Datacamp Python

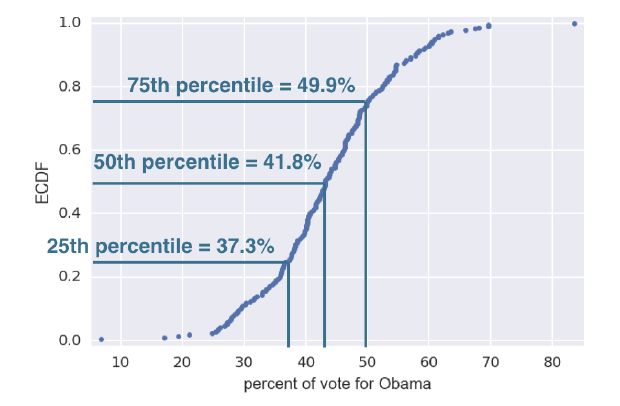
# Statistical Thinking in Python Part 1

## Exploratory Data Analysis

### ECDF

Empirical Cumulative Distribution Function is the distribution function associated with the empirical measure of a sample. The x axis are the observations, sorted from the smallest to the largest. The y axis is the percentage of observation smaller than or equal to the specified value.

For example



25% of observations has value less than or equal to 37.3

50% of observations has value less than or equal to 41.8

75% of observation has value less than or equal to 49.9

## Quantitative exploratory data analysis

Mean and median

Percentiles, outliers and box plots

Variance and standard deviation

Covariance and the Pearson correlation coefficient

## Thinking probabilistically – Discrete variables

Probability allows us to describe uncertainty.

### Binomial Random Variable

In order for a variable to be a binomial random variable,

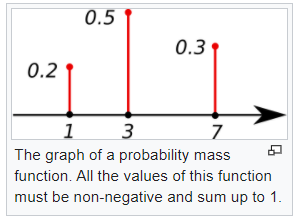
* Each trail must be independent
* Each trail can be called a “success” or a “failure”
* There area a fixed number of trials
* The probability of success on each trial is constant

### Bernoulli trial

An experiment that has two options, “success” and “failure”.

### Probability Mass Function (PMF)

A probability mass function is a function that gives the probability that a discrete random variable is exactly equal to some value.



P(X = 1) = 0.2

P(X = 3) = 0.5

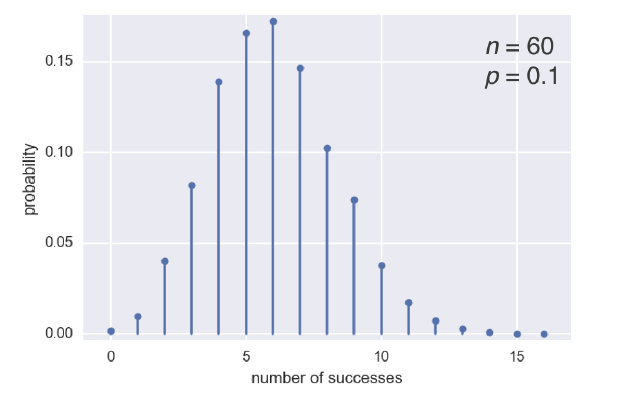
P(X = 7) = 0.3

The discrete random variable X has 0.5 probability to have the value 3.

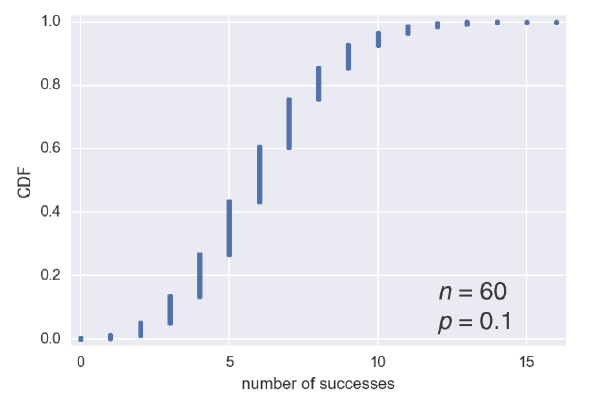
### Binomial Distribution

The number r of success in n Bernoulli trials with probability of p of success, is Binomially distributed.

