Data Science with Python

Tran Vu Khanh, PhD

University of Wollongong

Statistical and Data Techniques in Data Science

Outline

- Statistical Description
  - ✓ Central Tendencies
  - ✓ Dispersion
  - ✓ Shape of Data
  - ✓ Covariance and Correlation
  - ✓ Distribution
- Statistical Inference

Statistical Description

What is statistics?

Science of data collection, summarization, analysis and interpretation

Statistical analysis is at the core of data science. Using statistics,

- we can learn about the distribution of the data,
- how much of variance there is between values, and
- how values for one feature of the data seem to influence other values.

Descriptive versus Inferential Statistics:

- Descriptive Statistic: Data description (summarization) such as center, variability and shape.
- Inferential Statistic : Drawing conclusion beyond the sample studied, allowing for prediction.

Statistical Description of Data

Statistics describes a numeric set of data by its

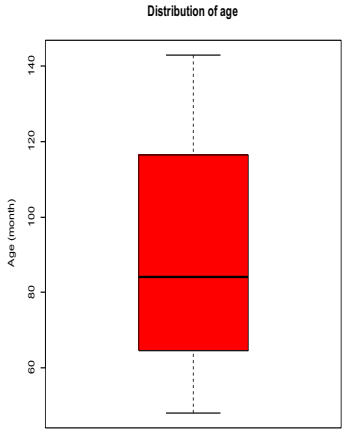
- Center (mean, median, mode etc)
- Variability (standard deviation, range etc)
- Shape (skewness, kurtosis etc)

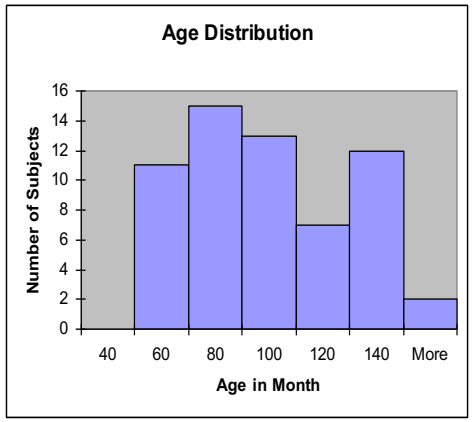
Statistics describes a categorical set of data by

- Frequency, percentage or proportion of each category

How statistics help us

By simply looking at the data, we fail to produce any informative account to describe the data, however, statistics produce a quick insight into data using graphical and numerical statistical tools





Mean	90.41666667
Standard Error	3.902649518
Median	84
Mode	84
Standard Deviation	30.22979318
Sample Variance	913.8403955
Kurtosis	-1.183899591
Skewness	0.389872725
Range	95
Minimum	48
Maximum	143
Sum	5425
Count	60

Statistics Vocabulary

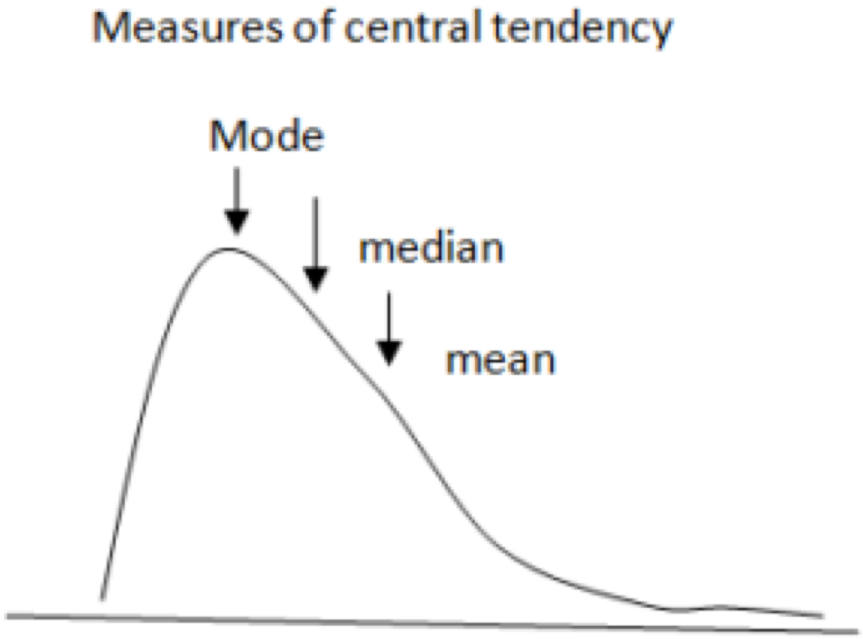
Population: The entire group one desires information about

Sample: A subset of the population taken because the entire population is usually too large to analyze. Its characteristics are taken to be representative of the population

Variables are the quantities measured in a sample. They may be classified as:

- **Quantitative** i.e. numerical
- **Continuous** (e.g. pH of a sample, patient cholesterol levels)
- **Discrete** (e.g. number of bacteria colonies in a culture)
- **Categorical**
- **Nominal** (e.g. gender, blood group)
- **Ordinal** (ranked e.g. mild, moderate or severe illness). Often ordinal variables are re-coded to be quantitative.

Central Tendencies



$$\text{Mean} = \frac{\text{sum of all values}}{\text{total number of values}}$$

$$\text{Median} = \text{middle value (when the data are arranged in order)}$$

$$\text{Mode} = \text{most common value}$$

Central Tendencies

- Center measurement is a summary measure of the overall level of a dataset
- Commonly used methods are mean, median, mode, quantiles.

Mean: Summing up all the observation and dividing by number of observations.

Example: Mean of 20, 30, 40 is (20+30+40)/3 = 30.

Formula:

Population

$$\mu = \frac{1}{N}\sum_{j=1}^{N} x_j$$

Sample

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

Central Tendencies

Median: The middle value in an ordered sequence of observations. That is, to find the median we need to order the data set and then find the middle value.

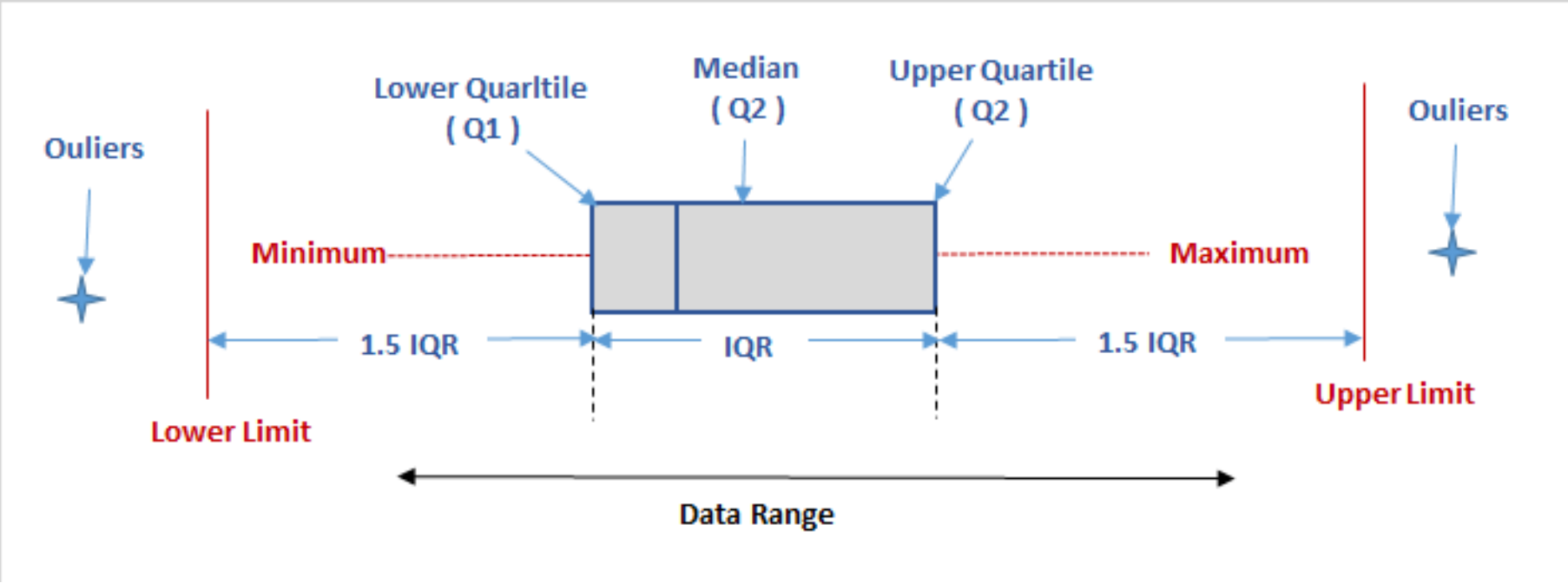
In case of an even number of observations the average of the two middle most values is the median.

For example, to find the median of {9, 3, 6, 7, 5},
- we first sort the data giving {3, 5, 6, 7, 9},
- then choose the middle value 6.

If the number of observations is even, e.g., {2, 3, 5, 6, 7, 9}, then the median is the average of the two middle values from the sorted sequence, in this case, (5 + 6) / 2 = 5.5.

Mode: The value that is observed most frequently. The mode is undefined for sequences in which no observation is repeated.

Central Tendencies



Central Tendencies

Quartiles: Quartiles are divided into four regions that cover the total range of observed values. Cut points for these regions are known as quartiles.

The first quartile (Q1) is the first 25% of the data. The second quartile (Q2) is between the 25^{th} and 50^{th} percentage points in the data. The upper bound of Q2 is the median. The third quartile (Q3) is the 25% of the data lying between the median and the 75% cut point in the data.

Q1 is the median of the first half of the ordered observations and Q3 is the median of the second half of the ordered observations.

An example with 15 numbers

3 6 7 11 13 22 30 40 44 50 52 61 68 80 94

Q1 Q2 Q3

The first quartile is Q1=11. The second quartile is Q2=40 (This is also the Median.) The third quartile is Q3=61.

Central Tendencies

Deciles: If data is ordered and divided into 10 parts, then cut points are called Deciles

Percentiles: If data is ordered and divided into 100 parts, then cut points are called Percentiles. 25^{th} percentile is the Q1, 50^{th} percentile is the Median (Q2) and the 75^{th} percentile of the data is Q3.

