Galvanize statistics short course
Day 1

Introductions
Motivation
Random process
Random variable
Probability
Combinatorics
Probability Distributions
Statistics

F. Burkholder (credit A. Richards, M. Drury, I. Corneillet)

Introductions

- About me
  - https://www.linkedin.com/in/frankburkholder/
- About you
  - name, background, reason for attending this workshop?
- Course resources:
  - Statistics short course: https://galvanizeopensource.github.io/stats-shortcourse/
  - Mathematics short course: https://galvanizeopensource.github.io/math-essentials-for-data-science/index.html
  - Another statistics presentation: https://github.com/GalvanizeOpenSource/statistics-workshop
  - Galvanize Data Science Prep: https://www.galvanize.com/data-science/prep

Motivation

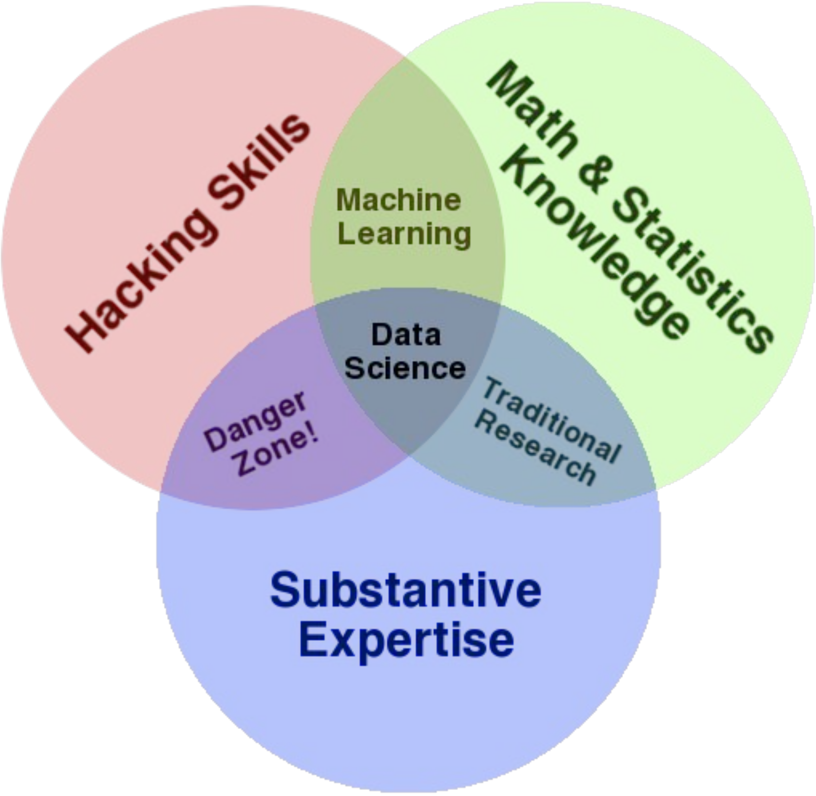
“Data Scientist: Person who is better at statistics than any software engineer and better at software engineering than any statistician.”—Josh Wills, Director of Data Eng. at Slack

“Data scientists are kind of like the new Renaissance folks, because data science is inherently multidisciplinary.”—John Foreman, Vice President of Product Management at MailChimp

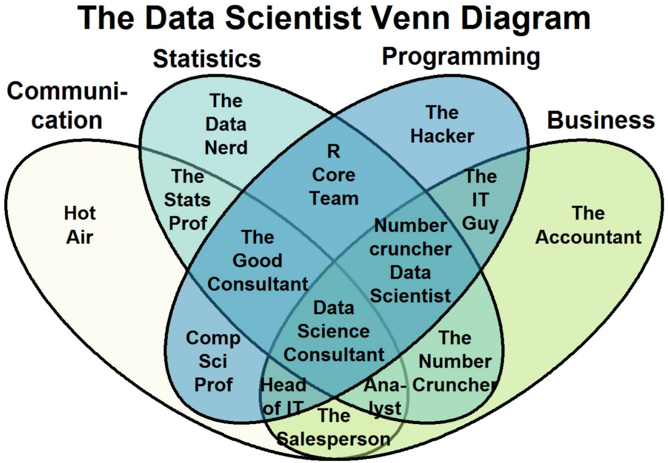
“The job of the data scientist is to ask the right questions. If I ask a question like ‘how many clicks did this link get?’ which is something we look at all the time, that’s not a data science question. It’s an analytics question. If I ask a question like, ‘based on the previous history of links on this publisher’s site, can I predict how many people from France will read this in the next three hours?’ that’s more of a data science question.”—Hilary Mason, Founder, Fast Forward Labs

https://www.kdnuggets.com/2017/05/42-essential-quotes-data-science-thought-leaders.html

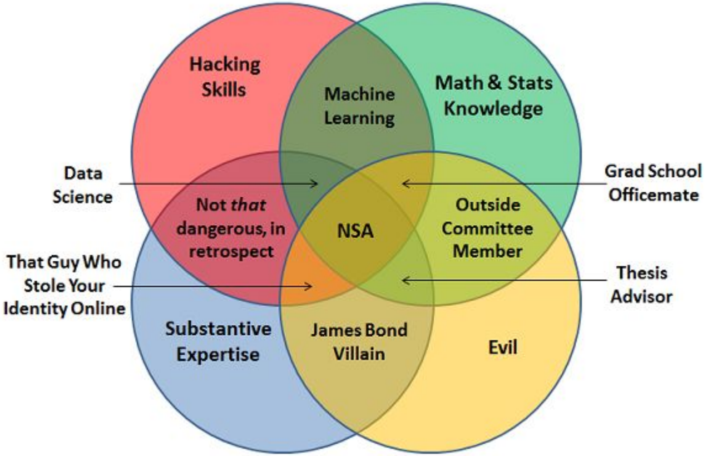
Motivation



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



http://rpubs.com/ptherond/DSc_Venn

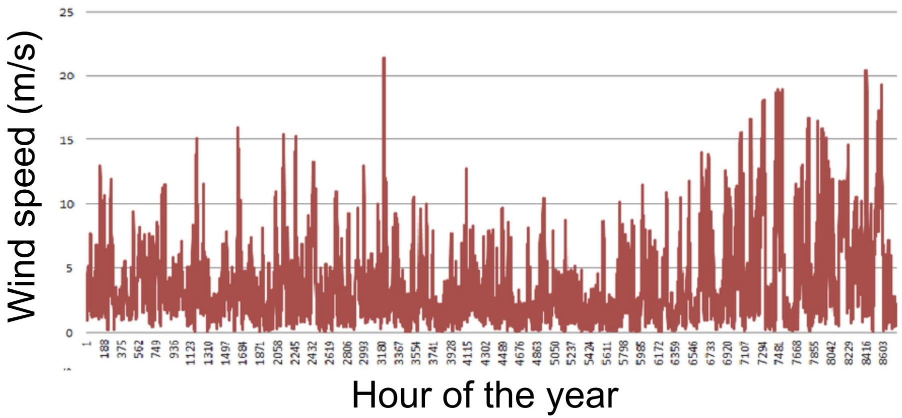


https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html

Random process

A process where the outcome is uncertain. It’s probabilistic (stochastic) rather than deterministic in nature.

