2018

Intro to Data Science - Data Exploration 3 (Week Four)

Grant Long
CUNY City College

NYC Tech-in-Residence Corps

Today's Agenda

1. Types of Data
2. Useful Statistical Distribution
3. Important Summary Statistics
4. Independence
5. Key Theorems

Week 3 Recap

- Elements of the ETL Process
- Processing Tools: Luigi, Airflow
- Handling Missing Data: Drop, Impute

HW Recap

1. Assignment 2 Notes
    - There are cells other than code. Try **markdown**!
    - Restart kernel and run all cells when you finish
    - Answer all questions for full credit
    - Collaboration is ok, copying is not. Disclose collaborators going forward.
2. How was DataCamp?
3. How do we feel about projects?

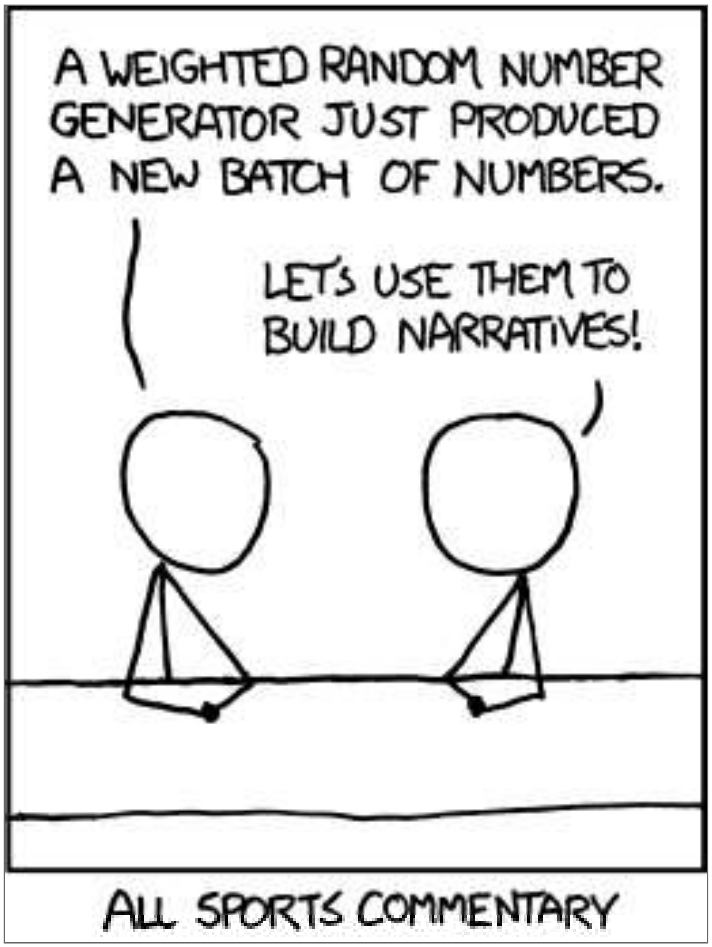
Who's Feeling Lucky?

sta·tis·tics

noun

The practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

Source


A WEIGHTED RANDOM NUMBER GENERATOR JUST PRODUCED A NEW BATCH OF NUMBERS.
LET'S USE THEM TO BUILD NARRATIVES!
ALL SPORTS COMMENTARY

xkcd

Types of Data

Boolean

Categorical

Continuous

Probability Distributions

A mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.

Source

A Few Important Distributions

Binomial

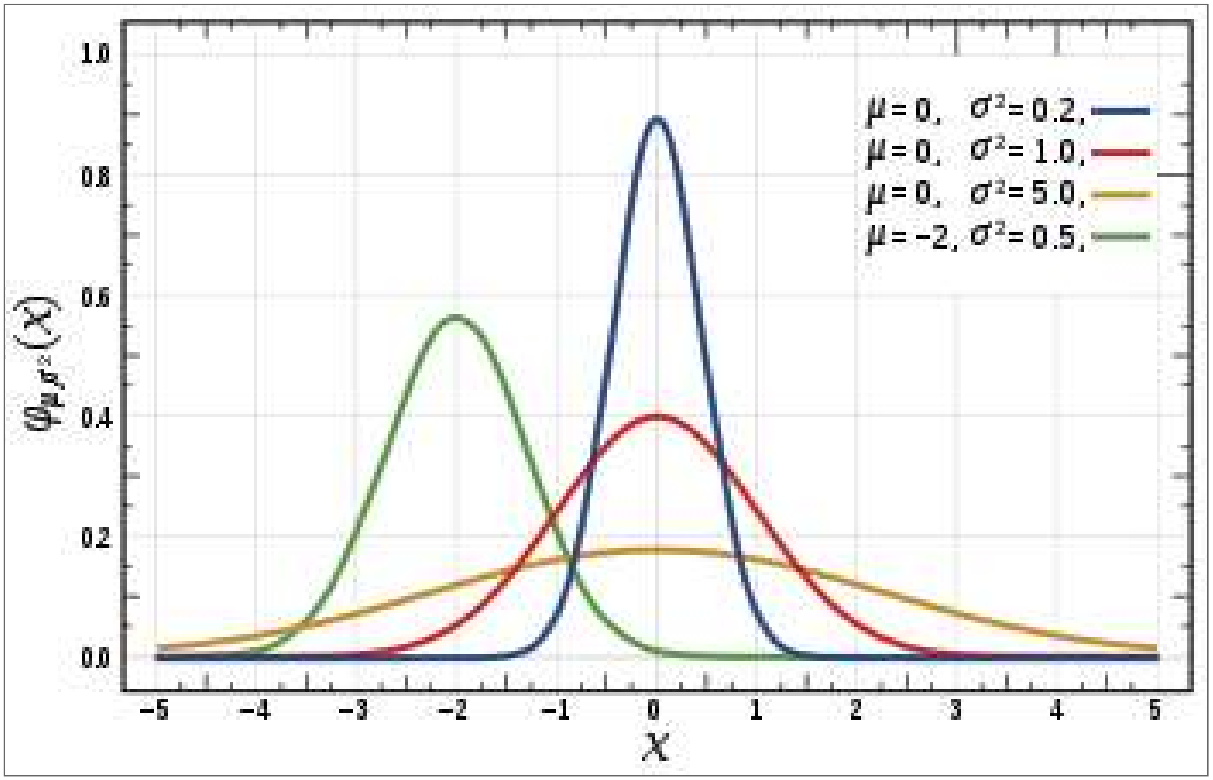
$$f(k, n, p) = \Pr(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

describes the likelihood for k successes over n trials with p probability of success where:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