Quantiles are cut points dividing the range of a probability distribution into continuous intervals.

Commonly,
Min=quantile(0%); Q1=quantile(25%); Q2=median; Q3=quantile(75%); Max =quantile(100%)

Five Number Summary: The five number summary of a distribution consists of the smallest (Minimum) observation, the first quartile (Q1), the median(Q2), the third quartile, and the largest (Maximum) observation written in order from smallest to largest.

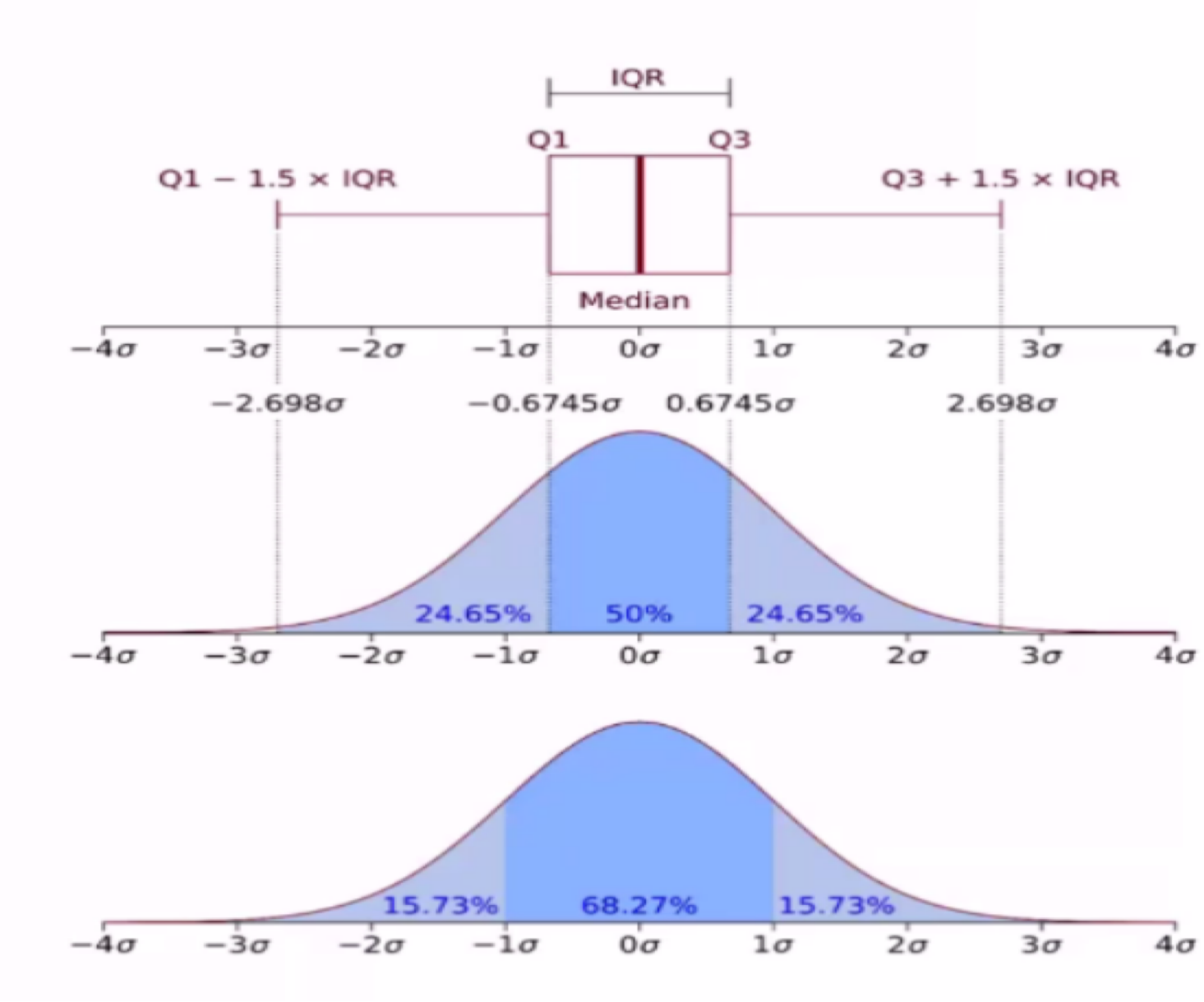
Central Tendencies

Box Plot: A box plot is a graph of the five number summary. The central box spans the quartiles. A line within the box marks the median. Lines extending above and below the box mark the smallest and the largest observations (i.e., the range). Outlying samples may be additionally plotted outside the range.

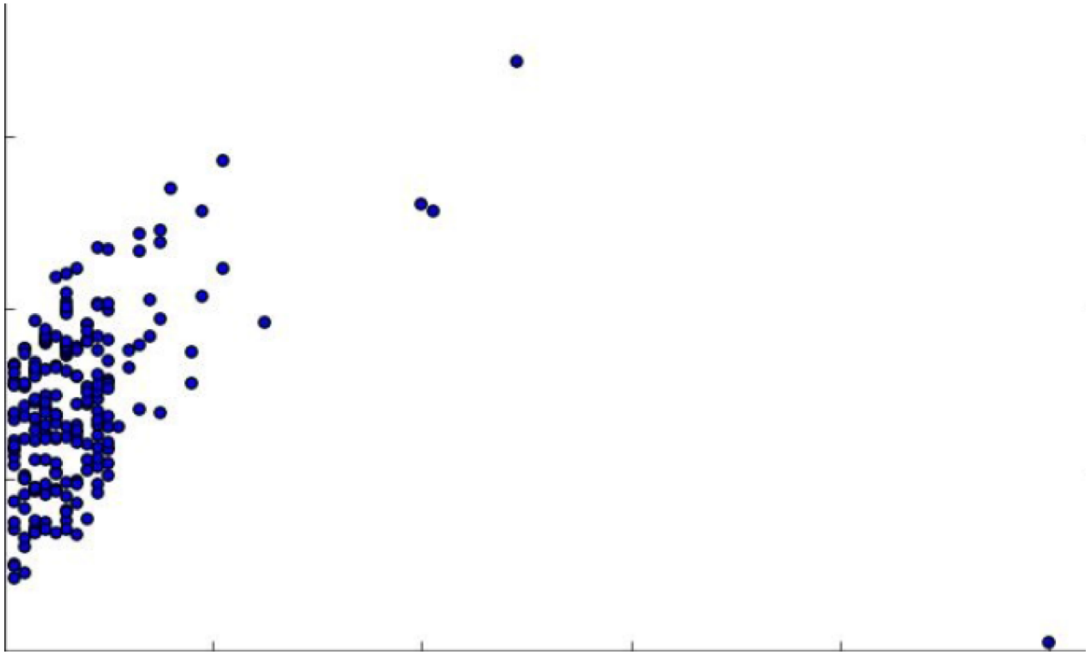
The **interquartile range (IQR)** formula is the first quartile quartile subtracted from the third quartile: **$IQR = Q_3 - Q_1$.**

Outliers:

- Outliers are value that are far from the central tendency
- Outliers might be caused by errors in collecting or processing the data, or they might be correct but unusual measurements.
- It is always a good ideas to check for outliers, and sometimes it is useful and appropriate to discard them
- Variance, Standard deviation, covariance, correlation can be very sensitive to outliers.


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

Central Tendencies



Example of outliers

Dispersion

- Variability (or dispersion) measures of how spread out data
- Commonly used methods: *range*, *variance*, *standard deviation*, *coefficient of variation etc*.

Range=max-min: The difference between the largest and the smallest observations. The range of 10, 5, 2, 100 is (100-2)=98. It's a crude measure of variability.

Variance: The variance of a set of observations is the average of the squares of the deviations of the observations from their mean.

Standard deviation (std)=square root of variance. Std measures how wide the data is spread around the mean

Formula:

Population

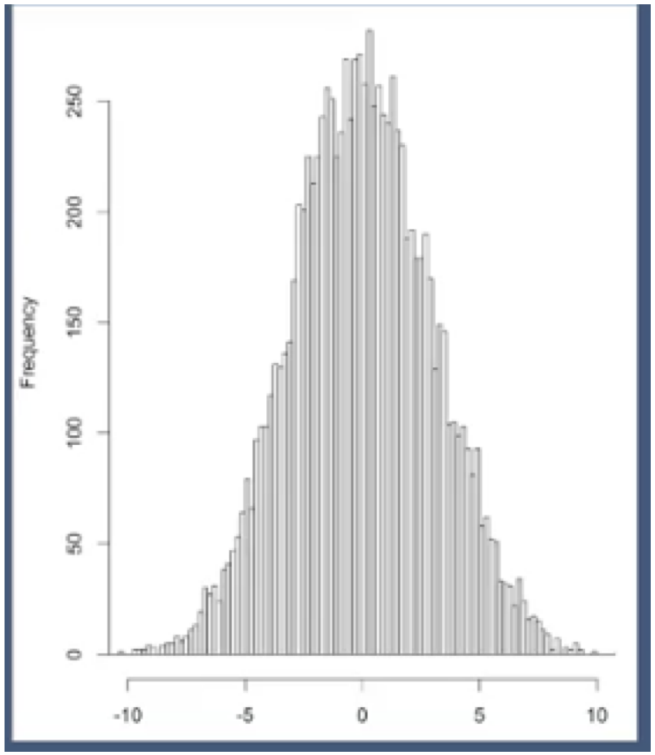
$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)^2$$