The wind speed in the diagram appears to be due to a random process.



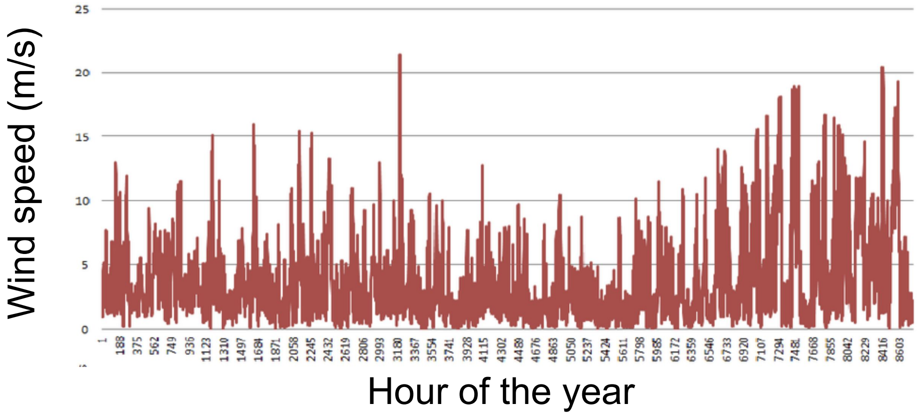
Random variable

A random variable X is something that can be used to generate outcomes in a way that probabilistic statements about the outcomes can be made.

Examples:
Say X is flipping a coin.
P(X = H) = 0.5 (fair coin)
P(X = H) = 0.3 (unfair coin)

Say X is the wind speed
P(X > 3 m/s and X < 12 m/s) = ?

A random variable is a variable
that depends on a random process.



Probability

The study of patterns in a random process in which **the characteristics of a random process are known**.

The goal is often understanding what the **output of the random process** is.

- Expected value
- Minimum, maximum
- Distribution of possible values

For example:

Let’s say we have a fair coin: p(Heads) = 0.5
After 20 flips, how likely are we to get 5 H, 7 H, 10 H, 20 H?

Statistics

The study of a random process where some characteristics of it are unknown.

The goal is often to figure out (infer) characteristics of the random process given data that the process made.

For example:

> Let’s say we have a coin, and we don’t know if it’s fair or not. We flip it 20 times and get 8 heads. How likely is that result if it’s a fair coin?

Probability vs. Statistics

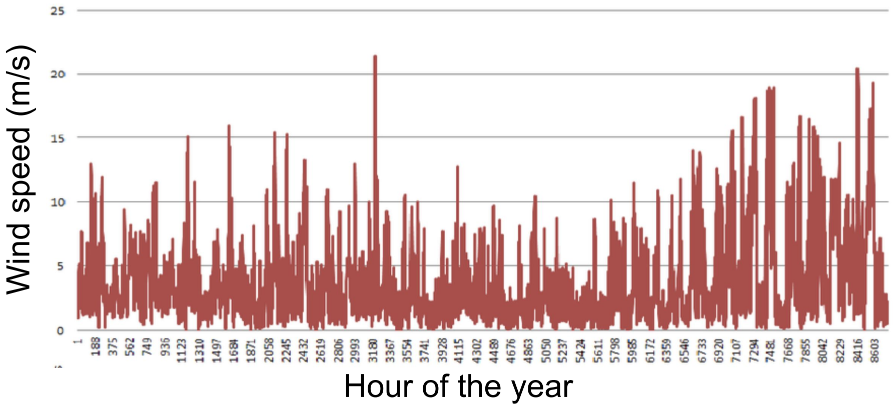
Statistics: Have data, infer what random process made the data.
Probability: Know the random process, what data will it make?

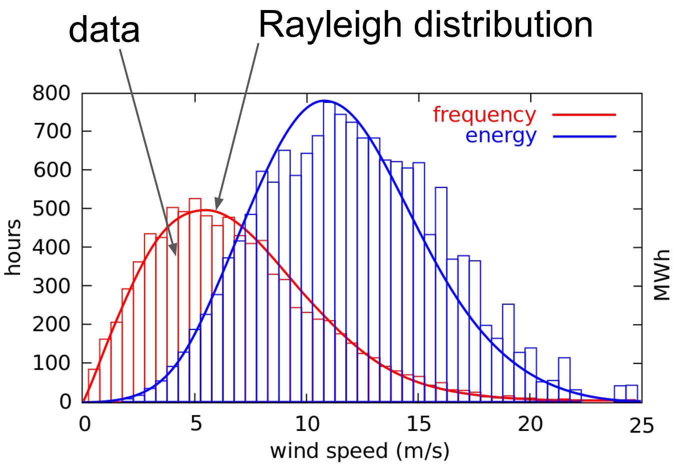
data -> infer process	Statistics
know process -> data	Probability

Solving a problem: using Probability and Statistics

Question: How much energy is our wind farm going to generate?

1. Gather wind speed data at the site.
2. Use statistics to infer the distribution (Rayleigh) and its parameters (alpha) from the wind speed data.
3. Use probability to determine the likely distribution of wind speeds using the Rayleigh distribution.
4. Relate wind speed to energy generated.





wikipedia - Wind power

You try: Probability, Statistics, or both?

1. You were rolling a 6-sided die and you rolled five 6s in a row. Do you think it's fair?
2. With a fair die, how likely is it to roll five 6s in a row?
3. You got an average of 4 robocalls per day last week. How many do you think you'll get today?
4. In a mobile phone manufacturing line, 1 phone is found to be defective per 1000 phones. You watched the process and saw 1500 phones go by until 1 was rejected. Was the failure rate incorrect?
5. You're a teacher grading your students' tests:
   a. You plot a histogram of the results and find that they're well described by a normal distribution with a mean of 75%, standard deviation of 5%.
   b. Another teachers' top student scores 100% on the exam. Is she a better teacher than you are?

