Probability From Scratch

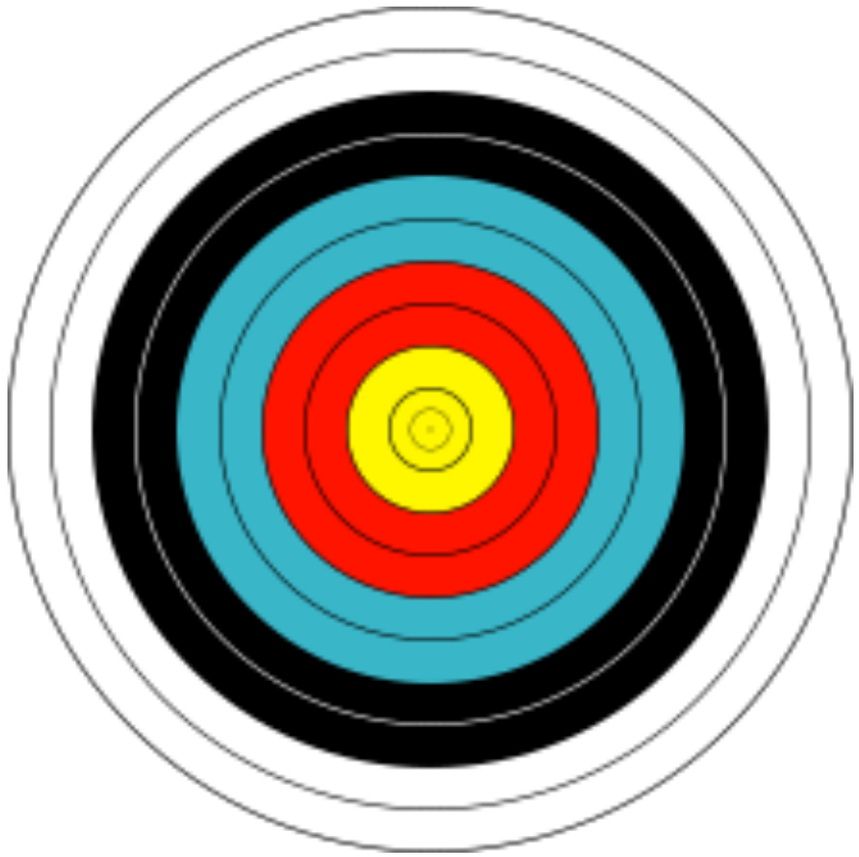
Bruno Gonçalves

www.data4sci.com/newsletter

https://github.com/DataForScience/Probability

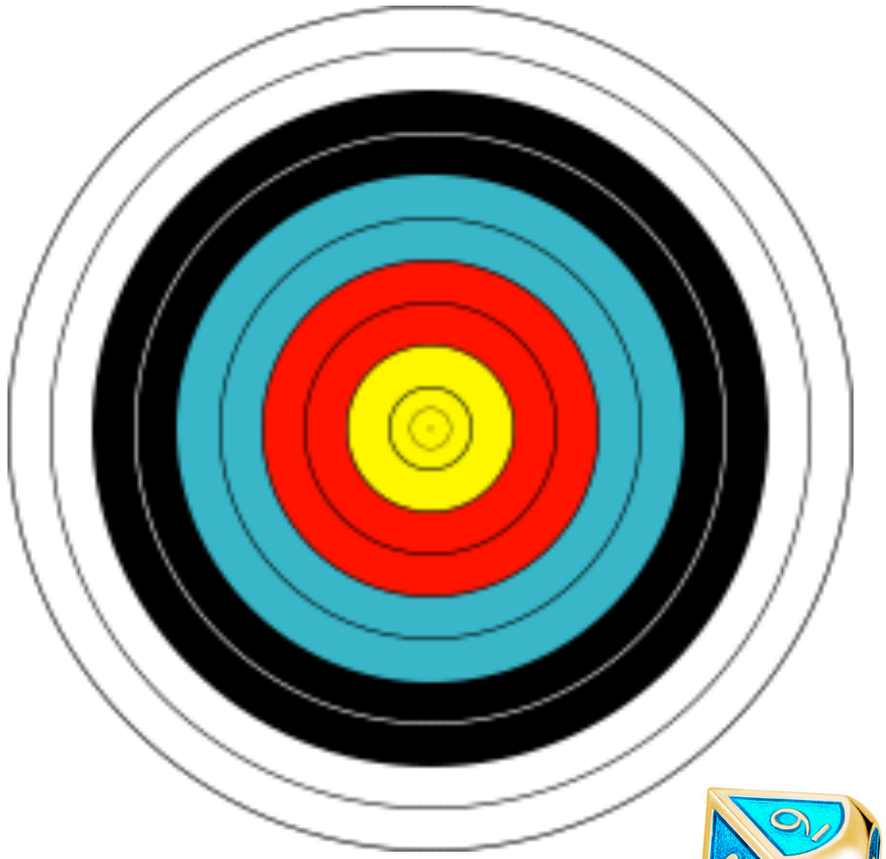
Lesson I:
Basic Definitions and Intuition

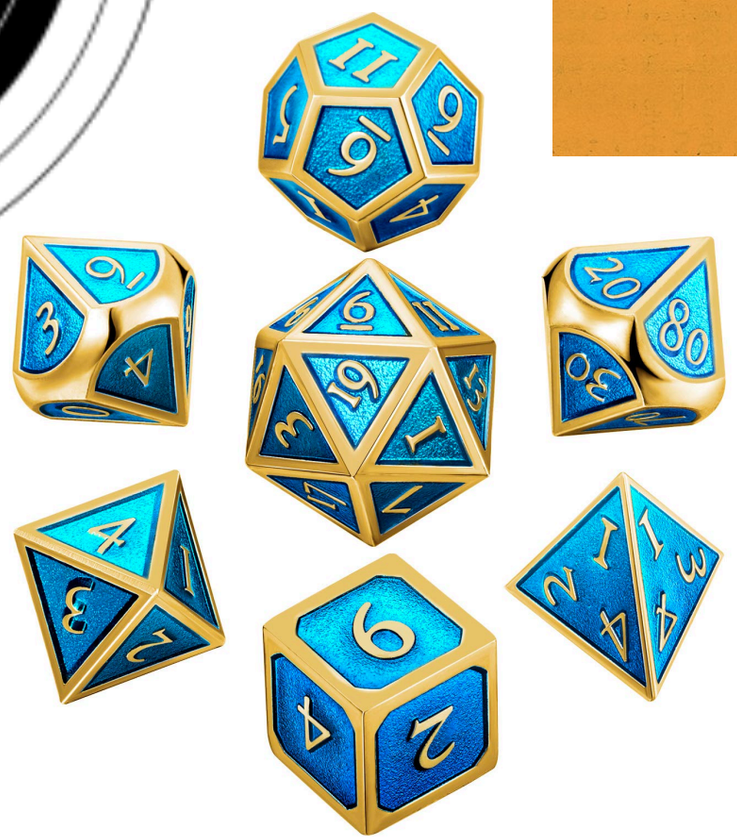
Randomness



Randomness / Uncertainty

https://images-na.ssl-images-amazon.com/images/I/81ZDkjk0NAL._SL1500_.jpg





Kolmogorov's Probability Axioms

https://en.wikipedia.org/wiki/Probability_axioms

- Axiom 1: Probability is a real number greater or equal to 0.
- Axiom 2: Total probability is equal to 1.
- Axiom 3: Probability of mutually exclusive events is the sum of the probabilities.

Kolmogorov's Probability Axioms

https://en.wikipedia.org/wiki/Probability_axioms

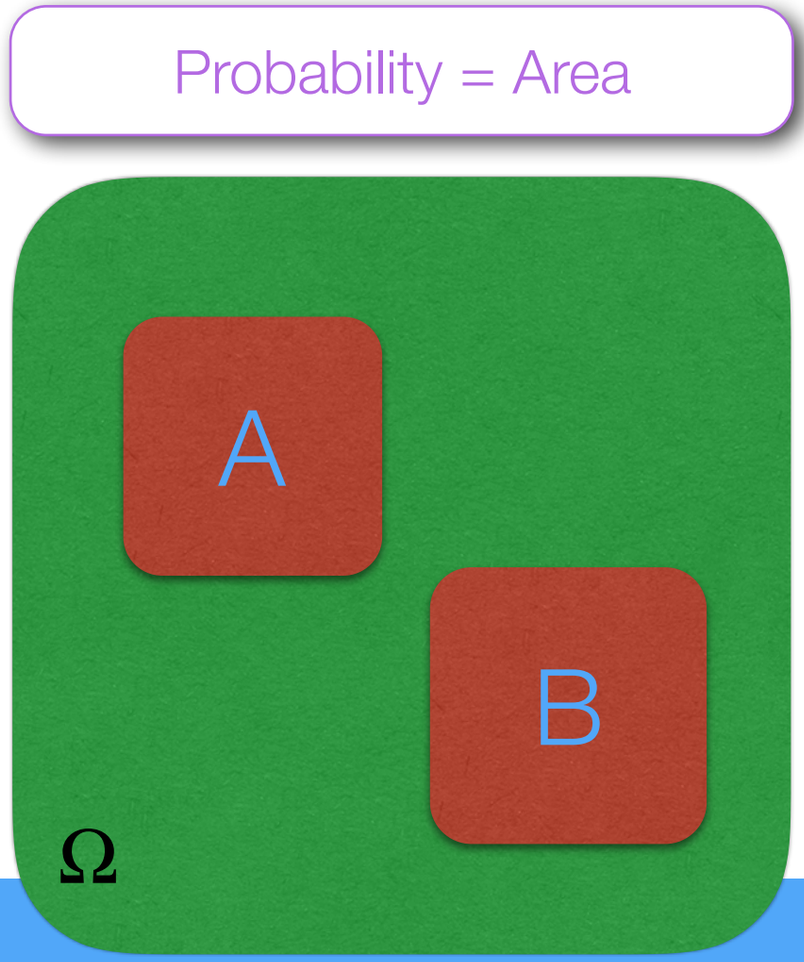
- Axiom 1: Probability is a real number greater or equal to 0.
- Axiom 2: Total probability is equal to 1.
- Axiom 3: Probability of mutually exclusive events is the sum of the probabilities.

Probability = Area

Kolmogorov's Probability Axioms

https://en.wikipedia.org/wiki/Probability_axioms

- Axiom 1: Probability is a real number greater or equal to 0.
- Axiom 2: Total probability is equal to 1.
- Axiom 3: Probability of mutually exclusive events is the sum of the probabilities.



$$0 \leq P(A) \leq 1$$

$$P(\Omega) \equiv 1$$

$$P(A, B) = P(A) + P(B)$$

Kolmogorov's Probability Axioms

https://en.wikipedia.org/wiki/Probability_axioms

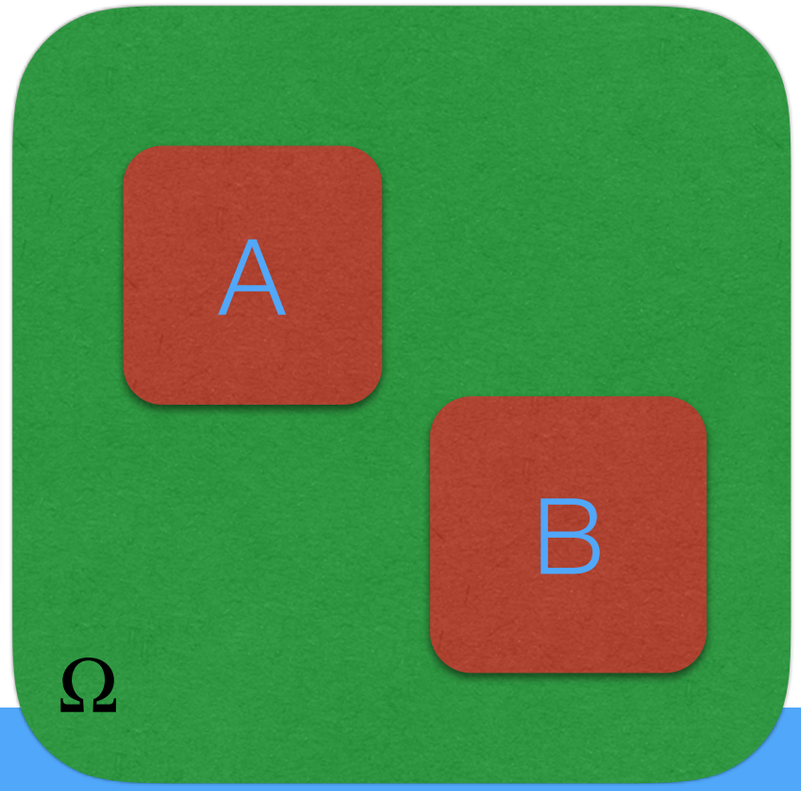
- Axiom 1: Probability is a real number greater or equal to 0.
- Axiom 2: Total probability is equal to 1.
- Axiom 3: Probability of mutually exclusive events is the sum of the probabilities.