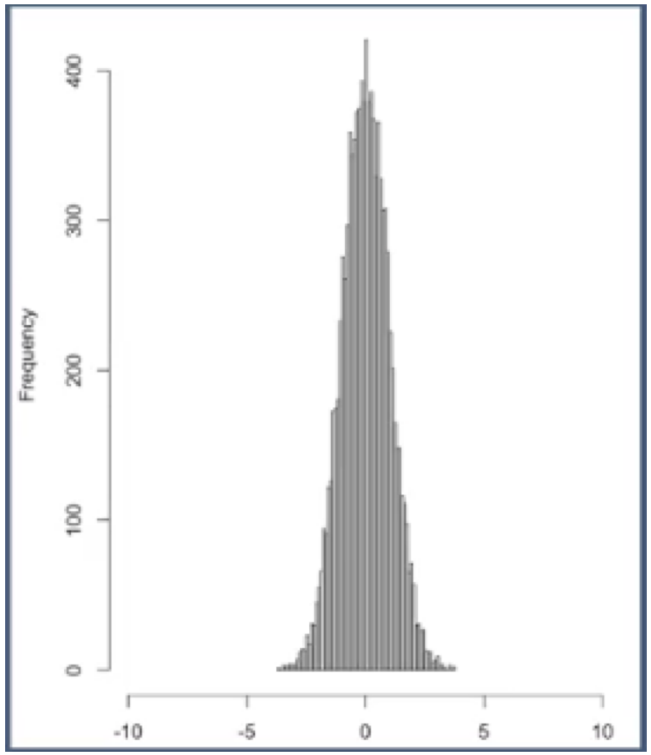
Sample:

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

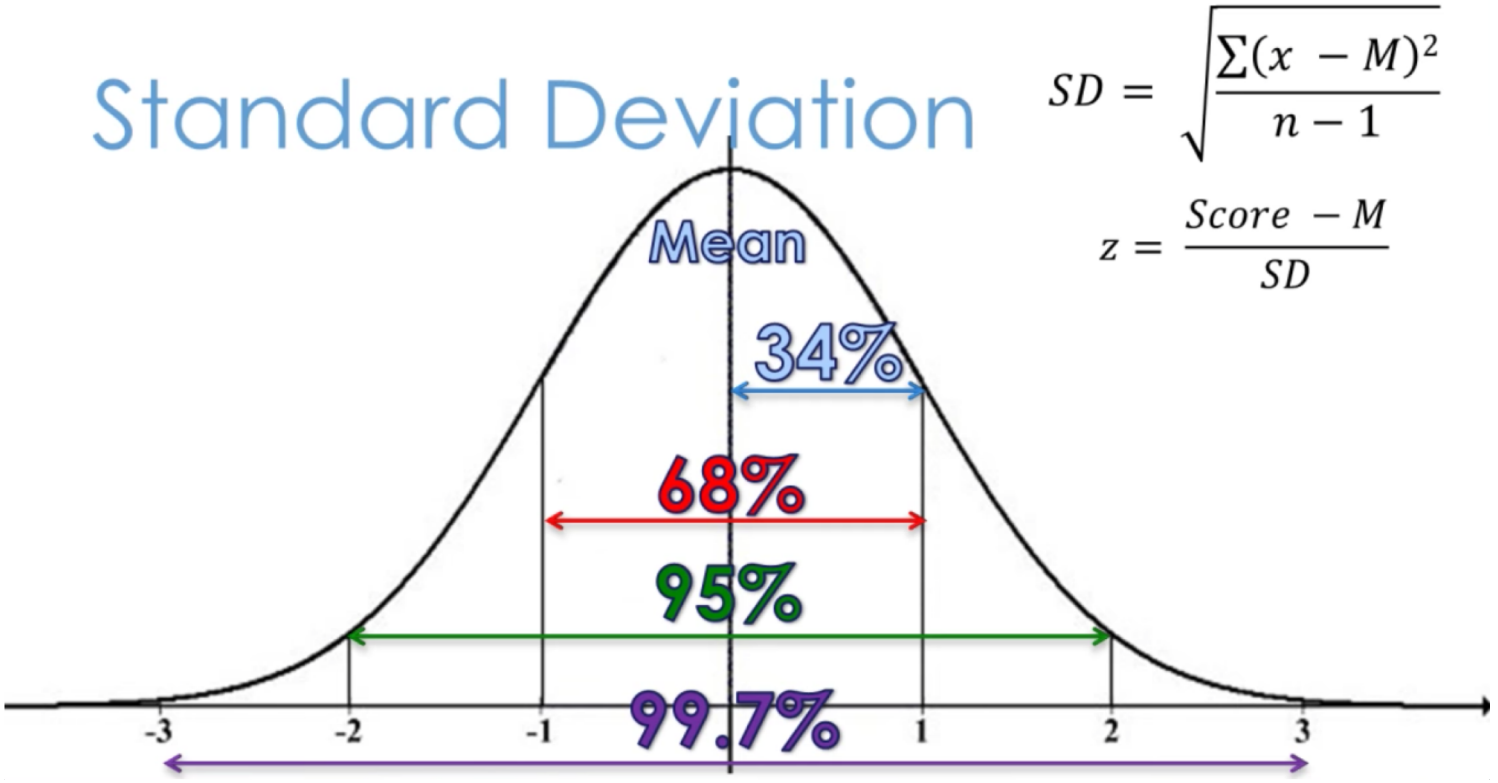
Dispersion

Standard deviation measures how wide the data is spread around the mean





Dispersion


Standard Deviation
$SD = \sqrt{\frac{\sum(x - M)^2}{n - 1}}$
$z = \frac{Score - M}{SD}$
Mean
34%
68%
95%
99.7%
-3
-2
-1
1
2
3

Dispersion

Coefficient of Variation: The standard deviation of data divided by it's mean. It is usually expressed in percent.

Coefficient of Variation formula:

CV for a population:

$$CV = \frac{\sigma}{\mu} * 100\%$$

CV for a sample:

$$CV = \frac{s}{\bar{x}} * 100\%$$

Shape of Data

Shape of data is measured by Skewness and Kurtosis

Skewness: How asymmetric data is spread around the mean

Formula:

Population

$$skewness = \frac{1}{N}\frac{\sum_{j=1}^{N}(x_j - \mu)^3}{\sigma^3}$$

Sample:

$$skewness = \frac{n}{(n-1)(n-2)}\frac{\sum_{j=1}^{n}(x_j - \bar{x})^3}{s^3}$$

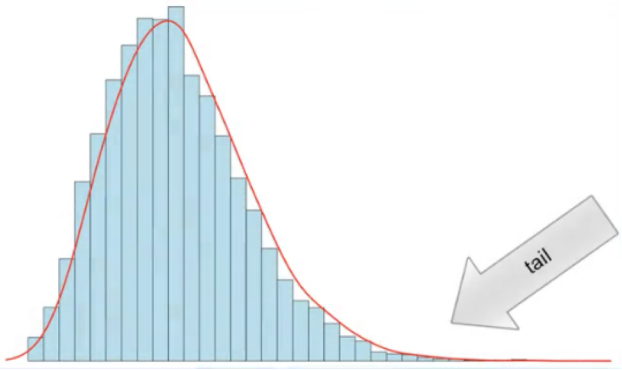
Shape of Data

Skewness: How asymmetric data is spread around the mean

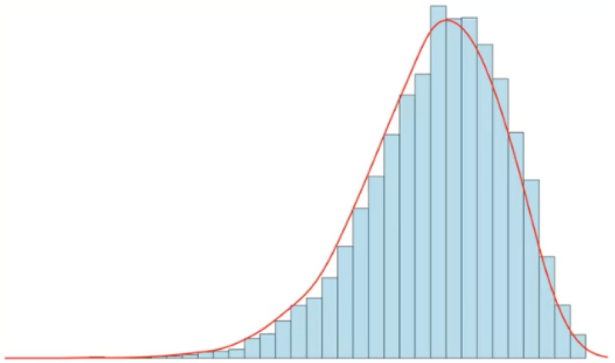
Positive or right skewed: Longer right tail

Negative or left skewed: Longer left tail

A normal distribution has a skew of 0.