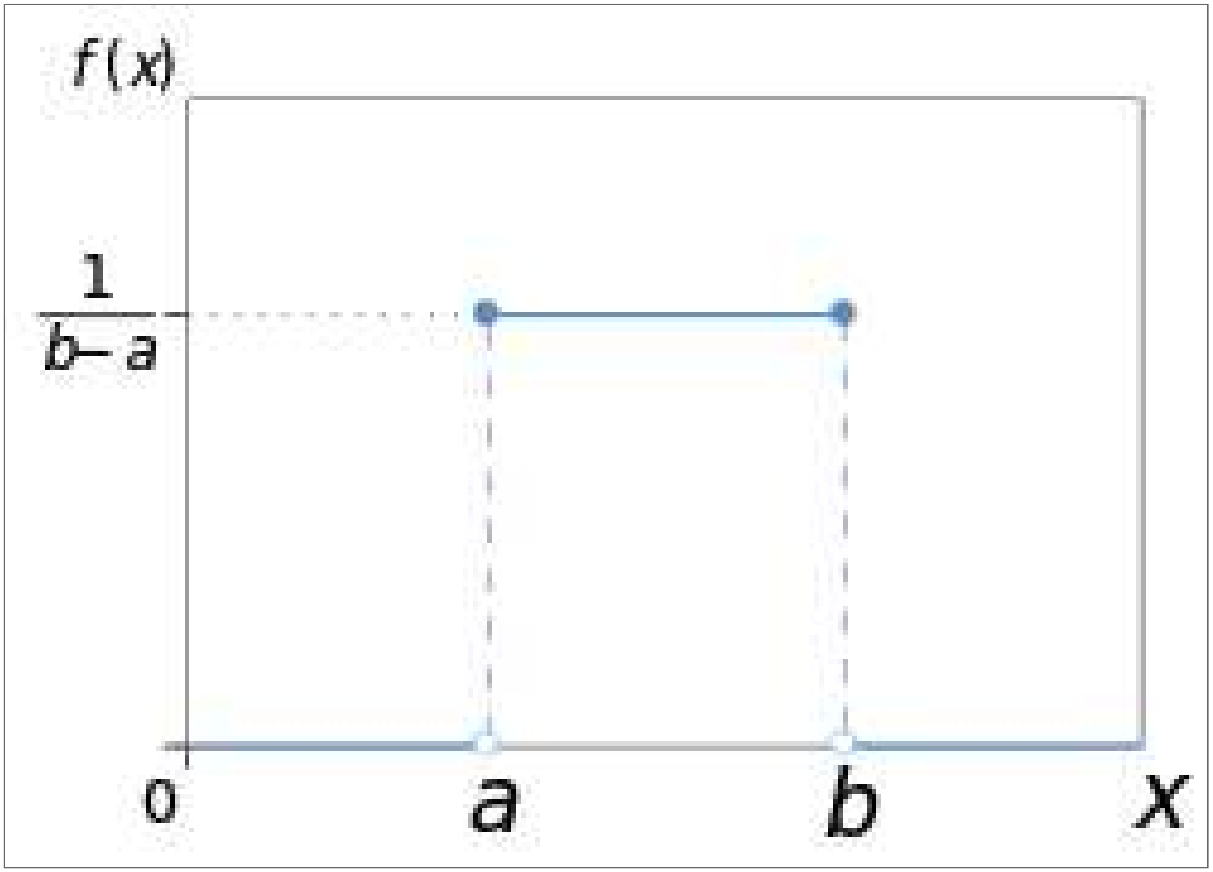
Wikipedia

Normal



Wikipedia

Uniform

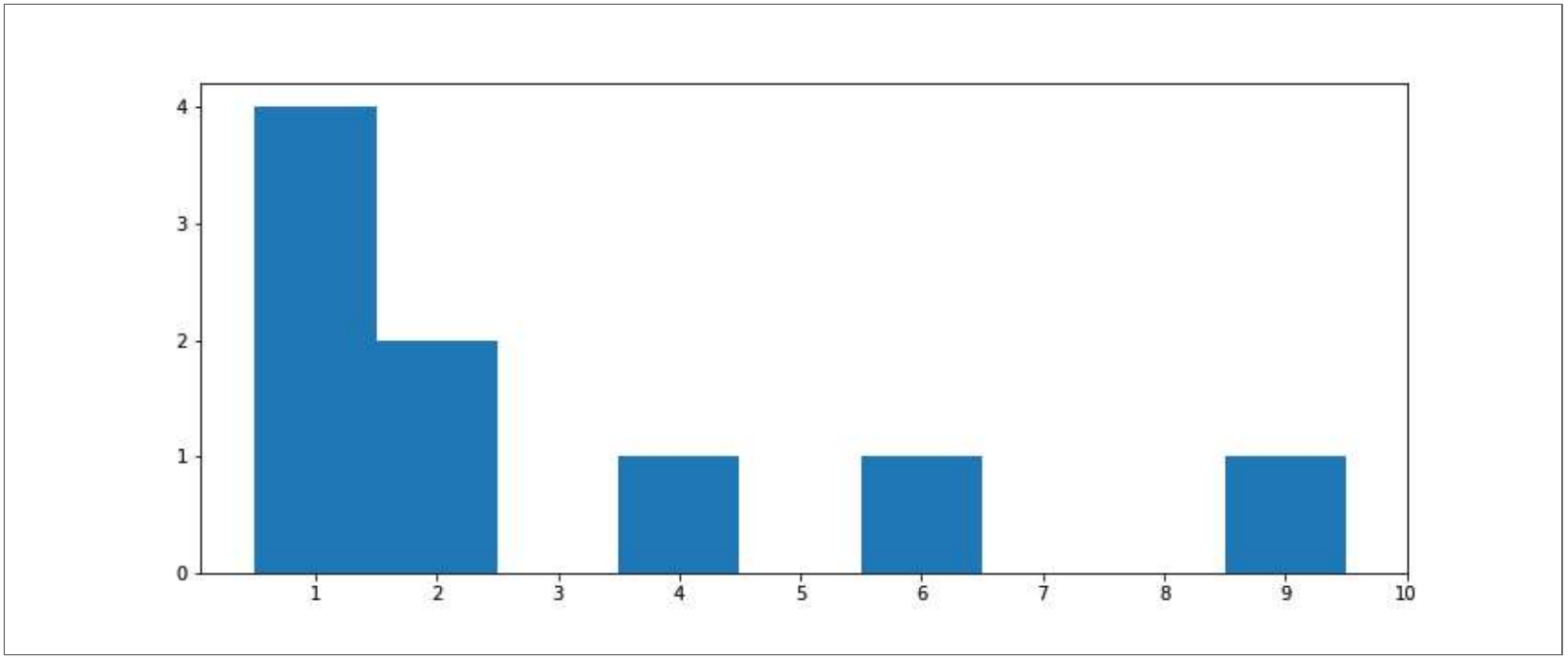


