4} = 0$$

So that won't work. How about we use absolute values?

$$\frac{|4| + |4| + |-4| + |-4|}{4} = \frac{4 + 4 + 4 + 4}{4} = 4$$

Why squared difference?

That looks good (and is the Mean Deviation), but what about this case:

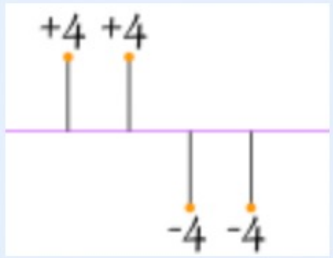


$$\frac{|7| + |1| + |-6| + |-2|}{4} = \frac{7 + 1 + 6 + 2}{4} = 4$$

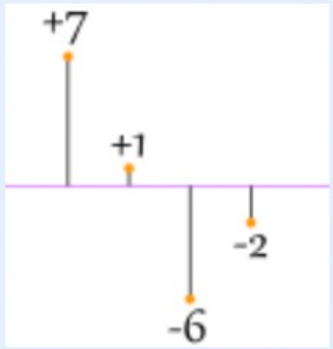
Oh No! It also gives a value of 4, Even though the differences are more spread out.

Why squared difference?

So let us try squaring each difference (and taking the square root at the end):



$$\sqrt{\left(\frac{4^2 + 4^2 + (-4)^2 + (-4)^2}{4}\right)} = \sqrt{\left(\frac{64}{4}\right)} = 4$$

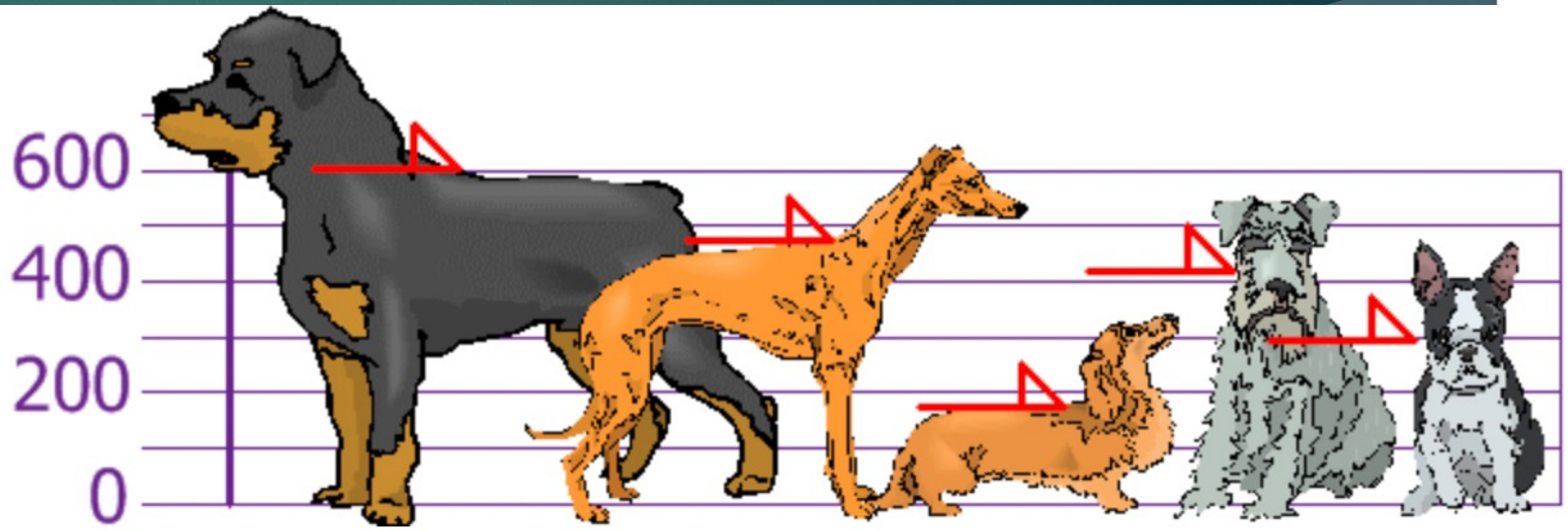


$$\sqrt{\left(\frac{7^2 + 1^2 + (-6)^2 + (-2)^2}{4}\right)} = \sqrt{\left(\frac{90}{4}\right)} = 4.74\ldots$$