Probability = Area

Prob(A) = Area A

Total Area = 1

Prob(A, B) = Area A + Area B

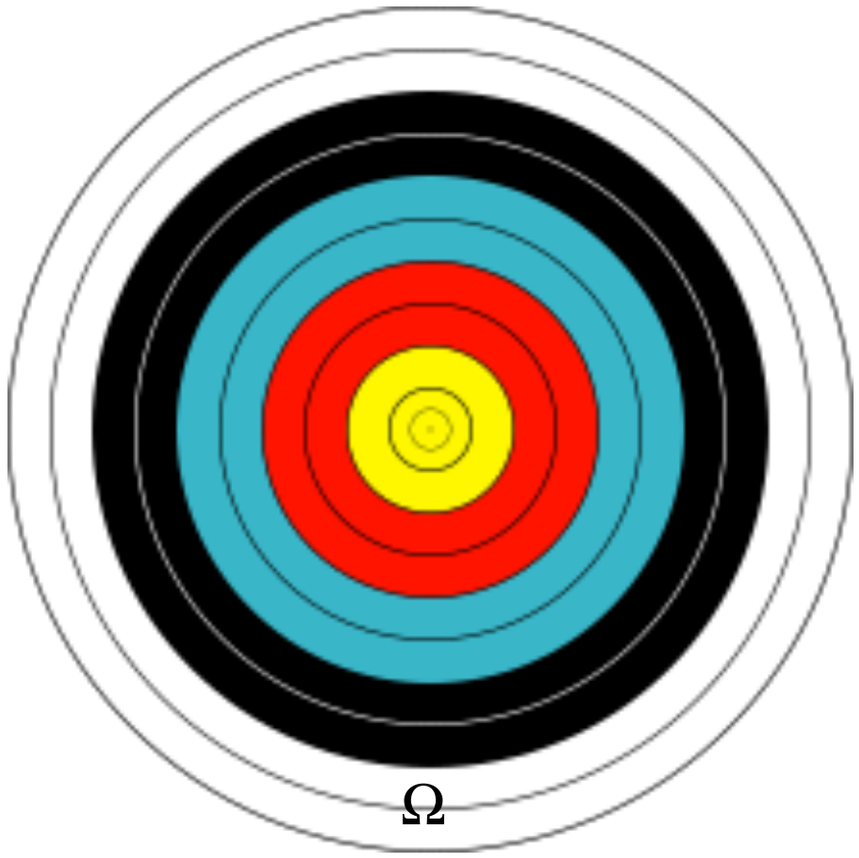


$$0 \leq P(A) \leq 1$$

$$P(\Omega) \equiv 1$$

$$P(A, B) = P(A) + P(B)$$

Probability = Frequency



Rolling Dice

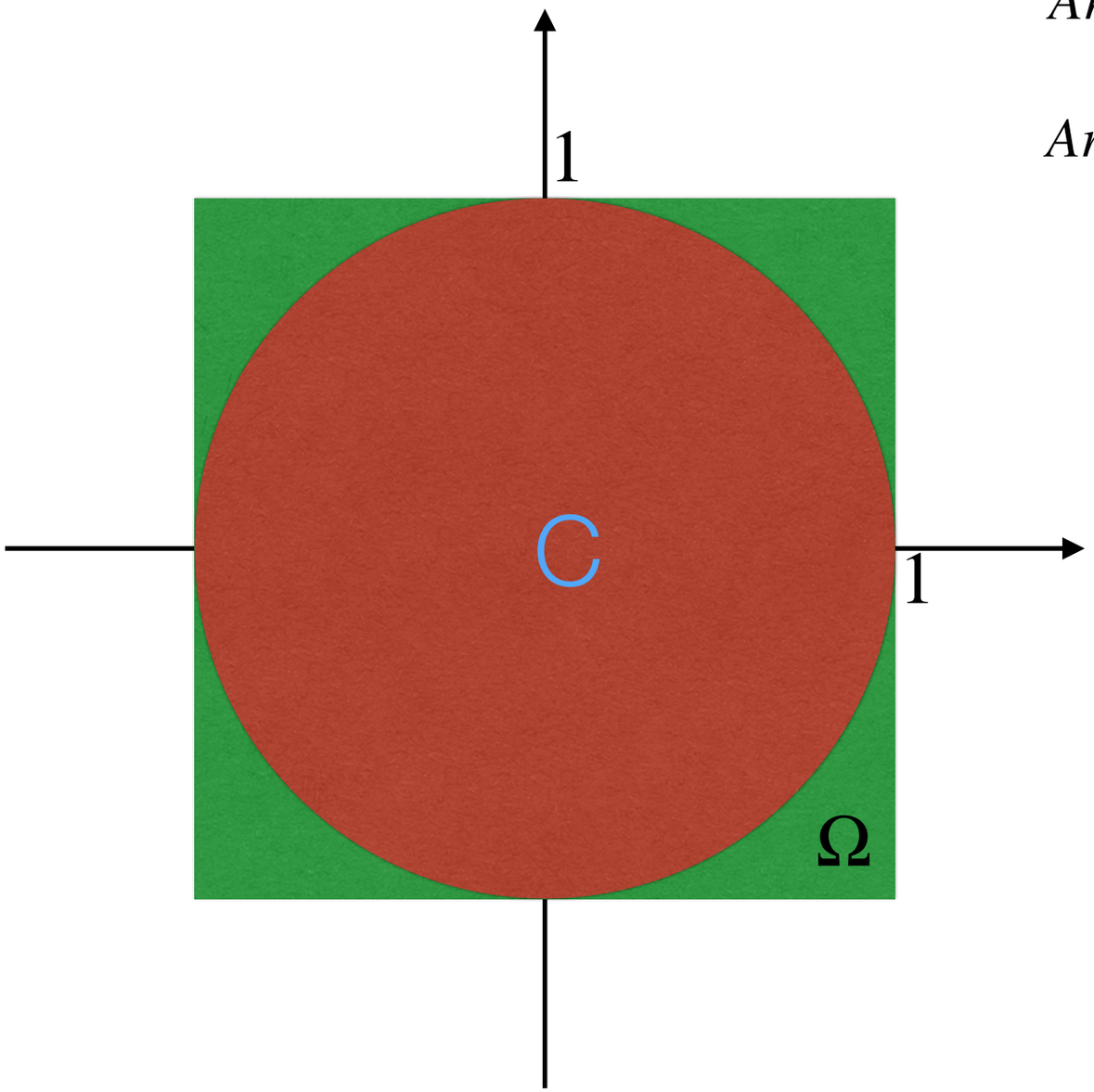
- 6 sided die

- 6 possibilities

- Each side equally likely: $P(x) = \frac{1}{6},\ x \in [1,6]$

1	2	3
4	5	6 Ω

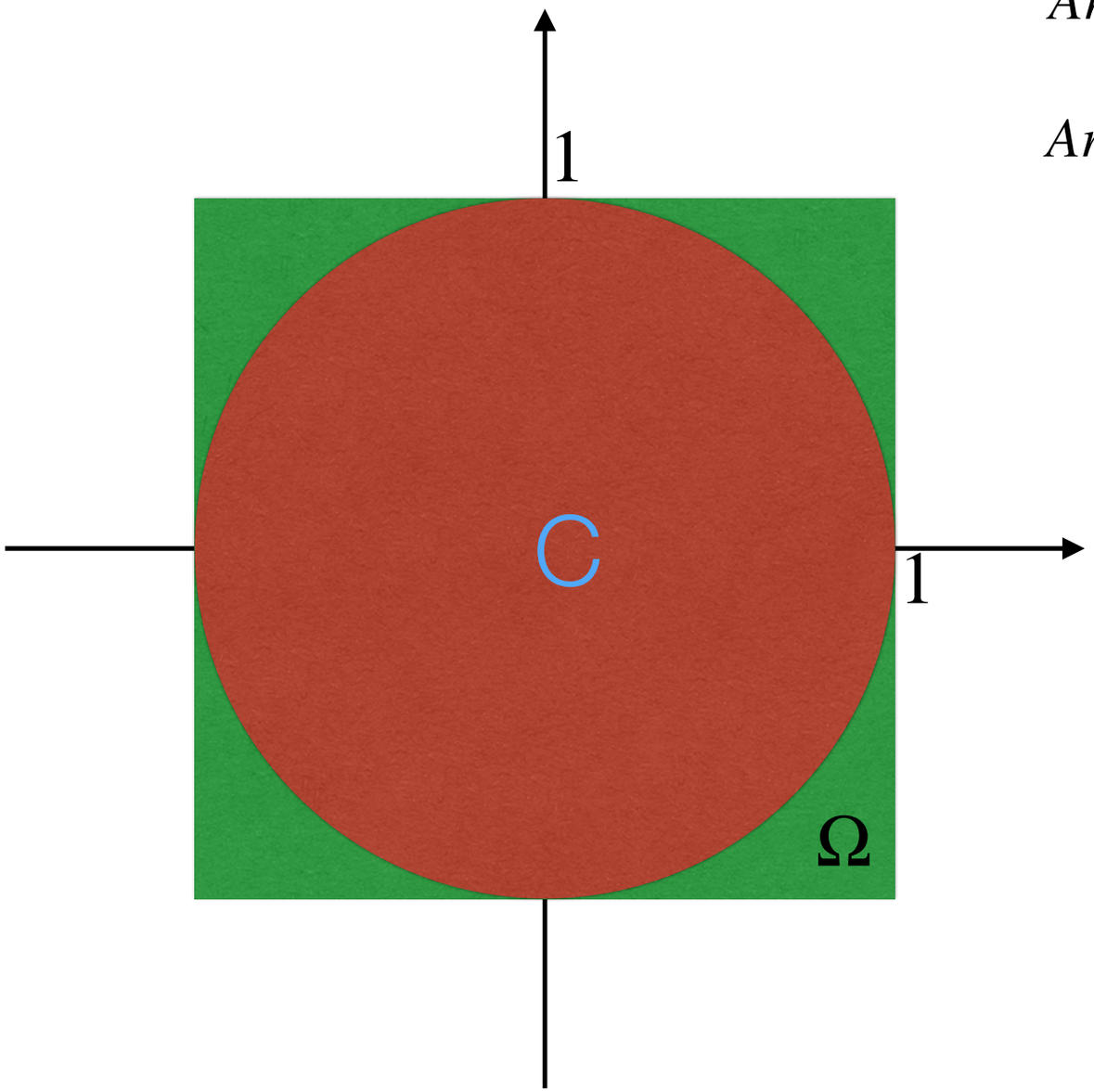
Continuous Version



$$Area\left(C\right) = \pi$$

$$Area\left(\Omega\right) = 4$$

Continuous Version

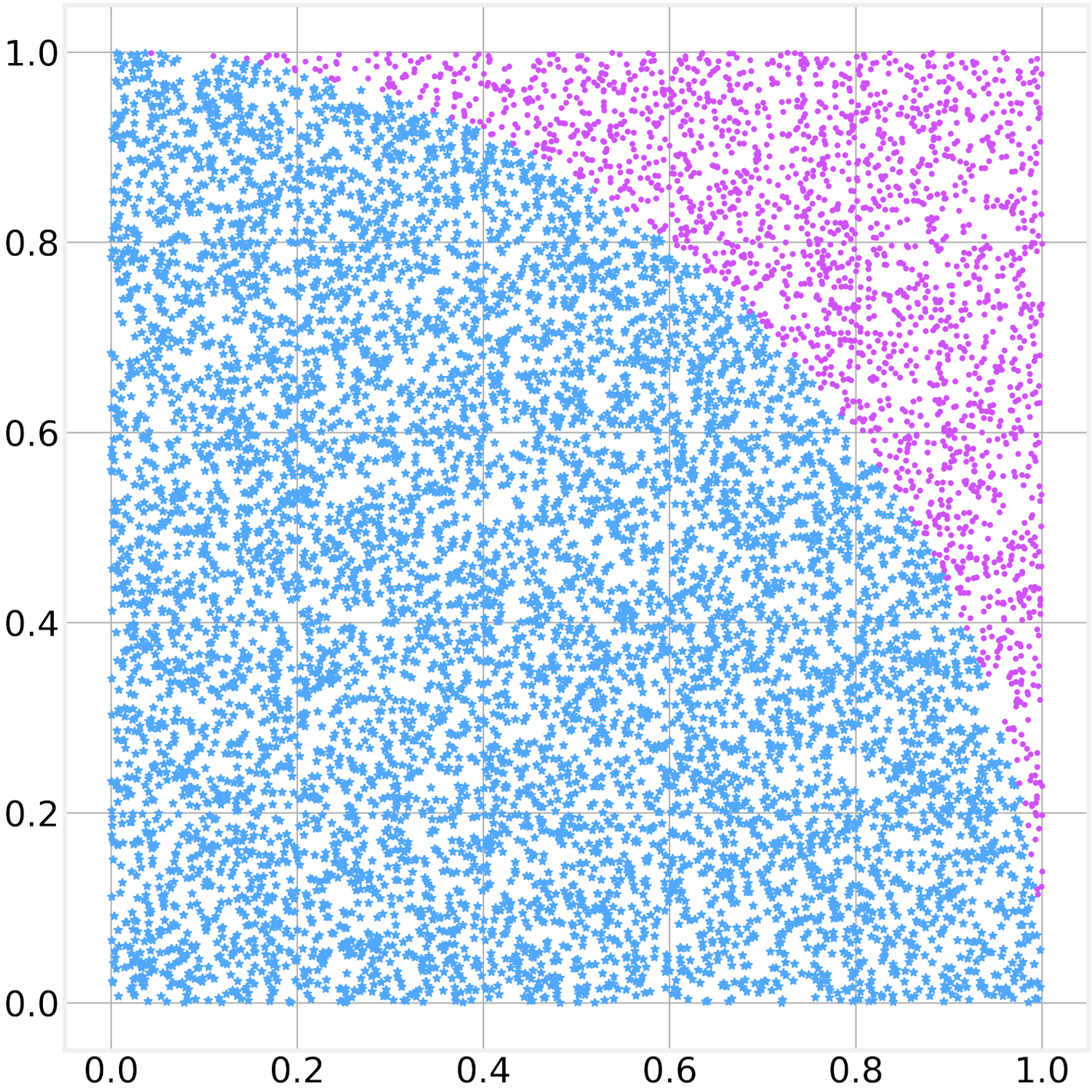


$$Area\left(C\right) = \pi$$

$$Area\left(\Omega\right) = 4$$

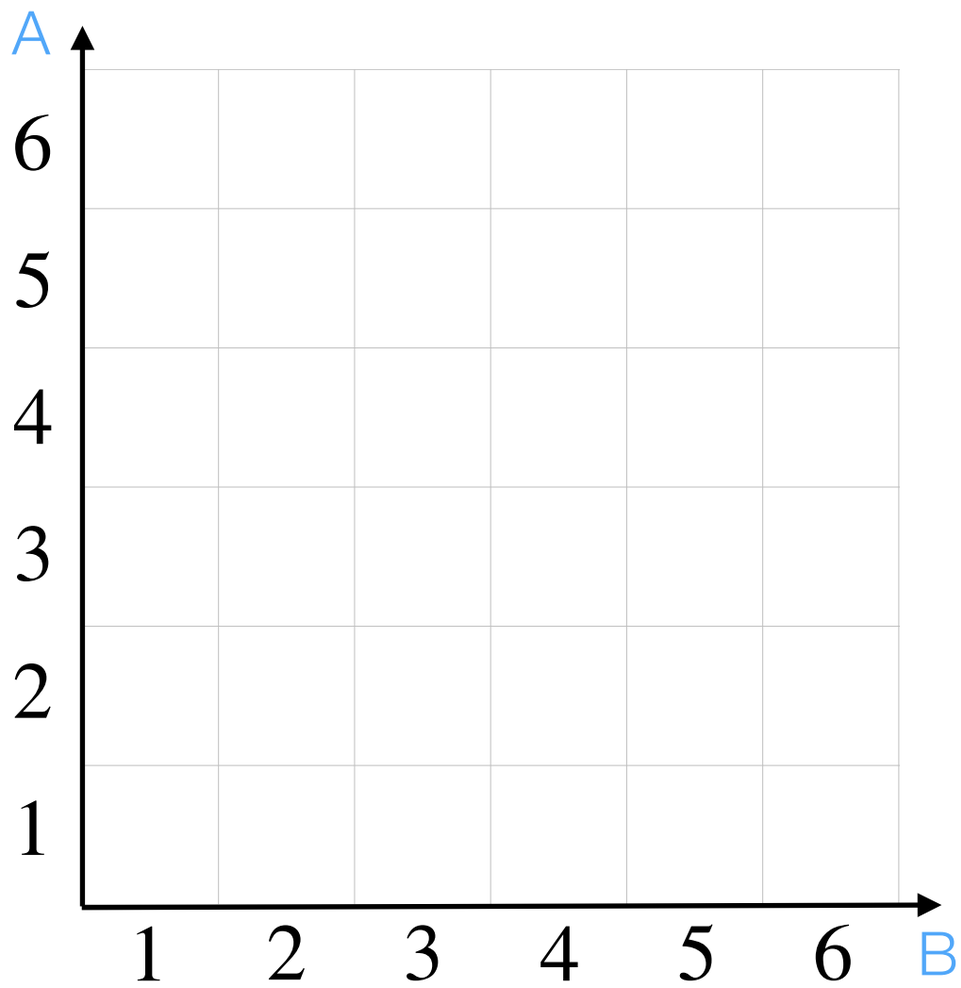
$$P\left(C\right) = \frac{Area\left(C\right)}{Area\left(\Omega\right)} = \frac{\pi}{4}$$

Continuous Version



Sequences of Events

- A happens, then B happens
  - I roll a 2 and a 6
  - I roll an odd number followed by an even number

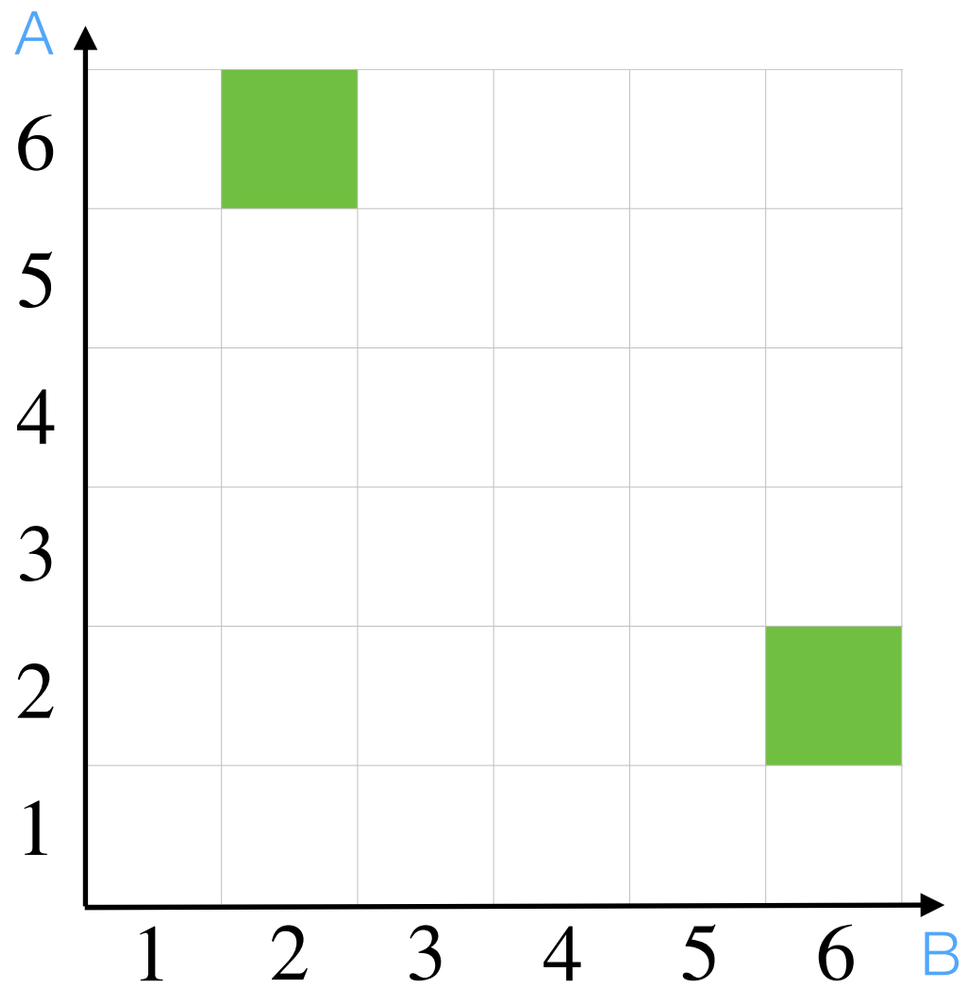


Multiply

Probabilities

Sequences of Events

- A happens, then B happens
  - I roll a 2 and a 6
  - I roll an odd number followed by an even number



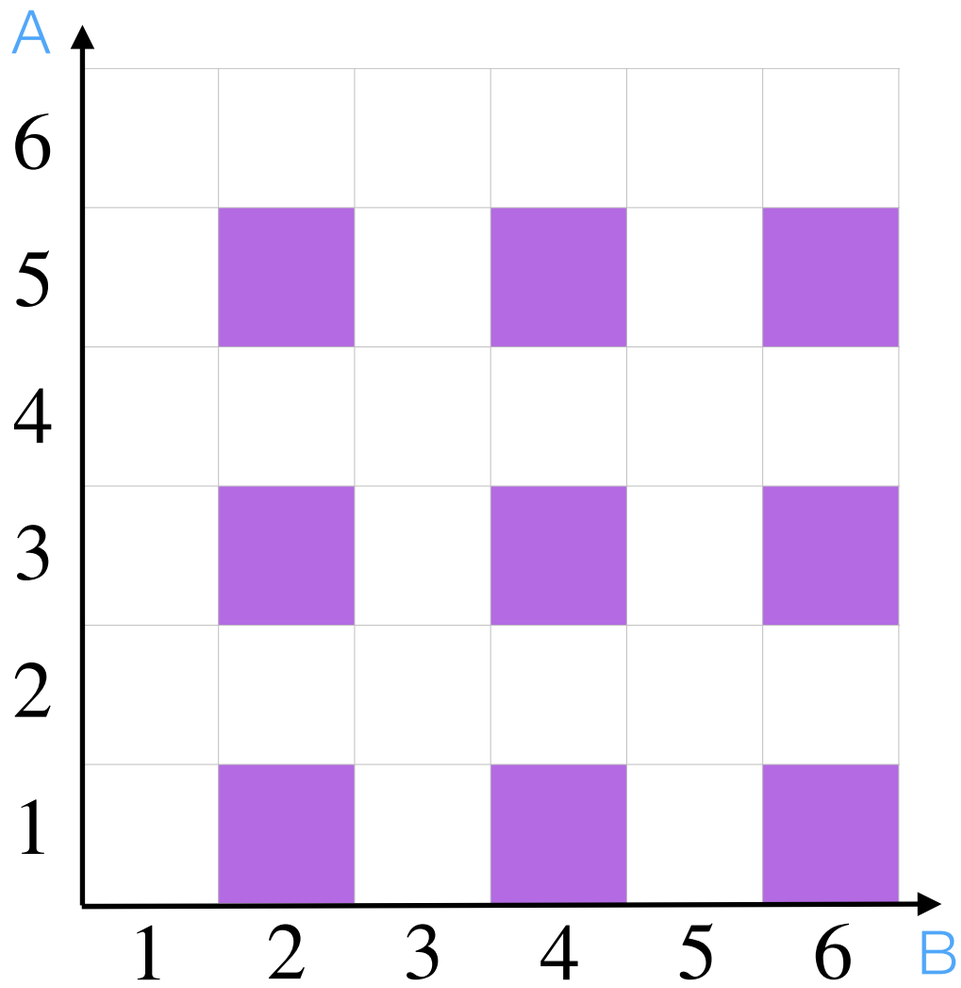
$$P\left(2,6\right) = \frac{2}{36}$$

Multiply

Probabilities

Sequences of Events

- A happens, then B happens
  - I roll a 2 and a 6
  - I roll an odd number followed by an even number



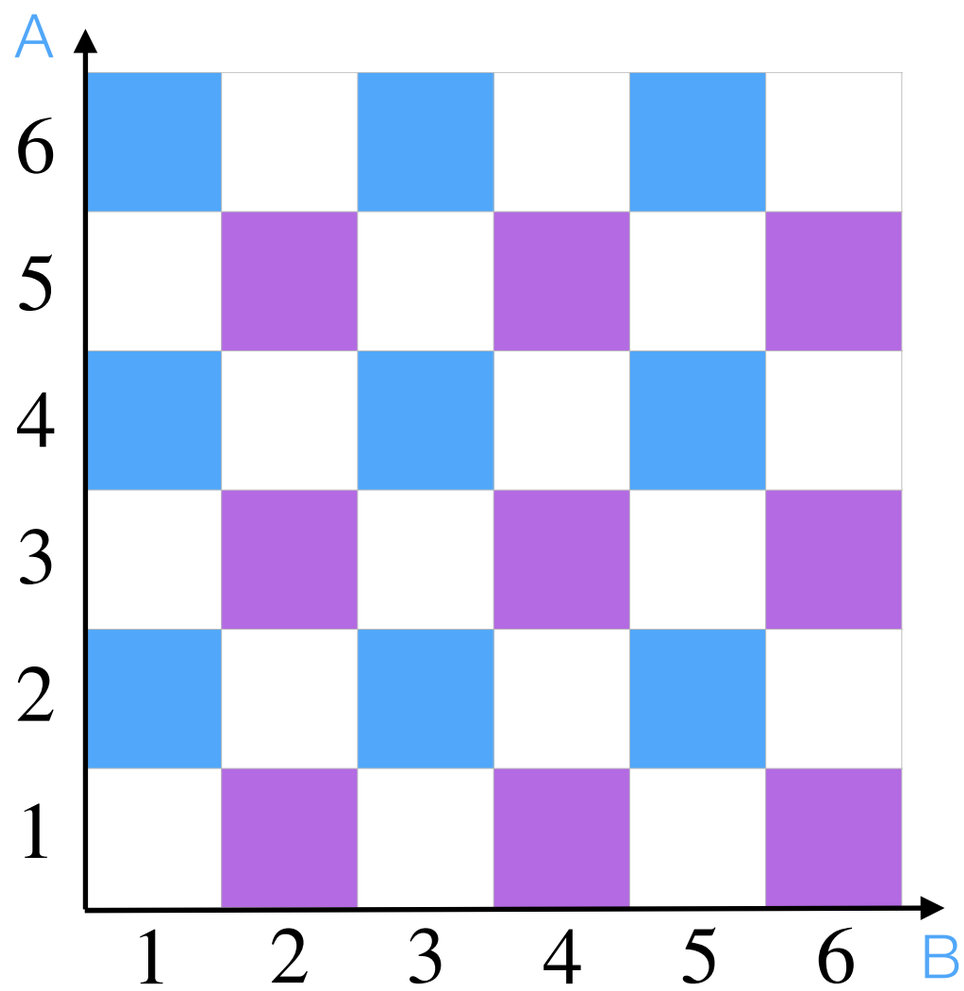
$$P\left(odd, even\right) = \frac{9}{36}$$

Multiply

Probabilities

Sequences of Events

- A happens, then B happens
  - I roll a 2 and a 6
  - I roll an odd number and an even number (two possibilities)



$$P\left(even, odd\right) = \frac{9}{36}$$

$$P\left(odd, even\right) = \frac{9}{36}$$

Multiply

Probabilities

General Procedure

- Enumerate all possible outcomes

- Calculate the "Area" of each outcome

- The Probability of a given outcome is the fraction of the total Area it occupies

- Histogram: Observed Frequency of each outcome $N(X)$

- Probability Distribution: Probability associated with every possible outcome $P(X)$

- Cumulative Probability Distribution: Probability associated with outcomes smaller or equal to each outcome $P(X \leq x)$

Combinatorics

- To enumerate all possible outcomes we often have to make use of ideas from Combinatorics:

- Permutations: The total number of possible sequences of n different elements:

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 1$$

- Combinations: The total number of ways of grouping N elements into two groups of size k and $N-k$:

$$C_k^N = \frac{N!}{k!(N-k)!}$$

- With just these two ideas we can easily calculate the number of possible outcomes in many situations:

  - Number of possible playing card shuffles:

$$52! = 80658175170943878571660636856403766975289505440883277824000000000000$$

  - Number of ways of getting 3 heads and 2 tails when flipping 5 coins:

$$C_3^5 = \frac{5!}{3!2!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \; 2 \cdot 1} = \frac{5 \cdot 4}{2 \cdot 1} = 10$$

Combinatorics - Coin flips

- Can we calculate the Probability distribution of the outcome of flipping 5 coins that come out heads with probability p?

- The probability of getting N_h heads and $N - N_h$ tails is:

$$p\left(N, N_h\right) = p_h^N \left(1-p\right)^{N-N_h}$$

- As we need to get N heads with probability p each time and $N - N_h$ tails with probability $1 - p$ each time

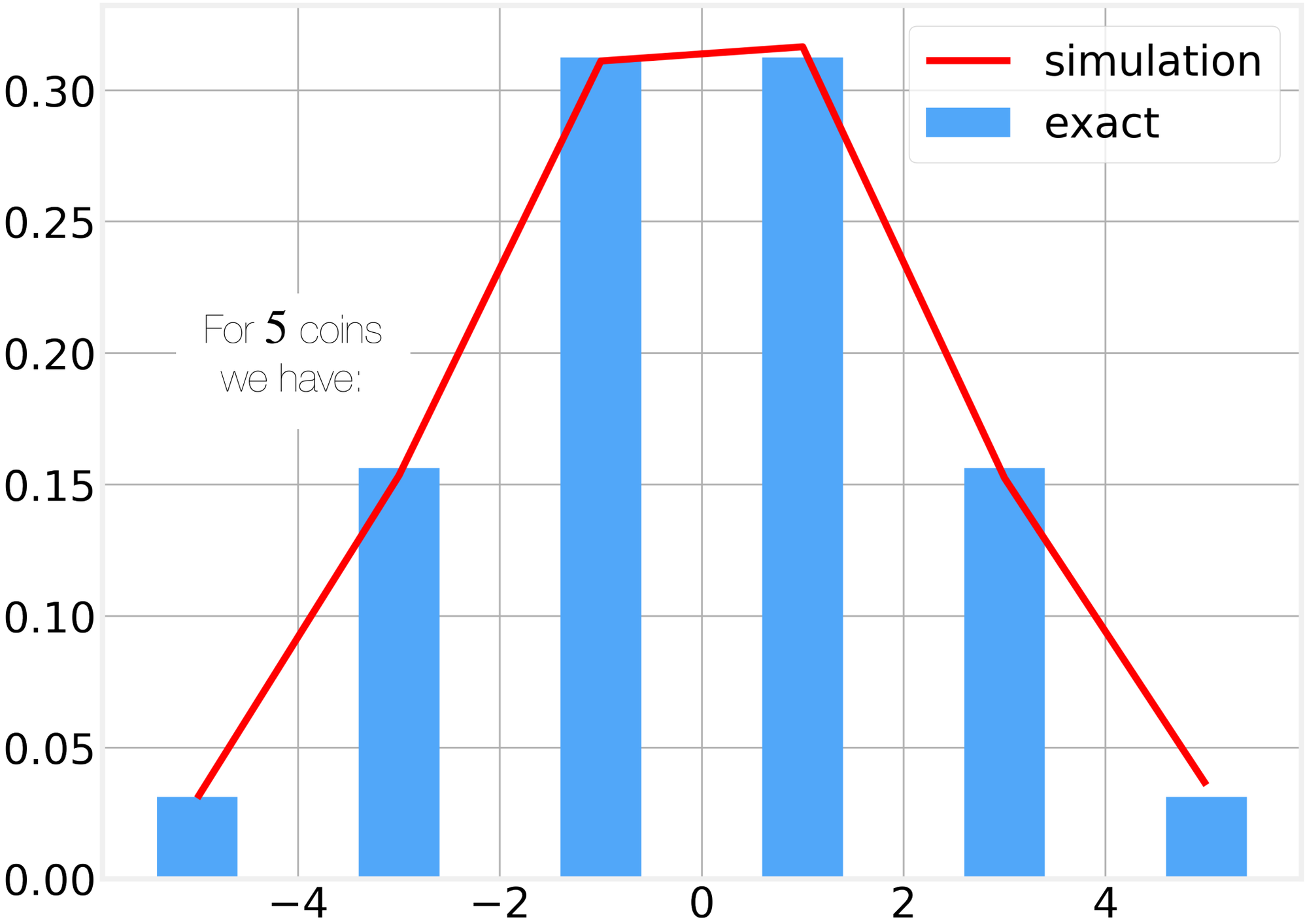
- From combinatorics we also know that the number of ways to get N_h heads out of total N flips is:

$$C_{N_h}^N = \frac{N!}{N_h!\left(N-N_h\right)!}$$

- Therefore

$$P\left(N, N_h\right) = \frac{N!}{N_h!\left(N-N_h\right)!} p_h^N \left(1-p\right)^{N-N_h}$$

Combinatorics - Coin Flips



Binomial Distribution

- The probability of getting k successes with n trials of probability p (k heads in n coin flips):

$$P_B\left(k,n,p\right) = \frac{n!}{k!\left(n-k\right)!}p^k\left(1-p\right)^{n-k}$$

- The mean value is:

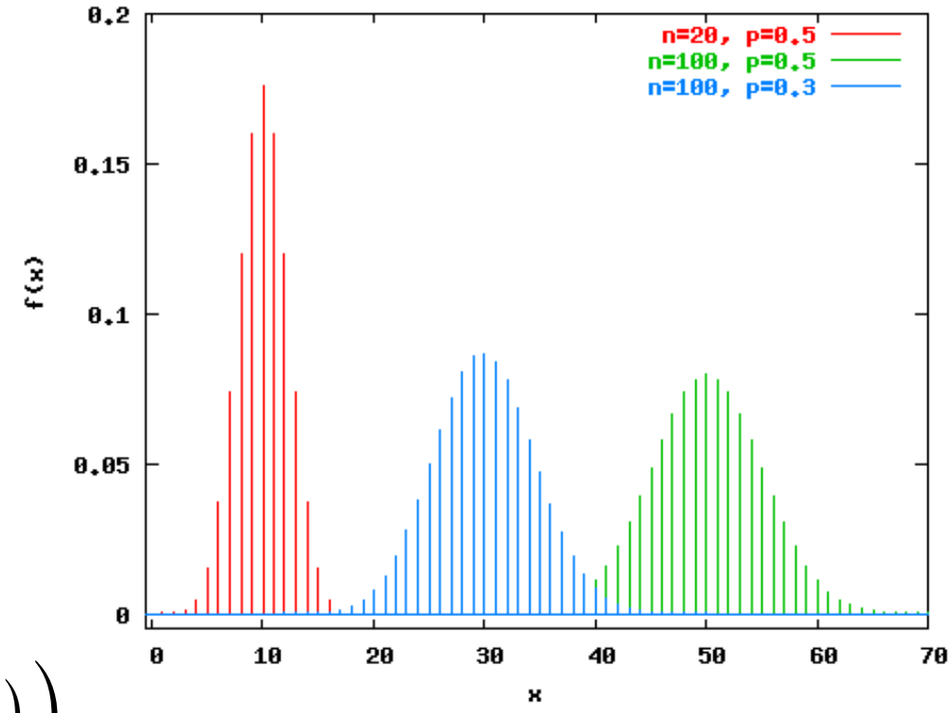
$$\mu = np$$

- and the variance:

$$\sigma^2 = np\left(1-p\right)$$

- and for sufficiently large n:

$$P_B\left(k,n,p\right) \sim P_N\left(np, np\left(1-p\right)\right)$$

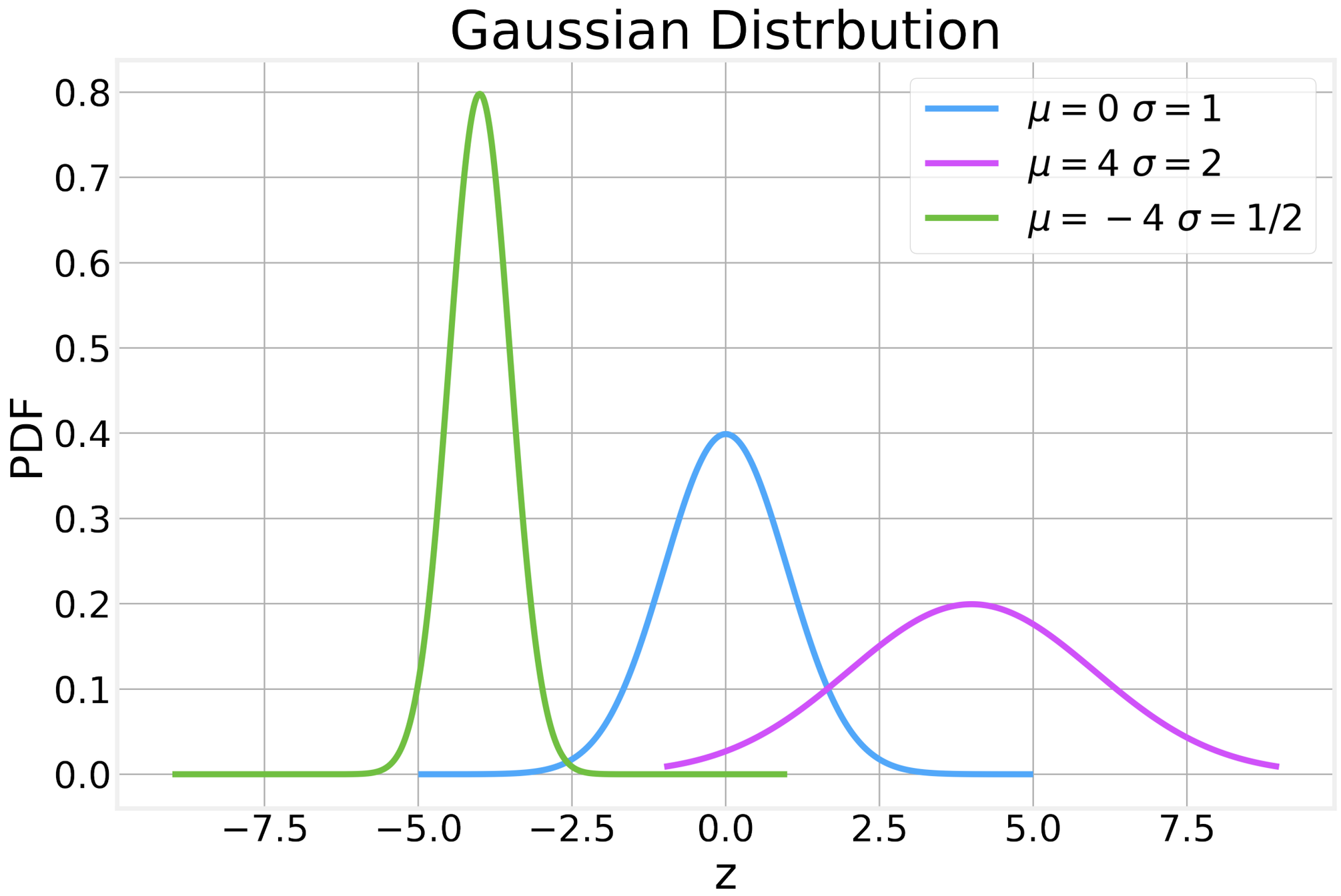


SciPy

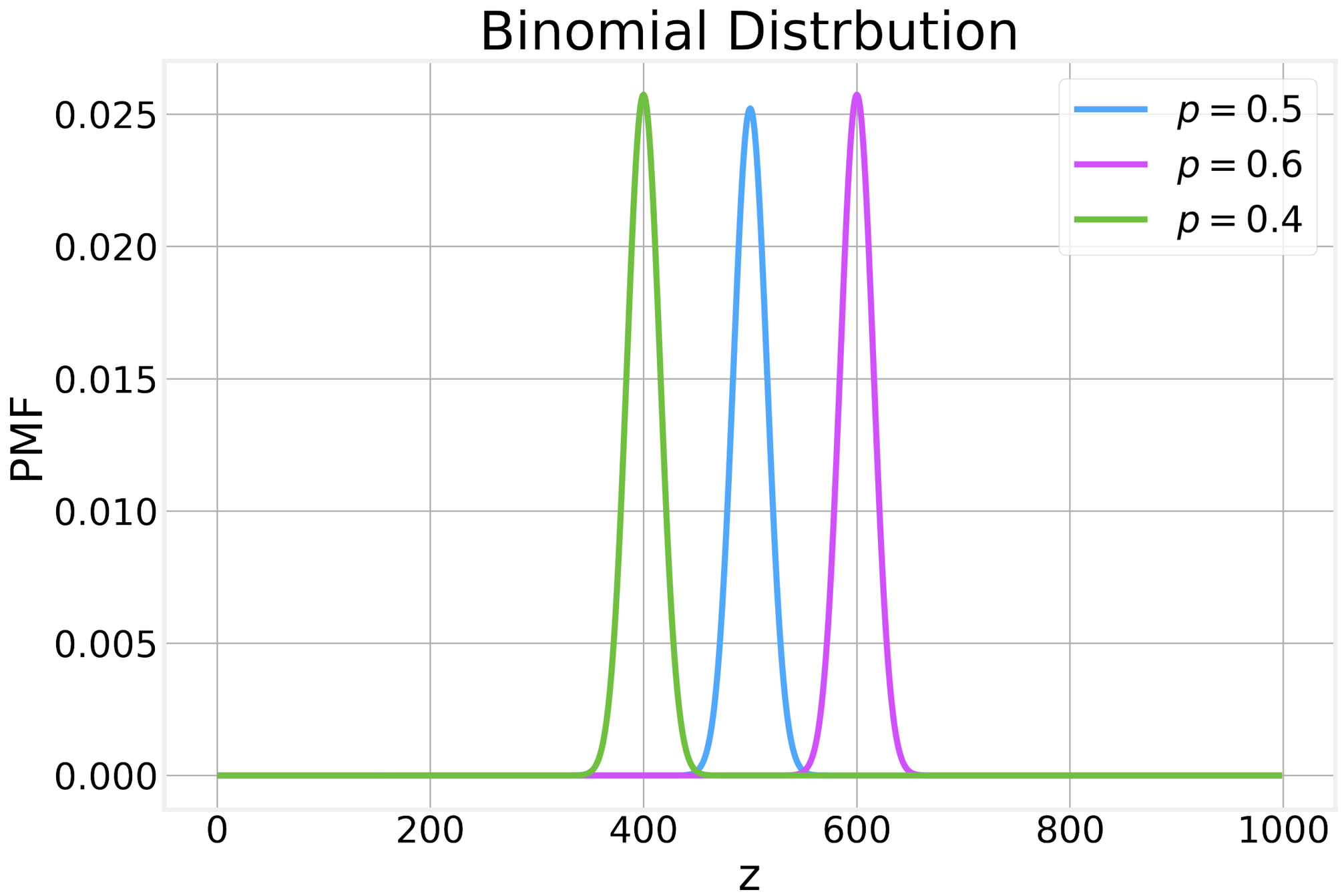
docs.scipy.org/doc/scipy/reference/stats.html

- The scipy.stats module provides a unified interface to many standard distributions
- In particular:
  - scipy.stats.norm() - Normal/Gaussian Distribution
  - scipy.stats.binom() - Binomial Distribution
  - scipy.stats.possion() - Poisson Distribution
  - among many others
- Once instantiated, each distribution provides some commonly used functions:
  - rvs() - Random variates.
  - pmf() - Probability mass function.
  - cdf() - Cumulative distribution function.
  - fit() - Parameter estimates for generic data.
  - median()/mean()/var()/std() -Median/Mean/Variance/Standard Deviation

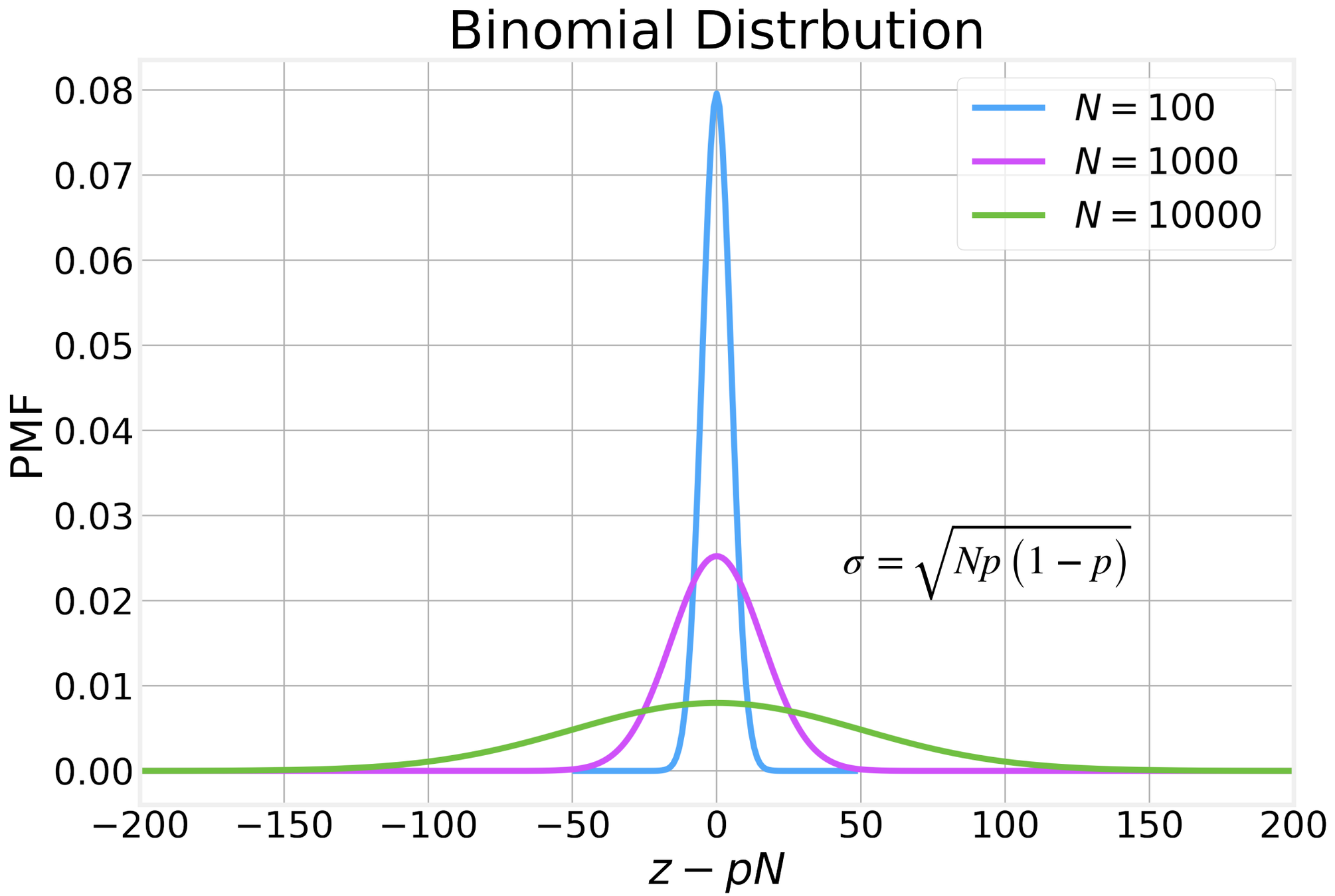
Normal Distribution



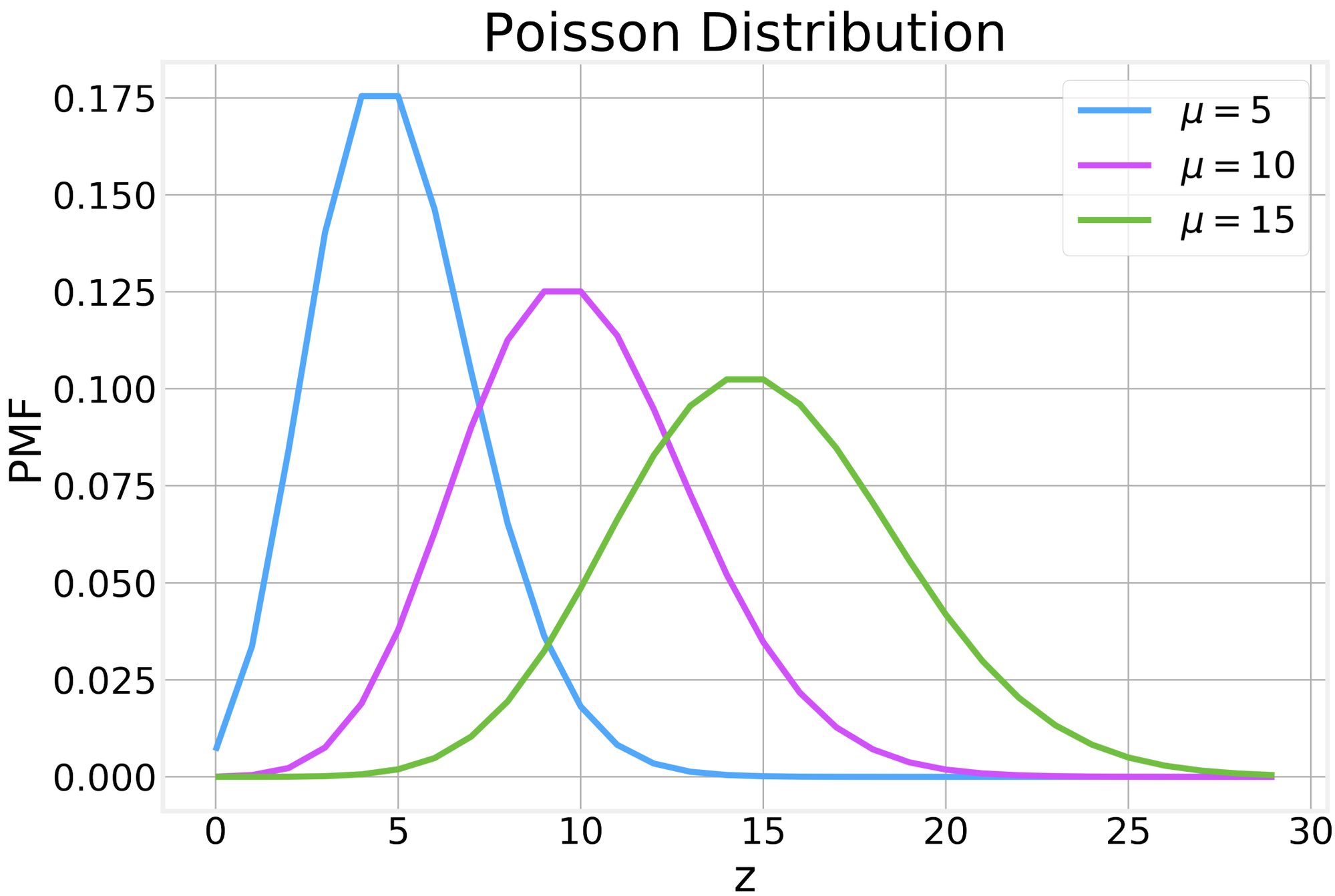
Binomial Distribution



Binomial Distribution



Poisson Distribution



Averaging and Expectations

- A n-sided die has a uniform probability $\frac{1}{n}$ of landing on any of it's sides.

- Let's call the value seen after roll i of the die, x_i

- After 10 rolls of, say, a 6-sided die, we might have:

$$x_i = [4,6,4,3,5,1,1,5,2,4]$$

- The behavior of this variable is stochastic, but what about the behavior of functions of this random variable?

- For example, the average:

$$\mu_N \equiv \langle \tilde{x} \rangle_N = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- In this specific example, the average is 3.5 as expected, but if, say, rolls 6 and 7 had been 6s instead of 1s, the average value would be 4.5.

Averaging and Expectations

- If we use 10000 rolls of the dice, we find that the average is:

$$\langle \tilde{x} \rangle_{10000} = 3.4888$$

- But if we repeat the same "experiment" 10 times, we will find 10 different values:

$\langle \tilde{x} \rangle_{10000}$
3.4888
3.5111
3.4686
3.4893
3.5233
3.4941
3.4975
3.5276
3.4948
3.4775

- So what is the correct value?

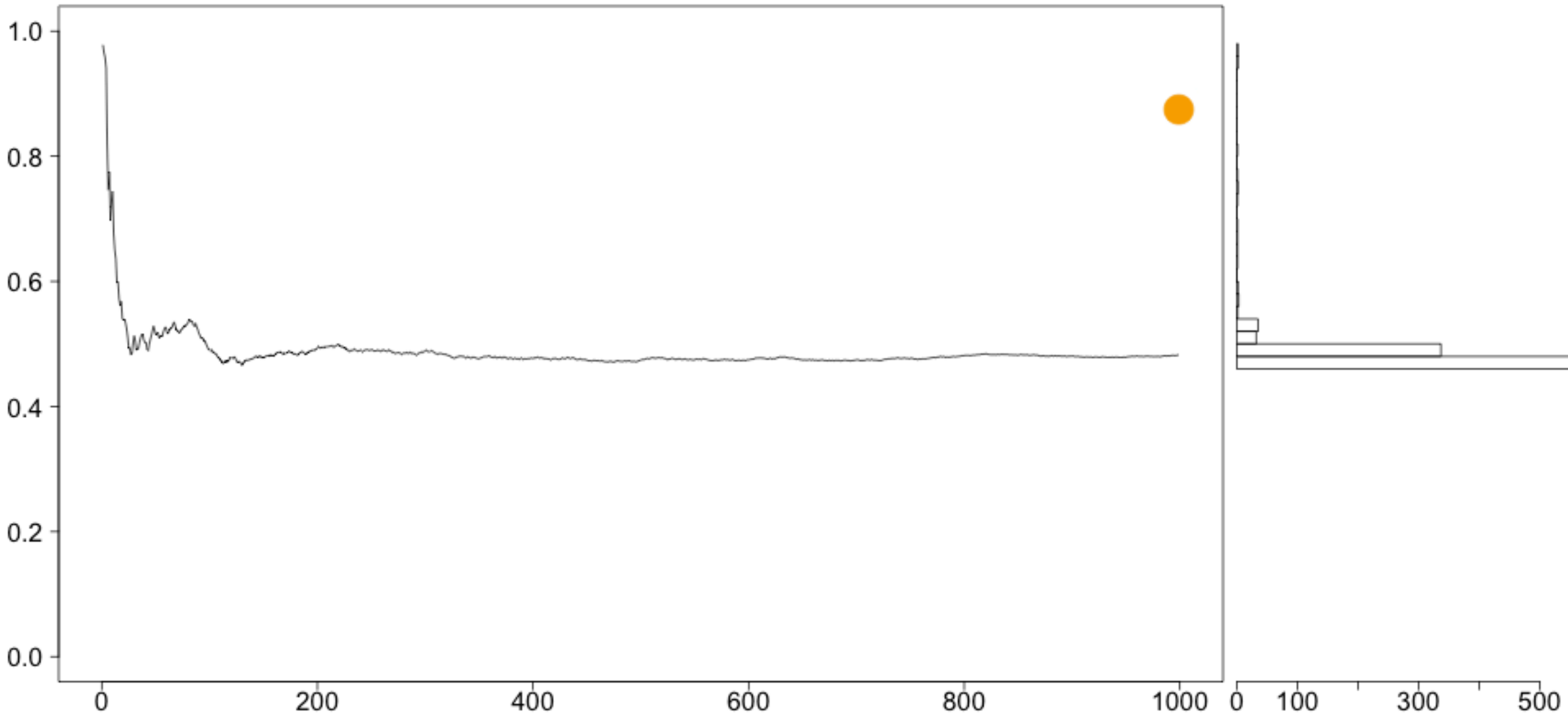
- For any given set of dice rolls, we are only estimating the true value.

- Estimated values are usually denoted with a ~ or ^ over the variable

- In general, the higher the number of rolls we consider in a given realization, the better our estimate of the average.

- The true, or expected value, is the one obtained after an infinite number of realizations. This number can be estimated with just a bit of algebra

Law of Large Numbers



http://youtu.be/08ZjT7GhENI

Averaging and Expectations

- As we saw:

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- This can be rewritten as:

$$\mu_N = \sum_{\alpha=1}^{n} \frac{N_\alpha}{N} x_\alpha$$

- Where α denotes all n possible values of the variable x_i or, in other words, the values on the sides of the die.