Combinatorics

Combinatorics is an area of mathematics primarily concerned with counting, both as a means and an end in obtaining results. - Wikipedia

The basic problem solving skill you need to solve problems in probability is counting.

$$P(\text{event}) = \frac{\#\text{ of desired outcomes}}{\#\text{ of total outcomes}}$$

Combinatorics

- How many ways are there to arrange four letters of the alphabet?
- How many ways are there to arrange four different letters of the alphabet?
- How many ways are there to arrange 25 books on a bookshelf?
- How many hands are full houses?
- How many hands have three-of-a-kind?
- How many hands have three-of-a-kind and are not also full houses?

Basic counting principle

If a task can be accomplished as a series of steps, then the number of outcomes of the task is the product of the number of outcomes for each individual step.

How many ways to arrange four letters of the alphabet?

- Pick the first letter
  - How many options?
- Pick the second letter
  - How many options? (assume you can re-pick letters, i.e. sample with replacement)
- Pick the third letter
  - How many options?
- Pick the fourth letter
  - How many options?

How many ways to arrange four letters of the alphabet?

- Pick the first letter
  - 26
- Pick the second letter
  - 26
- Pick the third letter
  - 26
- Pick the fourth letter
  - 26

Number of ways (outcomes): $26 \times 26 \times 26 \times 26 = 456,976$

How many ways to arrange four letters of the alphabet?

- Pick the first letter
    - How many options?
- Pick the second letter
    - How many options? (assume you can't re-pick letters, i.e. sampling without replacement)
- Pick the third letter
    - How many options?
- Pick the fourth letter
    - How many options?

How many ways to arrange four letters of the alphabet?

- Pick the first letter
  - How many options?
- Pick the second letter
  - How many options? (assume you can't re-pick letters, i.e. sampling without replacement)
- Pick the third letter
  - How many options?
- Pick the fourth letter
  - How many options?

$$26 \times 25 \times 24 \times 23 = 358,800$$

Permutations: ordered selections without replacement

The number of **ordered** selections of **k** objects **without replacement** from a population for **n** objects is called the **number of permutations** of **k** objects taken from **n**.

$$P(n,k) = \underbrace{n \times (n-1) \times (n-2) \times \ldots \times (n-k+1)}_{k \text{ factors}} = \frac{n!}{(n-k)!}$$

Permutations example

You have 25 books on a bookshelf; how many ways are there to arrange these books in any order?

what’s:

n?

k?

P(n, k)?

Permutations example

You have 25 books on a bookshelf; how many ways are there to arrange these books in any order?

what’s:

n? 25

k? 25

P(n, k)? 25! / (25 - 25)! = 1.551e+25

Permutations example

You have 25 books in a box. You pick 5 of them at random and arrange them in any order on the bookshelf. How many different arrangements are possible?

what’s:

n?

k?

P(n, k)?

Permutations example

You have 25 books in a box. You pick 5 of them at random and arrange them in any order on the bookshelf. How many different arrangements are possible?

what's:

n? 25

k? 5

P(n, k)? 25!/(25 - 5)! = 6,375,600

Arrangements & Permutations - you try

1. How many different car number plates are possible with 3 digits followed by 3 letters?

2. 4 girls and 3 boys are sitting on a bench.
   a. How many different ways could they sit if there were no restrictions?

   b. How many different ways could they sit if the ordered needed to alternate girl-boy? (starting with girl)

3. In a 100 m dash with 10 athletes, how many different ways could they finish first, second, and third?

Arrangements & Permutations - you try

1. How many different car number plates are possible with 3 digits followed by 3 letters?

   10 * 10 * 10 * 26 * 26 * 26

2. 4 girls and 3 boys are sitting on a bench.
   a. How many different ways could they sit if there were no restrictions?

      7!

   b. How many different ways could they sit if the ordered needed to alternate girl-boy? (starting with girl)

      4! * 3!

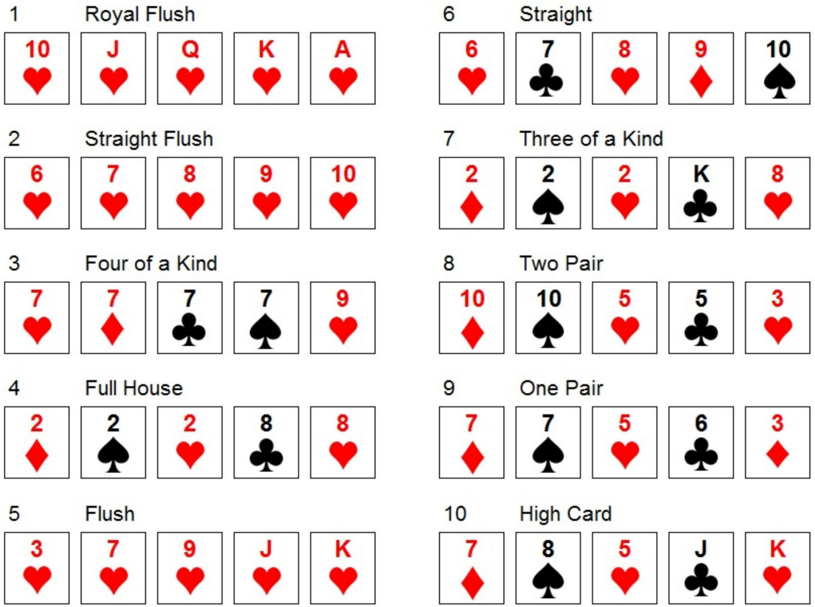
3. In a 100 m dash with 10 athletes, how many different ways could they finish first, second, and third?

   10! / (10-3)!

Combinations - where order doesn’t matter

How many hands are possible when drawing from a standard (52-card) deck?

(a hand is an **unordered** collection of five cards)



https://allmathconsidered.wordpress.com/2017/05/23/the-probabilities-of-poker-hands/

Combinations - where order doesn't matter

To choose an unordered list of five cards, first choose five cards, then determine all the ways to order them:

$$\overbrace{\#\text{ of ordered hands}}^{52\times 51\times 50\times 49\times 48} =$$

$$\#\text{ of unordered hands} \times \underbrace{\#\text{ of ways to order a hand}}_{5\times 4\times 3\times 2\times 1}$$

Combinations - where order doesn't matter

$$\#\text{ of unordered hands} = \frac{\overbrace{\#\text{ of ordered hands}}^{52\times 51\times 50\times 49\times 48}}{\underbrace{\#\text{ of ways to order a hand}}_{5\times 4\times 3\times 2\times 1}}$$

Combinations - where order doesn’t matter

$$\#\ \text{of unordered selections} = \frac{\#\ \text{of ordered selections}}{\#\ \text{of ways to order a single selection}}$$

E.g.,

$$\#\ \text{of unordered hands} = \frac{\#\ \text{of ordered hands}}{\#\ \text{of ways to order a hand}}$$

Combinations

The number of **unordered** selections of **k** objects **without replacement** from a population for **n** objects is called the **number of combinations** of **k** objects taken from **n**.