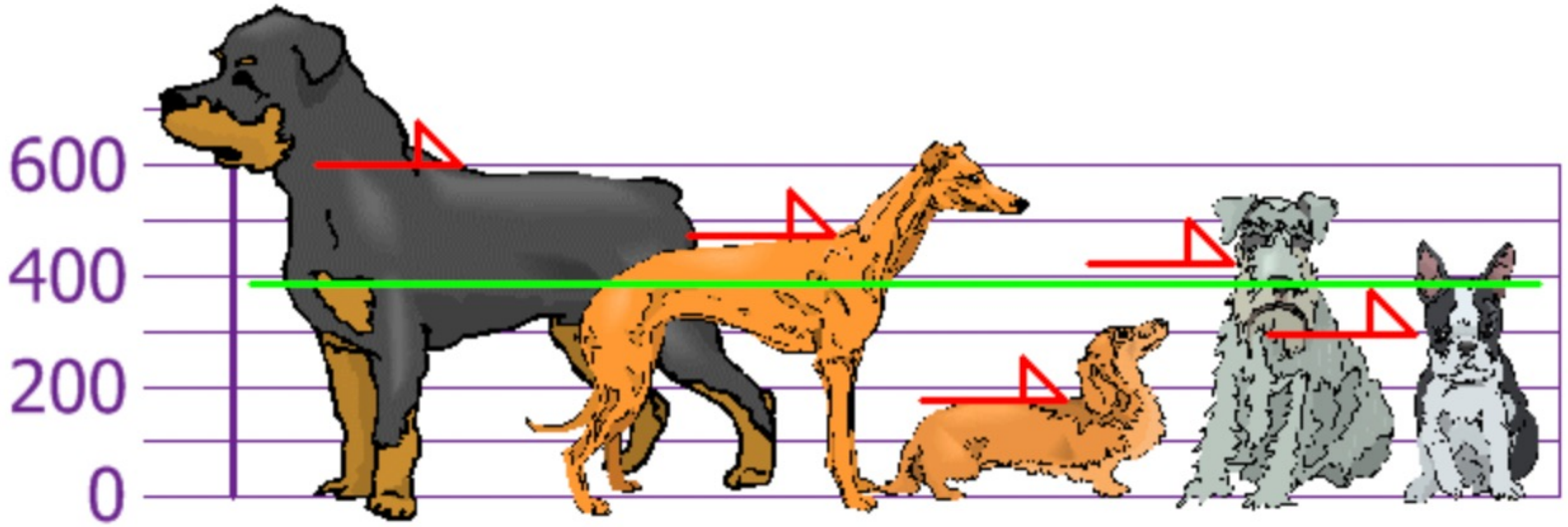
That is nice! The Standard Deviation is bigger when the differences are more spread out ... just what we want.

Back to Variance: An Example

- You and your friends have just measured the heights of your dogs (in millimetres):
- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.
- Find out the Mean, the Variance, and the Standard Deviation.