- If we notice that $\frac{N_\alpha}{N}$ is just our estimate of the value of the probability of observing α we can write:

$$\mu = \sum_{\alpha=1}^{n} p_\alpha x_\alpha$$

- where p_α is the true probability and $\mu \equiv \langle x \rangle$ is the true average value of the average, or the expected value of x

Central Limit Theorem

- As $N \to \infty$ the random variables:

$$\sqrt{N}\left(\mu_N - \mu\right)$$

- with:

$$\mu_N = \frac{1}{N}\sum_i x_i$$

Central Limit Theorem

- As $N \to \infty$ the random variables:

$$\sqrt{N}\left(\mu_N - \mu\right)$$

- with:

$$\mu_N = \frac{1}{N}\sum_i x_i$$

- converge to a normal distribution:

$$\mathcal{N}\left(0,\sigma^2\right)$$

Central Limit Theorem

- As $N \to \infty$ the random variables:

$$\sqrt{N}\left(\mu_N - \mu\right)$$

- with:

$$\mu_N = \frac{1}{N}\sum_i x_i$$

- converge to a normal distribution:

$$\mathcal{N}\left(0,\sigma^2\right)$$

- after some manipulations, we find:

$$\mu_N \sim \mu + \frac{\mathcal{N}\left(0,\sigma^2\right)}{\sqrt{N}}$$

The estimation of the mean converges to the true mean with the square root of the number of samples

Central Limit Theorem

- As $N \to \infty$ the random variables:

$$\sqrt{N}\left(\mu_N - \mu\right)$$

- with:

$$\mu_N = \frac{1}{N}\sum_i x_i$$

- converge to a normal distribution:

$$\mathcal{N}\left(0,\sigma^2\right)$$

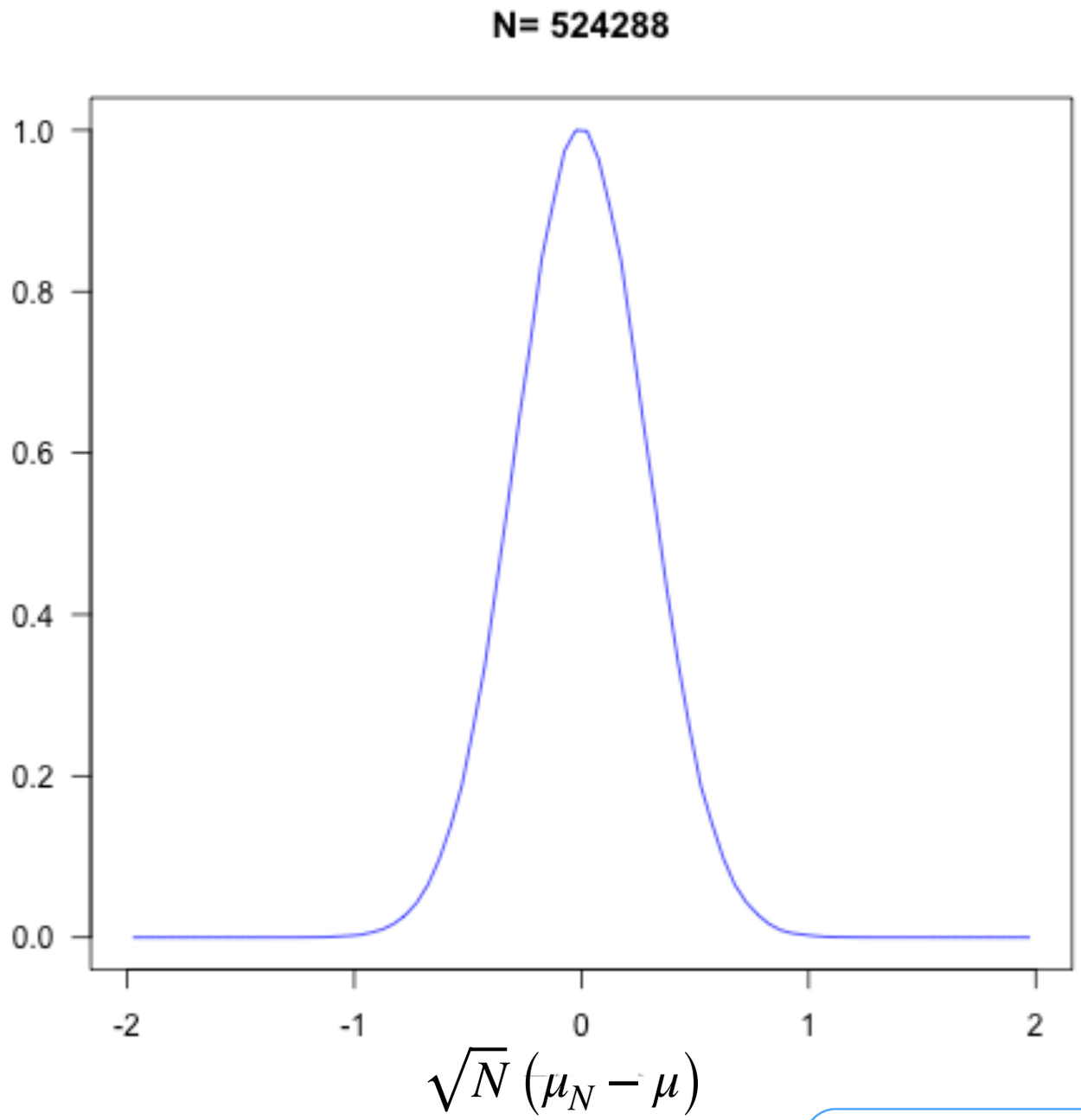
Error Estimate

- after some manipulations, we find:

$$\mu_N \sim \mu + \frac{\mathcal{N}\left(0,\sigma^2\right)}{\sqrt{N}} \to SE = \frac{\sigma}{\sqrt{N}} \qquad \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \mu\right)^2$$

The estimation of the mean converges to the true mean with the square root of the number of samples

Central Limit Theorem



http://youtu.be/08ZjT7GhENI

Gaussian Distribution

- The probability of observing value x from a normal distribution centered at μ and with variance σ^2:

$$P_N\left(x,\mu,\sigma\right)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

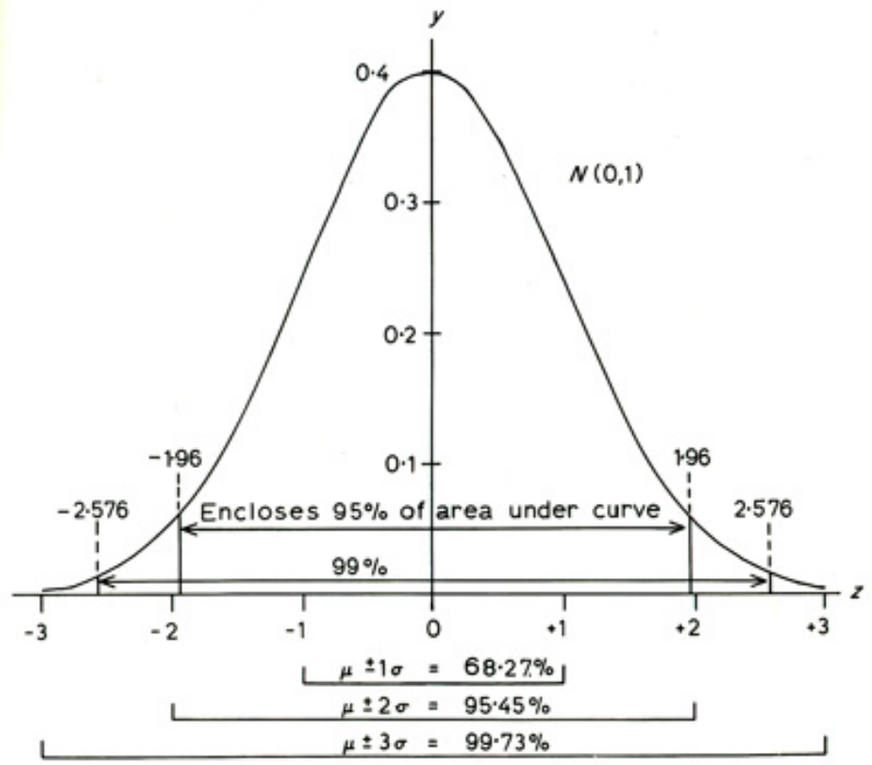
- The mean value is:

$$\mu=\frac{1}{N}\sum_i x_i$$

- and the variance:

$$\sigma^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\mu\right)^2$$

- and for sufficiently large n:



Experimental Measurements

- Experimental errors commonly assumed gaussian distributed

- Many experimental measurements are actually averages:

    - Instruments have a finite response time and the quantity of interest varies quickly over time

    - Stochastic Environmental factors

    - Etc

MLE - Fitting a theoretical function to experimental data

- In an experimental measurement, we expect (CLT) the experimental values to be normally distributed around the theoretical value with a certain variance. Mathematically, this means:

$$P\left(y - f\left(x\right)\right) \approx \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left(y - f\left(x\right)\right)^2}{2\sigma^2}\right]$$

- where $\boldsymbol{y}$ are the experimental values and $\boldsymbol{f\left(x\right)}$ the theoretical ones. The likelihood is then:

$$\mathscr{L} = -\frac{N}{2}\log\left[2\pi\sigma^2\right] - \sum_i \left[\frac{\left(y - f\left(x_i\right)\right)^2}{2\sigma^2}\right]$$

- Where we see that to maximize the likelihood we must minimize the sum of squares

Least Squares Fitting

MLE - Linear Regression

- Let's say we want to fit a straight line to a set of points:

$$y = w \cdot x + b$$

- The Likelihood function then becomes:

$$\mathscr{L} = -\frac{N}{2}\log\left[2\pi\sigma^2\right] - \sum_i \left[\frac{\left(y - w \cdot x_i - b\right)^2}{2\sigma^2}\right]$$

- With partial derivatives:

$$\frac{\partial \mathscr{L}}{\partial w} = \sum_i \left[2x_i\left(y_i - w \cdot x_i - b\right)\right]$$

$$\frac{\partial \mathscr{L}}{\partial b} = \sum_i \left[\left(y_i - w \cdot x_i - b\right)\right]$$

- Setting to zero and solving for $\hat{w}$ and $\hat{b}$:

$$\hat{w} = \frac{\sum_i \left(x_i - \langle x \rangle\right)\left(y_i - \langle y \rangle\right)}{\sum_i \left(x_i - \langle x \rangle\right)^2}$$

$$\hat{b} = \langle y \rangle - \hat{w}\langle x \rangle$$

MLE - Coin Flips

- Biased coin with unknown probability of heads (p)
- In a sequence of N flips, the likelihood of N_h heads and $N_t = N - N_h$ tails is proportional to:
- or simply:

$$\mathscr{L} = \log\left[\frac{N!}{N_h!N_t!}\right] + \log\left[p^{N_h}\left(1-p\right)^{N-N_h}\right]$$

$$\mathscr{L} \propto N_h \log\left[p\right] + \left(N - N_h\right)\log\left[1-p\right]$$

Ignoring the combinatorial factor!

- Taking the derivative:

$$\frac{\partial \mathscr{L}}{\partial p} = \frac{N_h}{p} - \frac{N-N_h}{1-p}$$

- Setting to zero and solving for p:

$$p = \frac{N_h}{N}$$

- which is how we estimated the probability above

Probability distributions

- We already saw three probability distributions:
    - Uniform - Dice
    - Binomial - Dice
    - Gaussian - Errors

- Common programing langues (Python, R, C++, Matlab, etc) have a wide variety of random number generators either built in or as add on packages.

- But how can we generating numbers following a specific (possibly empirical) distribution?

- First we must define the cumulative distribution:

$$P(X \leq x)$$

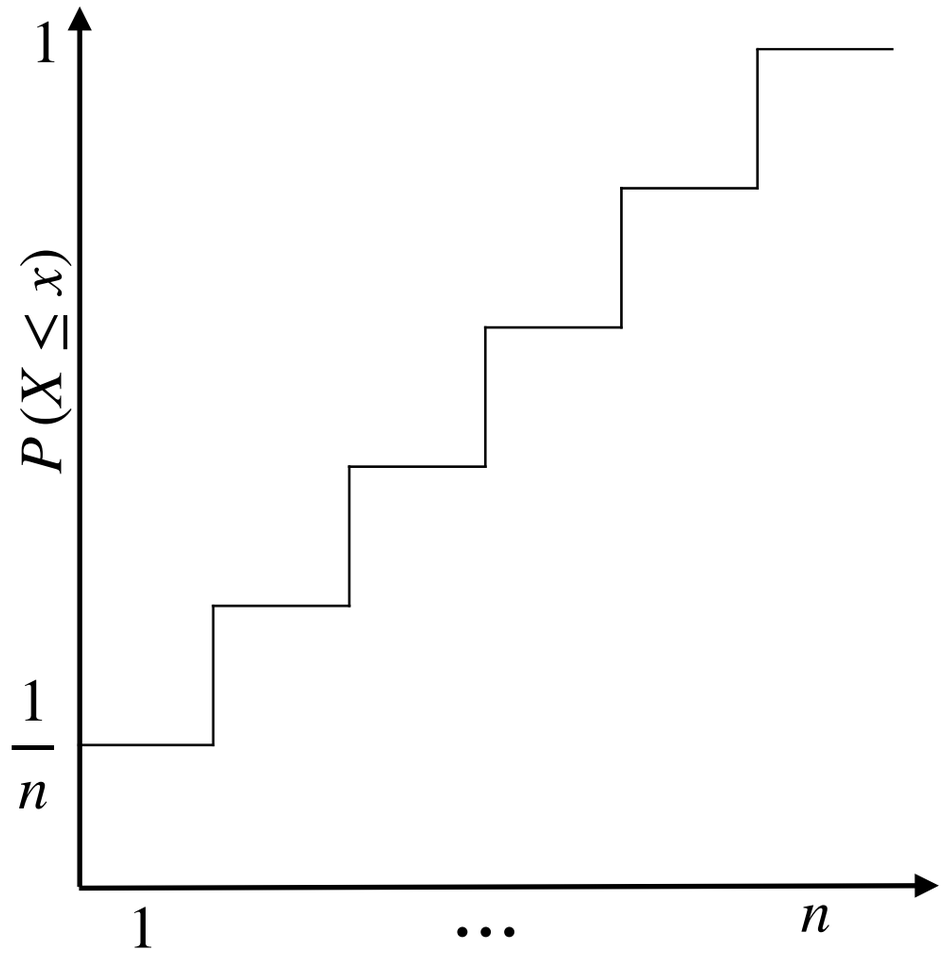
- representing the probability of observing a value smaller than some threshold.

Uniform Distribution

Probability Distribution

Cumulative Distribution

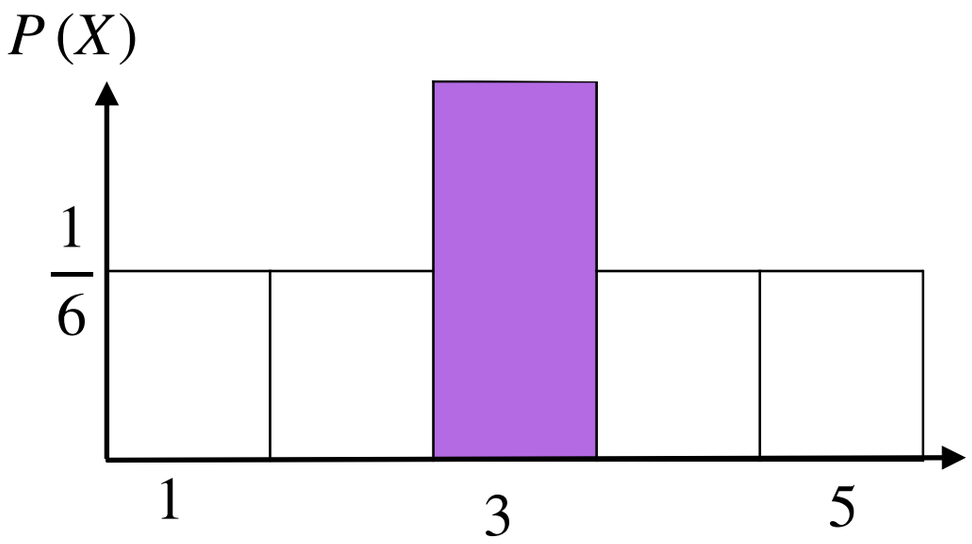
In a standard uniform distribution (a fair die) all outcomes are equally likely



Non-Uniform Distribution

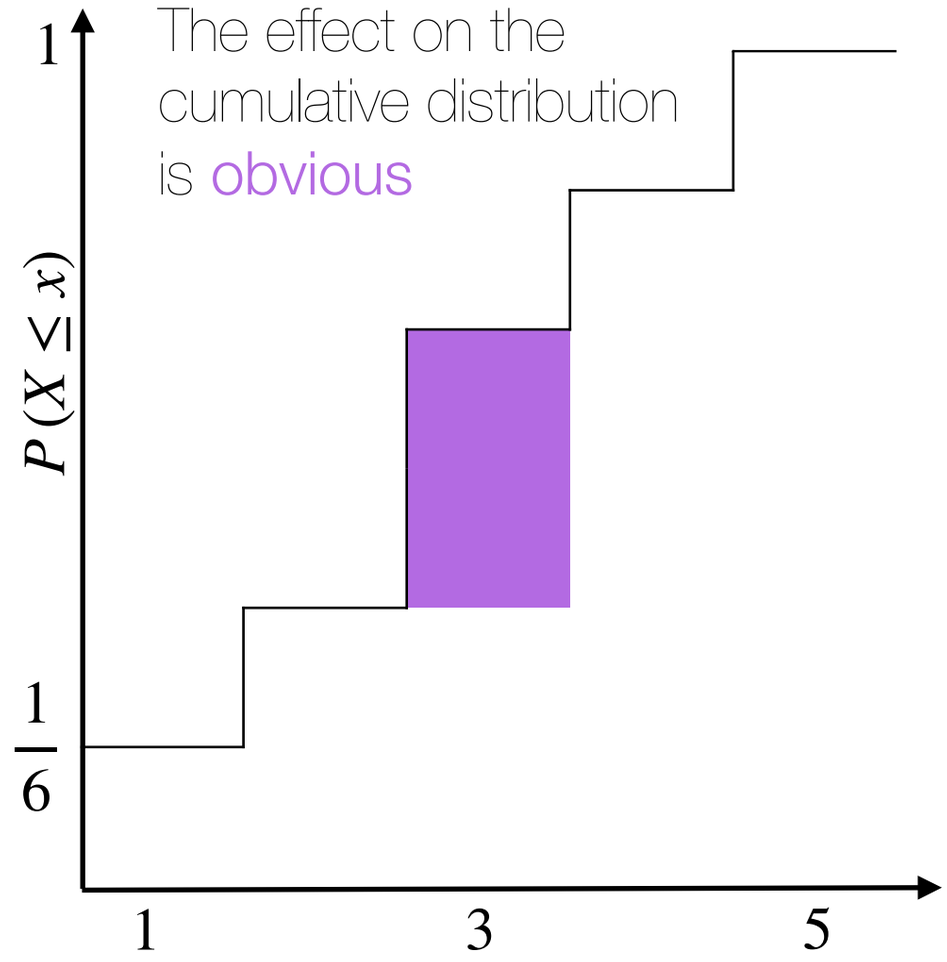
Probability Distribution

Now we have a funny die with two sides labeled as 3



Cumulative Distribution

The effect on the cumulative distribution is obvious



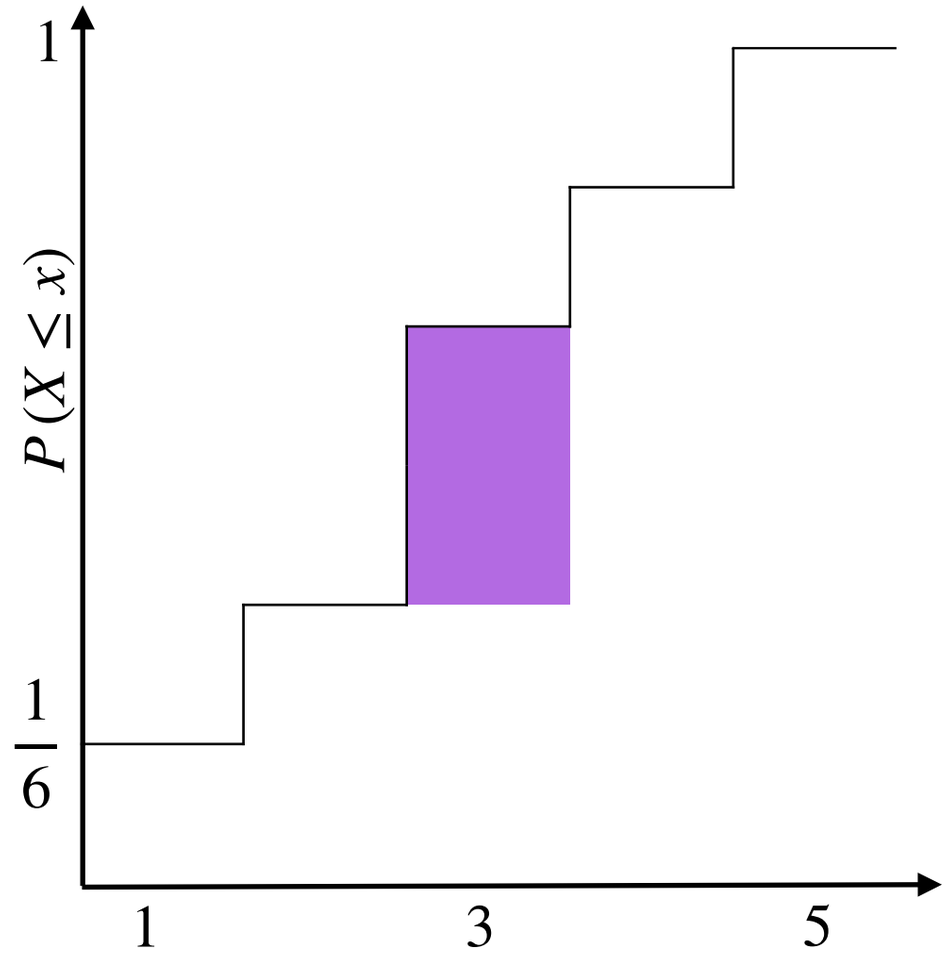
Non-Uniform Distribution

- The step larger step we see in cumulative distribution givens some clues as to how we might generate random numbers following this distribution

- What if our blindfolded monkey is throwing darts along the y axis, where will they hit on the x axis?

- Naturally, the bin with the largest step will receive the largest number of darts!

Cumulative Distribution



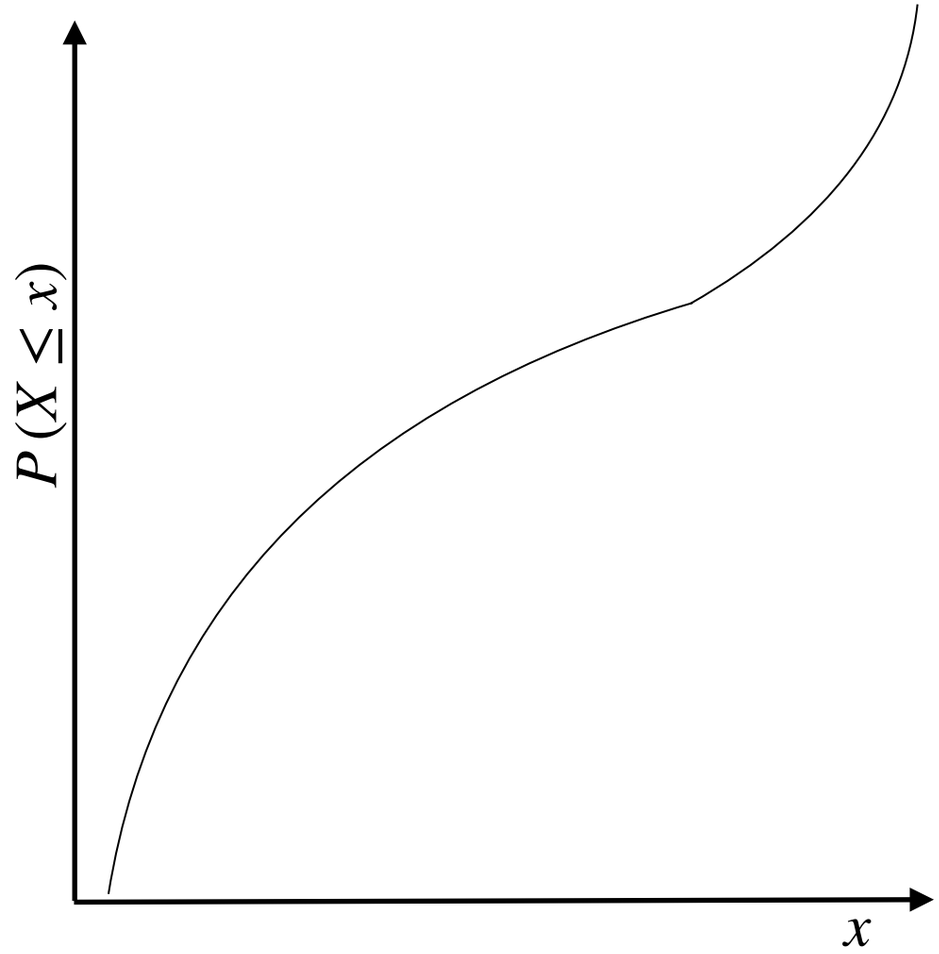
Generating arbitrary distributions

- Mathematically, this procedure is known as function inversion

- We "invert" the function $y = f(x)$ to find what would be the value of x that would produce a specific value of y

- If the analytical expression of the cumulative distribution is known we can invert it analytically, otherwise, we can do it numerically

Cumulative Distribution



Population Sizes

https://en.wikipedia.org/wiki/Lincoln_index

- So far we've known everything about the range of values we might encounter, the number of sides on the die, etc.
- But what if our goal is precisely to estimate the number of possible elements?
- "mark and recapture" is one strategy used in ecology to estimate population sizes:
  - visit a site and mark a number K of individuals (say, rabbits) that are the released
  - After the first visite a fraction $\frac{K}{N}$ of the total population has been marked
  - On a second visit capture a number n of individuals if k of those capture are marked we can assume

$$\frac{k}{n} \approx \frac{K}{N}$$

  - And obtain:

$$\hat{N} = \frac{Kn}{k}$$

- This is known as the Lincoln Index estimator and it assumes that the probability of any animal being chosen is $\frac{1}{N}$ in both visits

Lesson II:
Bayesian Statistics

Kolmogorov's Probability Axioms

https://en.wikipedia.org/wiki/Probability_axioms

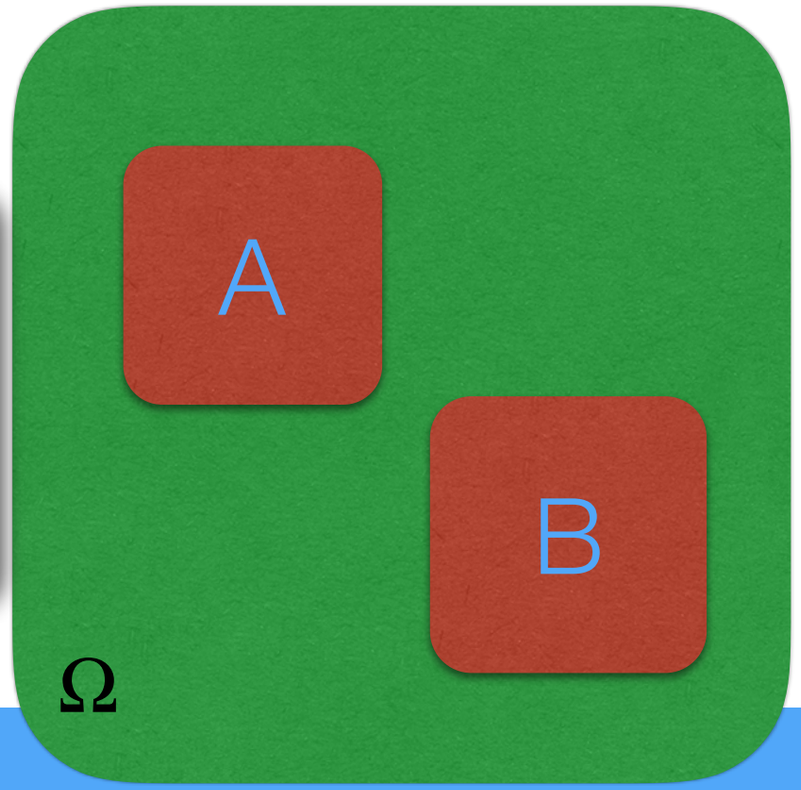
- Axiom 1: Probability is a real number greater or equal to 0.
- Axiom 2: Total probability is equal to 1.
- Axiom 3: Probability of mutually exclusive events is the sum of the probabilities.