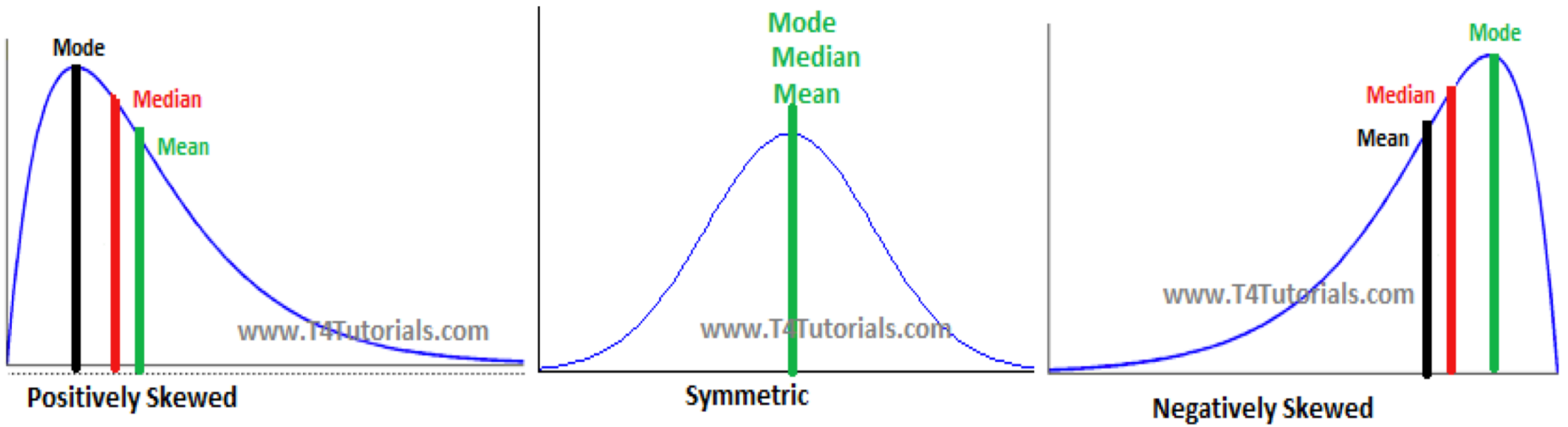


Positive skew

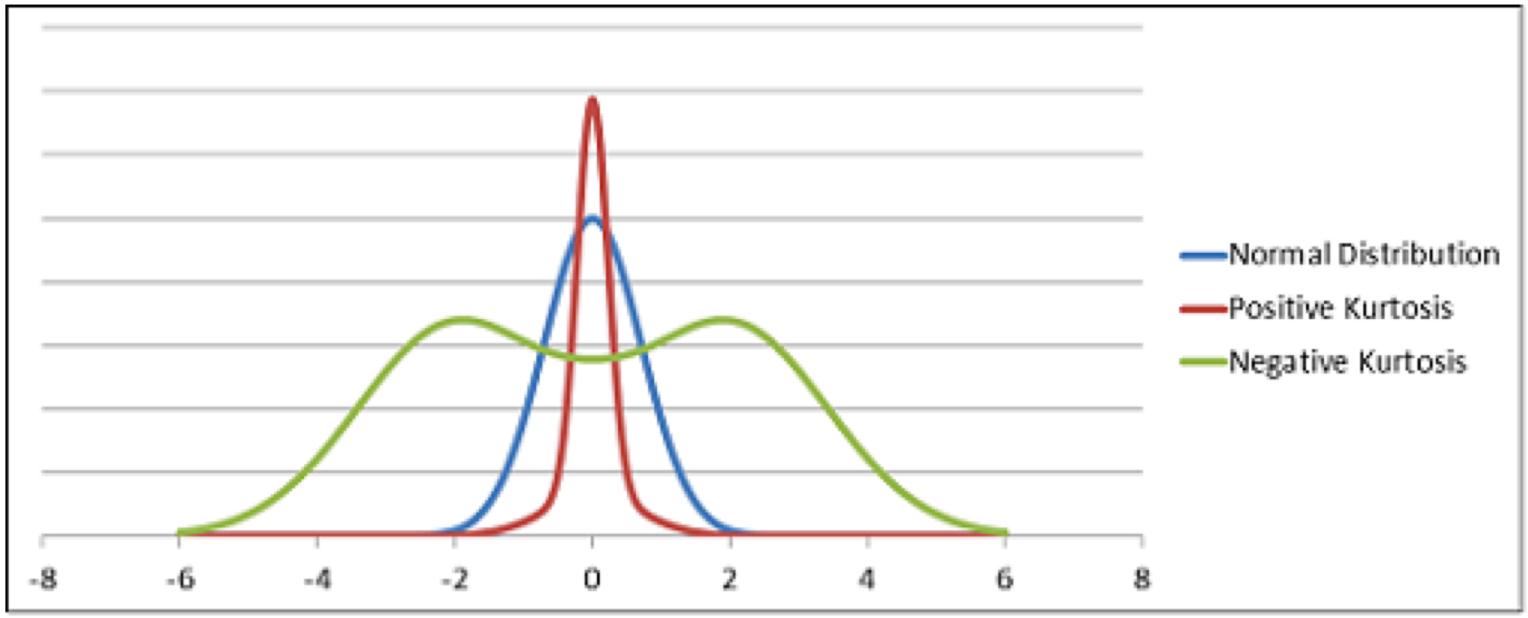


Negative skew

Shape of Data



Shape of Data



Shape of Data

Shape of data is measured by Skewness and Kurtosis

Kurtosis is a measure of peakedness. The higher of the kurtosis means the more outliers are present and the longer the tails of the distribution in the histogram are.

The kurtosis of normal distribution is 0.

Population

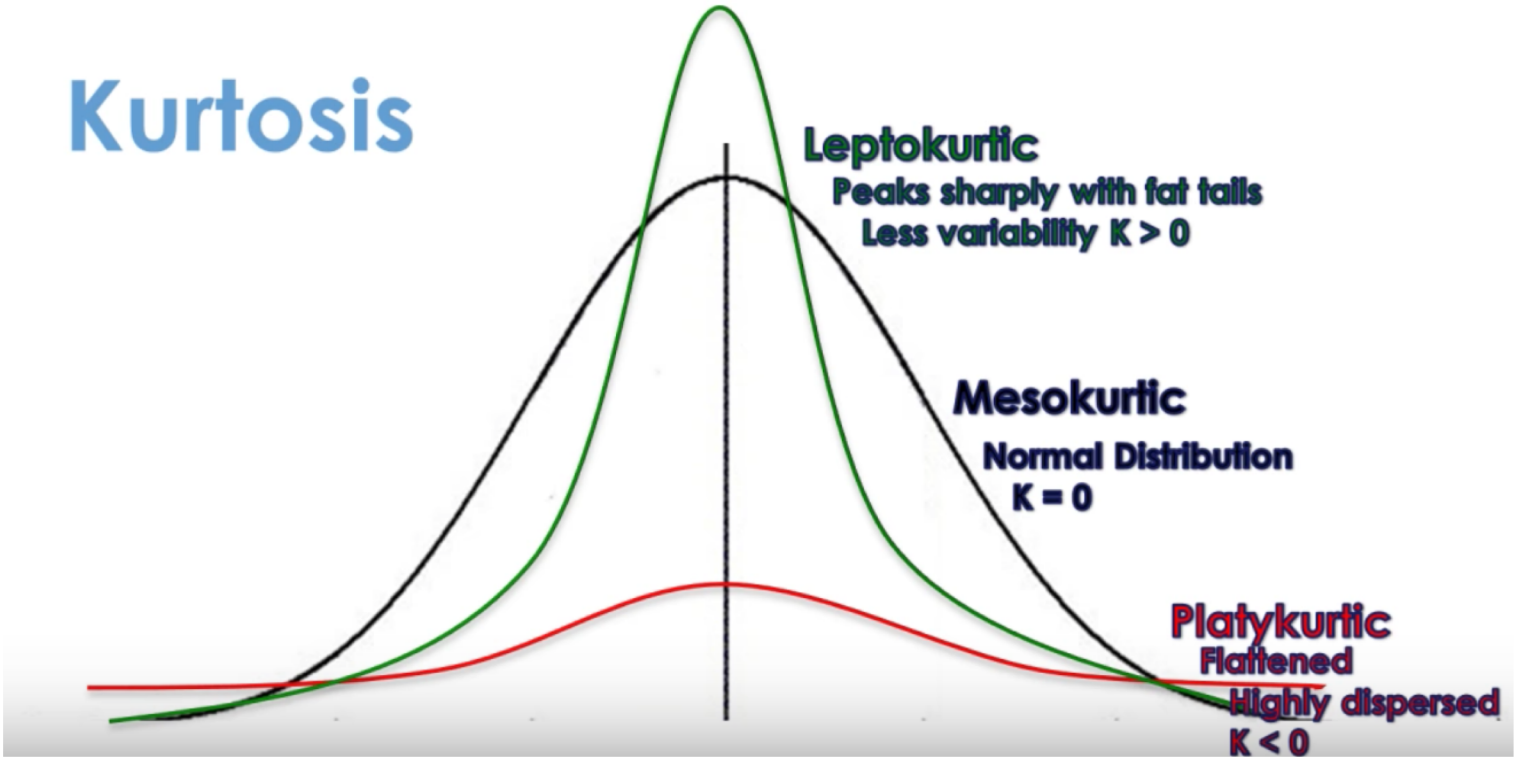
$$kurtosis = \frac{1}{N}\frac{\sum_{j=1}^{N}(x_j - \mu)^4}{\sigma^4}$$

Sample:

$$kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\frac{\sum_{j=1}^{n}(x_j - \bar{x})^4}{s^4}$$

Note: excess kurtosis =(simple) kurtosis -3

Shape of Data



Covariance and Correlation

Covariance is a measure of the tendency of two variables to vary together

Population Covarian Formula:

$$covaviance(X, Y) = \frac{1}{N} \sum_{j=1}^{N} (x_j - \mu(X))(y_j - \mu(Y))$$

Sample Covarian Formula:

$$covariance(X, Y) = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y})$$

Covariance can be hard to interpret, for a couple of reasons:

- Its units are product of the input's units, which can be hard to make sense of, e.g. kilogram-meters.
- Covariance can be large or small might depend only one variable.

One solution of these problems are to divide the deviation by standard deviation (std).

Covariance and Correlation

The correlation is unitless and always lies between -1 (perfect anti-correlation) and 1 (perfect correlation). A number like 0.25 represents a relatively weak positive correlation.

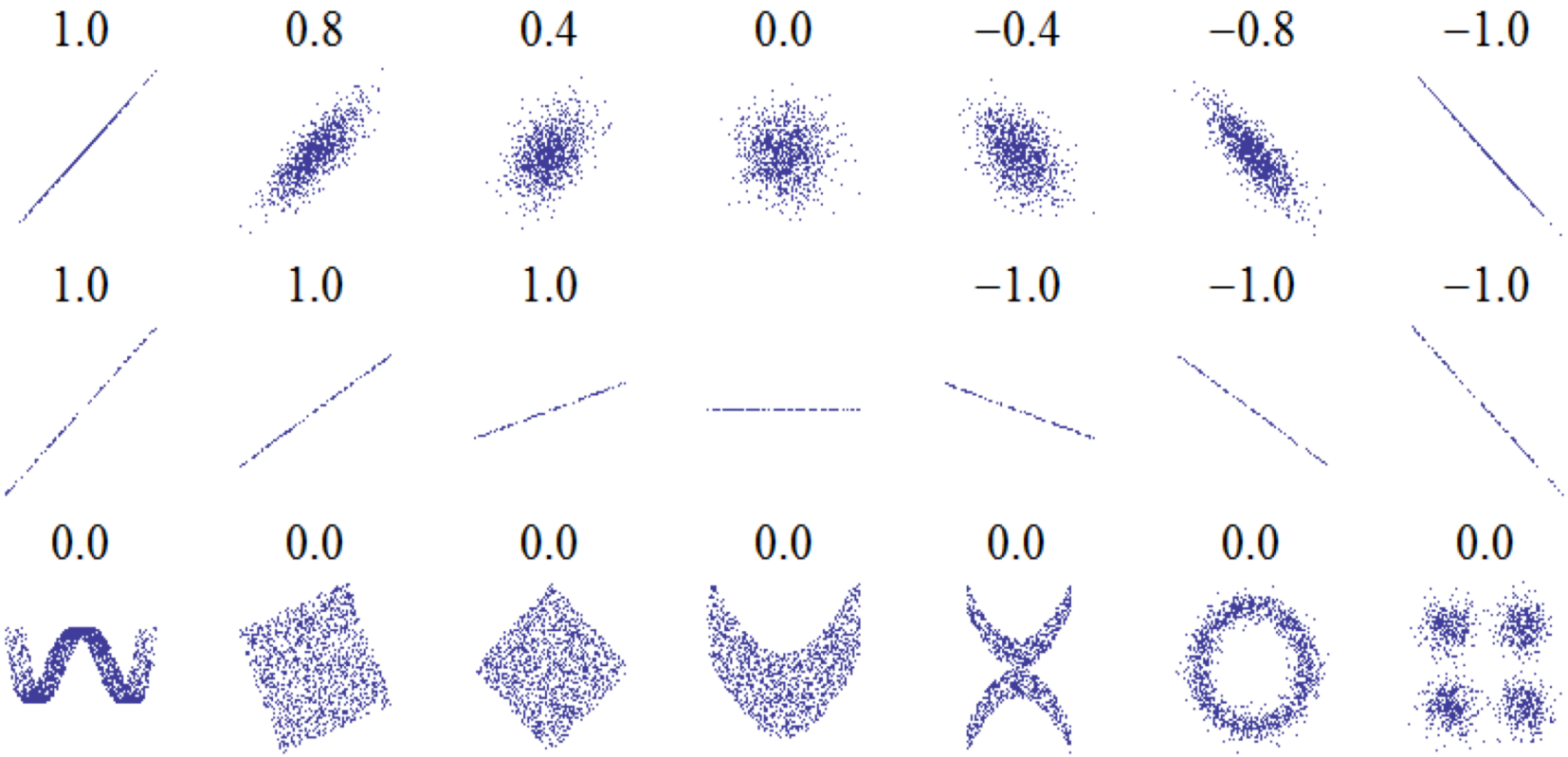
It is defined by

$$\rho = \frac{\text{covariance}(X, Y)}{\text{standard\_deviation}(X) * \text{standard\_deviation}(Y)}$$