Wikipedia

How to Describe Distributions

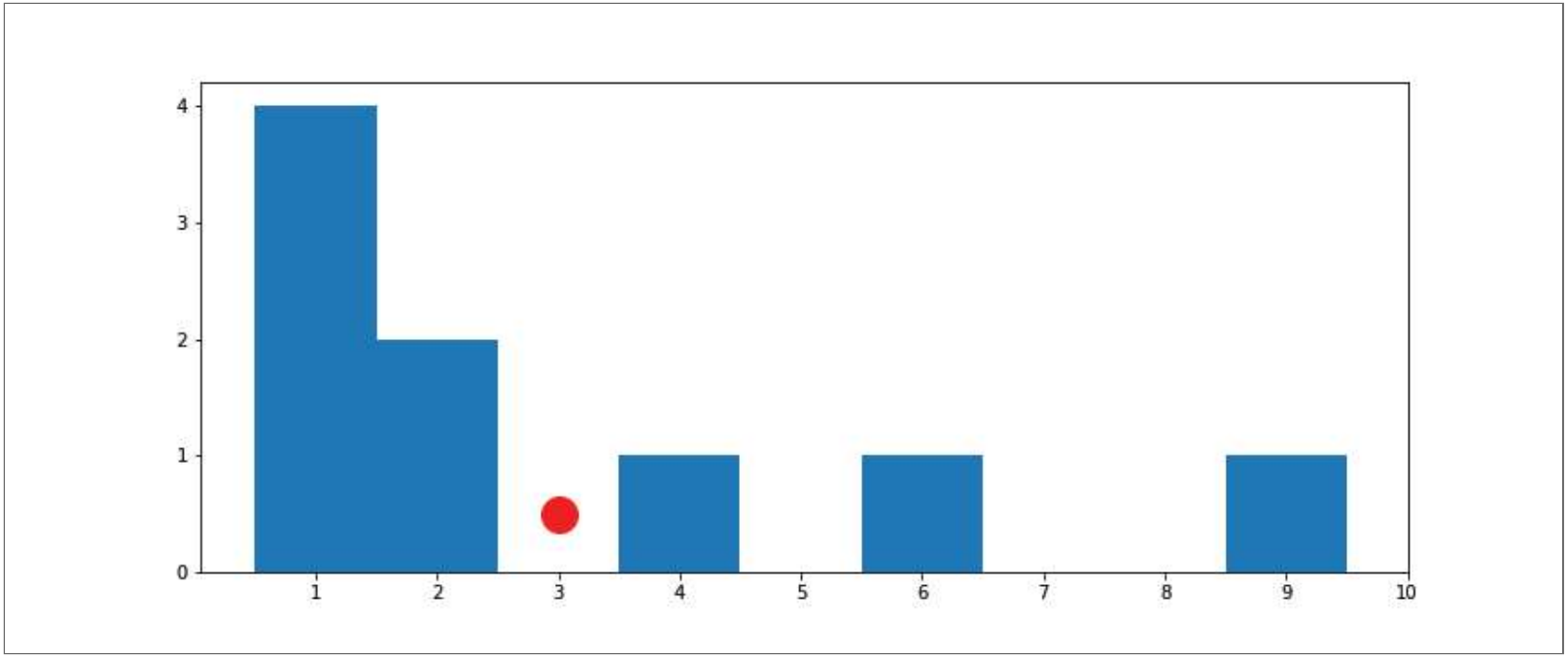
Central Tendency

[1, 1, 1, 1, 6, 2, 4, 2, 9]