Commonly called the binomial coefficient, or n choose k

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

Combinations example 1

How many different combinations of players results from a team of 8 players, where only 5 can play at a time on the court?

Combinations example 1

How many different combinations of players results from a team of 8 players, where only 5 can play at a time on the court?

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

8! / (5! * (8 - 5)!) = 56

Combinations example 2

How many hands are full houses? (a full house is a hand with a pair of the same value and a three-of-a-kind of the same value.)

Combinations - how many full houses?

How can we accomplish this task as a step-by-step process?

- Pick the rank (the number) of the three-of-a-kind
    - How many options?
- Pick three suits of that rank
    - How many options?
- Pick the rank the number of the pair
    - How many options?
- Pick two suits of that rank
    - How many options?

Combinations - how many full houses?

How can we accomplish this task as a step-by-step process?

- Pick the rank (the number) of the three-of-a-kind
  - How many options?
- Pick three suits of that rank
  - How many options?
- Pick the rank the number of the pair
  - How many options?
- Pick two suits of that rank
  - How many options?

$$C(13,1) \times C(4,3) \times C(12,1) \times C(4,2) = 13 \times 4 \times 12 \times 6 = 3,744$$

Combinations - you try

A committee of 5 people is to be chosen from 6 men and 4 women.

1. How many committees are possible?

2. There’s a desire to keep the ratio of men to women the same. How many committees are possible with 3 men and 2 women?

3. A woman majority committee is desired. How many woman-majority committees are possible?

Combinations - you try

A committee of 5 people is to be chosen from 6 men and 4 women.

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

1. How many committees are possible?

   10! / (5! * (10 - 5)!) = 252

2. There's a desire to keep the ratio of men to women the same. How many committees are possible with 3 men and 2 women?

   6! / (3! * (6 - 3)!) * 4! / (2! * (4 - 2)!) = 20 * 6 = 120

3. A woman majority committee is desired. How many woman majority committees are possible?

   3-2 and 4-1 both majority.

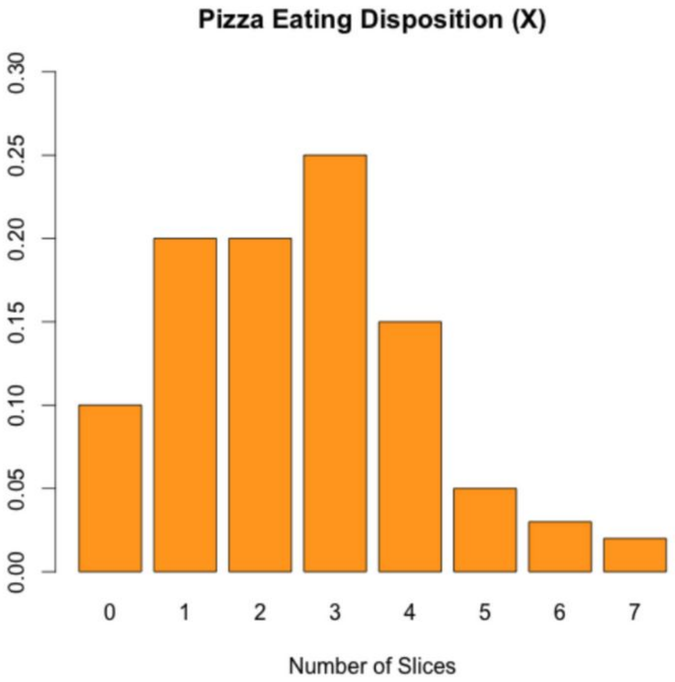
   4! / (3! * (4 - 3)!) * 6! / (2! * (6-2)!) + 4! / (4! * (4 - 4)!) * 6! / (1! * (6-1)!) = 66

Probability Distributions

A **probability distribution** is a mathematical function that provides the **probabilities of occurrence** of different **possible outcomes** in an experiment.

In more technical terms, the probability distribution is a description of a random phenomenon in terms of the probabilities of events. - Wikipedia

An example of a probability distribution:

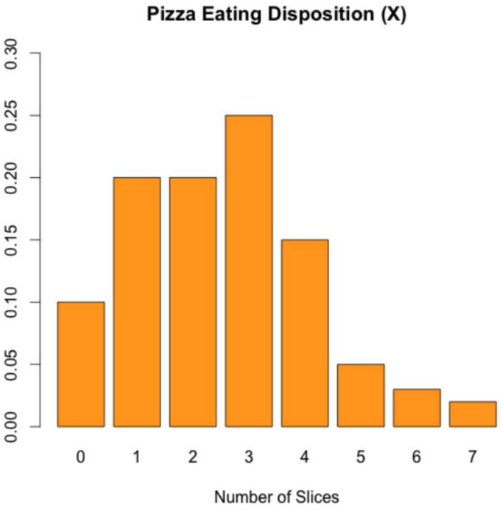


Probability Distributions

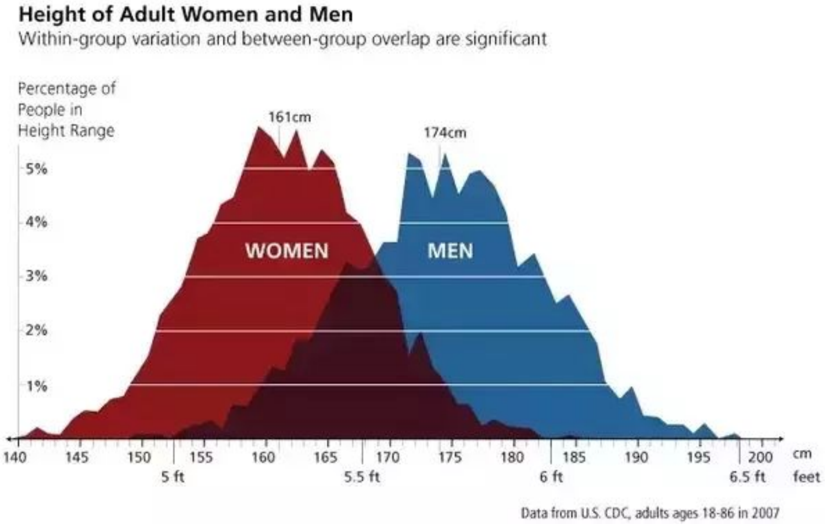
Probability Distributions are classified into two categories:

- discrete – producing outcomes that can be mapped to integers (such as 1, 2, …) or outcomes with a fixed interval between them (0.35, 0.45, 0.55, …)
- continuous – producing real-valued outcomes (such as 3.14… or 2.71…)

Discrete



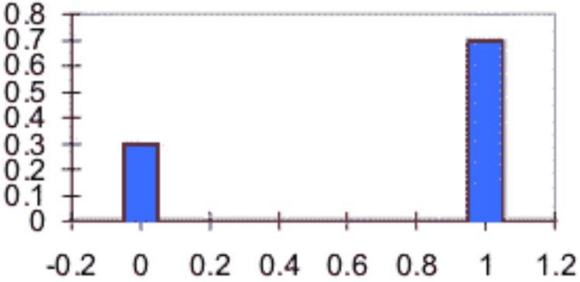
Continuous



Two types of probability distributions

Discrete

defined at fixed intervals
in the support



probability mass function

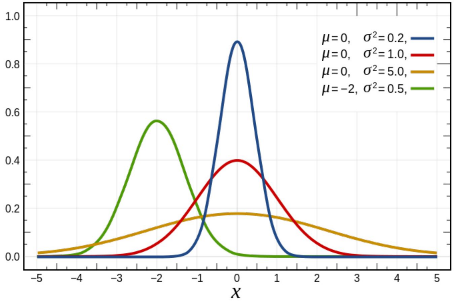
(pmf)

can read probabilities directly
from distribution

$$Pr(X = x)$$

Continuous

defined for all values
in the support



probability density function

(pdf)

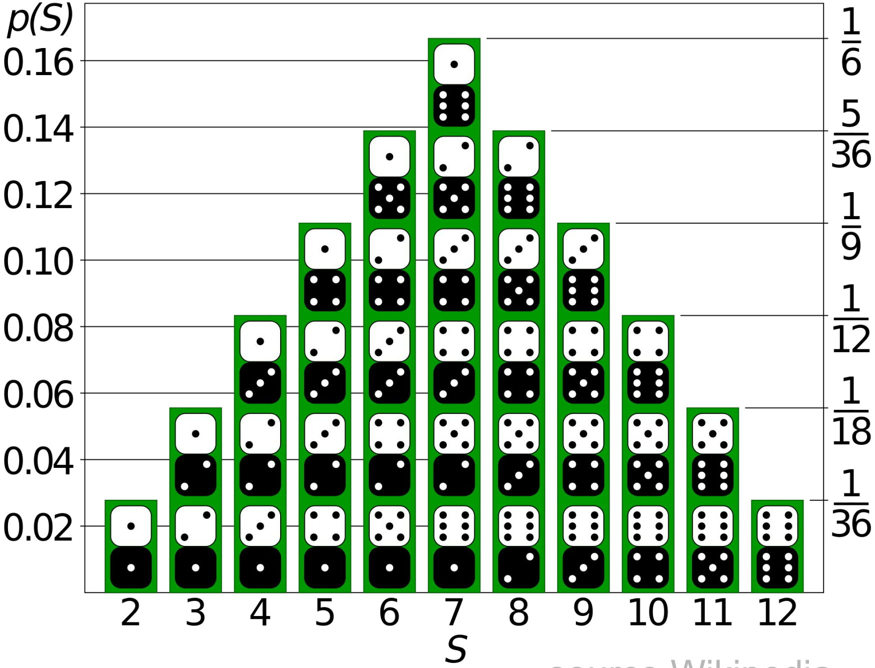
relative probabilities on y-axis
get probabilities in an interval E by integrating:

$$Pr(X = x \in E) = \int_E f(X = x)\, dx$$

Samples taken from probability dist. are i.i.d.

i.i.d: **i**ndependent and **i**dentically **d**istributed.

Summed rolls from two fair 6-sided dice:



9, 4, 8, 5, 2, 3, 9, 7, 6, 12, etc...

These rolls are i.i.d.

source Wikipedia

Essential Discrete Probability Distributions

A **Bernoulli** models one attempt that results in either a success of a failure. Such an attempt is called a Bernoulli trial, with a success typically recorded as a 1 and a failure typically recorded as a 0.

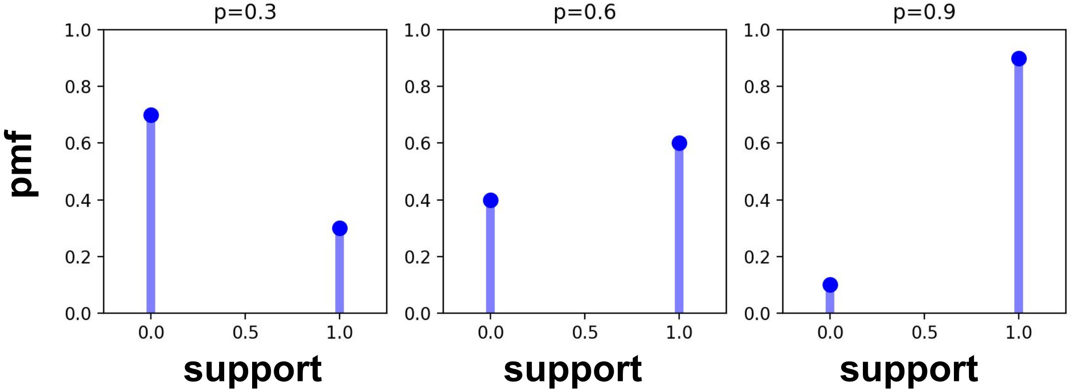
Probability mass function (pmf): $Pr(X=1)=p,\ \text{and}\ Pr(X=0)=1-p$

where *p* is the probability of success.

