MEASURES OF DISPERSION

MEASURES OF DISPERSION

A measure of dispersion can be seen as a measure of **RISK**. If I take a randomly selected data point from my dataset (or in the future), how far can I get from the typical/average value?

- Variance
- Standard deviation
- Range
- Coefficient of Variation
- Etc.

DEFINITIONS

- Let us suppose we have a numeric univariate dataset with N values,

  $X = \{x_1, x_2, \ldots x_N\}$, and that the arithmetic mean is known and denoted by symbol μ.

$$Variance = \frac{1}{N}\sum_i (x_i - \mu)^2$$

$$\text{Standard Deviation or } \sigma = \sqrt{\frac{1}{N}\sum_i (x_i - \mu)^2} = \sqrt{Variance}$$

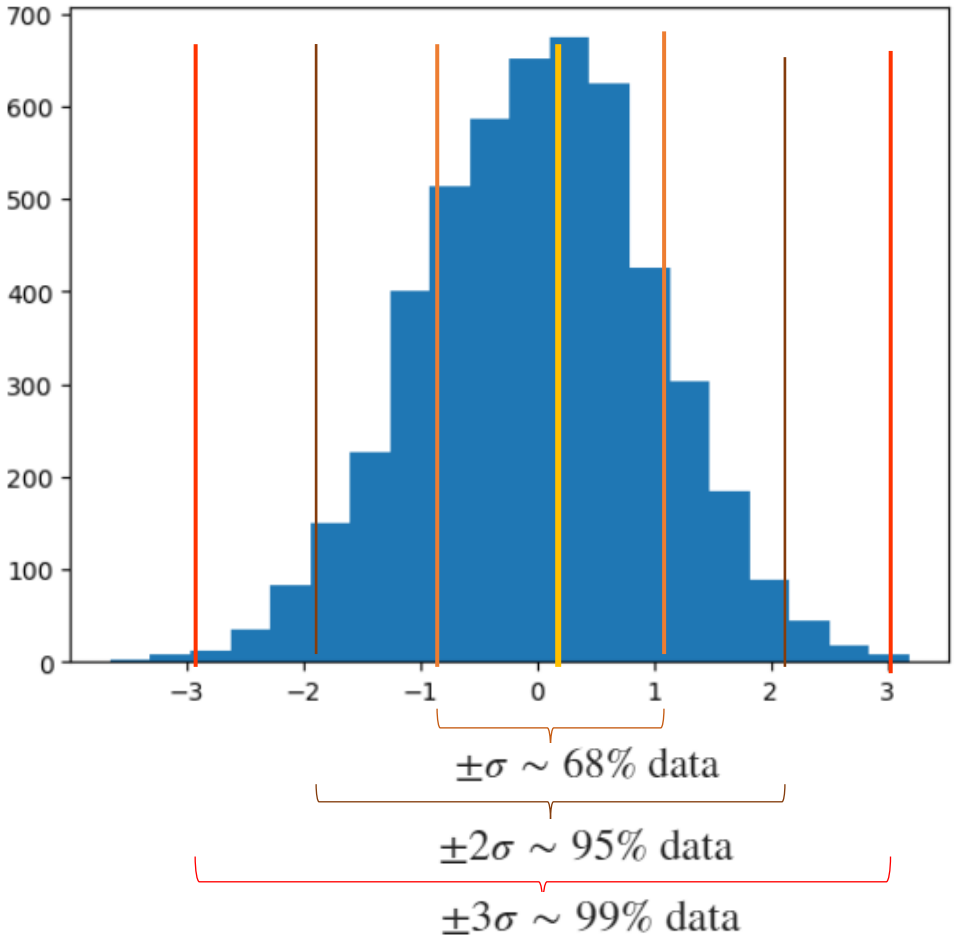
$$\text{Range} = \max(X) - \min(X)$$

$$\text{Coefficient of Variation} = \frac{\sigma}{\mu}$$

STANDARD DEVIATION: EXAMPLE AND REASONING

The following frequency table and histograms correspond to the raw data in file *data_lecture3[data1]*

	Bin Boundaries	Counts
0	[-3.65 , -3.31]	2.0
1	[-3.31 , -2.97]	8.0
2	[-2.97 , -2.63]	12.0
3	[-2.63 , -2.29]	35.0
4	[-2.29 , -1.94]	82.0
5	[-1.94 , -1.6]	149.0
6	[-1.6 , -1.26]	226.0
7	[-1.26 , -0.92]	401.0
8	[-0.92 , -0.58]	513.0
9	[-0.58 , -0.23]	586.0
10	[-0.23 , 0.11]	651.0
11	[0.11 , 0.45]	674.0
12	[0.45 , 0.79]	625.0
13	[0.79 , 1.13]	426.0
14	[1.13 , 1.48]	304.0
15	[1.48 , 1.82]	184.0
16	[1.82 , 2.16]	89.0
17	[2.16 , 2.5]	44.0
18	[2.5 , 2.84]	18.0
19	[2.84 , 3.19]	9.0




```
mean=np.mean(values)
variance = np.var(values)
std = np.std(values)
Rg =  np.max(values) - np.min(values)
coefvar = std/mean
print ('mean =',round(mean,3))
print('Variance =',round(variance,3))
print('Standard Deviation =', round(std,3))
print('Range =',round(Rg, 3))
print('Coefficient of Variation =',round(coefvar,3))
```

```
mean = 0.006
Variance = 0.991
Standard Deviation = 0.995
Range = 6.84
Coefficient of Variation = 174.122
```

Subtle point in the definitions

For statistical and profound reasons, there are 2 competing definitions of variance and standard deviation for samples of finite size.

a)

$$Variance = \frac{1}{N}\sum_i (x_i - \mu)^2$$

$$\sigma = \sqrt{\frac{1}{N}\sum_i (x_i - \mu)^2}$$

b)

$$\text{Variance} = \frac{1}{N-1}\sum_i (x_i - \mu)^2$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_i (x_i - \mu)^2}$$

The difference is always small specially for large datasets (N > 30)

However, some Python libraries/functions implement a) and some other implements b).
Usually this is not an issue, but do not be surprised if you see slightly different results depending on the function that you are using!

Example:
`numpy.std` and `numpy.var` use a)
but `numpy.cov` uses b)

```
print(np.cov(x,x)[0,0])
print(np.var(x))
```

```
1.016063103836798
1.0059024727984303
```

Python practice measures of dispersion

- Let's compute some measure of location for the numeric data in file data_lecture3.xlsx