Probability = Area

Prob(A) = Area A

Total Area = 1

Prob(A or B) = Area A + Area B



$$0 \leq P(A) \leq 1$$

$$P(\Omega) \equiv 1$$

$$P(A \ or \ B) = P(A) + P(B)$$

Kolmogorov's Probability Axioms

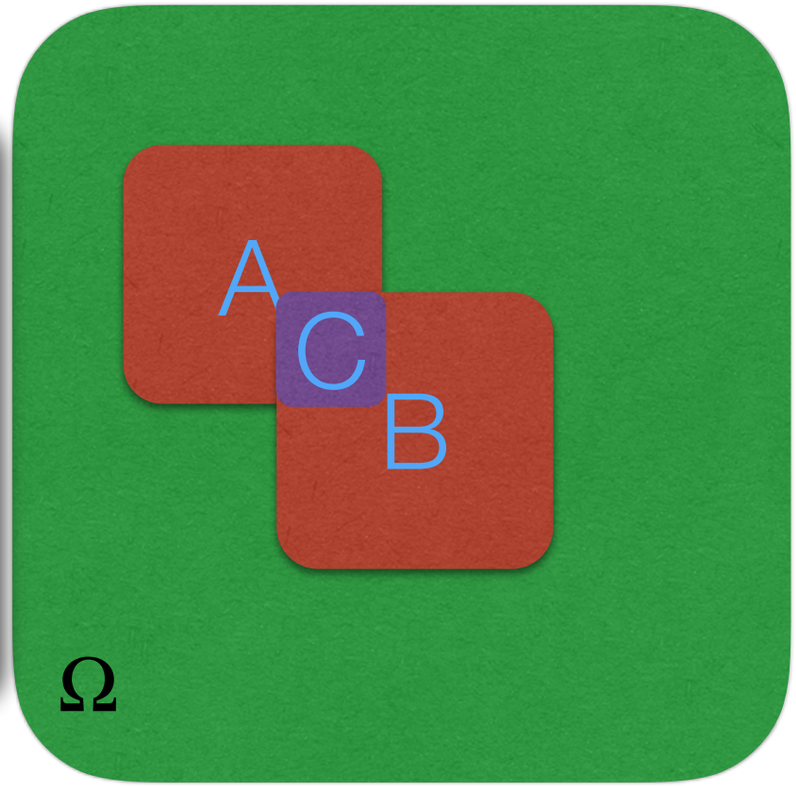
https://en.wikipedia.org/wiki/Probability_axioms

Probability = Area

Prob(A) = Area A

Total Area = 1

Prob(A or B) = Area A + Area B - Area C



$$0 \leq P(A) \leq 1$$

$$P(\Omega) \equiv 1$$

$$P(A \ or \ B) = P(A) + P(B) - P(C)$$

$$P(C) = P\left(A \ and \ B\right) = \text{overlap of A and B}$$

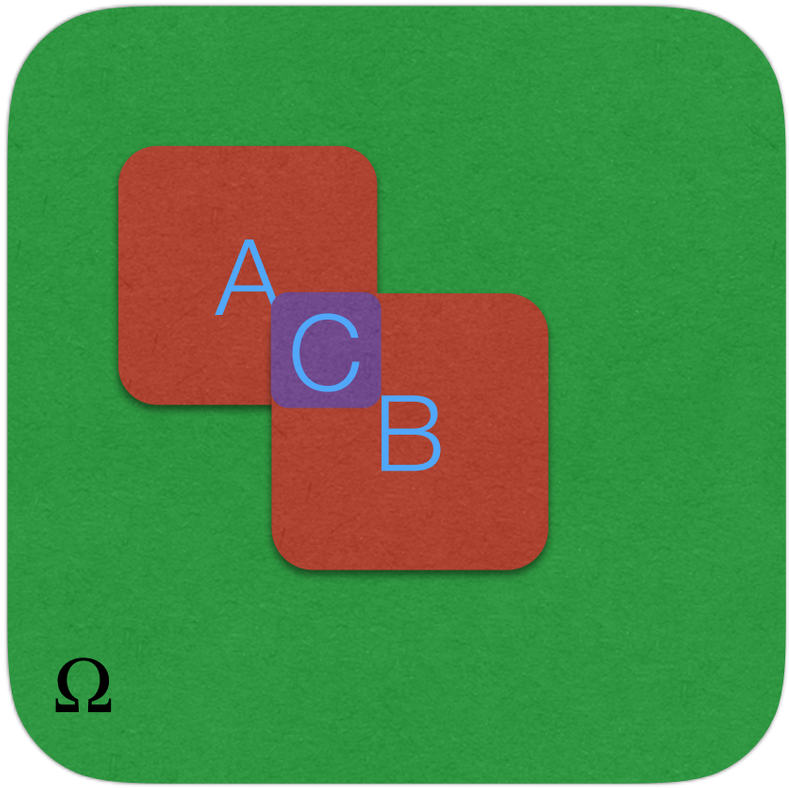
What's the probability that I'm in B given that I'm in A?

Conditional Probability

- What is the Probability of A given B?

- What is the probability that I'm in A if I know that I'm in B?

- What is the fraction of B that is overlapped by A?



- Normalize the area of the overlap (C) by the area you're conditioning on (B).

- The conditional probability of A given B is defined as:

$$P(A|B) = \frac{P(C)}{P(B)}$$

- We just have to figure out how to calculate $P(C)$

Conditional Probability

- Let's play a game by flipping two 2 coins.
- I Win if I get at least one Head and Lose otherwise
- There's 4 possibilities.
- What's the probability of a Win?

$$P(W) = \frac{3}{4}$$

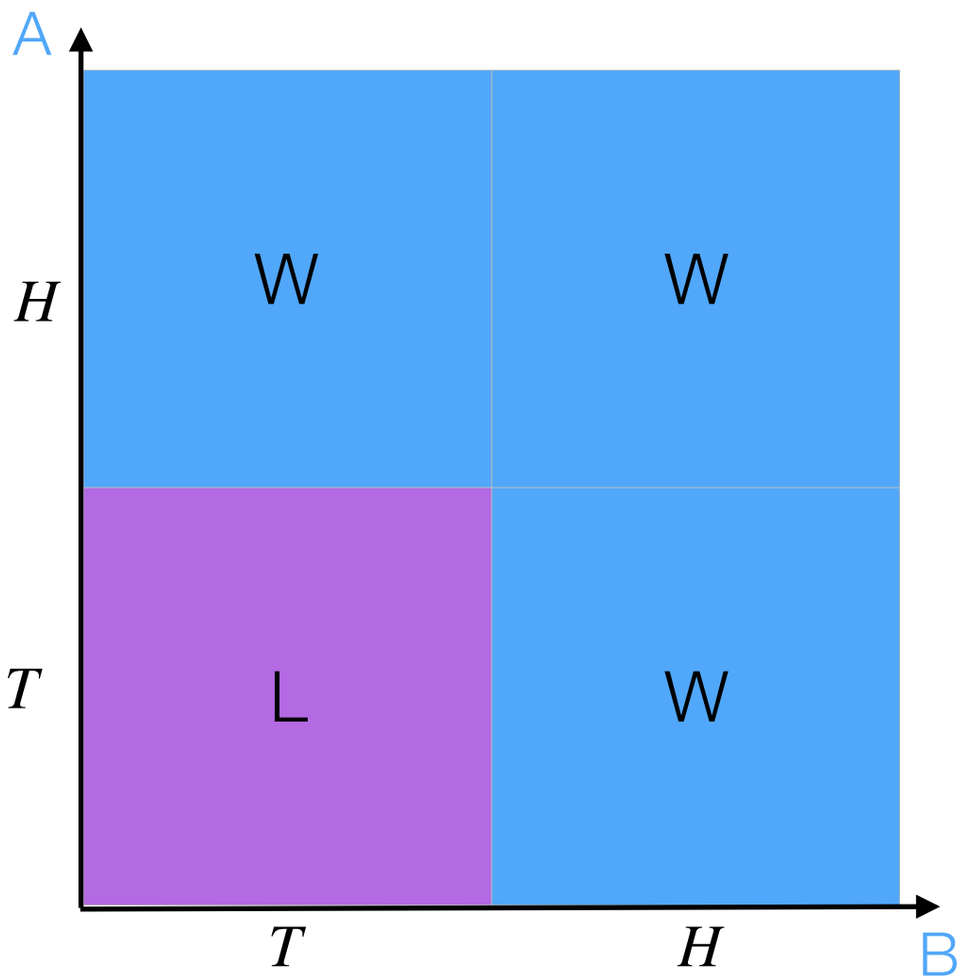
- What's the probability of a Win given that the second roll is Heads:

$$P(W|H) = \frac{2}{2} = 1$$

- What is the probability that the second rol is Heads given a win:

$$P(H|W) = \frac{2}{3}$$

- Symmetric conditional Probabilities can be different. The process of conditioning reflects extra information that we have available



Bayes Theorem

- We already saw that:

$$P(A|B) = \frac{P(C)}{P(B)}$$

- Conversely:

$$P(B|A) = \frac{P(C)}{P(A)}$$

- From which we can write:

$$P(C) = P(A|B)\,P(B)$$

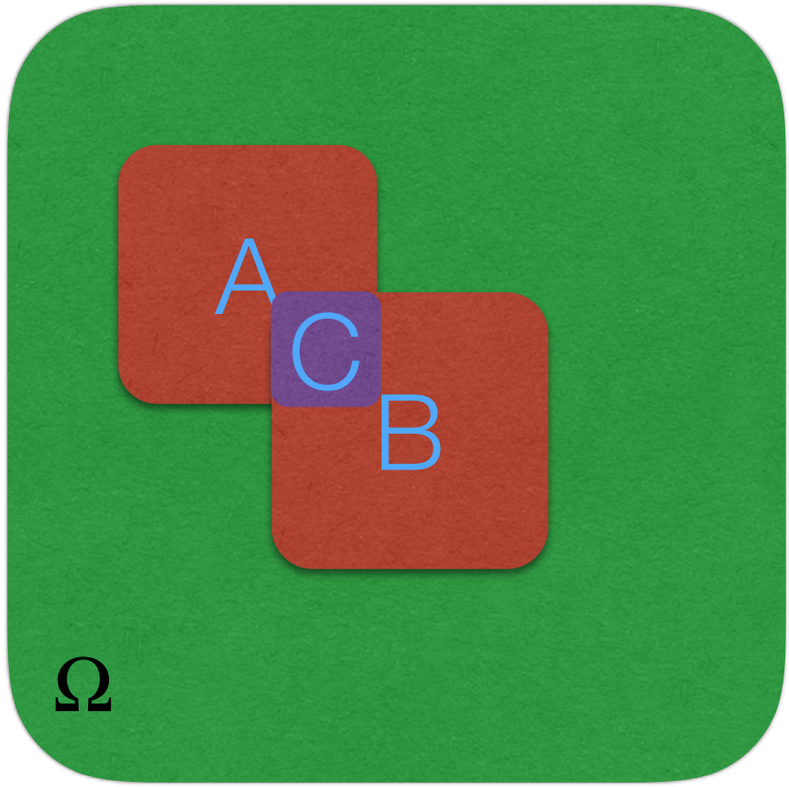
$$P(C) = P(B|A)\,P(A)$$

- Or, in other words:

$$P(A|B)\,P(B) = P(B|A)\,P(A)$$

- And finally:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$



Bayes Theorem

Bayes Theorem

- Now we can understand the previous example a bit better: A

$$P\left(H|W\right) = \frac{P\left(W|H\right)P\left(H\right)}{P\left(W\right)}$$

- We already know that:

$$P\left(W|H\right) = 1$$

- since a single heads is sufficient to give us a win.

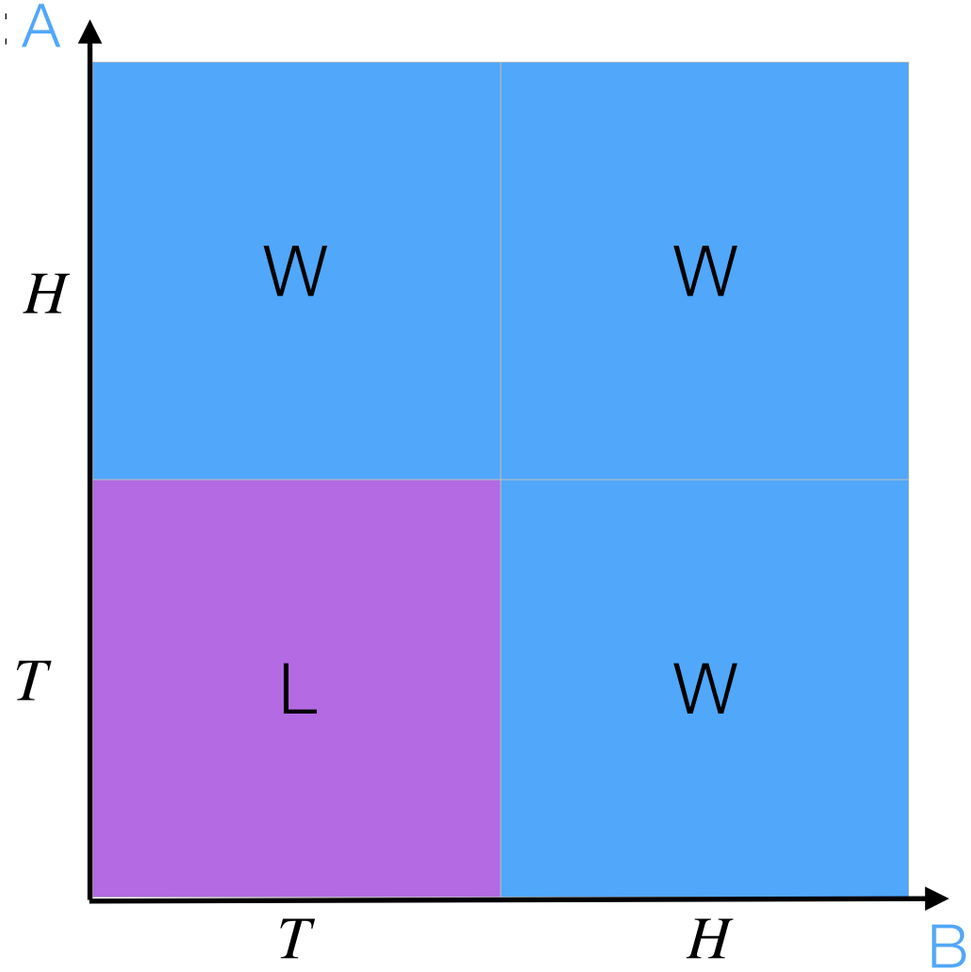
- We also know that

$$P\left(W\right) = \frac{3}{4} \qquad P\left(H\right) = \frac{1}{2}$$

- Therefore:

$$P\left(H|W\right) = \frac{P\left(W|H\right)P\left(H\right)}{P\left(W\right)} = \frac{1 \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}$$

- As we had already seen



Bayes Theorem

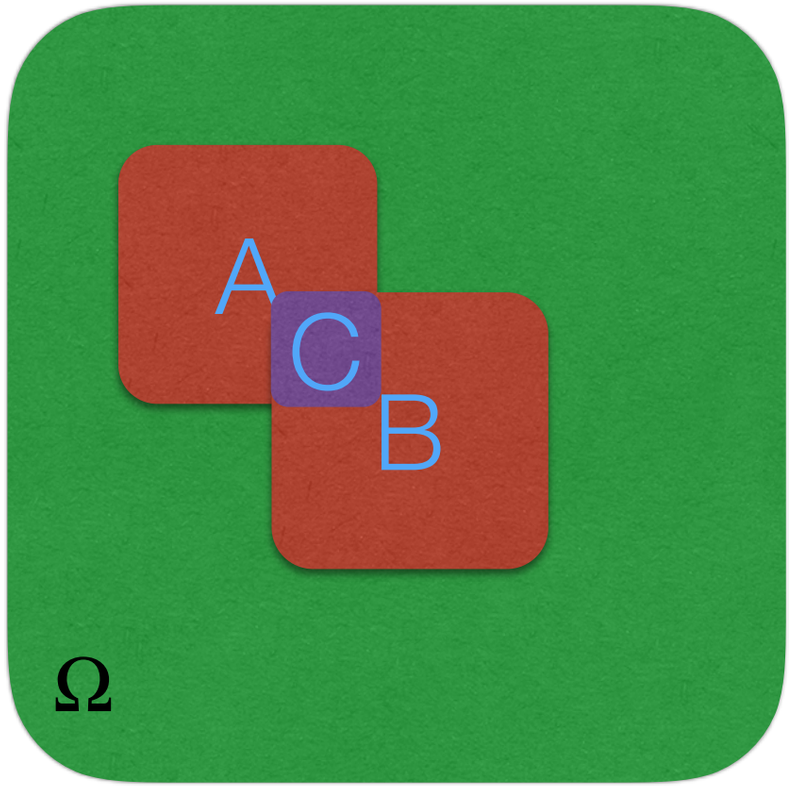
- A simple way of remembering this formula is to remember that:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- And simply work it out from the two ways of defining $P(C)$

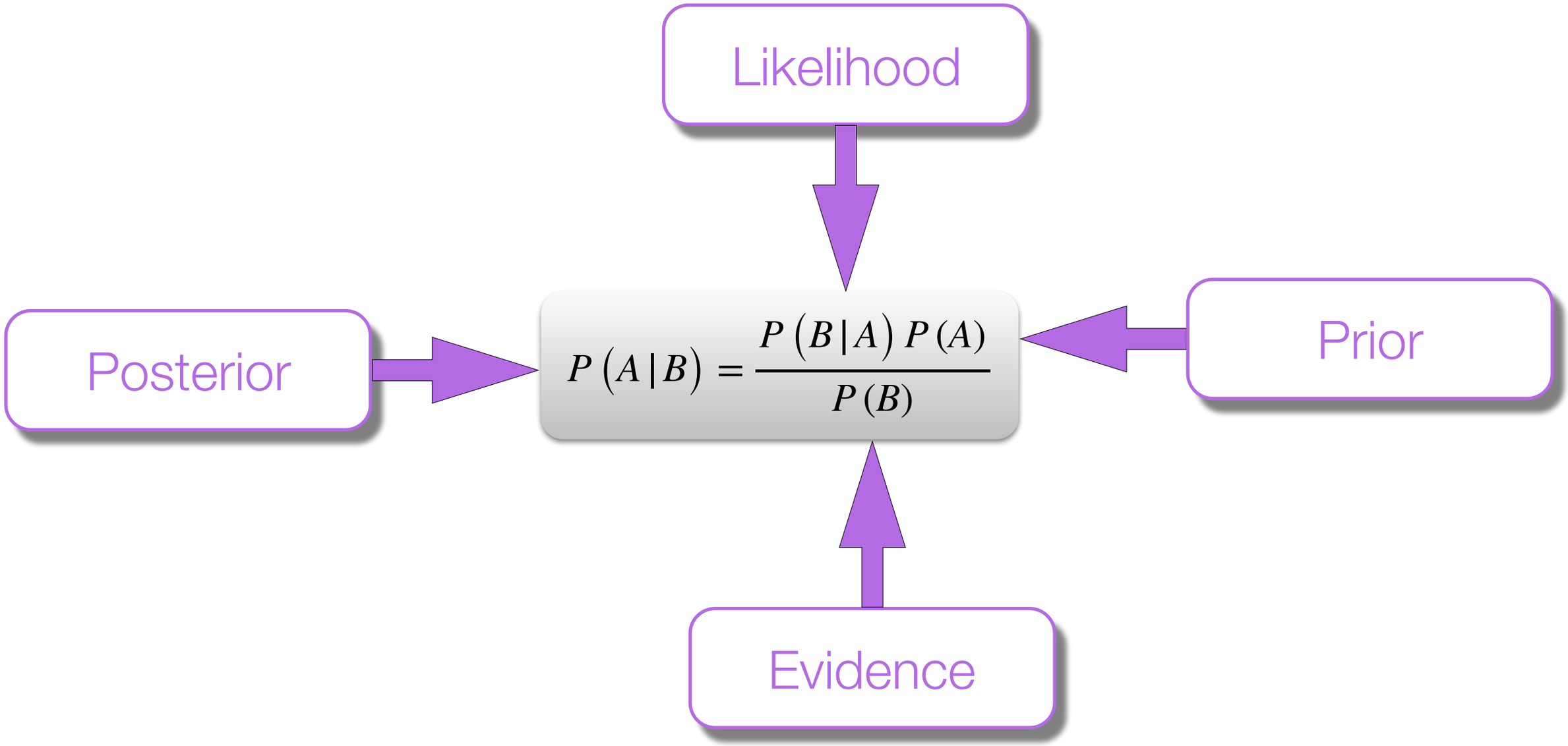
$$P(C) \equiv P(C)$$

- Despite it's simplicity, Bayes' Theorem is extremely powerful and resulted in the flourishing of a whole new branch of statistics, Bayesian Statistics



- The process of conditioning reflects the inclusion of added information. You can Bayes' Theorem as a way of updating your belief about a given situation in the presence of new information

Bayes Theorem - Terminology



Bayes Theorem - Terminology

Likelihood
Prob Data Given the Hypothesis

Posterior
Prob Hypothesis
Given the Data

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Prior
Prob Hypothesis

Evidence
Prob Data

Medical Tests

Your doctor thinks you might have a rare disease that affects 1 person in 10,000. A test that is 99% accurate comes out positive. What's the probability of you having the disease?

Bayes Theorem:

$$P\left(disease|positive\ test\right) = \frac{P\left(positive\ test|disease\right)P\left(disease\right)}{P\left(positive\ test\right)}$$

Total Probability:

$$P\left(positive\ test\right) = P\left(positive\ test|disease\right)P\left(disease\right)$$
$$+P\left(positive\ test|no\ disease\right)P\left(no\ disease\right)$$

Finally:

$$P\left(disease|positive\ test\right) = 0.0098$$

Medical Tests

Your doctor thinks you might have a rare disease that affects 1 person in 10,000. A test that is 99% accurate comes out positive. What's the probability of you having the disease?