The following is the Binomial PMF



And the Binomial CDF



### The Poisson process

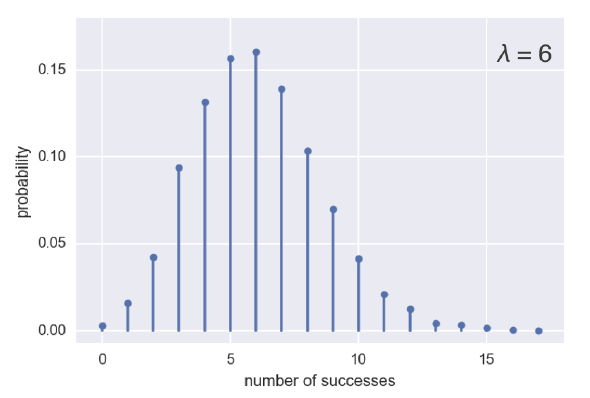
A Poison process describes the number of times an event occurs in a period of time, or in a particular area, or over some distance or within any other kind of measurement.

* The experiment counts the number of occurrences of an event over some other measurement.
* The mean is the same for each interval
* The count of events in each interval is independent of the other intervals
* The intervals don’t overlap

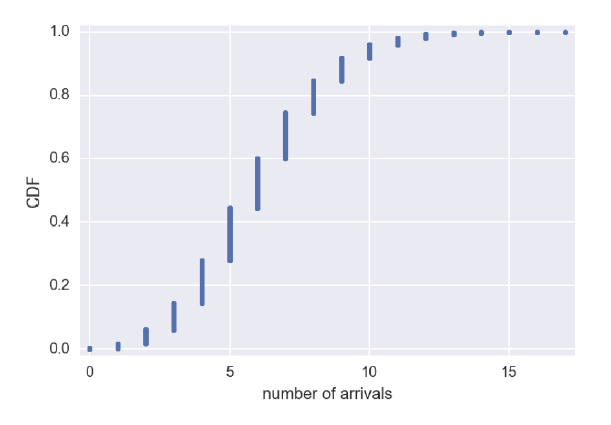
### Poisson distribution

The number r of arrivals of a Poisson process in a given time interval with average rate of ? arrivals per interval is Poisson distributed.