The mean (expectation) and variance are:

- $E[X]=p$
- $Var(X)=p(1-p)$

Example: flip a coin. (H or T)



Essential Discrete Probability Distributions

The **binomial** distribution defines the probability of observing exactly *k* successes out of *n* identical Bernoulli trials

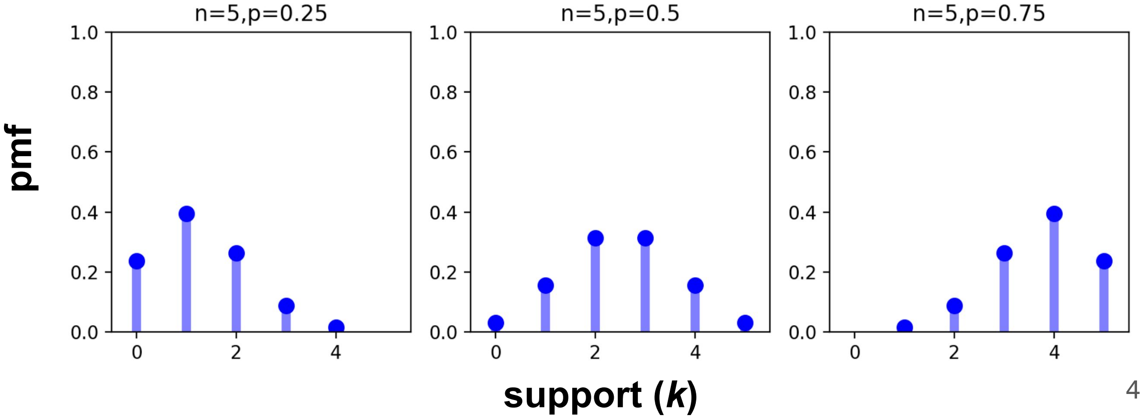
pmf: $$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ for } k \in \{0, 1, \ldots, n\}$$

where p is the success probability of the Bernoulli trial.

The mean (expectation) and variance are:

- np
- $np(1-p)$

Example: flips of a coin.
[H, T, T, H, H]



Essential Discrete Probability Distributions

The **geometric** distribution defines the probability of needing to perform *k−1* identical Bernoulli trials before a success is observed on the *kth* trial.

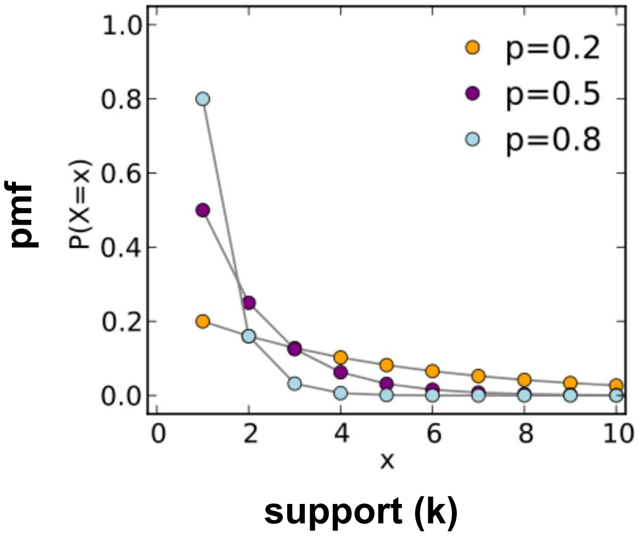
pmf: $$Pr(X = k) = (1 - p)^{k-1} p, \text{ for } k \in \{0, 1, \dots\}$$

where p is the probability of success.

The mean, variance are:

- $E[X] = \frac{1}{p}$
- $Var(X) = \frac{1-p}{p^2}$

Example: Number of times
you play the lottery until you finally win!



Essential Discrete Probability Distributions

The **Poisson** distribution models the number of times an event occurs (*k*) within a fixed time interval, given an average rate of occurrence.

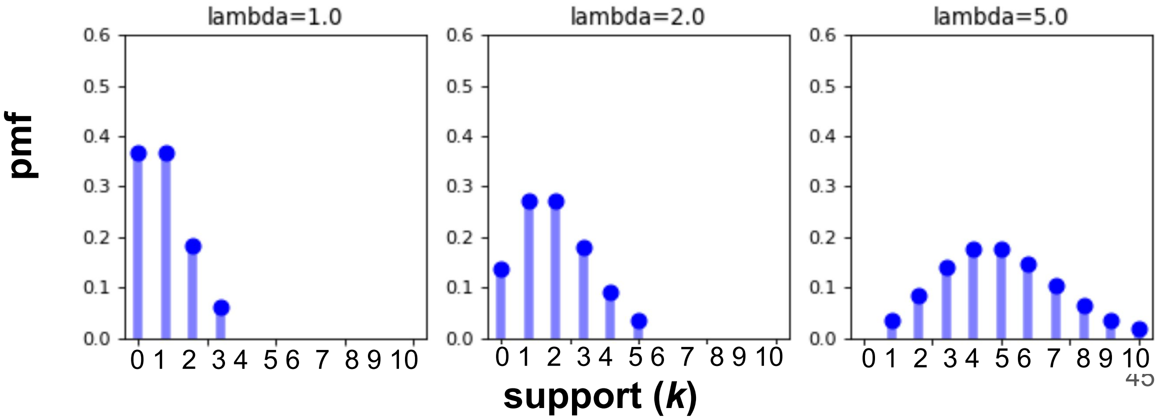
pmf: $$Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \text{ for } k \in \{0, 1, 2, \dots\}$$

where λ is the mean rate of occurrence.

The mean and variance are identical

- $E[X] = Var(X) = \lambda$

Example: Number of busses passing a street corner in 1 hr.



Essential Continuous Probability Distributions

The (continuous) **uniform** distribution generates completely random occurrences of equal value over a defined space. (can also be discrete)

Probability density function (pdf):

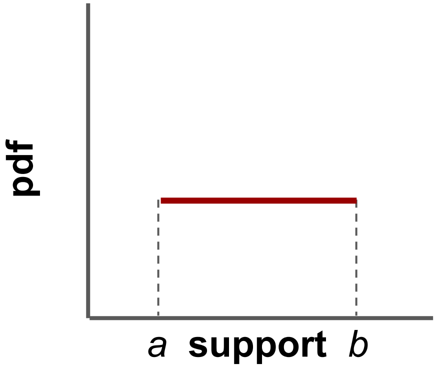
$$f(X = x) = \begin{cases} \frac{1}{b-a}, & \text{for } x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

where *a* and *b* are the limits of the support.

The mean, variance are:

- $E[X] = \frac{a+b}{2}$
- $Var(X) = \frac{(b-a)^2}{2}$

Example: A uniform random number between a and b.