Bayes Theorem:

$$P\left(disease|positive\ test\right) = \frac{P\left(positive\ test|disease\right)P\left(disease\right)}{P\left(positive\ test\right)}$$

Total Probability:

$$P\left(positive\ test\right) = P\left(positive\ test|disease\right)P\left(disease\right) + P\left(positive\ test|no\ disease\right)P\left(no\ disease\right)$$

Finally:

$$P\left(disease|positive\ test\right) = 0.0098$$

Base Rate Fallacy

Low Base Rate Value
+
Non-zero False Positive Rate

Medical Tests

Consider a population of 1,000,000 individuals. The numbers we should expect are:

	disease	no disease
positive	99	9,999
negative	1	989,901

$$P\left(disease|positive\ test\right) = \frac{TP}{TP+FP} = 0.0098$$

$$P\left(no\ disease|negative\ test\right) = \frac{TN}{TN+FN} = 0.99999$$

Medical Tests

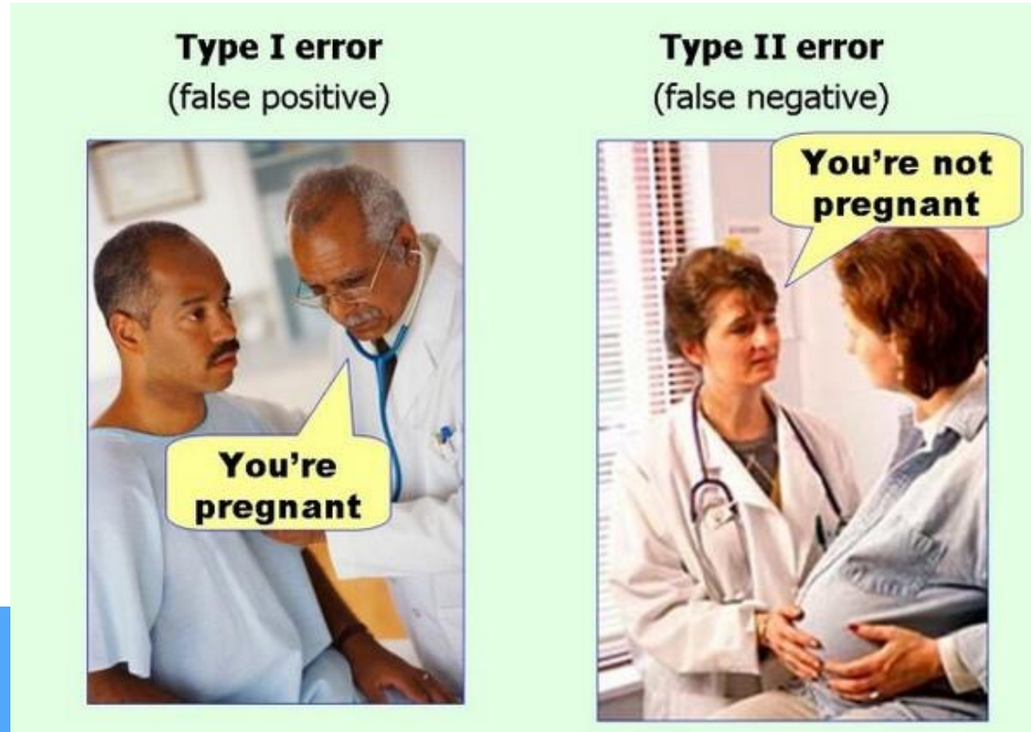
Consider a population of 1,000,000 individuals. The numbers we should expect are:

Marginals

	disease	no disease	
positive	99	9,999	10,098
negative	1	989,901	989,902
Marginals →	100	999,900	

$$P\left(disease|positive\ test\right) = \frac{TP}{TP + FP} = 0.0098$$

$$P\left(no\ disease|negative\ test\right) = \frac{TN}{TN + FN} = 0.99999$$



A second Test

Bayes Theorem still looks the same:

$$P\left(disease|positive\ test\right) = \frac{P\left(positive\ test|disease\right)P\left(disease\right)}{P\left(positive\ test\right)}$$

but now the probability that we have the disease has been updated:

$$P^{\dagger}\left(disease\right) = 0.0098$$

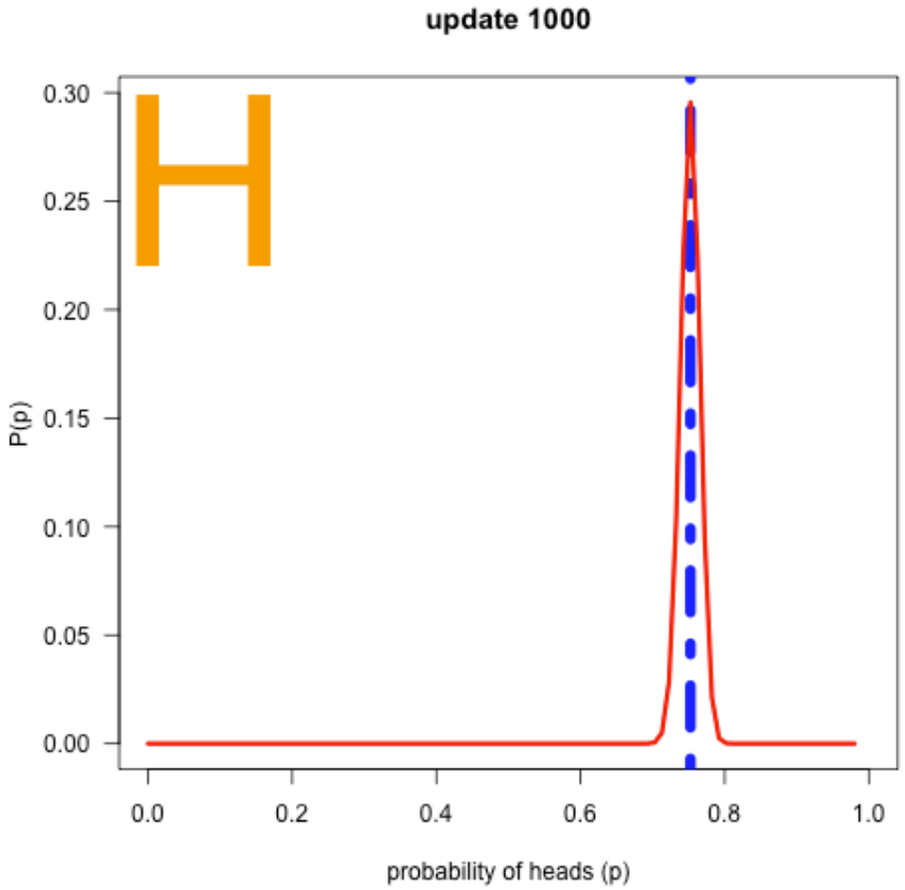
So this time we find:

$$P\left(disease|positive\ test\right) = 0.4949$$

Each test is providing new evidence, and Bayes theorem is simply telling us how to use it to update our beliefs.

Bayesian Coin Flips

- Biased coin with unknown probability of heads (p)
- Perform N flips and update our belief after each flip using Bayes Theorem



$$P\left(p\,|\,heads\right) = \frac{P\left(heads\,|\,p\right)P\left(p\right)}{P\left(heads\right)}$$

$$P\left(p\,|\,tails\right) = \frac{P\left(tails\,|\,p\right)P\left(p\right)}{P\left(tails\right)}$$

http://youtu.be/GTx0D8VY0CY

Beta Distribution

$$P_B(k,n,p) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

- Related to Binomial and has a very similar form:

$$P_\beta(x,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

- with $x \in [0,1]$ and $\alpha,\ \beta > 0$

- Can be thought of as modeling the probability of p for a given number of α heads and β tails

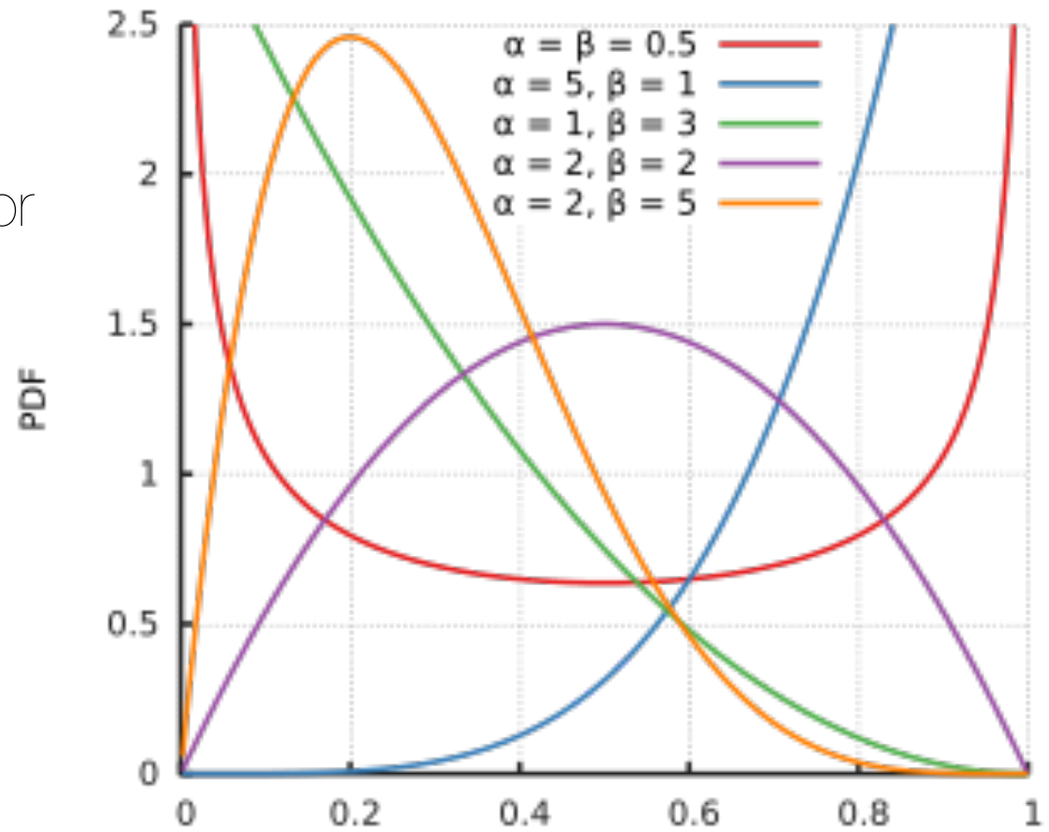
- $\Gamma(\alpha)$ is the continuous version of $\alpha!$

- The mean is:

$$\mu = \frac{\alpha}{\alpha+\beta}$$

- And the variance:

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$



Expectation Maximization

- Iterative algorithm to learn parameter estimates in models with unobserved latent variables

- Two steps for each iteration

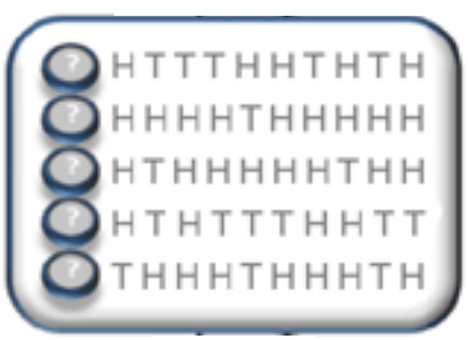
    - Expectation: Calculate the likelihood of the data given current parameter estimate

    - Maximization: Find the parameter values that maximize the likelihood

- Stop when the relative variation of the parameter estimates is smaller than some value

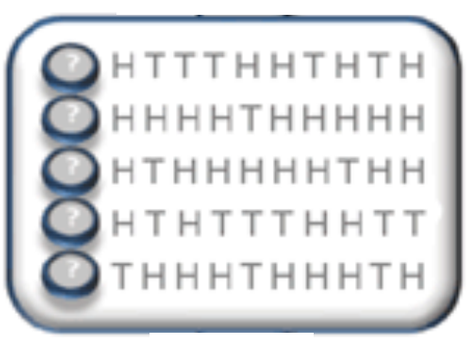
Expectation Maximization

Nature BioTech 26, 897 (2008)



Nature BioTech 26, 897 (2008)

Expectation Maximization

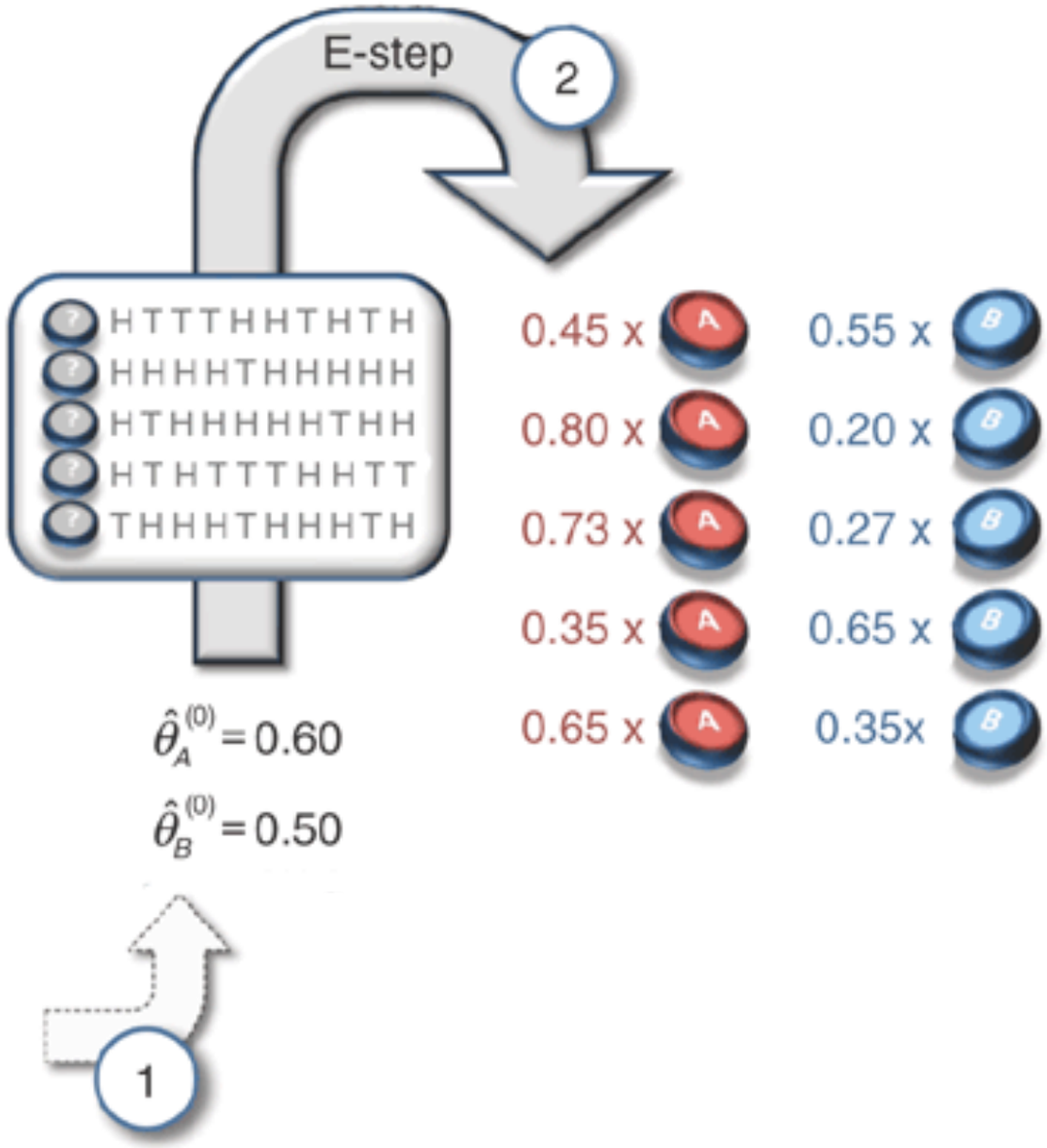


$$\hat{\theta}_A^{(0)} = 0.60$$

$$\hat{\theta}_B^{(0)} = 0.50$$

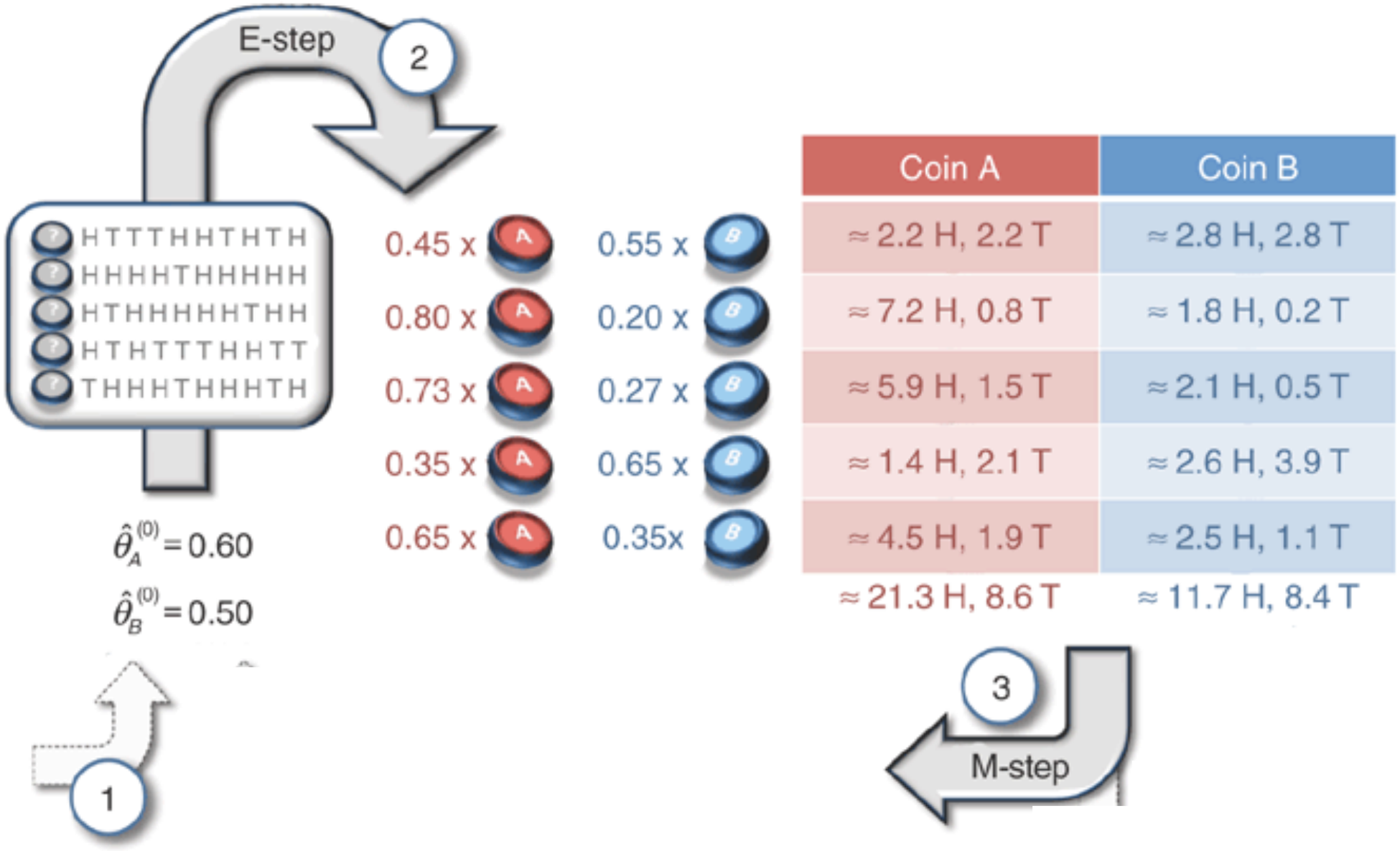
Nature BioTech 26, 897 (2008)

Expectation Maximization



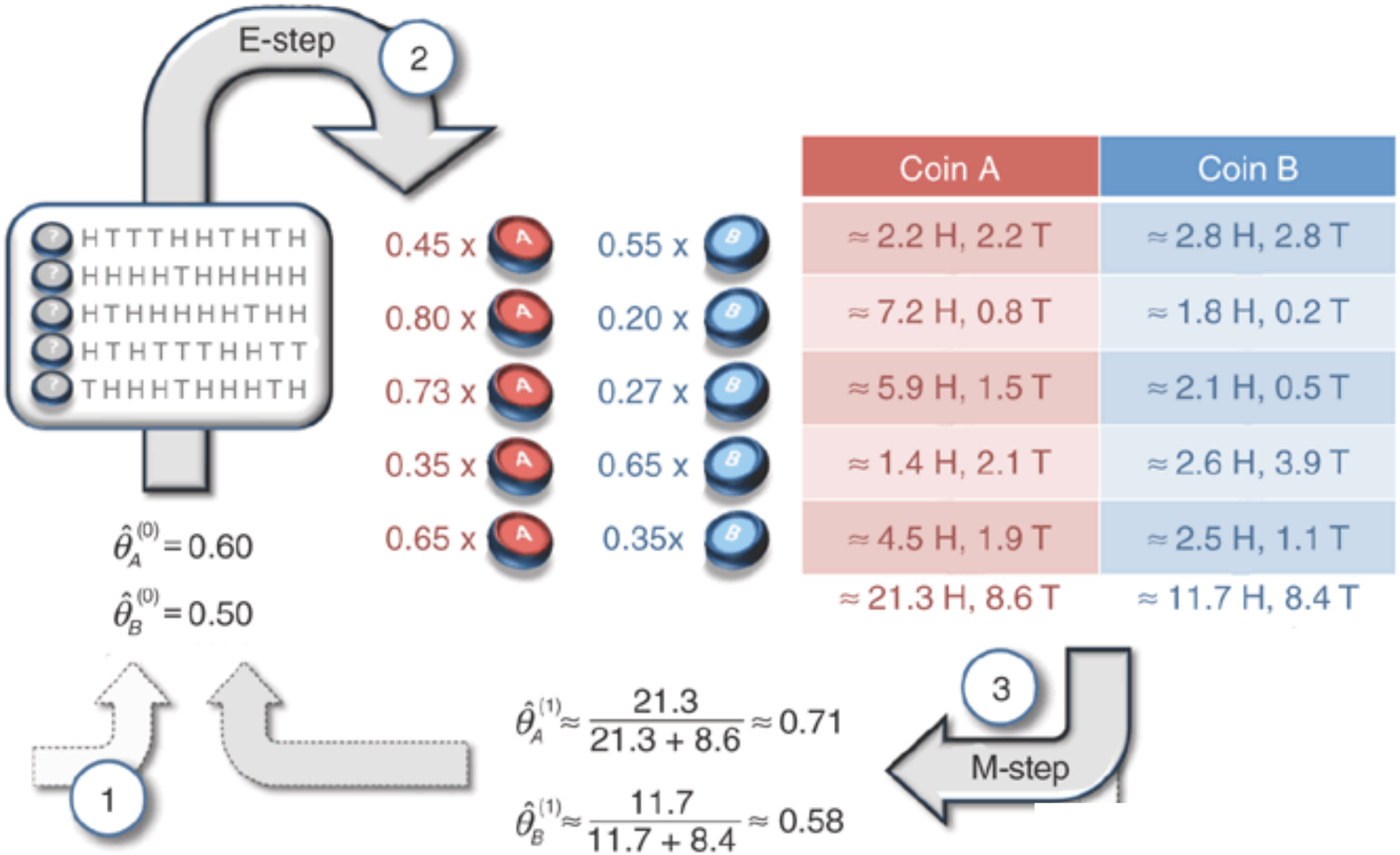
Nature BioTech 26, 897 (2008)