Poisson PMF



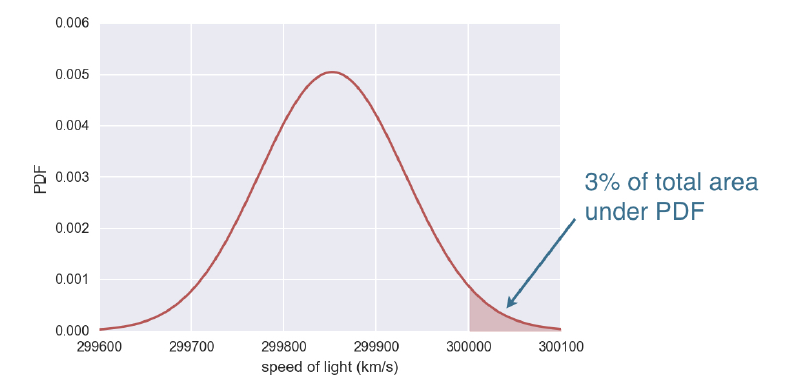
Poisson CDF



## Thinking probabilistically – Continuous variables

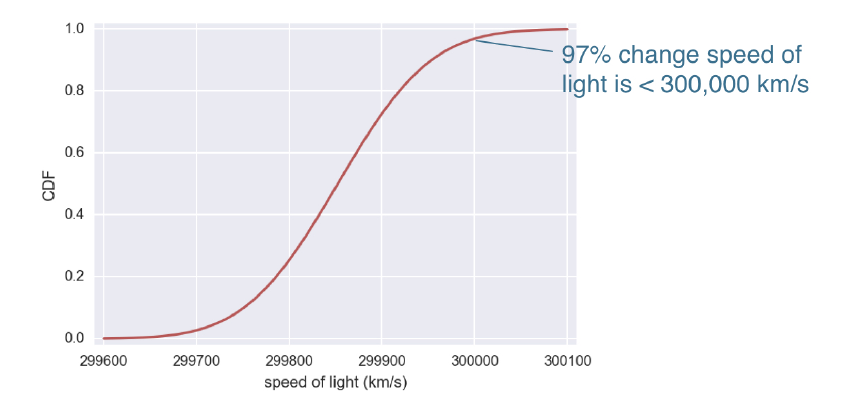
### Probability Density Function (PDF)

Continuous analog to the PMF. It describes the probability of observing a value of a continuous variable.



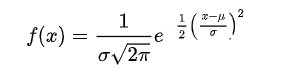
The probability of observing the light whose speed is greater than 300,000 km/s is 3%, the total area under the PDF

On CDF



### The Normal Distribution

The normal (or Gaussian) distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

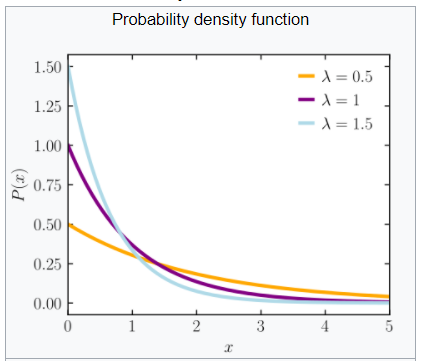


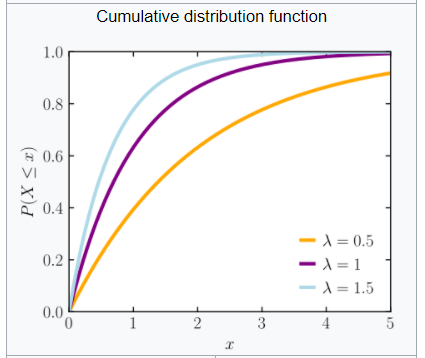
Where mu is the mean and sigma is the standard deviation

### The Exponential Distribution

The exponential distribution is the probability of the time between events in a Poisson point process.

The waiting time between arrivals of a Poison process is exponentially distributed.





# Statistical Thinking in Python Part 2

## Parameter estimation by optimization

Optimal parameters bring the model closest agreement with the data, given that the model we choose is the correct model.

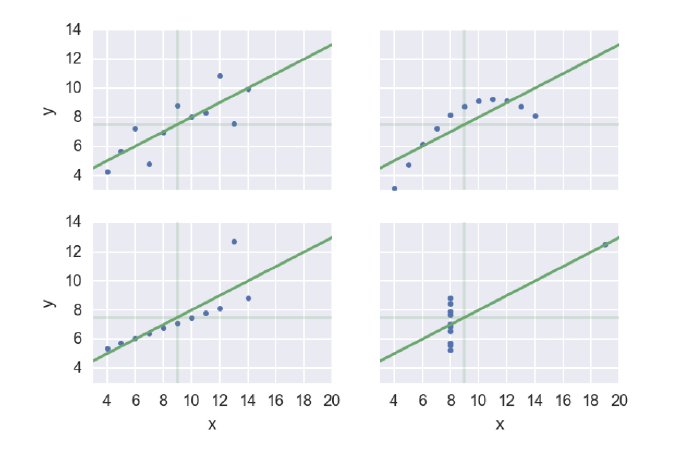
### Linear Regression

Linear regression is the process of using a linear model with point-intercept form to approximate the data.

Least squares is the process of finding the parameters for which the sum of the squares of the residuals is minimal.

numpy.polyfit(x, y, degree = 1) can be used for linear regression

### Anscombe’s quarter



The four data sets have the same mean value of x, mean value of y, residual and line-slope formulas.

Therefore, perform EDA first.

## Bootstrap confidence intervals

**Resampling** is the process of random selecting observations from the data set (with replacements) as if we are re-conducting the experiment.