Back to Variance: An Example

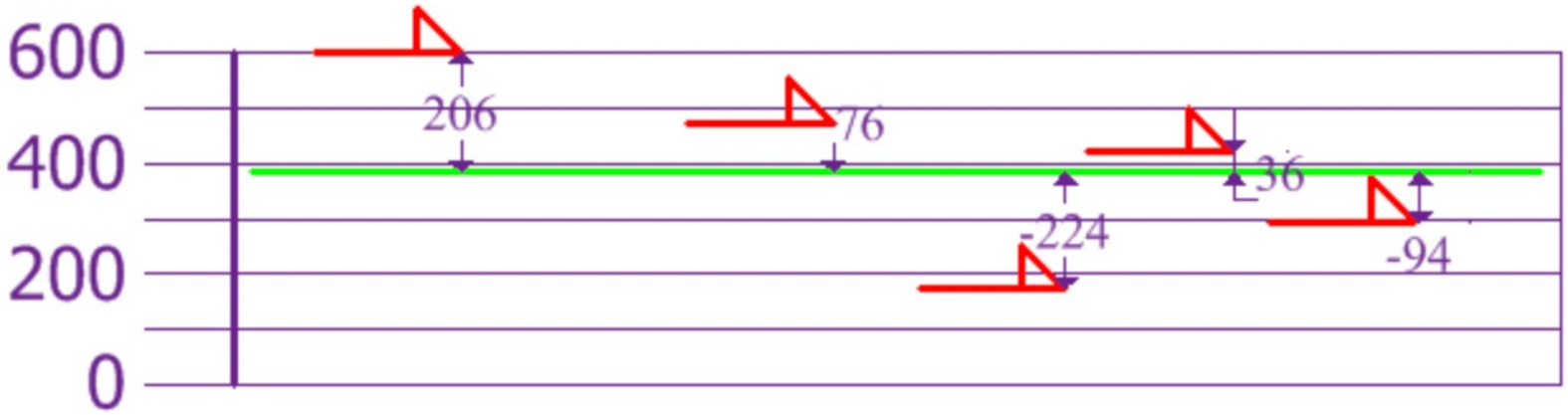


$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5}$$

$$= \frac{1970}{5}$$

$$= 394$$

- Now we calculate each dog's difference from the Mean:



Back to Variance: An Example

- To calculate the Variance, take each difference, square it, and then average the result:

Variance

$$\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

$$= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$$

$$= \frac{108520}{5}$$

$$= 21704$$

From this example: Standard Deviation

- And the Standard Deviation is just the square root of Variance, so:

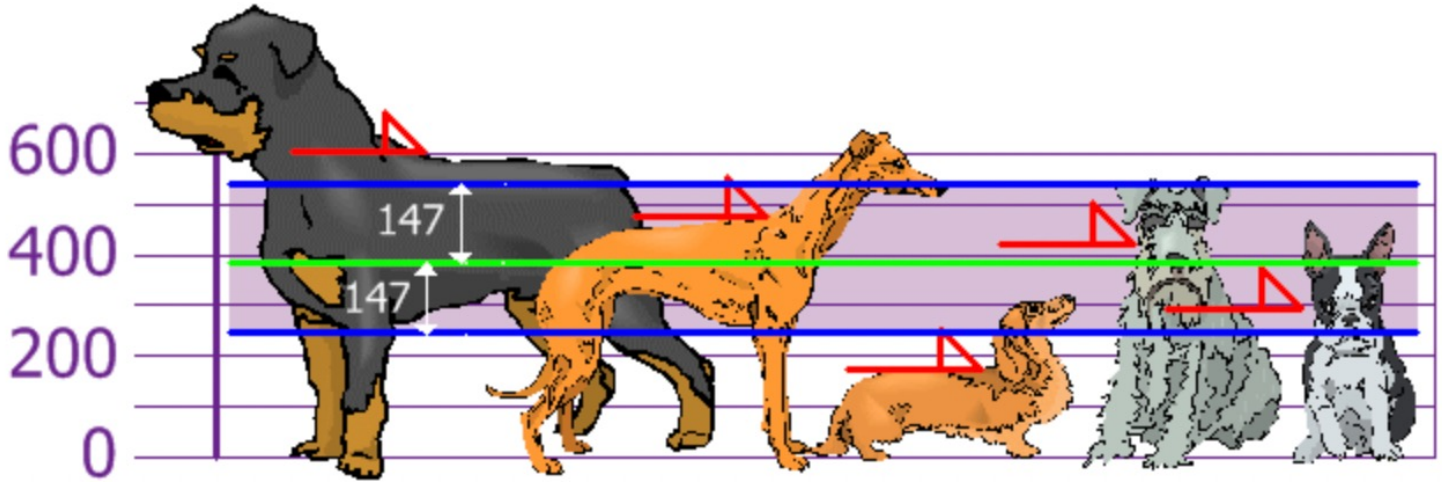
Standard Deviation

$\sigma = \sqrt{21704}$

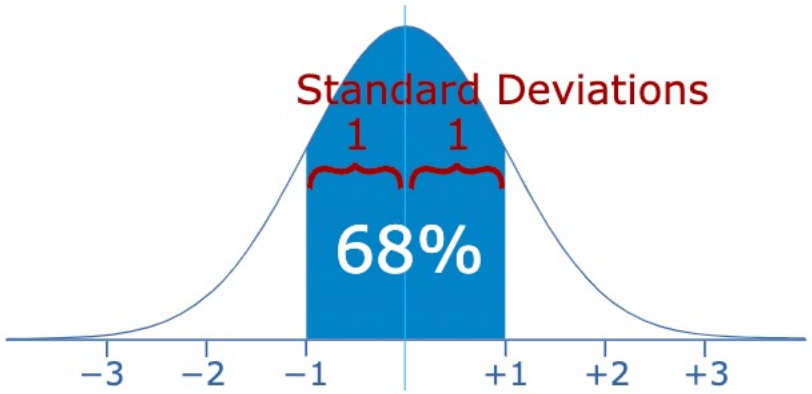
$= 147.32...$

$=$ **147** *(to the nearest mm)*

- And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean.
- So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.
- Rottweilers **are** tall dogs. And Dachshunds **are** a bit short, right?



- We can expect about 68% of values to be within plus-or-minus 1 standard deviation.



Population vs Sample data

- Our example has been for a **Population** (the 5 dogs are the only dogs we are interested in).
- But if the data is a **Sample** (a selection taken from a bigger Population), then the calculation changes!

When you have "N" data values that are:

- **The Population**: divide by **N** when calculating Variance (like we did)
- **A Sample**: divide by **N-1** when calculating Variance

- All other calculations stay the same, including how we calculated the mean.

Example: if our 5 dogs are just a **sample** of a bigger population of dogs, we divide by **4 instead of 5** like this:

- Sample Variance = 108,520 / **4** = **27,130**
- Sample Standard Deviation = √27,130 = **165** (to the nearest mm)

Think of it as a "correction" when your data is only a sample.

(Sample) variance

- The sum of squared deviations from the mean divided by n – 1 where n is the number of data values.
- Average of the squared deviations.

$$\text{Variance} = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

- Why the denominator is n-1 instead of n? To obtain the true estimate of the variance with regard to the population, it is divided by n-1 so that the estimated value would be little larger. This value it is known as the true estimate of the variance.

Practice Example

- Dataset 1 → 30 40 50 60 70 → Mean = 50, Median = 50, Variance = lower
- Dataset 2 → 0 25 50 75 100 → Mean = 50, Median = 50, Variance = higher
- Dataset 1
  - $(30-50)^2 + (40-50)^2 + (50-50)^2 + (60-50)^2 + (70-50)^2 / 5-1 = 250$
  - Standard Deviation = $\sqrt{250} = 15.8113$
- Dataset 2
  - $(0-50)^2 + (25-50)^2 + (50-50)^2 + (75-50)^2 + (100-50)^2 / 5-1 = 1562.5$
  - Standard Deviation = $\sqrt{1562.5} = 39.5284$

STANDARD deviation

- The square root of the variance.

$$\text{Standard deviation} = s = \sqrt{\text{Variance}}$$

- Standard deviation is preferred over the mean absolute deviation.