This value is called Pearson's correlation after Karl Pearson, an influential early statistician.

Covariance and Correlation

Examples of correlation



Distribution

The *distribution* of a statistical data set (or a population) is a listing or function showing all the possible values (or intervals) of the data and how often they occur.

- Discrete (Frequency) distribution
- Continuous distribution

Example of Frequency distribution: Consider a data set of 26 children of ages 1-6 years. Then the frequency distribution of variable ‘age’ can be tabulated as follows:

Frequency Distribution of Age

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2

Histograms with Matplotlib

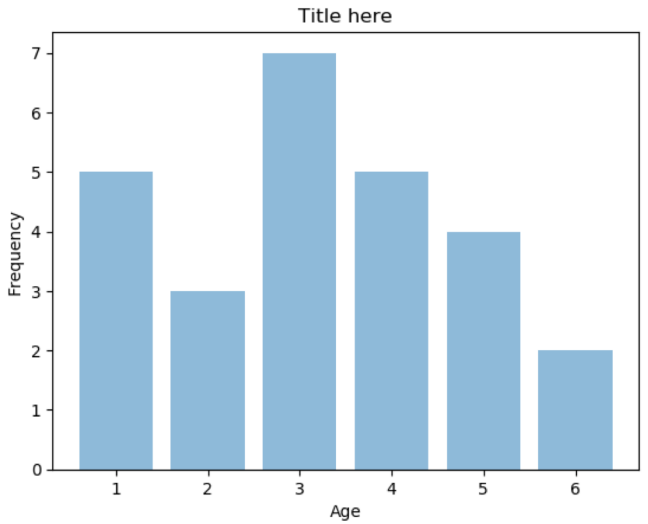
Code

```
x = ('1', '2', '3', '4', '5', '6')
y_pos = np.arange(len(x))
x = [5,3,7,5,4,2]
plt.bar(y_pos, x, align='center',
alpha=0.5)

plt.xticks(y_pos, x)
plt.ylabel('Frequency')
plt.xlabel('Age')
plt.title('Title here')

plt.show()
```

Output



Distribution

Bernoulli trial: An experiment that has two options: "success" (True) and "failure" (False).

```
np.random.seed(42)
random_numbers=np.random.random(size=10)
heads=random_numbers<0.5
print(heads)
print(np.sum(heads))
```

```
[ True False False False  True  True  True False False False]
4
```

Distribution

Binomial distribution: The number r of successes in n Bernoulli trials with probability p of success, is Binomially distributed

Example: The number r of heads in 4 coin flips with probability 0.5 of heads, is Binomially distributed.

```
np.random.binomial(4,0.5)
```

```
0
```

```
np.random.binomial(4,0.5,size=10)
```

```
array([3, 1, 1, 1, 1, 2, 2, 1, 2, 1])
```

```
np.random.binomial(4,0.5)
```

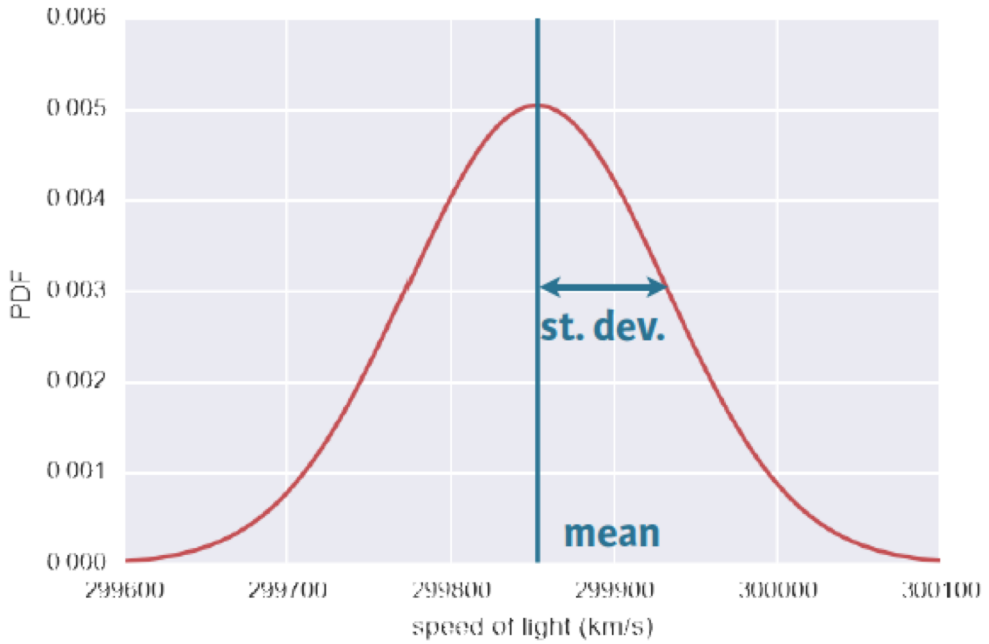
```
4
```

```
np.random.binomial(4,0.5,size=10)
```

```
array([1, 2, 2, 3, 1, 2, 2, 0, 2, 1])
```

Distribution

Normal distribution: Describes a continuous variable whose PDF has a single symmetric peak.