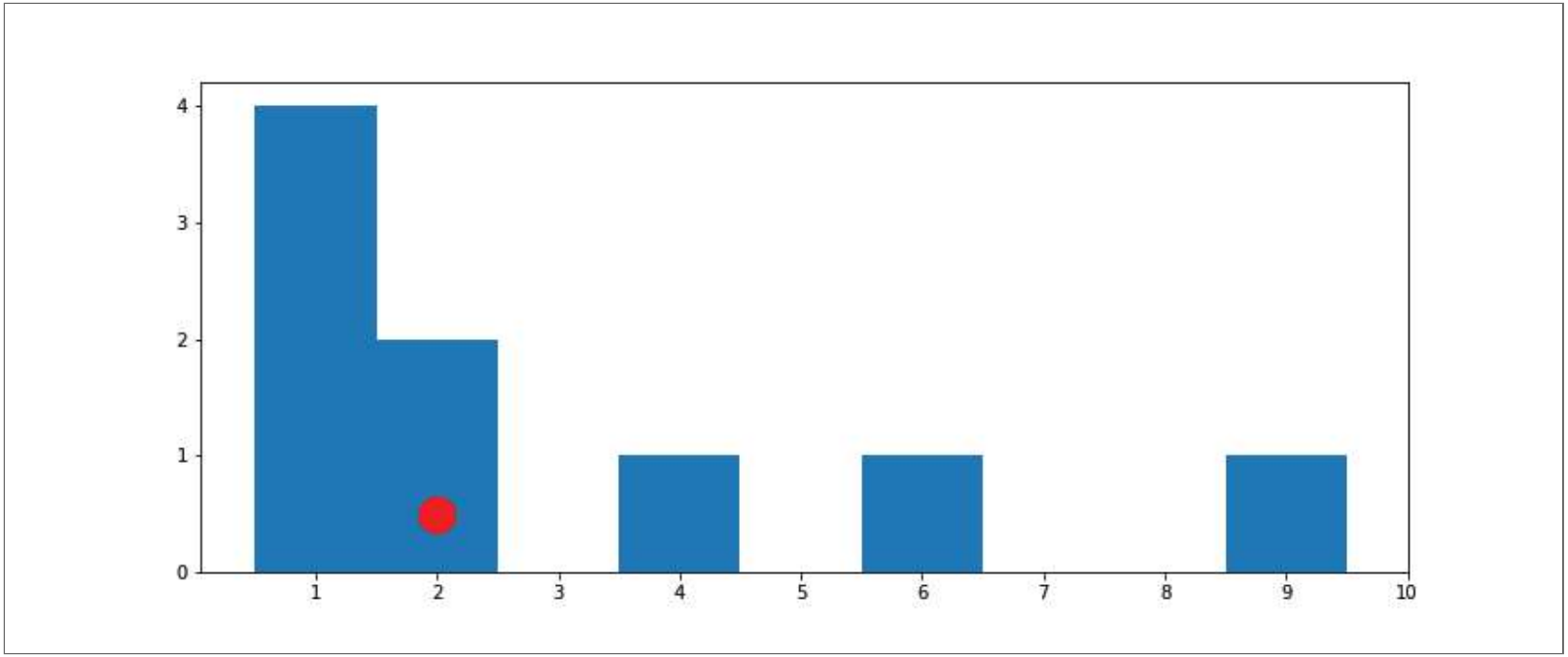
Central Tendency

Mean



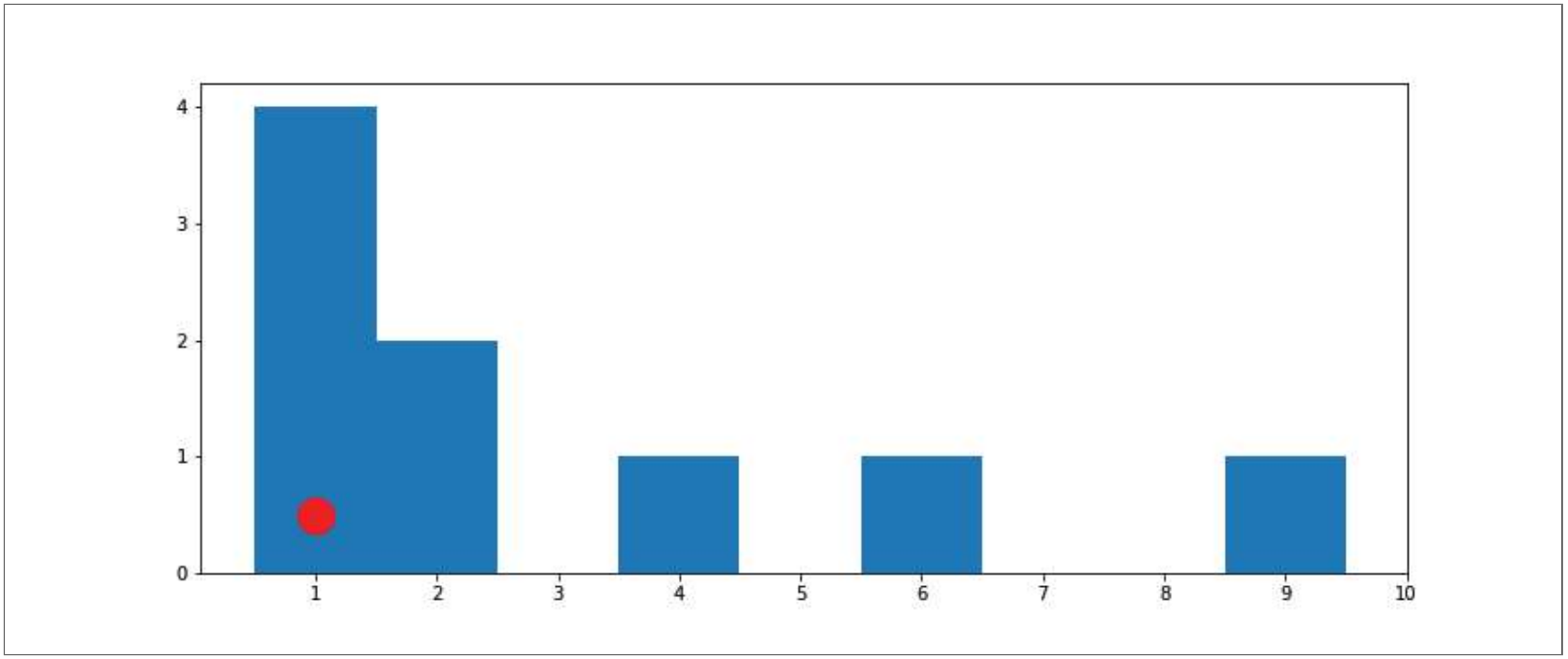
Central Tendency

Median



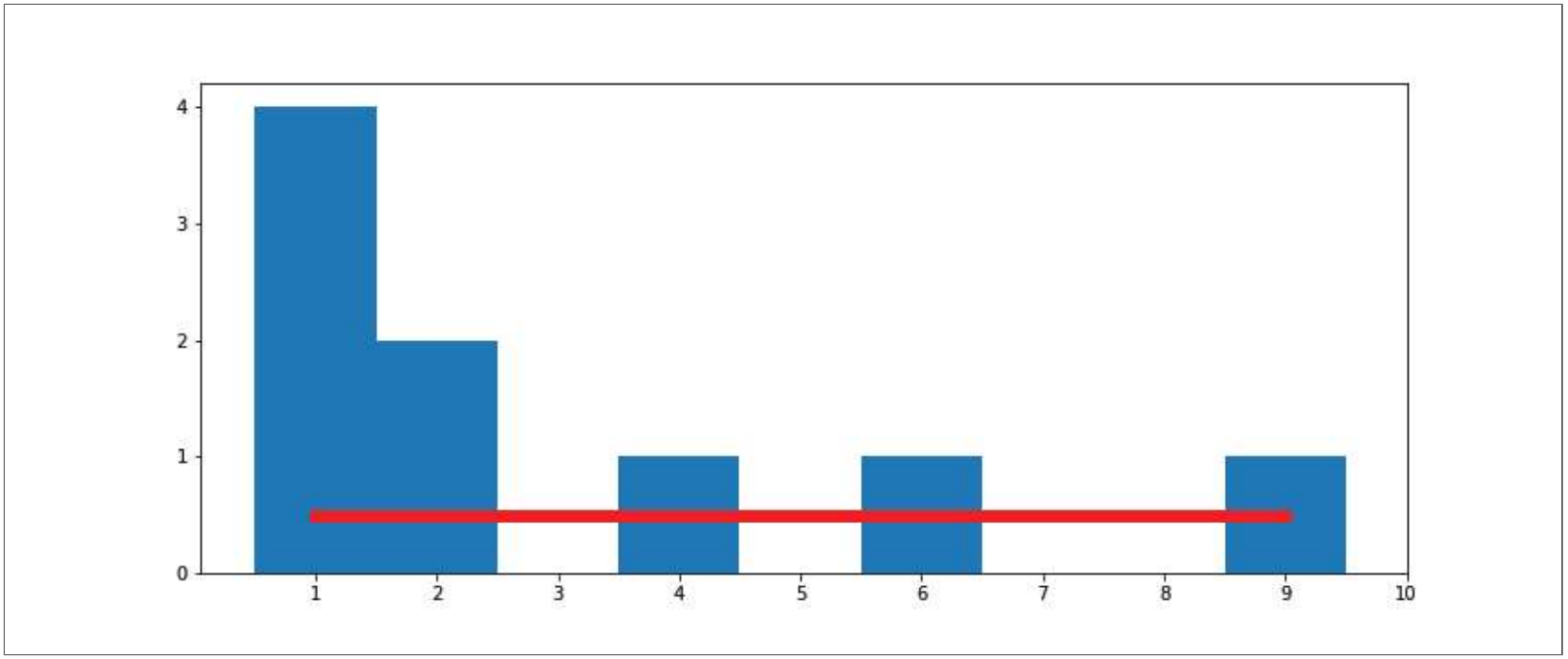
Central Tendency

Mode



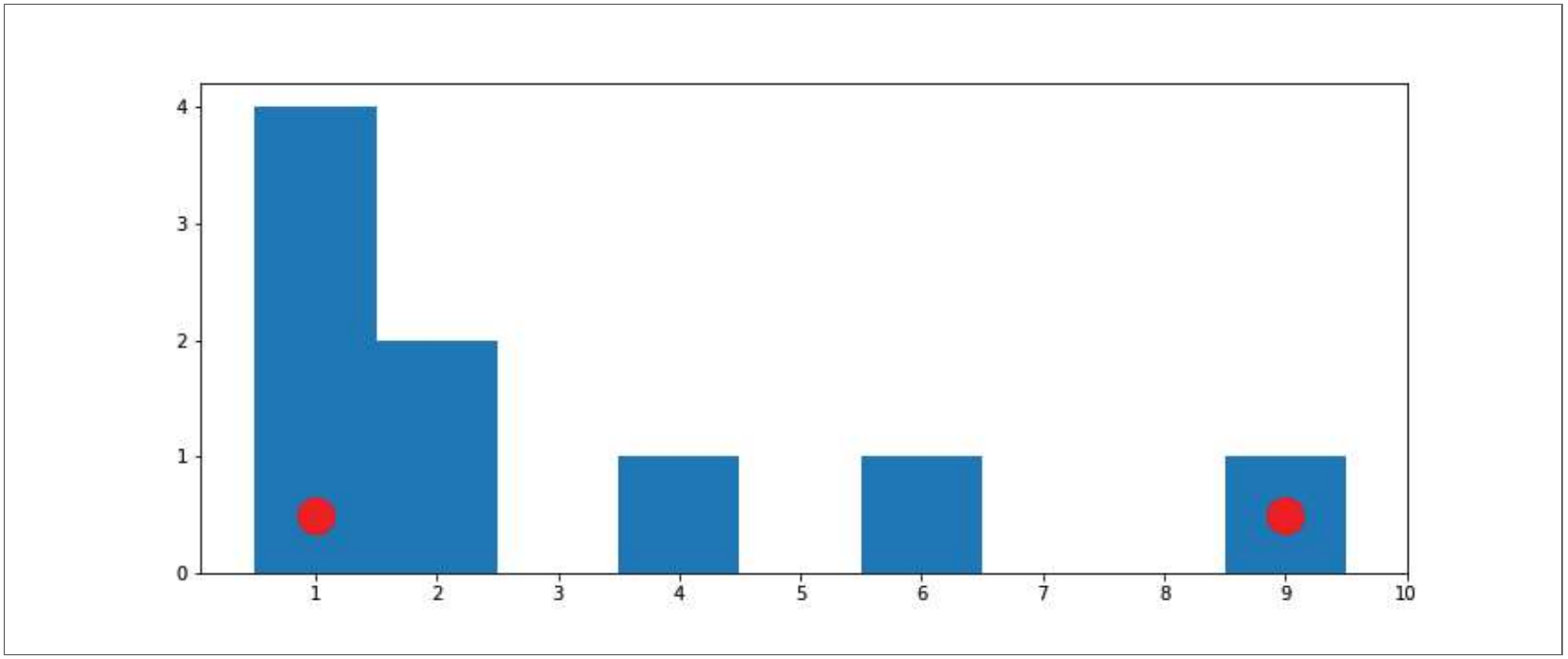
Variation

Range