Essential Continuous Probability Distributions

The **Gaussian** or **normal** distribution is a continuous probability distribution whose probability density function is defined as

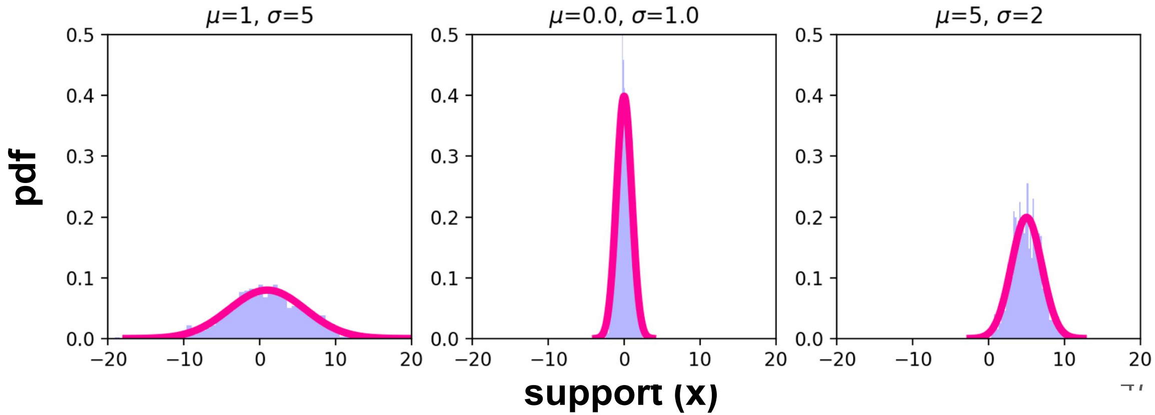
pdf: $$f(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ for } x \in (-\infty, \infty)$$

where *μ* and *σ* are the mean and variance.

The mean, variance are:

- $E[X] = \mu$
- $Var(X) = \sigma^2$

Example: Distribution of measurement errors



Essential Continuous Probability Distributions

The **exponential** distribution is useful for estimating “time to arrival” outcomes of Poisson processes.

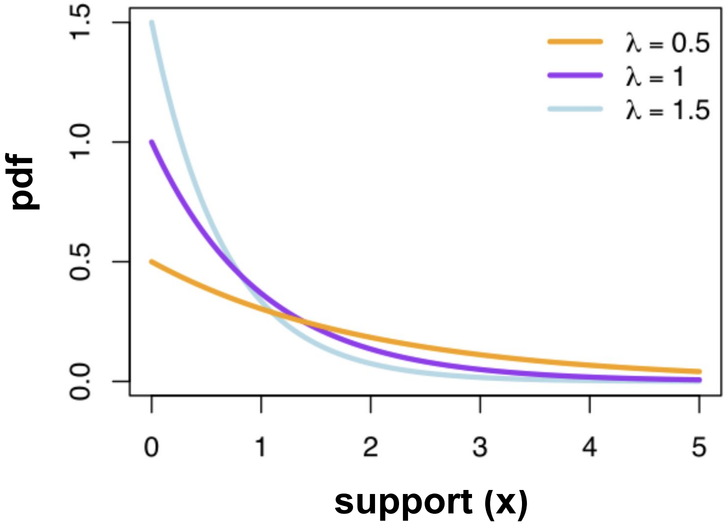
pdf: $$Pr(X = x) = \lambda e^{-\lambda x}, \text{ for } x \in (0, \infty)$$

where λ is the rate of the Poisson process.

The mean, variance are:

- $E[X] = \frac{1}{\lambda}$
- $Var(X) = \frac{1}{\lambda^2}$

Example: If 6 buses pass the bus stop per hour, how long will I wait for a bus?



Probability distributions: you try

What probability distribution would you use to model the following situations?

- What is the probability that a man’s height is between 5’9” and 6’0”?
- Deciding to go for a run or not.
- Determining how many days pass before you finally decide to go for a run.
- Determining how likely it is that you go for 10 runs in a month.
- Calculating which day of the month you buy new shoes.
- Assuming you run at a 9 minute mile average pace, determining how likely it is that you pass the 3 mile mark in a race in 25 minutes?

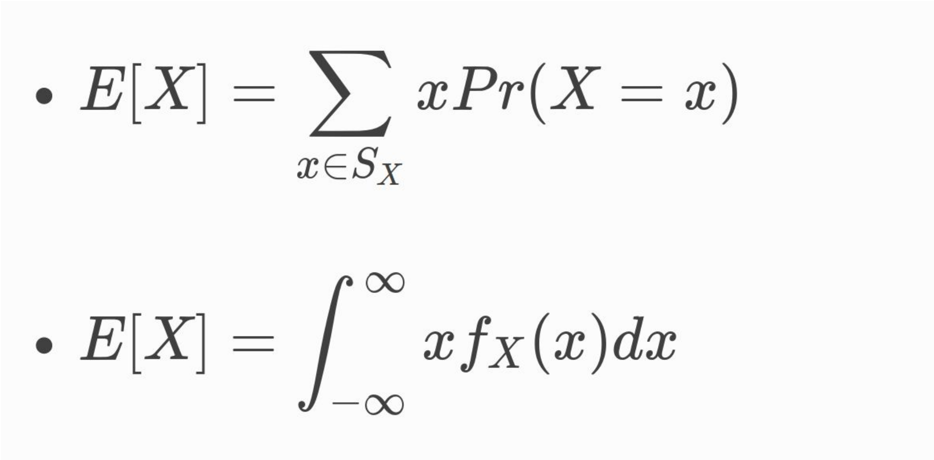
Statistics

Statistics is the sister discipline to probability in mathematics.

Statistics addresses the inverse problem of learning about probability distributions from data, as opposed to the forward problem of generating data from probability distributions.

The term statistic is also a definition – a statistic is a “mathematical calculation of some data”.

Statistics - Expectation



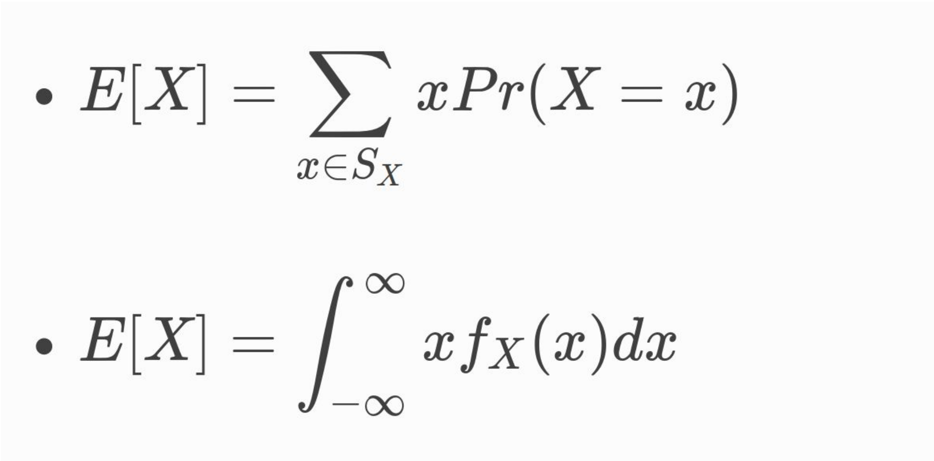
- $E[X] = \sum_{x \in S_X} x Pr(X = x)$
- $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$

discrete distribution

continuous distribution

What’s the expectation (expected value) of a fair six-sided die?