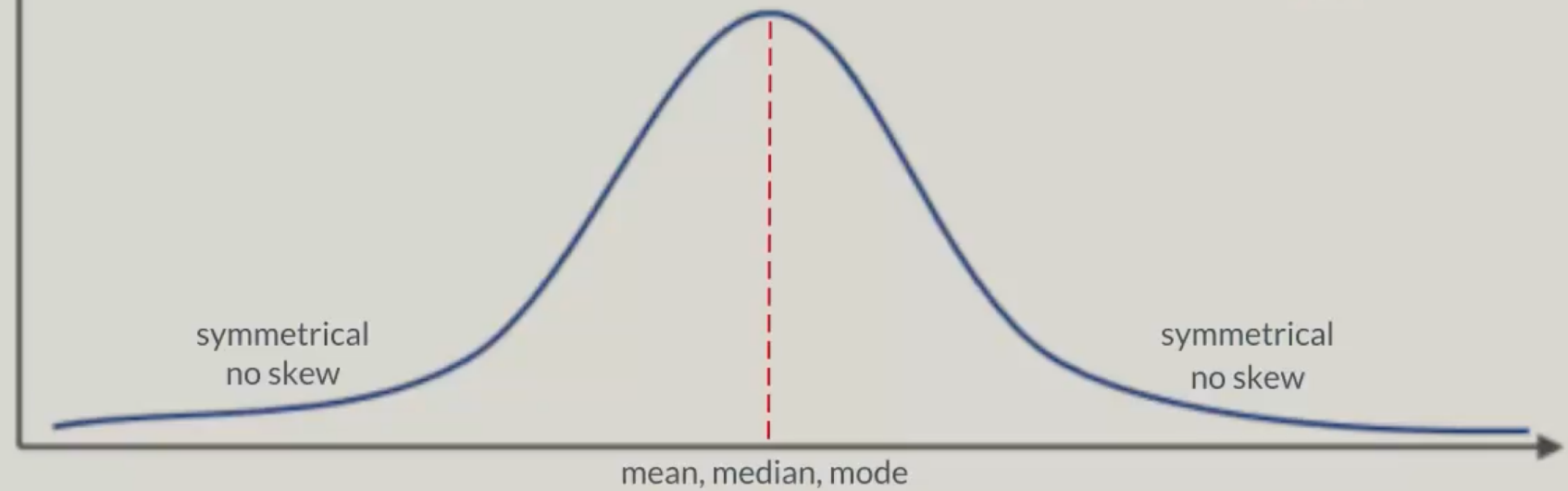
Normal distribution

$$N \sim (\mu, \sigma^2)$$

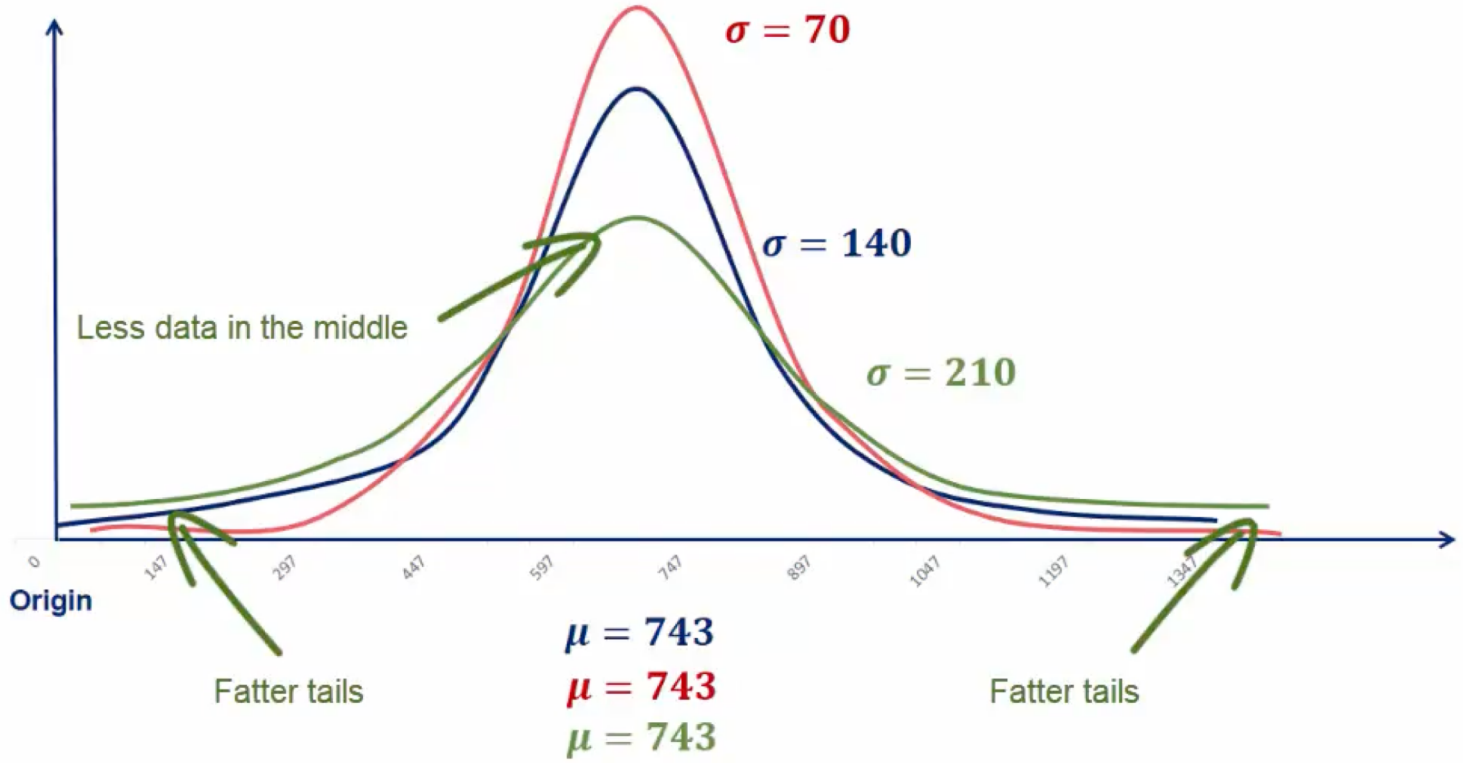
Normal Distribution Mean Variance

Gaussian distribution

Bell curve



Normal distribution. Controlling for the mean



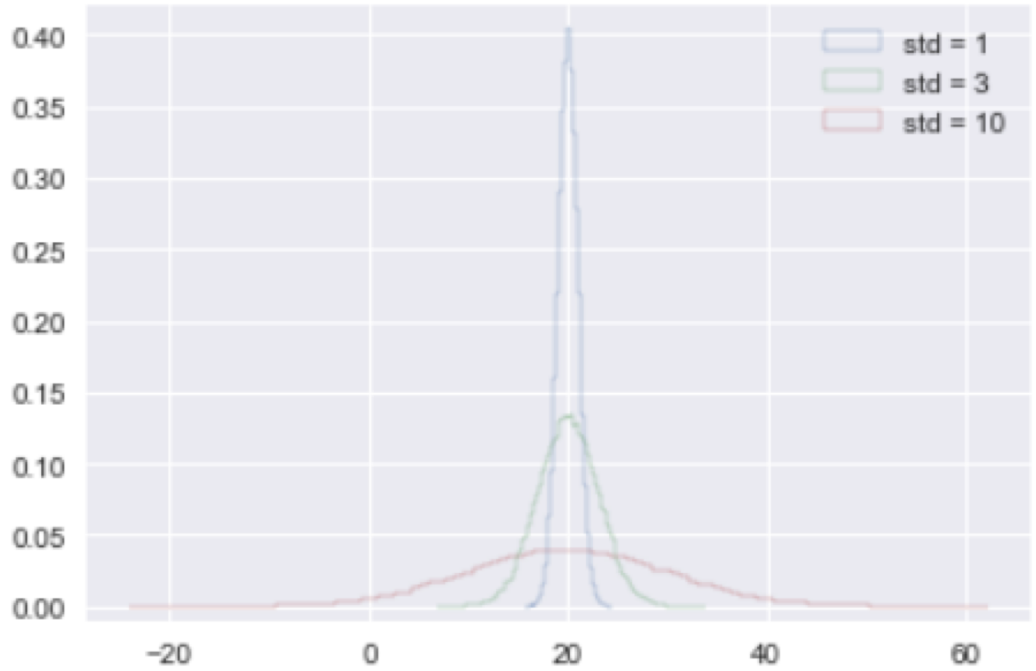
Distribution

```
samples_std1 = np.random.normal(20, 1, size=100000)
samples_std3 = np.random.normal(20, 3, size=100000)
samples_std10 = np.random.normal(20, 10, size=100000)

# Make histograms
_ = plt.hist(samples_std1, bins=100, normed=True, histtype='step')
_ = plt.hist(samples_std3, bins=100, normed=True, histtype='step')
_ = plt.hist(samples_std10, bins=100, normed=True, histtype='step')

# Make a legend, set limits and show plot
_ = plt.legend(('std = 1', 'std = 3', 'std = 10'))
plt.ylim(-0.01, 0.42)
plt.show()
```

Distribution



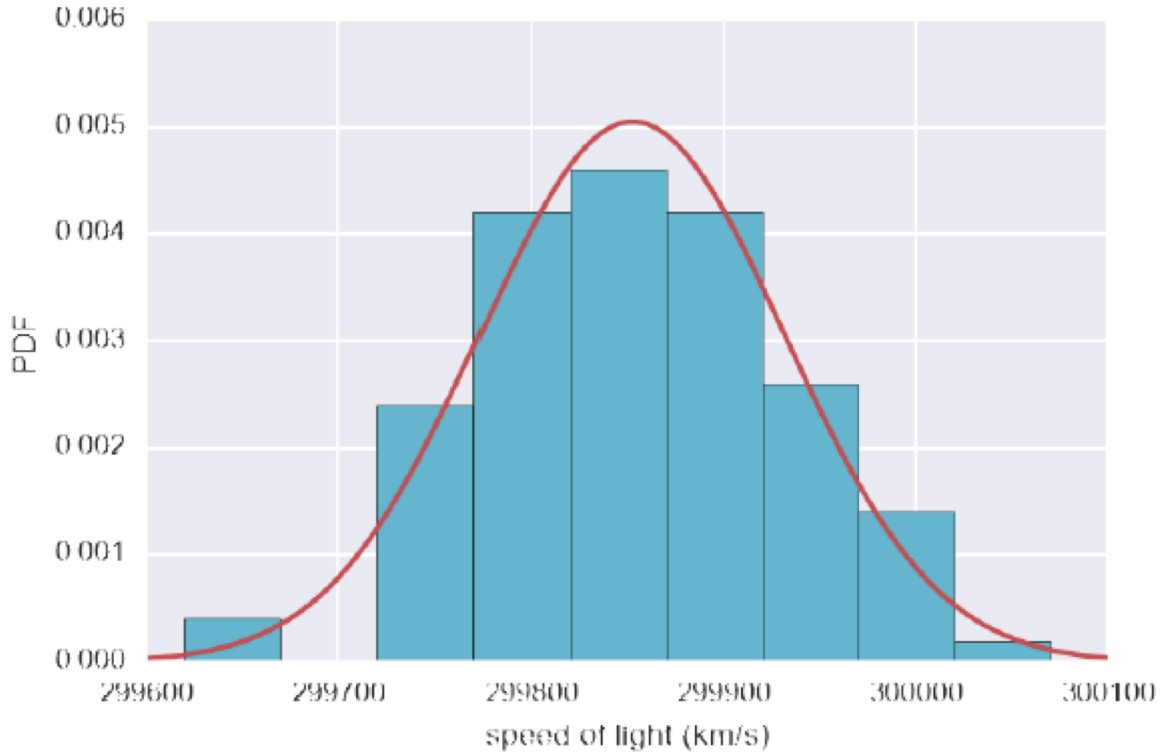
Distribution

Normal distribution

Parameter		Calculated from data
mean of a Normal distribution	≠	mean computed from data
st. dev. of a Normal distribution	≠	standard deviation computed from data