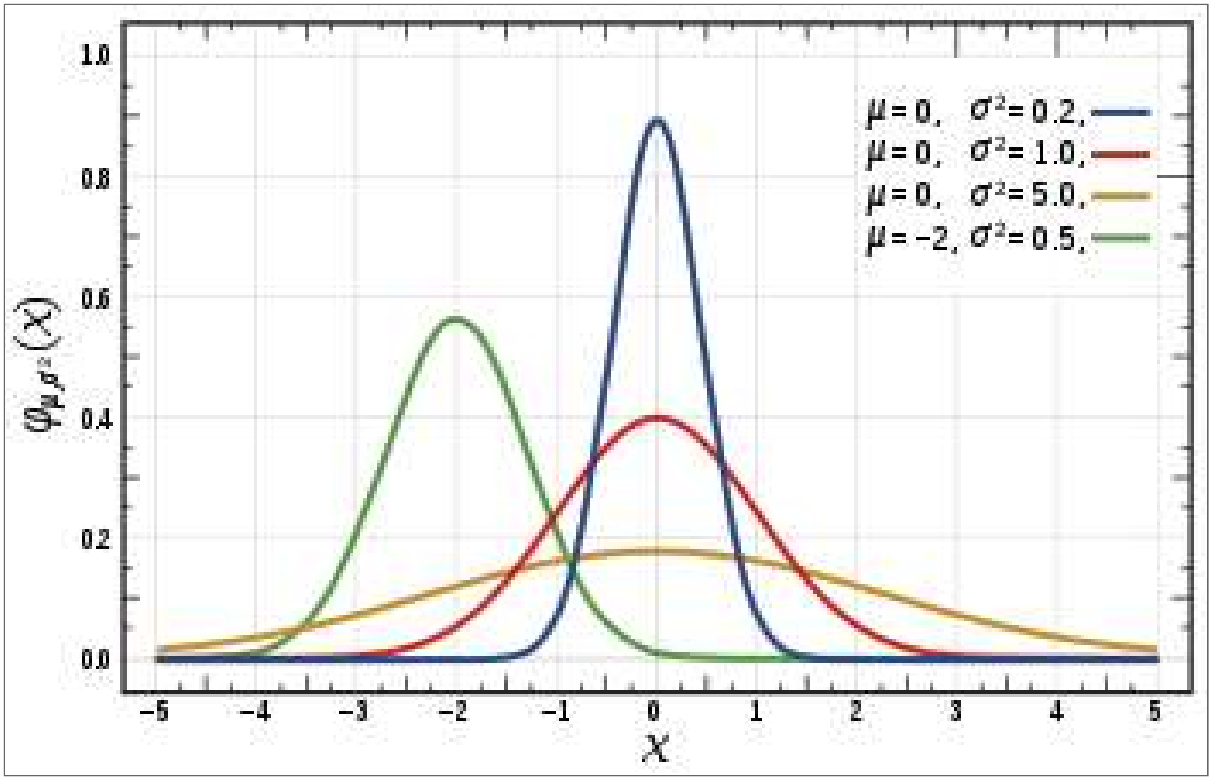
Variation

Min, Max



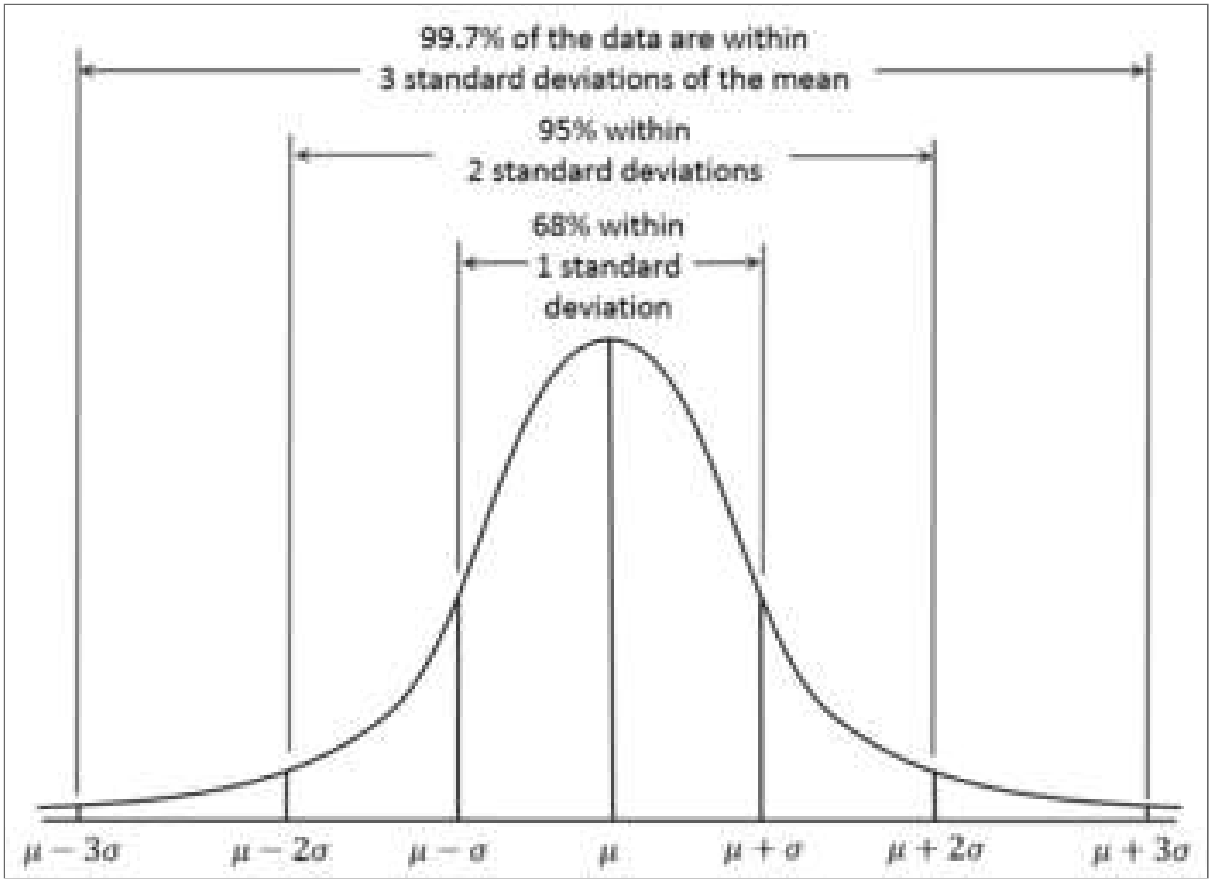
Variation

Variance, Standard Deviation



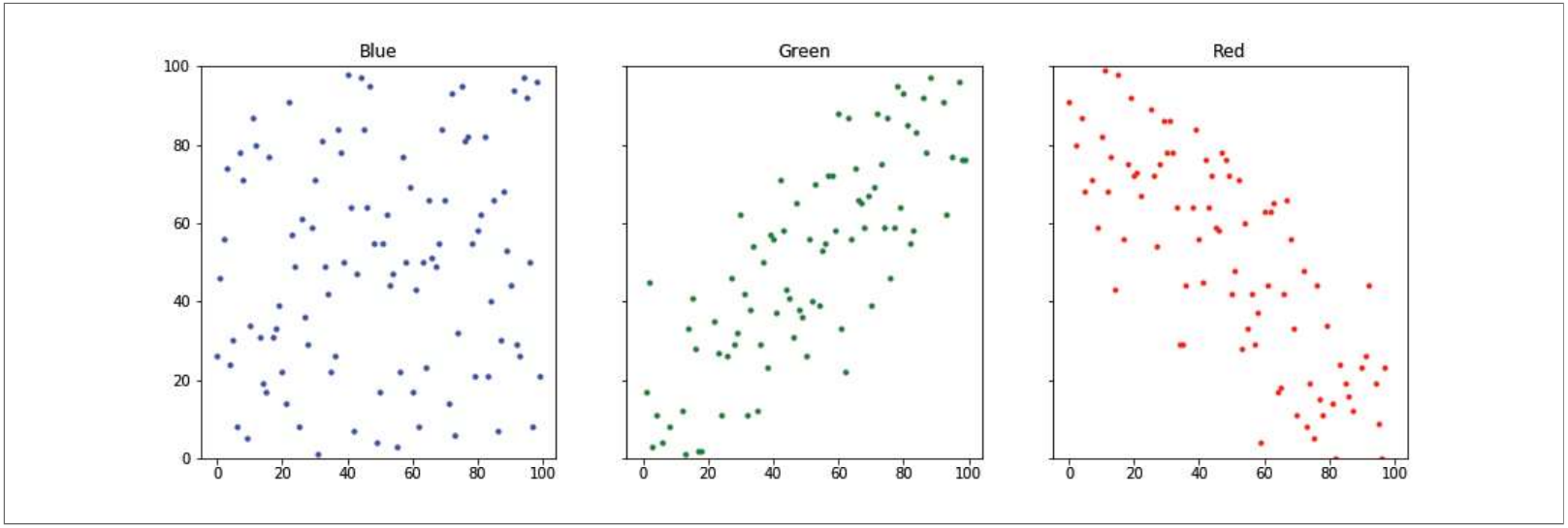
Variation

Percentiles



Dependence

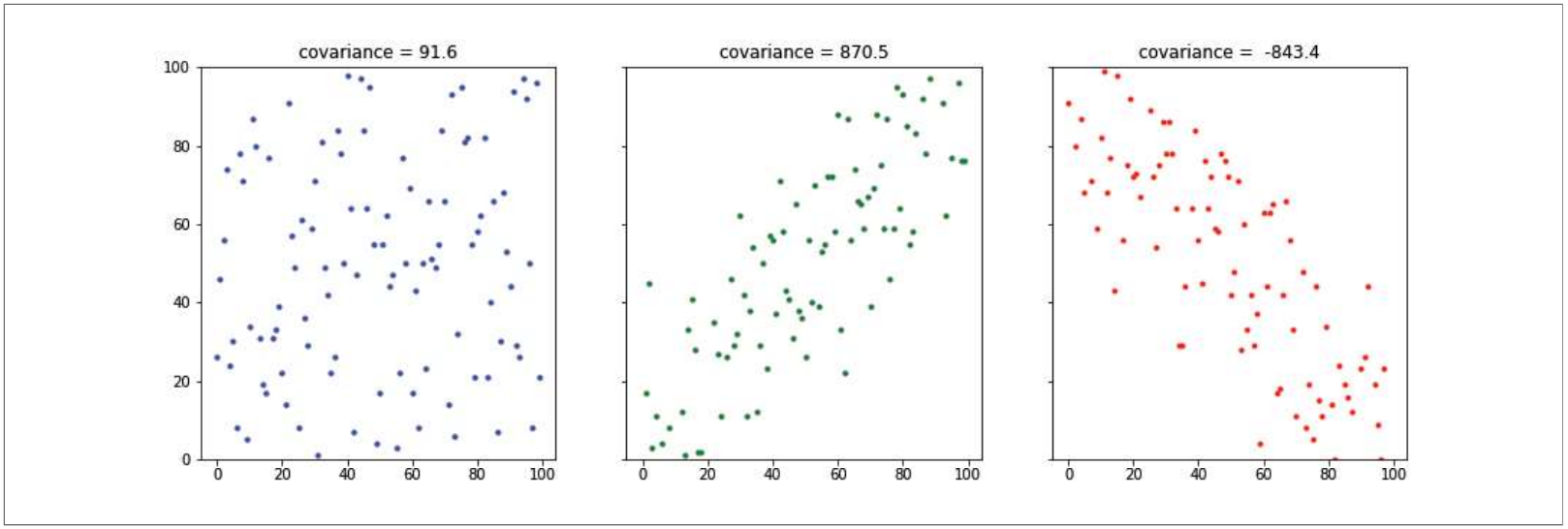