MEDIAN of MEDIAN ABOSULTE DEVIATION

- The median of the absolute value of the deviations from the median.

$$\text{Median absolute deviation} = \text{Median}\left(|x_1 - m|,\ |x_2 - m|,\ \ldots,\ |x_N - m| \right)$$

- Not influenced by the extreme values (Outliers).
- Dataset 1 → 30 40 50 60 70 → Median (20, 10, 0, 10, 20) → Median (0, 10, 10, 20, 20) = 10
- Dataset 2 → 0 25 50 75 100 →Median (50, 25 , 0, 25, 50) → Median (0, 25, 25, 50, 50) = 25

Summary of dispersion measures

	Mean	Median	Mean Abs. Dev.	Variance	Standard Dev.	Median of Median Abs. Dev
Dataset 1	50	50	12	250	15.8113	10
Dataset 2	50	50	30	1562.5	39.5284	25

Practice EXample

- **Find the mean, median, mode, range, variance, standard deviation, mean absolute deviation and median of the median absolute deviation for the following list of values:**

8, 9, 10, 10, 10, 11, 11, 11, 12, 13

→ **Mean = 10.5**

→ **Median = 10.5**

→ **Mode = 10, 11 (Bi-modal dataset)**

→ **Range = 13-8 = 5**

→ **Variance = 2.055556**

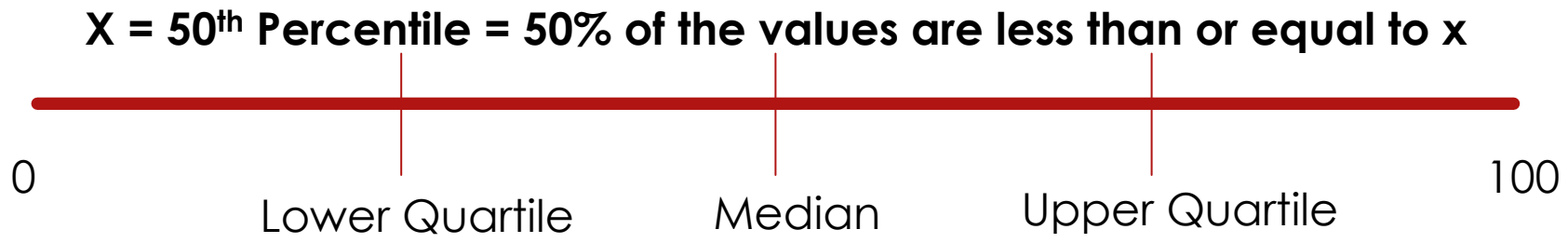
→ **Standard Deviation = 1.4337**

→ **Mean Absolute Deviation = 1.1**

→ **Median of the Median Absolute Deviation = 0.5**

percentile

- The value such that P percent of the values take on this value or less and (100–P) percent take on this value or more.
- The 25-th percentile is the lower quartile (Q1).
- The 50-th percentile is the median.
- The 75-th percentile is the upper quartile (Q3).
- Quantiles are the same as percentiles but are indexed by sample fractions rather than by sample percentage.
- 80^{th} Percentile = 0.8 Quantile



Finding Percentile

8, 9, 10, 10, 10, 11, 11, 11, 12, 13

- Total data points, n = 10
- 50^{th} percentile = 50% X 10 = 5^{th} value = 10
- Median = (10+11)/2 = 10.5

8, 9, 10, 10, 10, 11, 11, 11, 12, 13, 15, 16, 16, 17, 20

- Total data points, n = 15
- 50^{th} percentile = 50% X 15 = 7.5^{th} value = 11
- Median = 8^{th} value = 11
- 20^{th} percentile = 20% X 15 = 3^{rd} value = 10
- 75^{th} percentile = 75% X 15 = 11.25^{th} value = 15 + (16-15)X0.25 = 15.25
- 95^{th} percentile = 95% X 15 = 14.25^{th} value = 17 + (20-17)X0.25 = 17.75

Interquartile range

- The difference between the 75th percentile and the 25th percentile.
- IQR = Q3 – Q1
- Range = max - min

Useful resources

- Chapter 1, Practical Statistics for Data Scientists by Bruce and Bruce
- https://www.geeksforgeeks.org/python-pandas-dataframe/
- https://www.tutorialspoint.com/python_pandas/python_pandas_descriptive_statistics.htm
- https://medium.com/swlh/statistical-functions-of-pandas-2862c290053a
- https://pandas.pydata.org/pandas-docs/stable/reference/frame.html

Thank you