**Bootstrapping** is the use of resampled data to perform statistical inference.

A resampled array of data is a **bootstrap sample**.

A statistic computed from a resampled array is called **bootstrap replicate**.

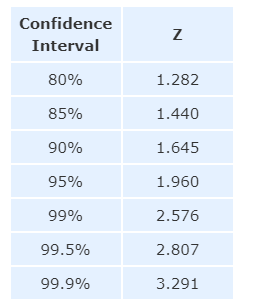
### Confidence interval of a statistic

If we repeat the measurements over and over again, p% of the observed values would lie within the p% confidence level.

**Example in theory**

The average height of men is 176cm, with a standard deviation of 20cm. What is the 95% confidence interval?

We find the z score of 2.5% and 97.5%



Plug in the z-value into the formula



And we have 175 +/- 6.2

Therefore, the interval [168.8, 181.2] includes 95% of the observations.

**Example in simulation**

****

We draw bootstrap replicas from the sample data

Take the percentiles to get the confidence interval.

### Pair bootstrap

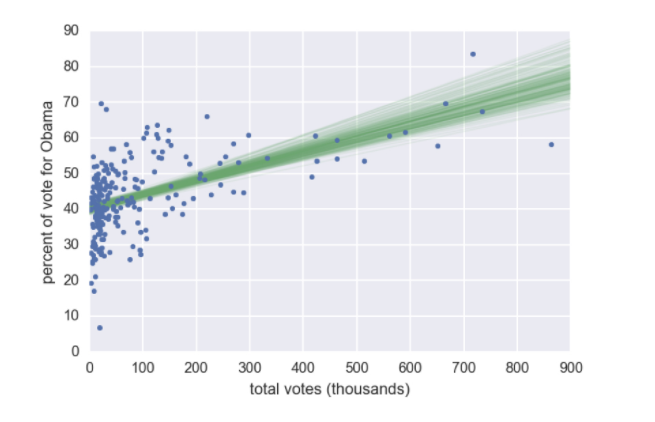
So far, we have been resampling with no assumption about the model or probability distribution of the underlying data, a.k.a, the resampling have been make on the data along.

If we want to perform statistical inference of a linear regression model, we have to consider two parameters, the slope and the intercept.

### Paris bootstrap for linear regression

* Resample data in pairs (using indices)
* Computer slope and intercept from resampled data
* Each slope and intercept is a bootstrap replicate
* Compute confidence intervals from percentiles of bootstrap replicates.

The results look like this



## Introduction to hypothesis testing

<https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/>

Example 1:

Percentage of vote to democratic party in counties among Ohio and Pennsylvania are similar.

* Use permutation to randomly reorder two arrays as if they are the same.

### P value

P value is used in the hypothesis testing to help you support or reject the null hypothesis: The p value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

### The Alpha value

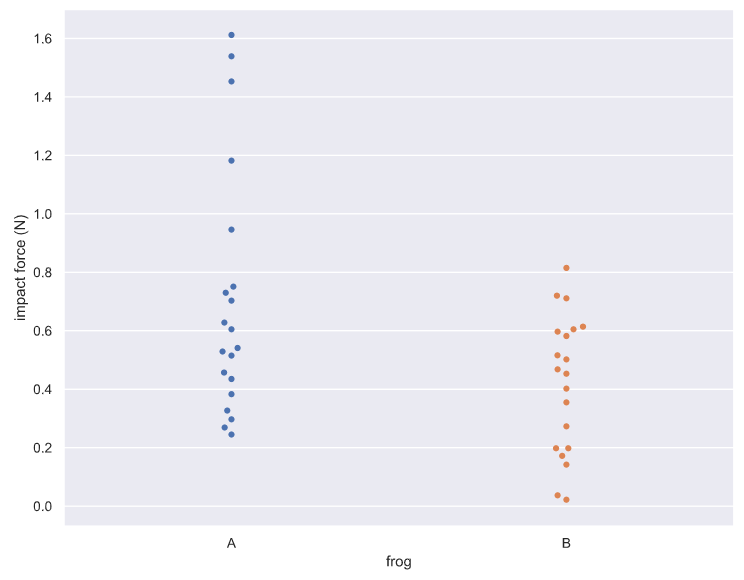
The alpha value is controller by the research and it relates to the confidence level. Depends on if you are running a one-tail test or a two-tail test, a = 1 – 10%, a = 1 – 10%/2

If the p value is extremely small or large and it falls into the area of reject, then we can reject the null hypothesis and approve the alternative hypothesis.

