Descriptive Statistics

Foundations of AI Academy

Descriptive Statistics

Quantitatively describe a set of data

Measures of Location – central tendency of data

Measures of Spread – how data are spread

Measures of Symmetry – shape of data distribution

Measures of Location

Mean – sum of values divided by number of values

Median – middle value in ordered list of values

Mode – most frequently occurring value

Measures of Location

Mean – sum of values divided by number of values

```
>>> np.mean([1,2,3]) ➔ 2.0
```

Median – middle value in ordered list of values

```
>>> np.median([1,2,3]) ➔ 2.0
```

Mode – most frequently occurring value

```
# Not in NumPy
```

Measures of Spread

Range – difference between largest and smallest values

Variance – how much the data varies from the mean

Standard Deviation – the "average" spread around the mean (square root of the variance)

Measures of Spread

Range – difference between largest and smallest values

```
>>> np.max(array) - np.min(array)
```

Variance – how much the data varies from the mean

```
>>> np.var([1,2,3]) ➔ 2.0
```

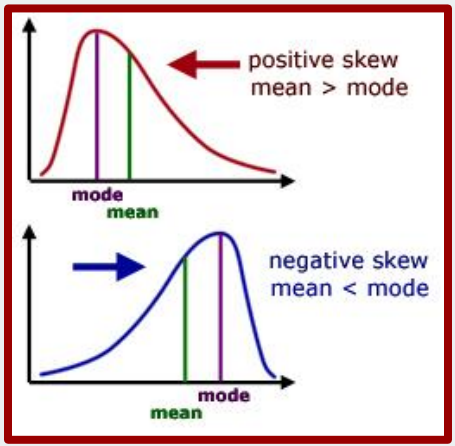
Standard Deviation – the "average" spread around the mean (square root of the variance)

```
>>> np.std([1,2,3]) ➔ 1.414…
```

Measure of Symmetry

Skewness – a measure of the lack of symmetry to the left and right of the center of a set of data

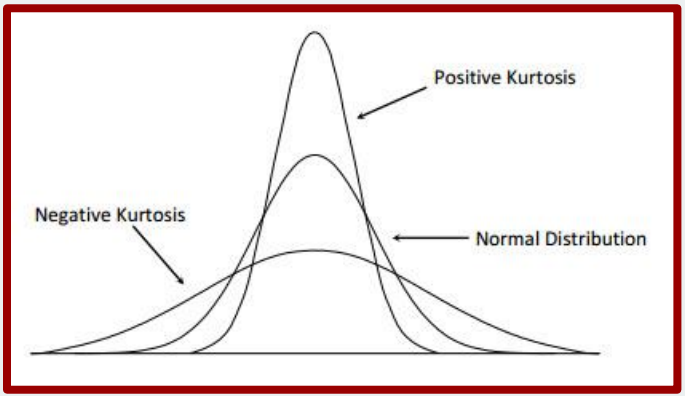
```
>>> scipy.stats.skew(array)
```



Measure of Symmetry

Kurtosis – a measure of the "tailedness," or the number of "outliers," of a data set, relative to a normal distribution

```
>>> scipy.stats.kurtosis(array)
```



Determining Normality

1. Check a histogram of your data

   Visual inspection will be the quickest way to determine if your data is **not normal**

2. Review the data's skewness

   This isn't an exact number, but the further away from **zero (0)**, the more non-normal the data.

3. Use the Kolmogorov-Smirnov (K-S) and Shapiro-Wilk (S-W) tests

   If the test is **not significant**, then your data is **normal**.

Checking for Normality

Kolmogorov-Smirnov (K-S)

```
>>> scipy.stats.kstest(array, 'norm')
```

Shapiro-Wilk (S-W)

```
>>> scipy.stats.shapiro(array)
```

Both return the test's and p-value.

Correlation

A measurement of the relationship(s) between data

Useful for

- Predicting one quantity from another
- May indicate a causal relationship

Correlation

A measurement of the relationship(s) between data

Useful for

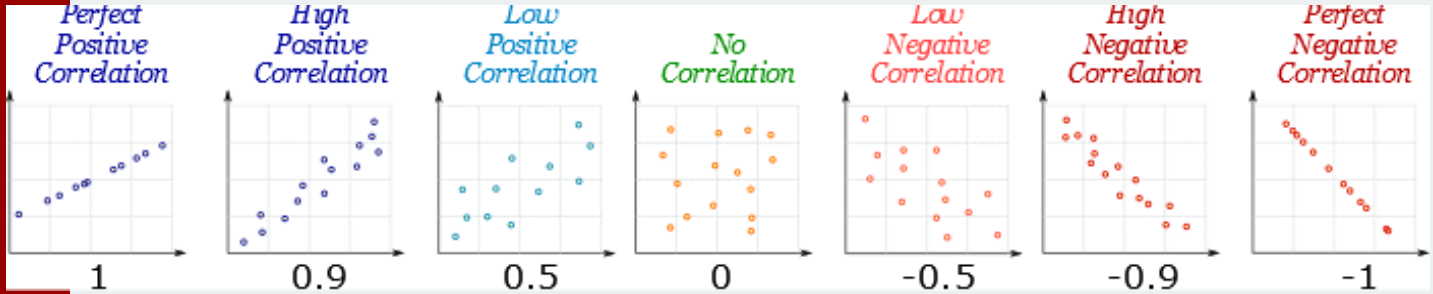
- Predicting one quantity from another
- **May** indicate a causal relationship

Correlation does not imply Causation!

Correlation

Positive when values increase together

Negative when one value increases as other decreases



Type of Correlation Measures

Pearson

linear association between continuous variables

Spearman's Rank

Special case of Pearson applied to ranked/sorted variables

Used for both continuous and discrete data

Kendall's Tau

used for discrete data only

Regression Analysis

Statistical process for estimating relationships

- Dependent variable
- One or more predictor variables

Linear regression

- Captures linear relationships
- Produces a linear function

Non-linear

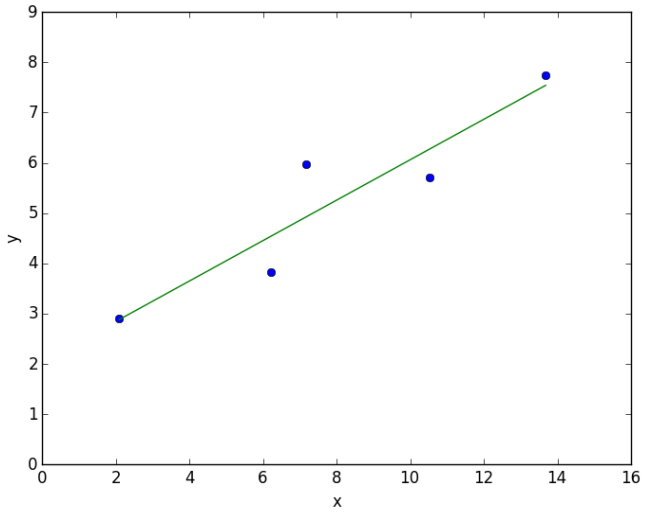
- Polynomial curve relationship

Simple Example

i	1	2	3	4	5
x	2.10	6.22	7.17	10.52	13.68
y	2.90	3.83	5.98	5.71	7.74

Simple Example

i	1	2	3	4	5
x	2.10	6.22	7.17	10.52	13.68
y	2.90	3.83	5.98	5.71	7.74



Scatter plot of data with hand-drawn approximation of line