Expectation Maximization



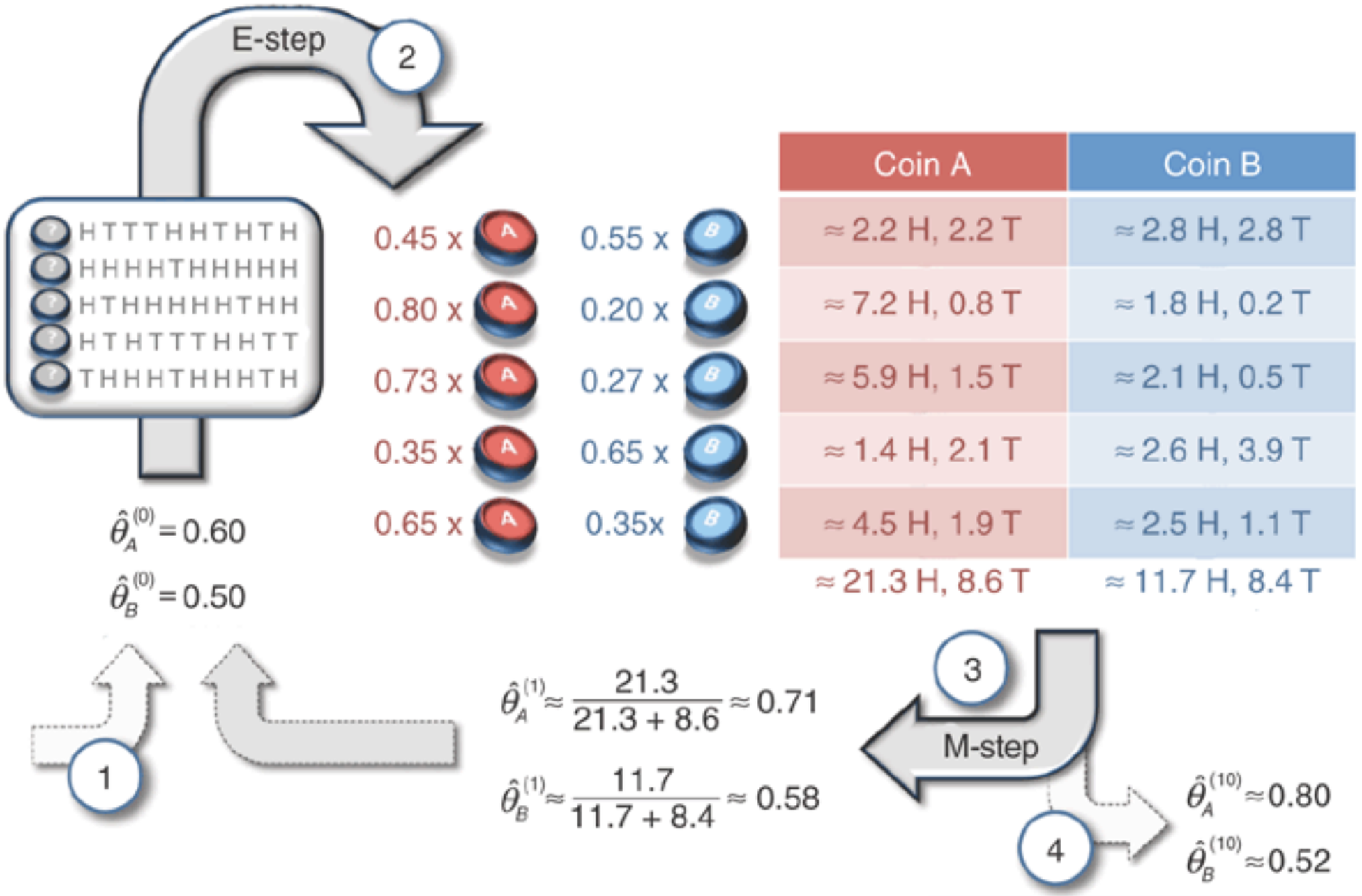
Nature BioTech 26, 897 (2008)

Expectation Maximization



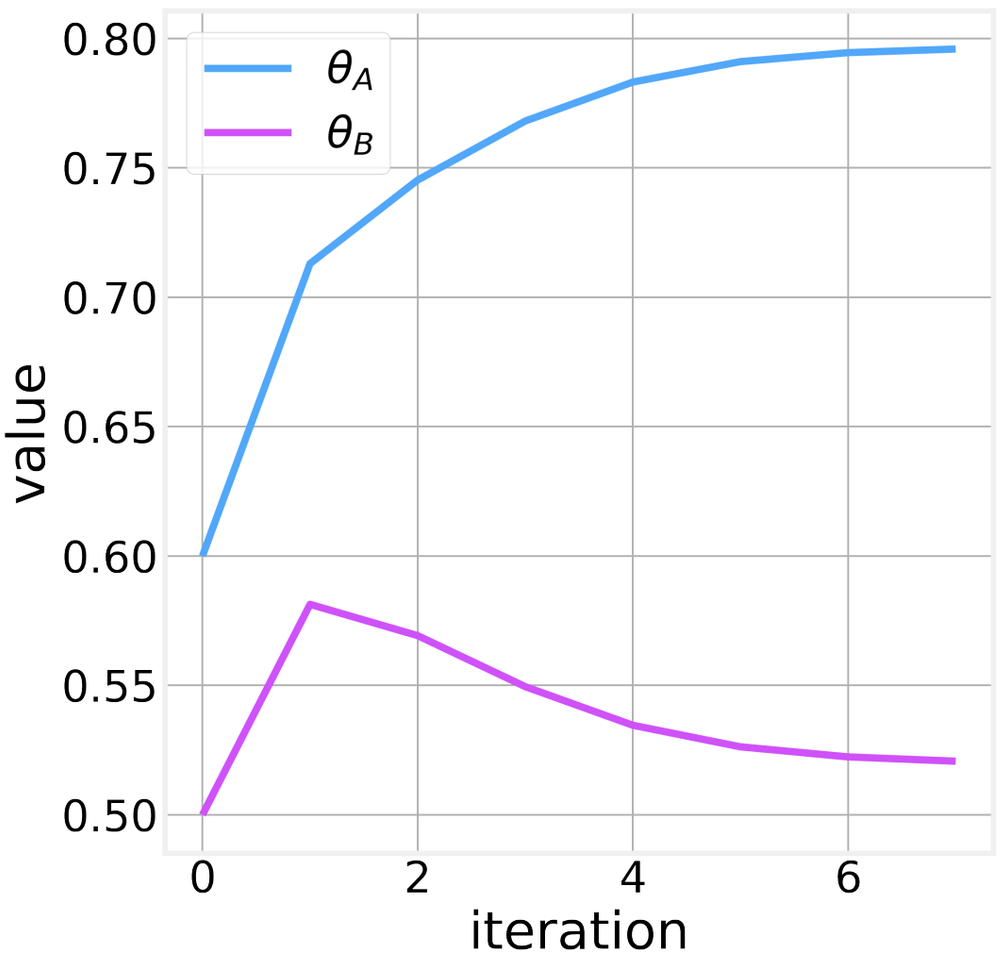
Nature BioTech 26, 897 (2008)

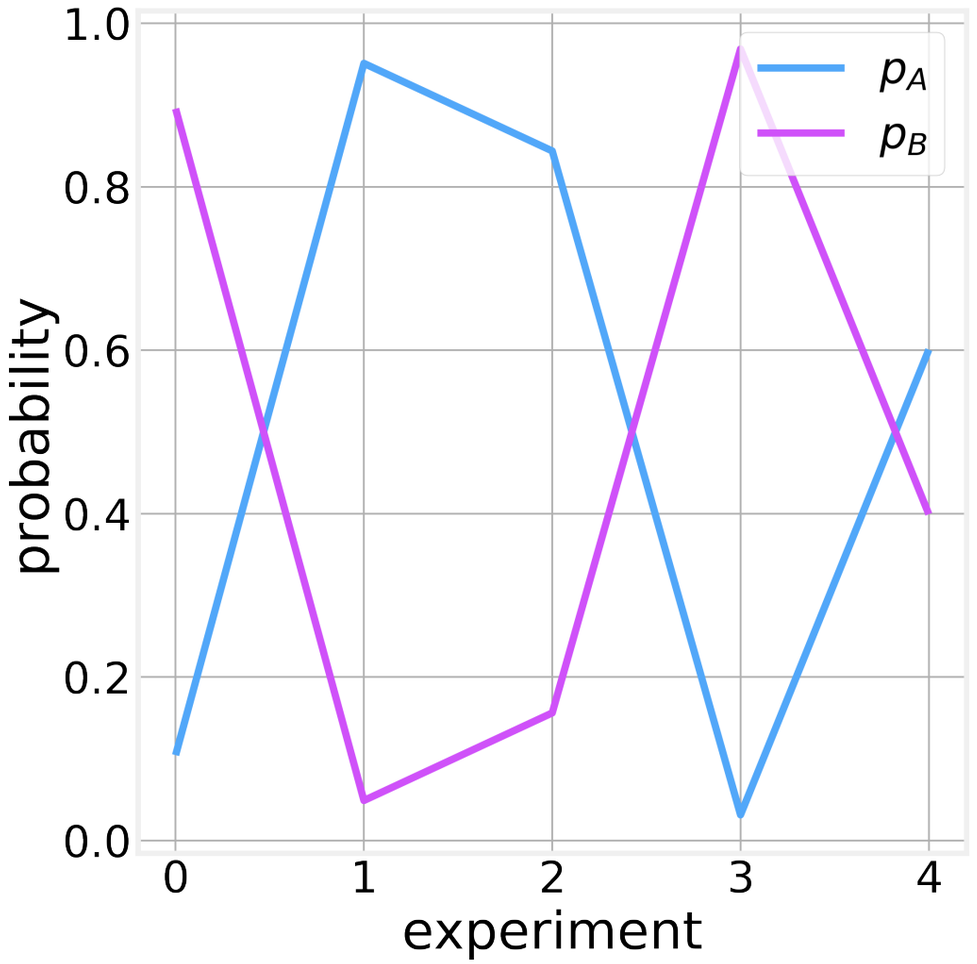
Expectation Maximization



Nature BioTech 26, 897 (2008)

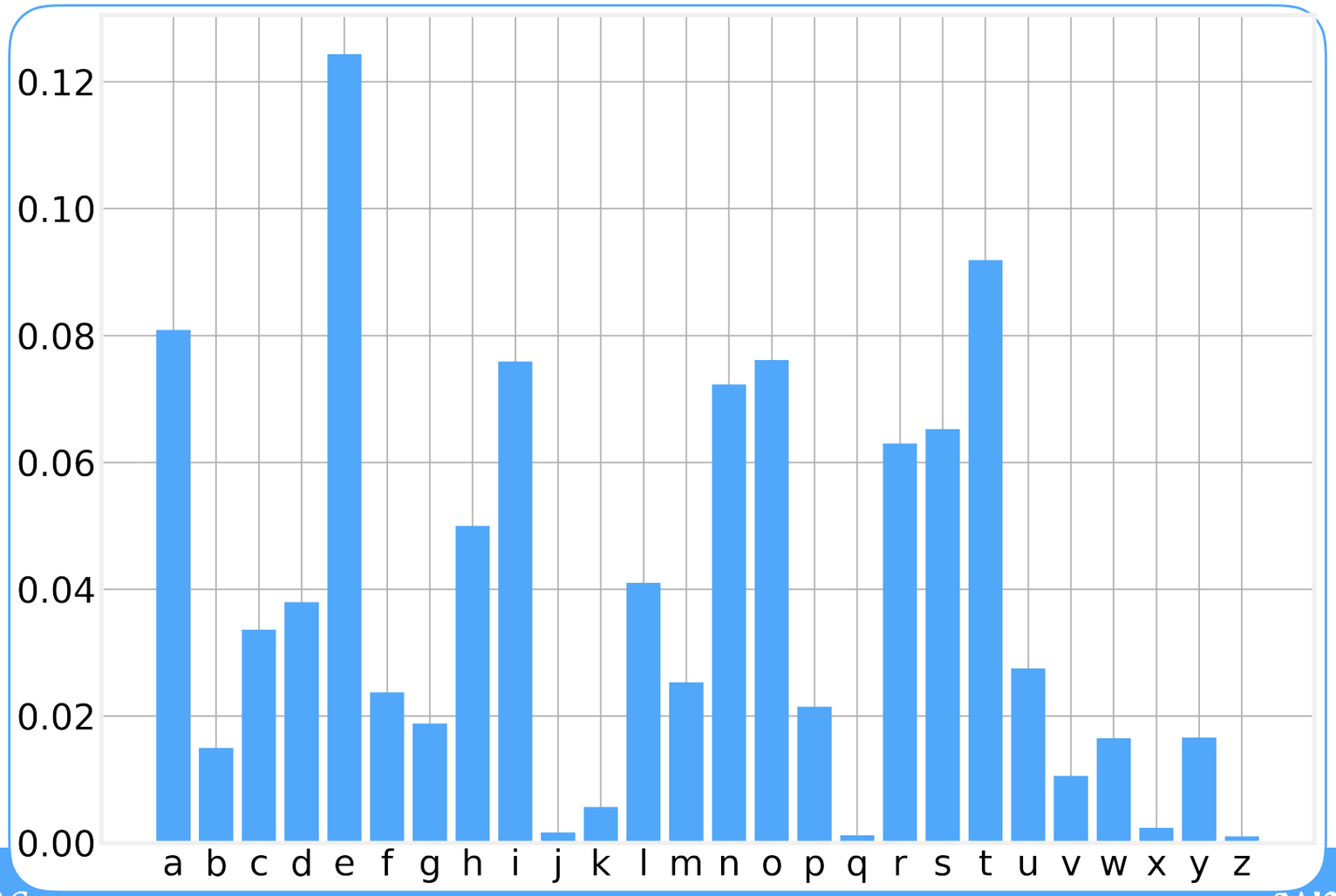
Expectation Maximization





Language Detection (Naive Bayes Classification)

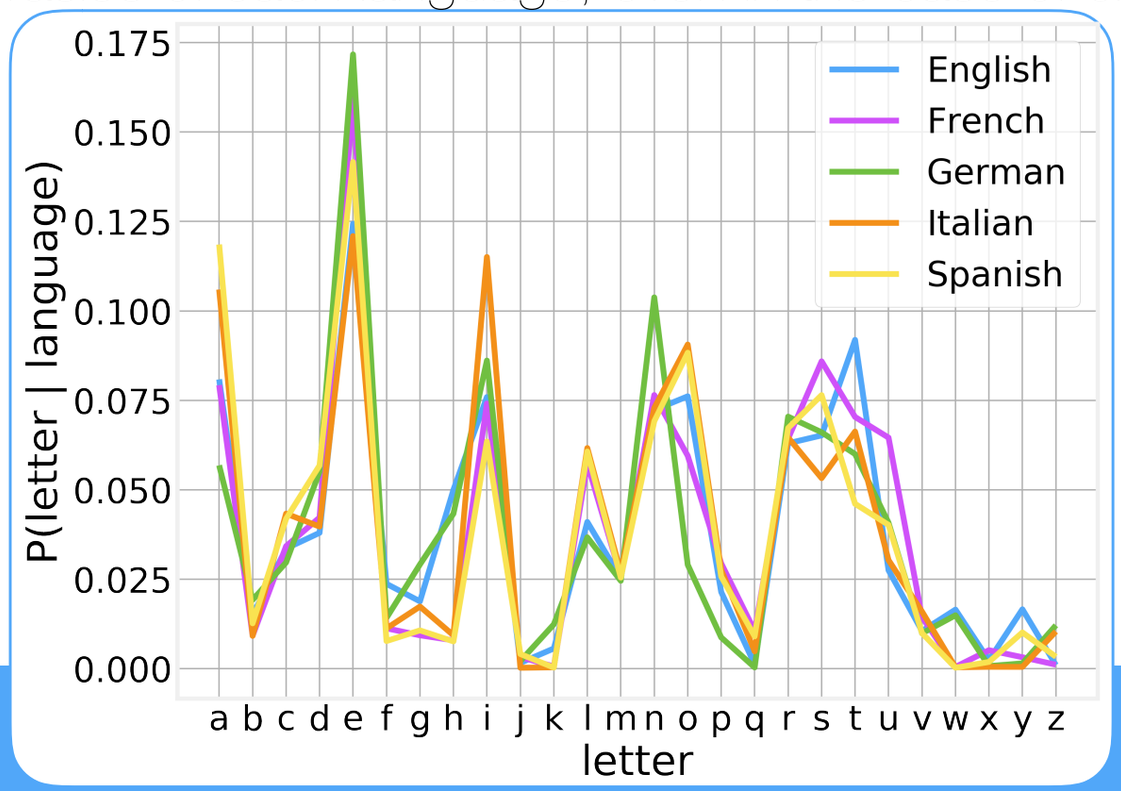
- Language detection allows us to preprocess our corpus so that we can focus on specific languages, sort documents by language, etc.
- Languages can be characterized by their character (letter) distribution.
- The character level distribution for English, obtained using Google Books 1-gram dataset is:



Character Distributions

- We measured the probability distribution of letters in the english language. In effect, we calculated: $P\left(letter|english\right)$
- The probability of seeing a specific letter given that the text is in English. If we do this for a few other languages we can have a table of the form: $P\left(letter|language\right)$

- Google Books covers several different languages, among which we find 5 different European languages: English, French, German, Italian and Spanish.
- Character distributions for different languages look different in at least a few of the characters due to the idiosyncrasies of each language, even in the case of closely related languages.



Conditional Probabilities

- Using these conditional probabilities, and Bayes Theorem, we can easily build a language detector. For that we just need to calculate:

$$P\left(language|text\right)$$

- Which we can rewrite as:

$$P\left(language|letter_1, letter_2, \cdots, letter_n\right)$$

- If we treat each letter independently, we obtain:

$$P\left(language|letter_1, letter_2, \cdots, letter_n\right) = \prod_i P\left(language|letter_i\right)$$

- This is known as the Naive Bayes Approach and is an obvious oversimplification: It completely ignores correlations present in the sequence of letters.

- All we have to do now is apply Bayes Theorem to our original table:

$$P\left(language|letter\right) = \frac{P\left(letter|language\right)P\left(language\right)}{P\left(letter\right)}$$

Naive Bayes

- And if we assume that all languages are equally probable (non-informative prior):

$$P\left(language\right) = \frac{1}{N_{langs}}$$

- Naive Bayes approaches (and many others) use terms of the form:

$$\prod_i P\left(A \,|\, B_i\right)$$

- which implies multiplying many small numbers. To avoid numerical complications, it is best to use, instead:

$$\sum_i \log P\left(A \,|\, B_i\right)$$

- Which is commonly referred to as the "Log-Likelihood". Our expression then becomes:

$$\mathscr{L}\left(language \,|\, letter_1, letter_2, \cdots, letter_n\right) = \sum_i \log\left[\frac{P\left(letter_i \,|\, language\right) P\left(language\right)}{P\left(letter_i\right)}\right]$$

Naive Bayes

- Or more simply:

$$\mathscr{L}\left(language|text\right) = \sum_i \log\left[\frac{P\left(letter_i|language\right)P\left(language\right)}{P\left(letter_i\right)}\right]$$

- And finally:

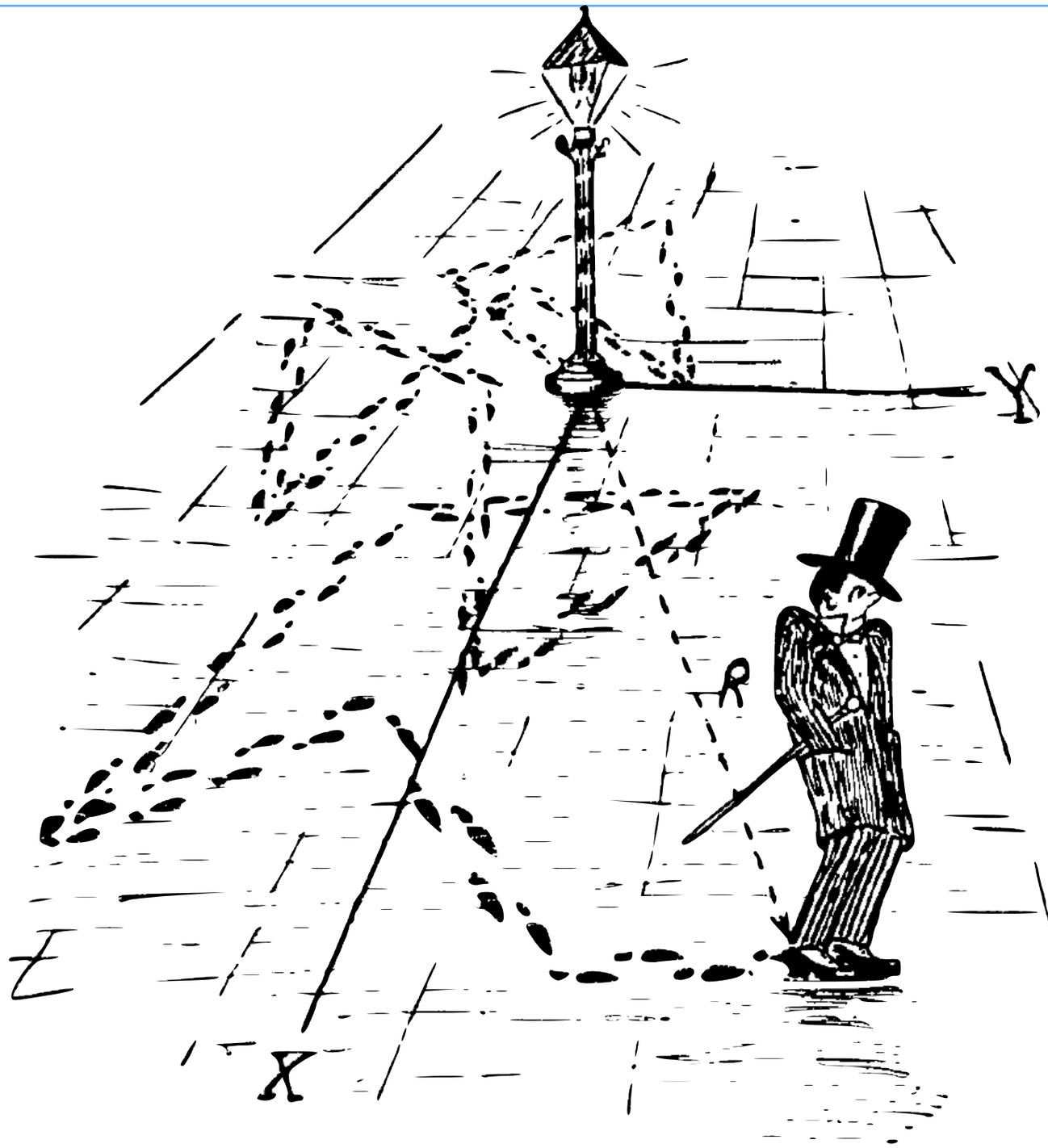
$$\mathscr{L}\left(language|text\right) = \sum_i \mathscr{L}\left(language|letter_i\right)$$

- Providing us with a quick and easy way to determine which language is more likely to be the correct one.