How to describe the relationship between two distributions?



formal definition

Dependence

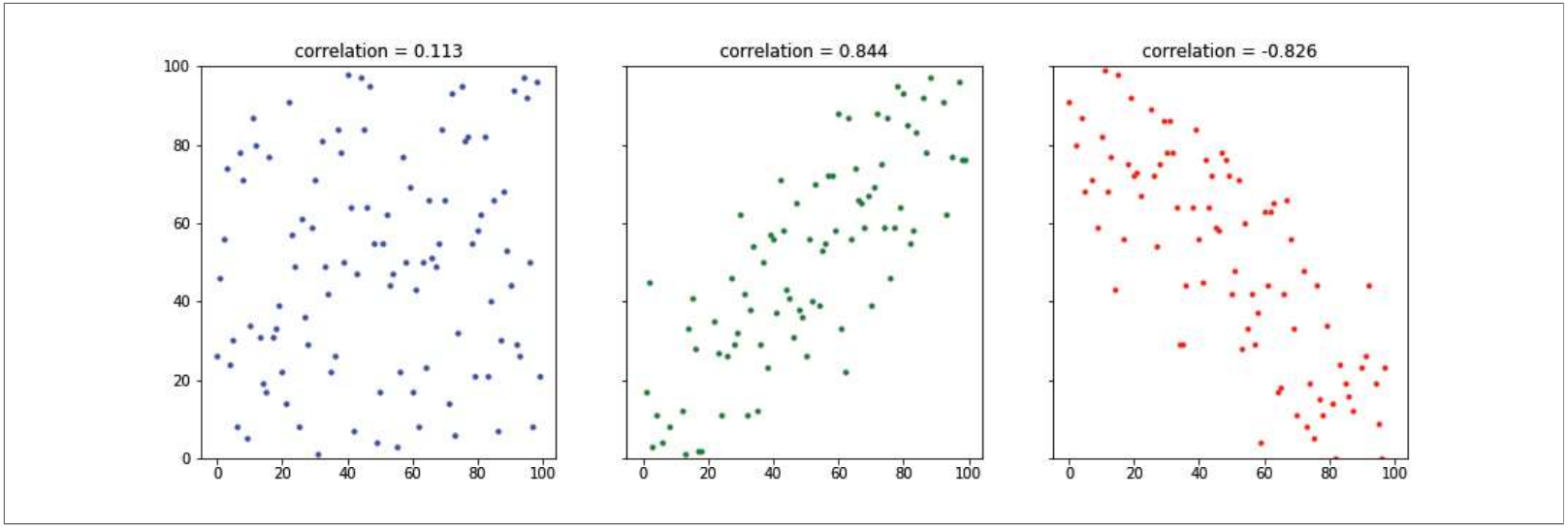
Covariance



formal definition

Dependence

Correlation



formal definition

Key Theorems

Law of Large Numbers

The average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

Key Theorems

Central Limit Theorem

When independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed.

Let's Code!

Wrap Up

1. Types of Data
2. Useful Statistical Distribution
3. Important Summary Statistics
4. Independence
5. Key Theorems

Reference: ***Data Science from Scratch***

Assignment 4: Due Monday, October 1 by 6:30pm

DataCamp's Statistical Thinking in Python (Part 2)

- The course should appear as assignment within your existing DataCamp account.
- Course takes 4+ hours, plan your time accordingly.