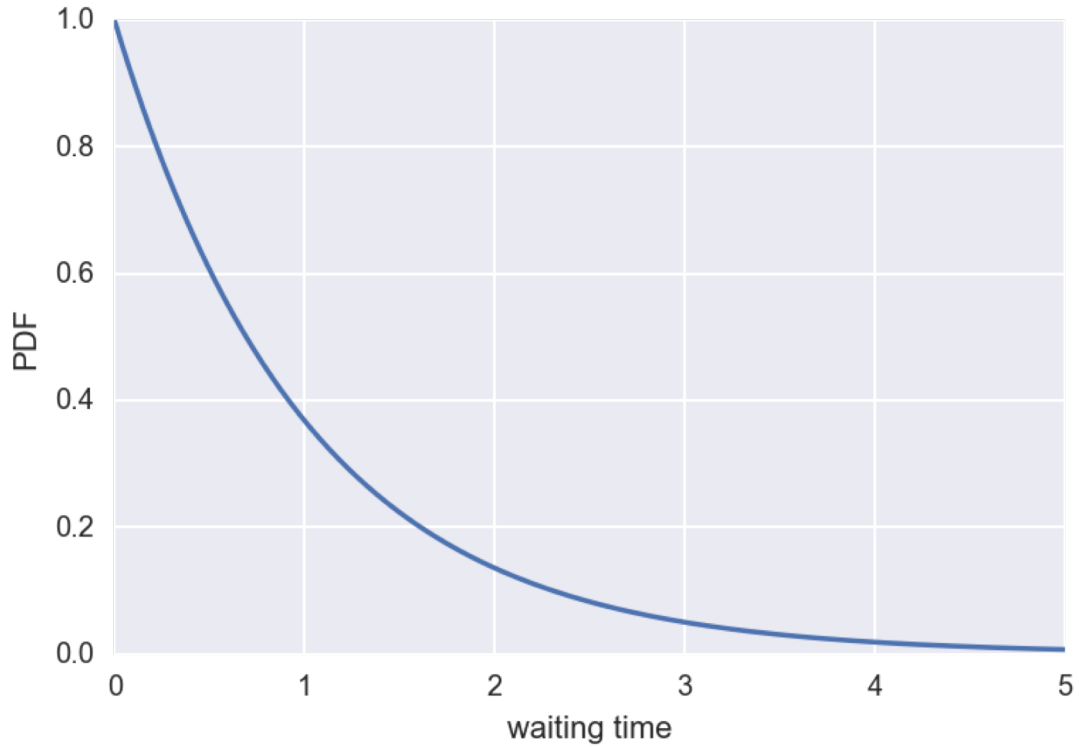
Distribution

Comparing data to Normal PDF



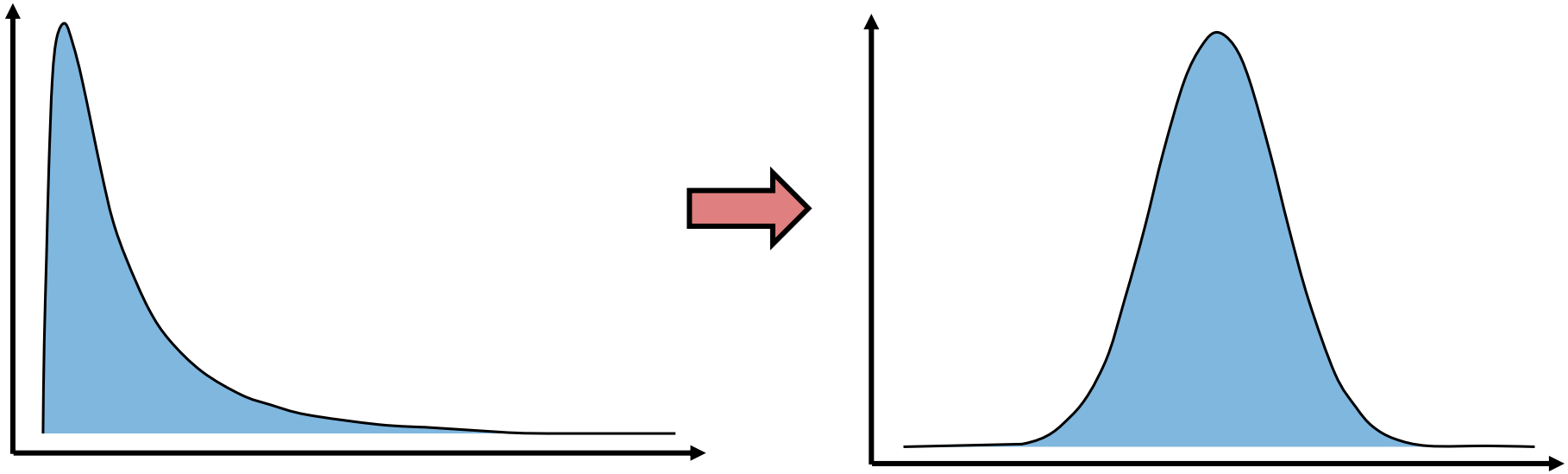
Distribution

The Exponential PDF



Distribution

Transformation of Data Distributions



```
from numpy import log

log_data = np.log(data)
```

Statistical Inference

Hypothesis testing

- Hypothesis: A statement we can test, e.g., a coin is biased, groups are different,

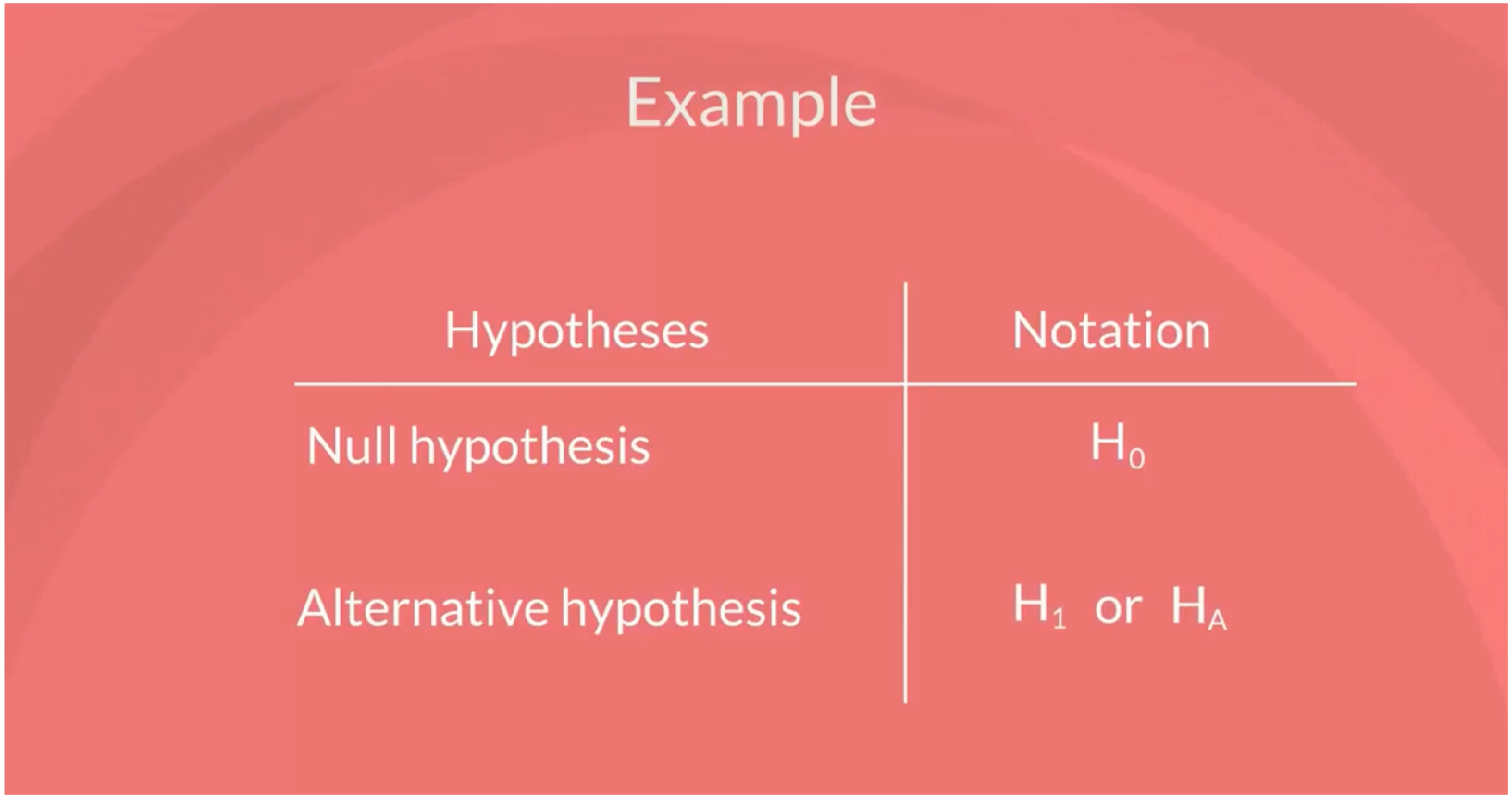
- We want to prove a hypothesis H_A , but its hard so we try to **disprove a null hypothesis H_0**.

  - Alternative hypothesis H_A : Our idea, e.g., a coin is biased, groups are different,
  - Null hypothesis H_0 : The alternative of our idea, e.g., a coin is unbiased, groups are the same.

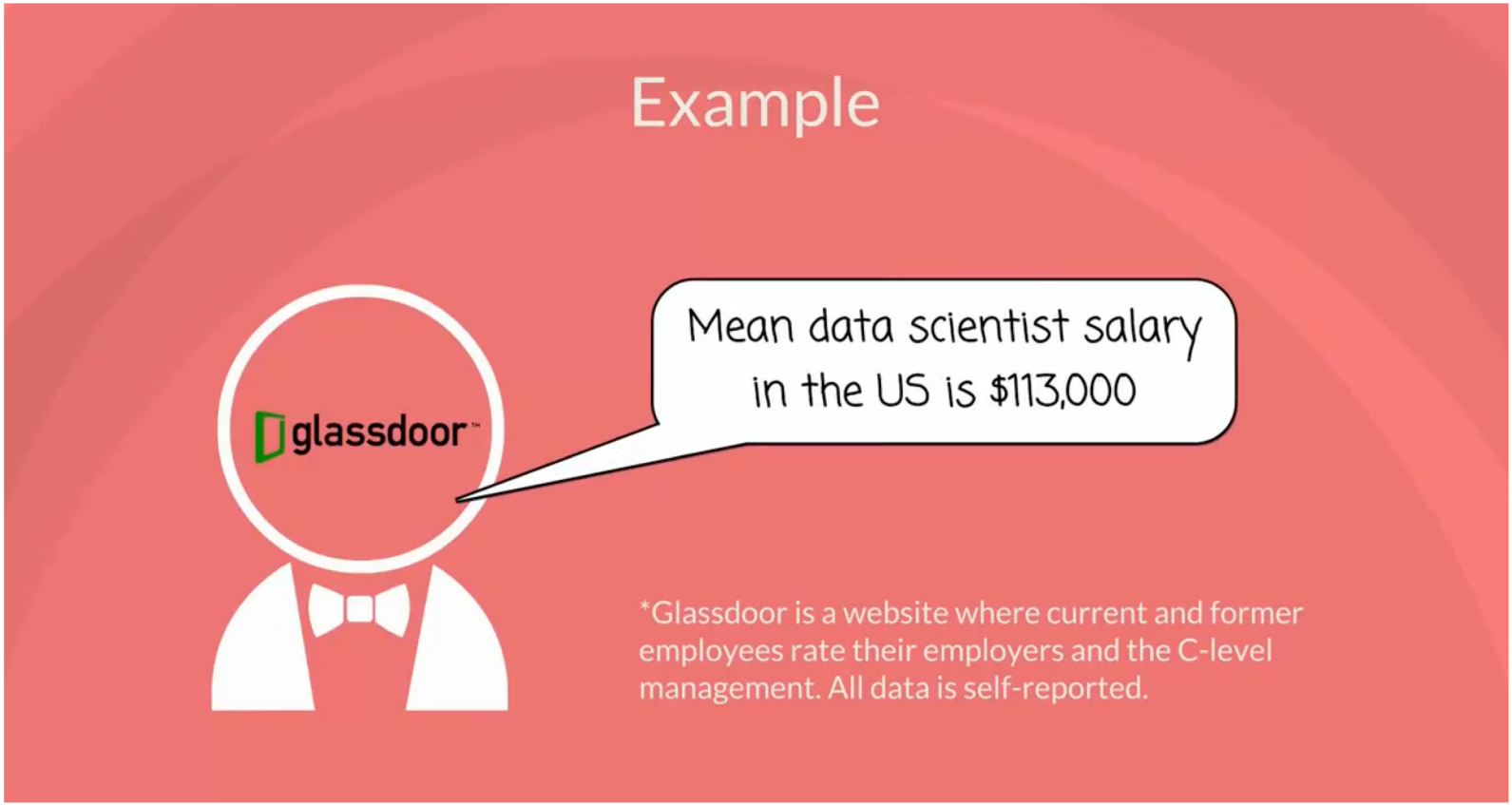
- The threshold probability is called a p-value. This directly controls the false positive rate (rate at which we expect to observe large s even if is H_0 true).