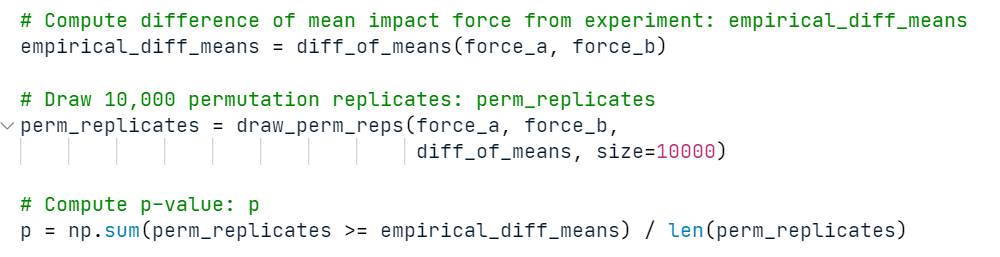
Example 1

Two frogs, one adult and one juvenile hit a wall with their tongue with a force. We are testing the hypothesis that the two frogs the same distribution of impact force.

First, use visual EDA



Then, draw permutations and calculate permutation replicates



Compute the p value, the percentage of permutation replicates that the difference of impact force is greater than the empirical difference

P = 0.0063.

The p value tells us that there is about a 0.6% chance that we would get the difference of means observed in the experiment if frogs were exactly the same.

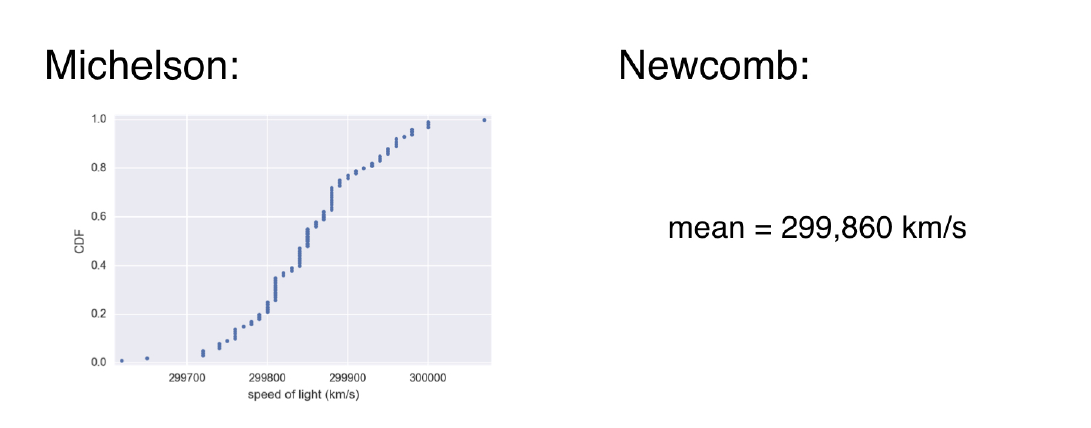
### Pipeline for hypothesis testing

* Clearly state the null hypothesis
* Define your test statistics (mean, property, difference etc / equal to, larger than)
* Generate many sets of simulated data assuming the null hypothesis is true
* Compute the test statistic for each simulated data set
* The p-value is the fraction of your simulated data sets for which the test statistic is at least as extreme as for the real data.

### Michelson and Newcomb

Michelson measure the speed of lights 100 times, Newcomb only have the team.

Question: Could Michelson got the dataset he had from the experiemnts, if the true mean of the light is of Newcomb’s



**Null hypothesis**

The true mean speed of light in Michelson’s experiments was actually Newcomb’s reported value.

## Hypothesis test examples

## Putting it all together