Statistics - Expectation



- $E[X] = \sum_{x \in S_X} x Pr(X = x)$
- $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$

discrete distribution

continuous distribution

What's the expectation (expected value) of a fair six-sided die?

$$1 * \frac{1}{6} + 2 * \frac{1}{6} + 3 * \frac{1}{6} + 4 * \frac{1}{6} + 5 * \frac{1}{6} + 6 * \frac{1}{6} = 3.5$$

Statistics - Variance

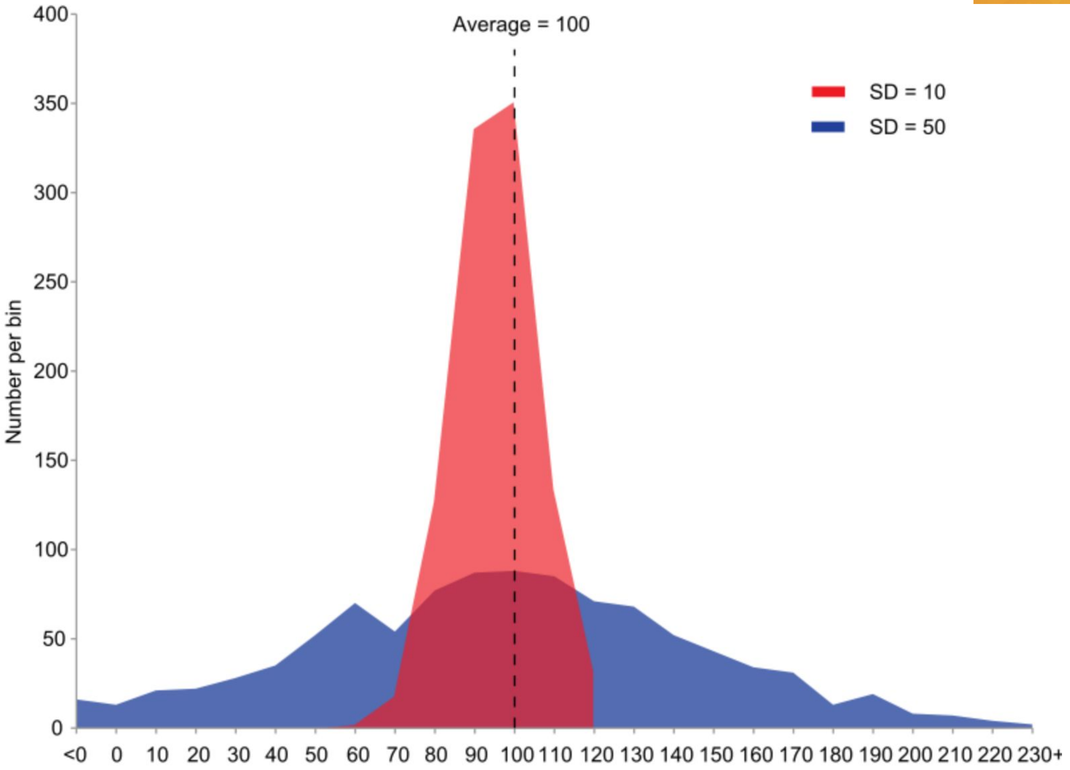
$$\mathrm{Var}(X) = \mathrm{E}\left[(X-\mu)^2\right]$$

The expected squared difference between the mean and each value in X.

$$\mathrm{Var}(X) = \sigma^2$$

The variance is the standard deviation squared.

It's a measure of the spread of your data from its mean.



source: Wikipedia

Statistics - Covariance

$$\operatorname{cov}(X,Y) = \mathbf{E}\left[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\right]$$

Covariance is a measure of the joint variability of two random variables.

The sign of the covariance therefore shows the tendency in the linear relationship between the variables: + they increase together, - as one increases the other decreases.

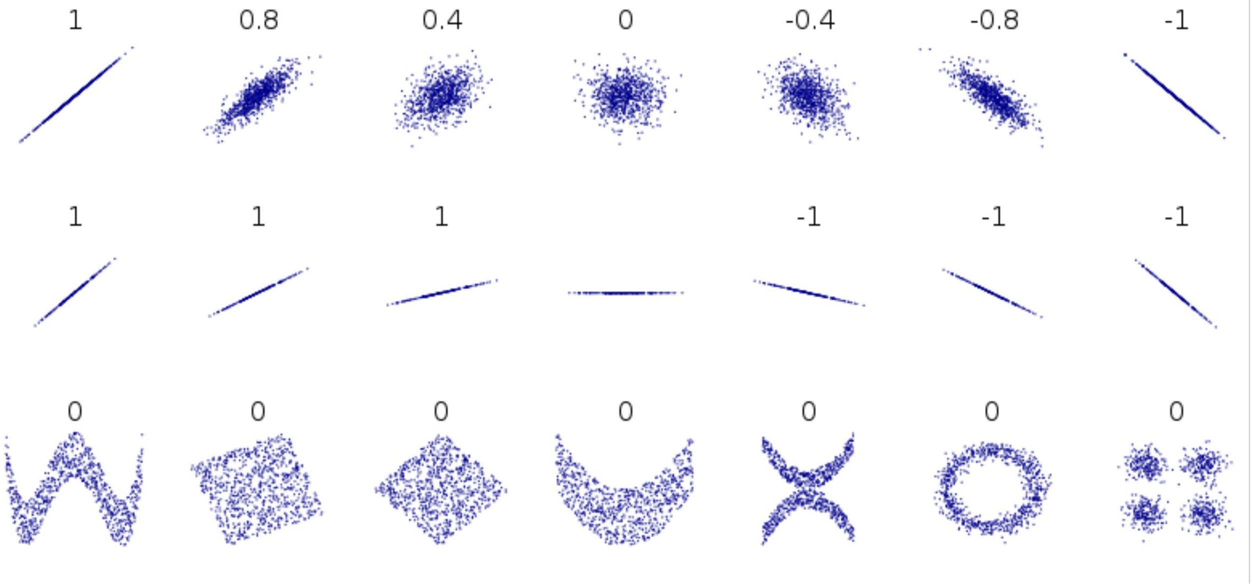
The magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the variables.

Normalized version: correlation coefficient.

Statistics - Correlation coefficient

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Commonly called Pearson's correlation coefficient. Examples:



Thank you!

See you on Day 2.