- A significance level is called p_given. The common values of p-given is 0.05, 0.02, 0.01, 0.005, 0.001
- If **p_value<p_given** we **accept H_A** (or **reject H_0**)

Hypothesis testing



Hypothesis testing



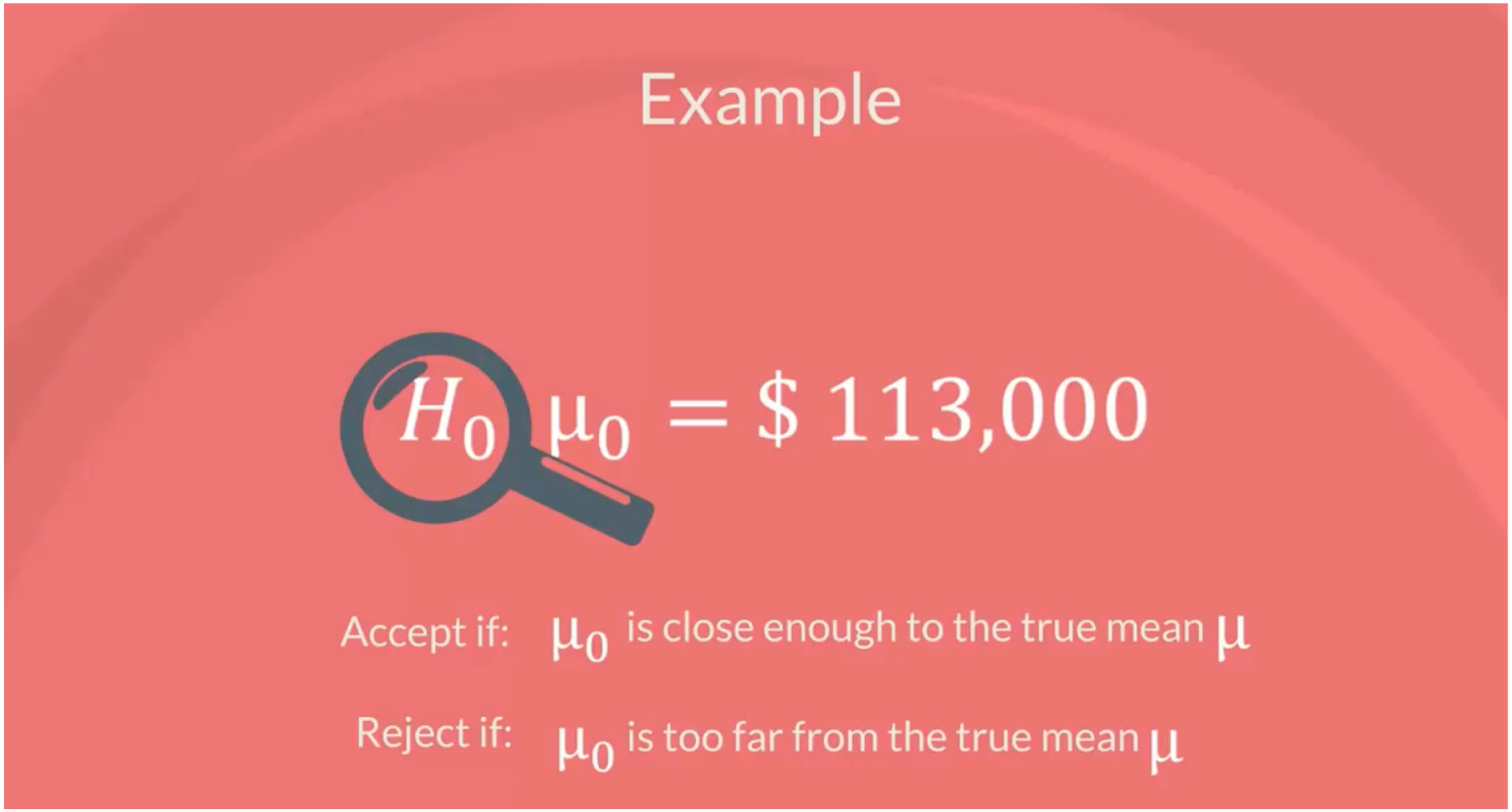
Hypothesis testing

Example

$$H_0: \mu_0 = \$\,113{,}000$$

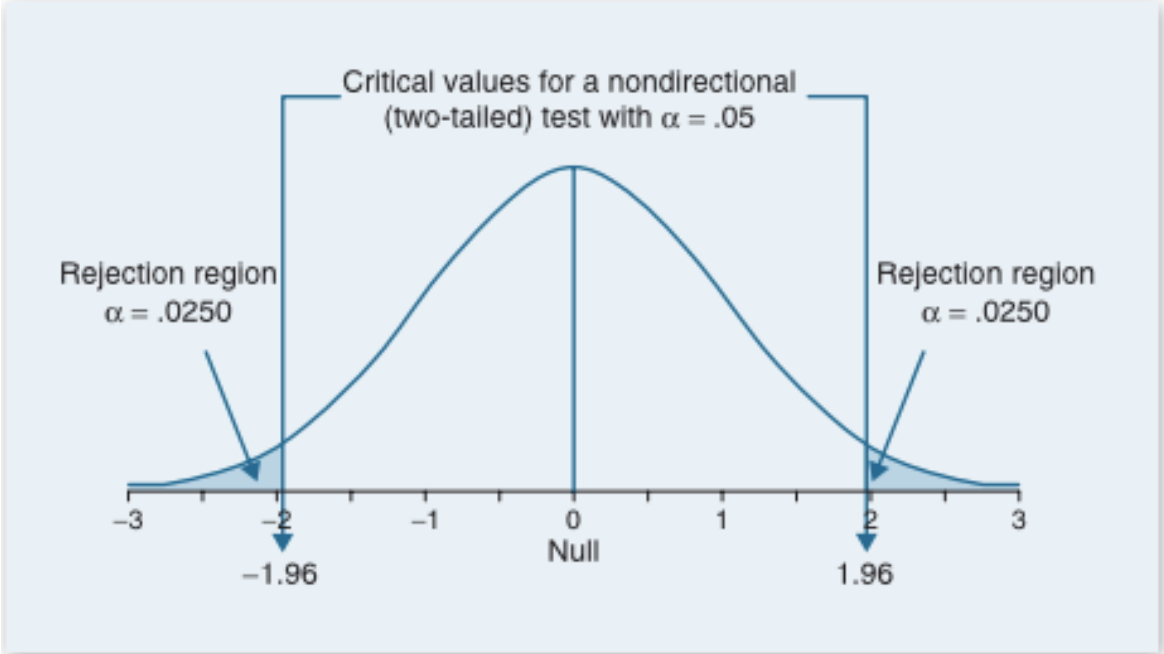
$$H_1: \mu_0 \neq \$\,113{,}000$$

Hypothesis testing



Hypothesis testing

Two-tailed Significance



When the p value is less than 5% ($p < .05$), we reject the null hypothesis

Hypothesis testing

Pearson correlation test

One sample t-test

Two sample (Student) t-test

ANOVA

Hypothesis testing

Pearson correlation test:

- Test association between two quantity variables. The test calculates a Pearson correlation coefficient and the p-value for testing non-correlation

- The p-value is the probability of seeing a t-statistic at least that far from 0 if the null hypothesis =“X and Y are dependent ” were true.

- Example: If X and Y are perfect correlation then p-value = 1.

Code:

```
import scipy.stats as stats
correlation, pvalue = stats.pearsonr(x,y)
```

Hypothesis testing

One sample t-test:

- The one-sample t-test is used to determine if a population man is a given number (null hypothesis).

- If X is normally distributed, or almost normally distributed, but not quite because of the presence of σ.

- t-statistic:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Code:

```
import scipy.stats as stats
t-statistic, pvalue = stats.ttest_1samp(x,\mu_0)
```

Hypothesis testing

Two sample t-test:

- The two-sample t-test is used to determine if two population means are equal.
- If X, Y are normally distributed. or **almost** normallv distributed.
- t-statistic:

$$t = \frac{\bar{x} - \bar{y}}{s \cdot \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

where

$$s = \sqrt{\frac{s_x^2(n_x - 1) + s_y^2(n_y - 1)}{n_x + n_y - 2}}$$

Code:

```
import scipy.stats as stats
t-stattistics, pvalue = stats.ttest_ind(x,y)
```

Hypothesis testing

ANOVA F-test:

- Analysis of variance (ANOVA) provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t -test to more than two groups.
- ANOVAs are useful for comparing (testing) three or more means (groups or variables) for statistical significance.
- It is conceptually similar to multiple two-sample t -tests, but is less conservative.

Code:

```
import scipy.stats as stats
fval, pval = stats.f_oneway(x,y,z)
```