Lesson III:
Random Walks and Markov Chains

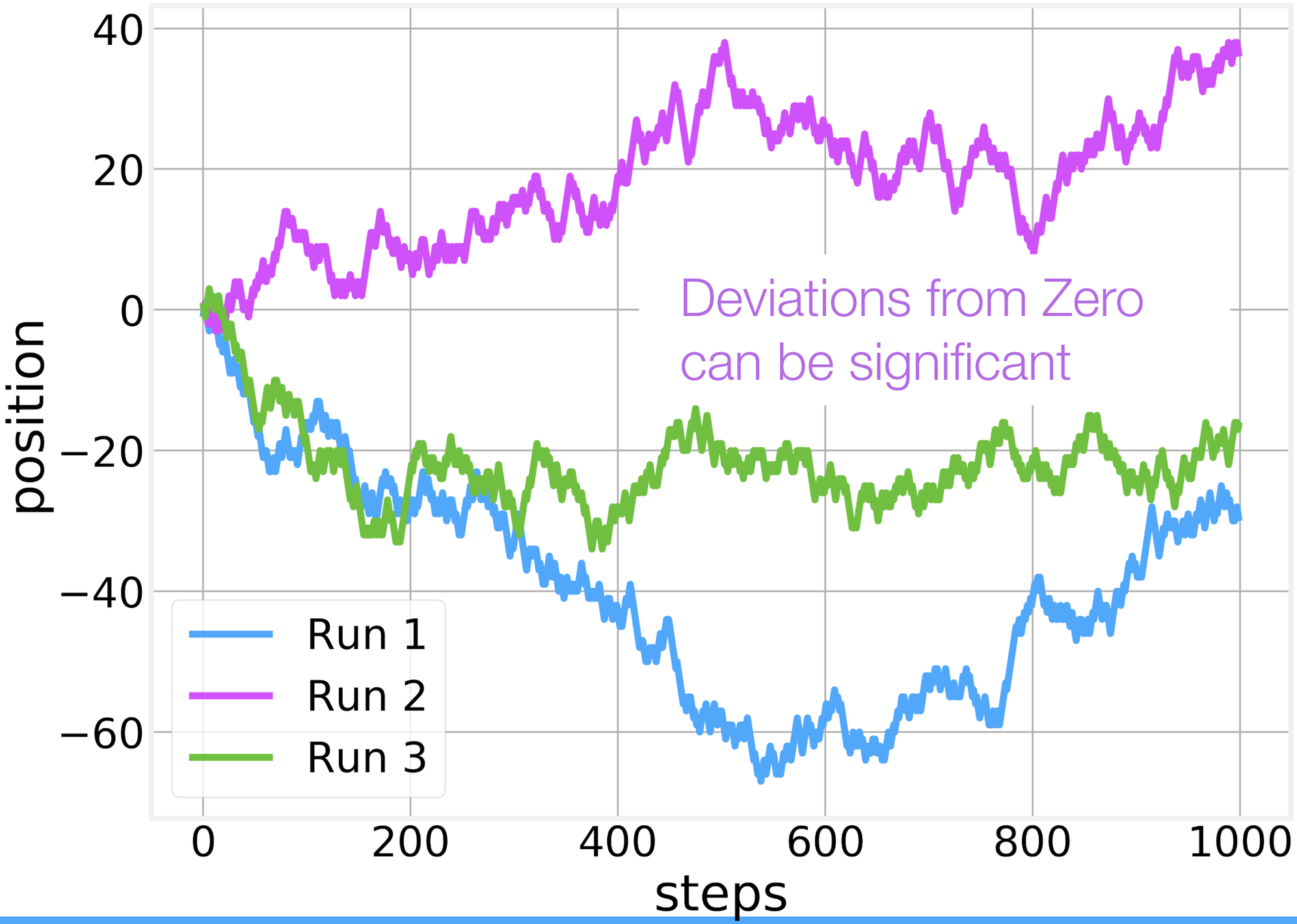
Random Walks

Illustration by George Gamow

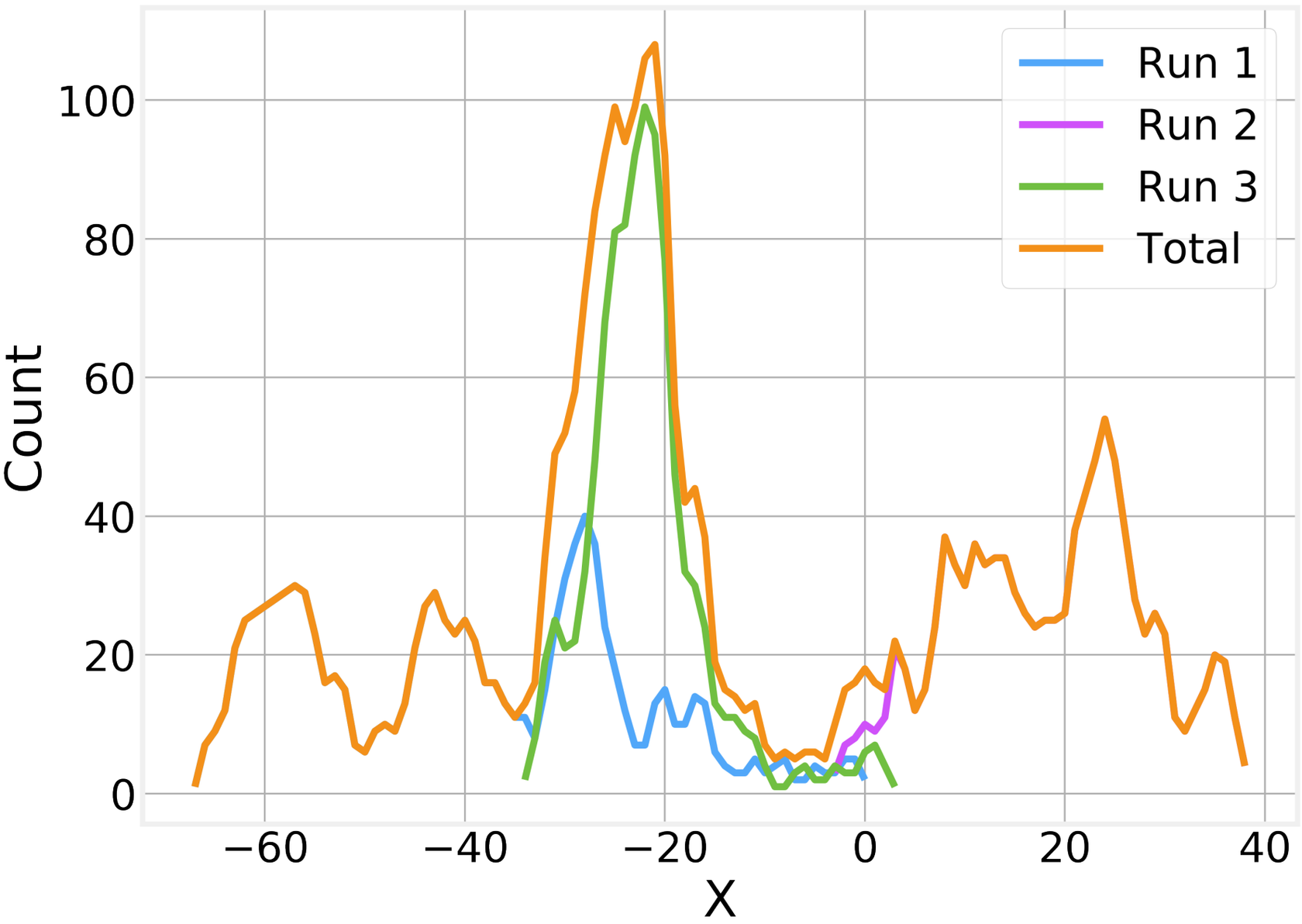


- At each step flip a coin
  - Heads: Move right
  - Tails: Move left
- If you start at position 0, do you ever reach position L?
- On average, we expect the position to be always close to 0.
- What if the coin is biased as in the previous example?

Random Walks

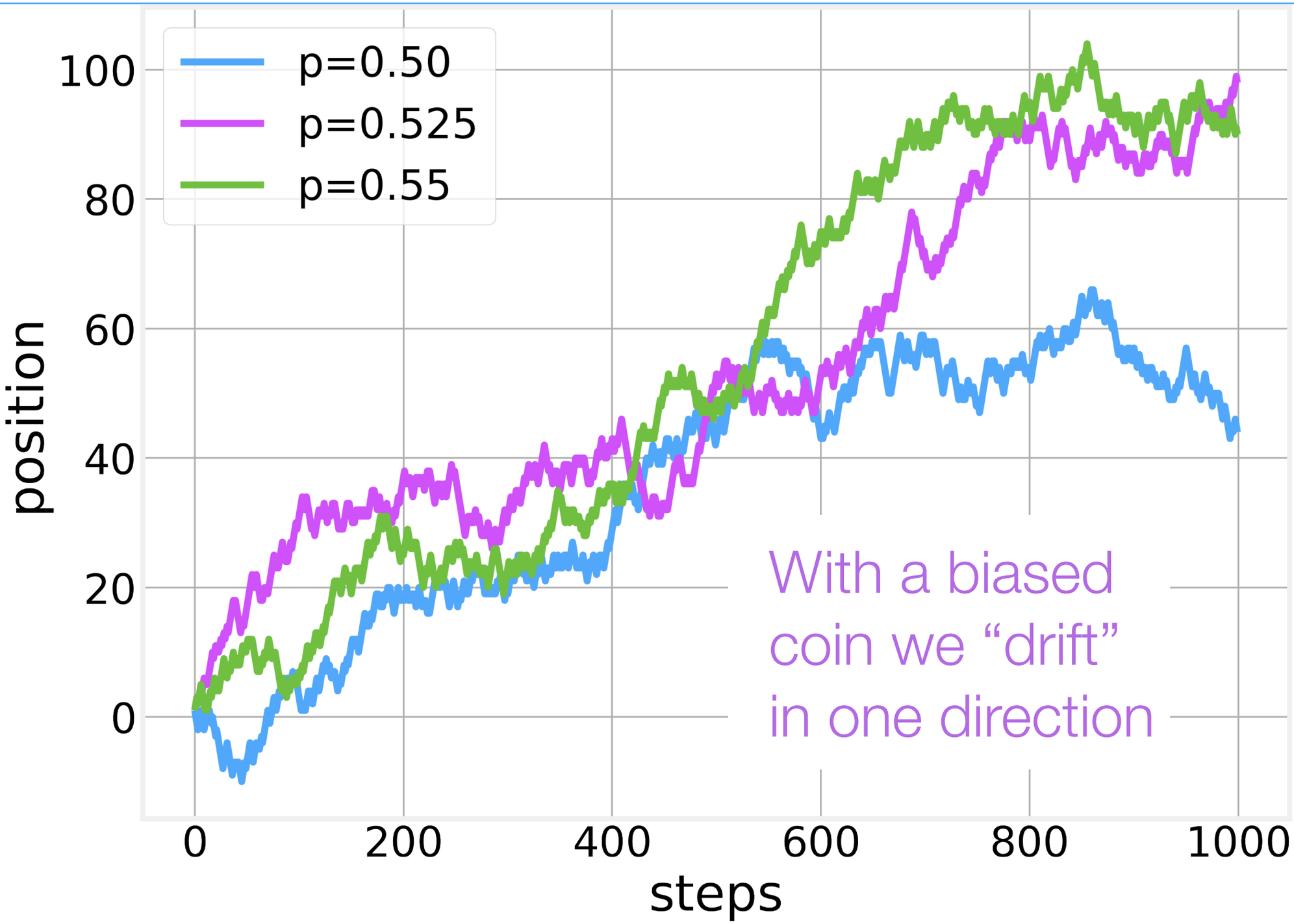


Visited Positions

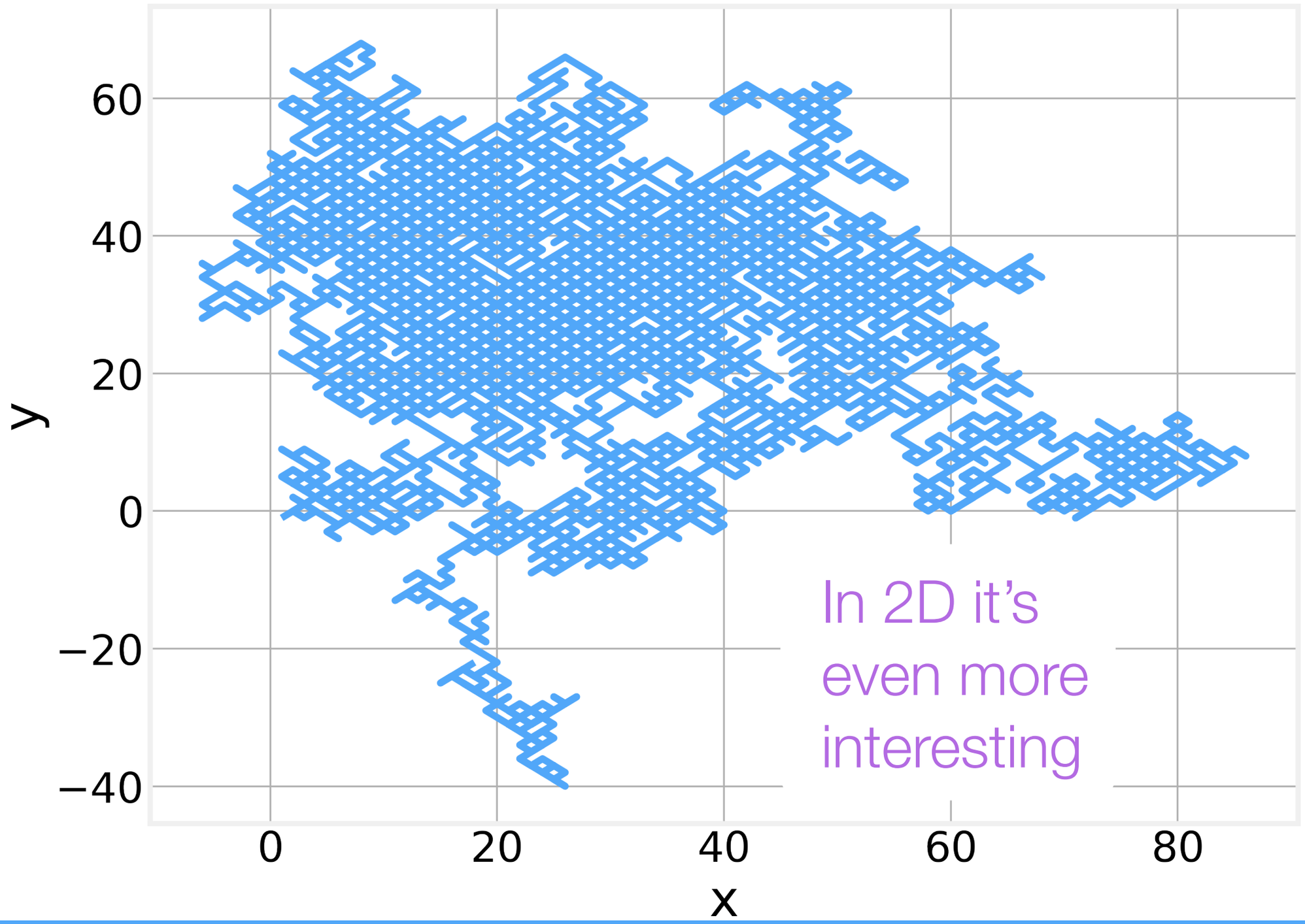


- The most visited positions, across multiple realizations, are the ones closest to the origin, 0
- This reflects the fact that the number of ways of reach 0 is much larger than the number of ways of reaching the extremes due to the combinatorial factor C_k^N
- 0 is better connected
- With a biased coin, the better connected location would be somewhere else due to drift

Random Walk with a drift



2 Dimensions



Lévy Flights

- Named after French Mathematician Paul Lévy

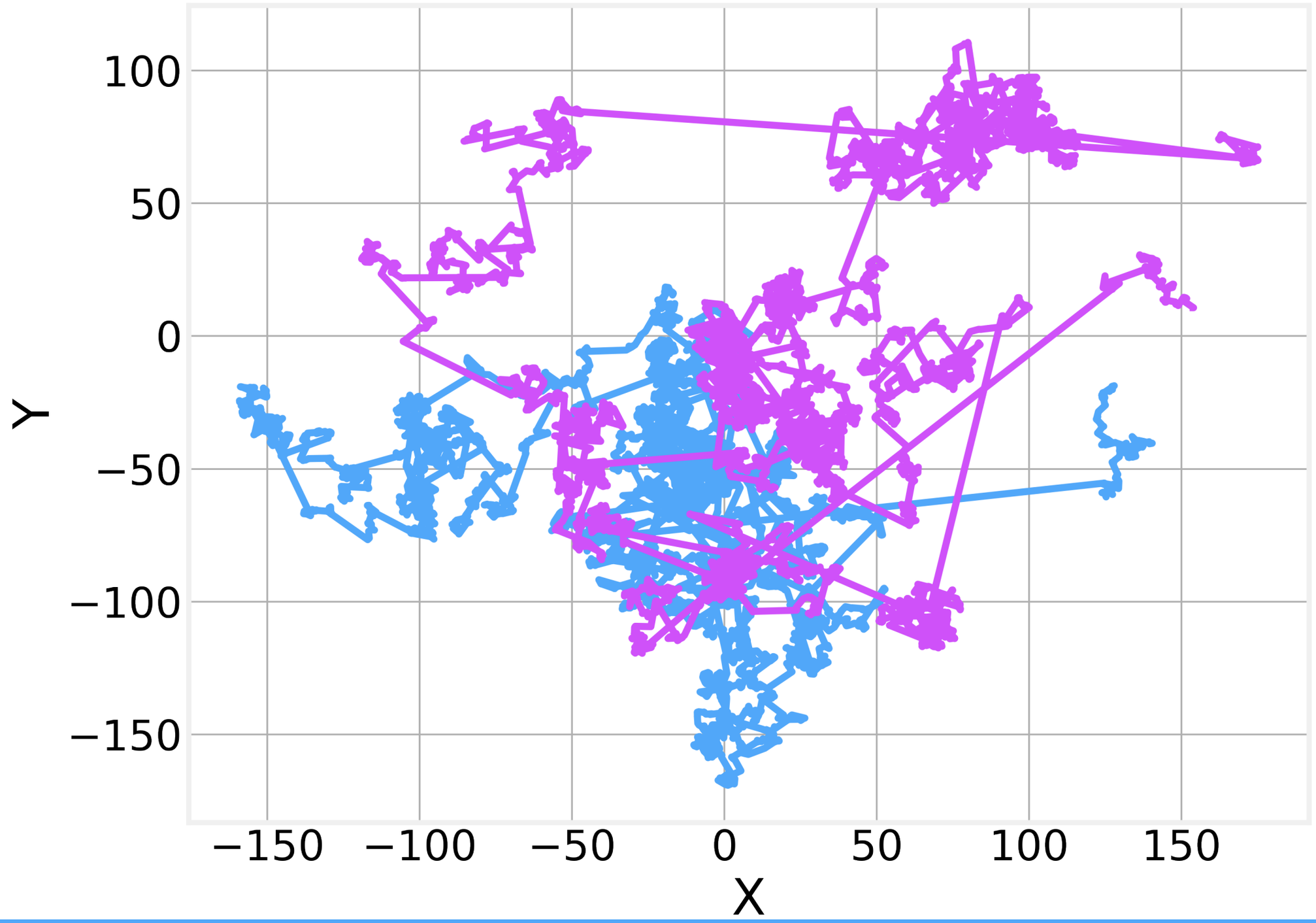
- 2D+ Random Walks where the step size is randomly chosen from a broad-tailed distribution

- Step direction is chosen homogeneously

- Important in chaos theory, finance, cryptography, etc…

- Simple model for foraging in nature - Most steps are small (exploitation) while some are very large (exploration)

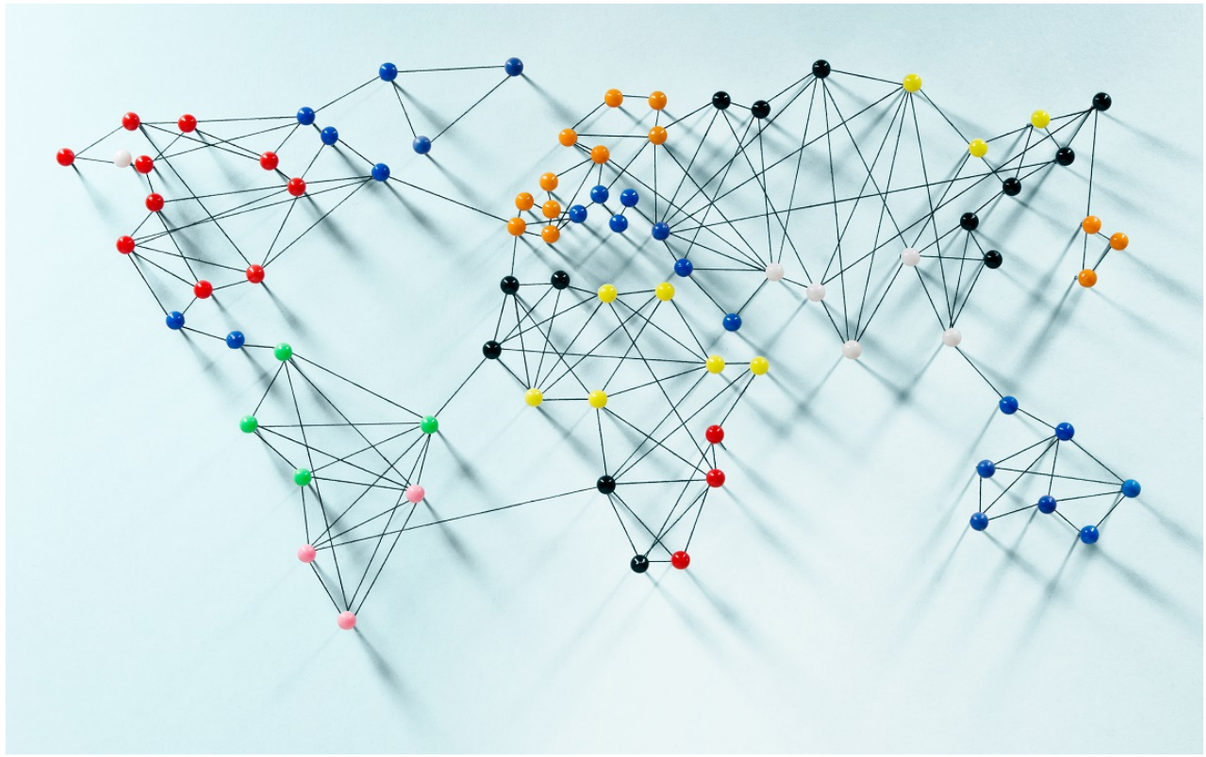
Lévy Flights



What if we're walking on a network?

https://www.analyticsvidhya.com/blog/2018/04/introduction-to-graph-theory-network-analysis-python-codes/

- Each possible location is a node (dot)
- Each path is an edge (line) connecting two nodes
- At each step i randomly choose among the k_i available edges



- Just a generalization of the previous 1D and 2D examples with fixed size steps
- Nodes can be cities, webpages, street intersections, etc and edges can be, respectively, airline connections, links, roads, etc

PageRank

Google

Manhattan

From Wikipedia, the free encyclopedia

Coordinates: 40°47′N 73°58′W

This article is about the New York City borough. For other uses, see Manhattan (disambiguation).

Manhattan (/mænˈhætən, mən-/), often referred to locally as ***the City***,[1] is the most densely populated of the five boroughs of New York City and its economic and administrative center, cultural identifier,[5] and historical birthplace.[6] The borough is coextensive with **New York County**, one of the original counties of the U.S. state of New York. The borough consists mostly of Manhattan Island, bounded by the Hudson, East, and Harlem rivers; several small adjacent islands; and Marble Hill, a small neighborhood now on the U.S. mainland, physically connected to the Bronx and separated from the rest of Manhattan by the Harlem River. Manhattan Island is divided into three informally bounded components, each aligned with the borough's long axis: Lower, Midtown, and Upper Manhattan.

Manhattan has been described as the cultural, financial, media, and entertainment capital of the world,[7][8][9][10] and the borough hosts the United Nations Headquarters.[11] Anchored by Wall Street in the Financial District of Lower Manhattan, New York City has been called both the most economically powerful city and the leading financial center of the world,[12][13][14][15][16][17] and Manhattan is home to the world's two largest stock exchanges by total market capitalization: the New York Stock Exchange and NASDAQ.[18][19] Many multinational media conglomerates are based in Manhattan, and the borough has been the setting for numerous books, films, and television shows. Manhattan real estate has since become among the most expensive in the world, with the value of Manhattan Island, including real estate, estimated to exceed US$3 trillion in 2013;[6][20] median residential property sale prices in Manhattan approximated US$1,600 per square foot ($17,000/m^2) as of 2018,[21] with Fifth Avenue in Midtown Manhattan commanding the highest retail rents in the world, at US$3,000 per square foot ($32,000/m^2) in 2017.[22]

Manhattan traces its origins to a trading post founded by colonists from the Dutch Republic in 1624 on Lower Manhattan; the post was named New Amsterdam in 1626. Manhattan is historically documented to have been purchased by Dutch colonists from Native Americans in 1626 for 60 guilders, which equals roughly $1038 in current terms.[23][24][25] The territory and its surroundings came under English control in 1664[25] and were renamed *New York* after King Charles II of England granted the lands to his brother, the Duke of York.[26] New York, based in present-day Manhattan, served as the capital of the United States from 1785 until 1790.[27] The Statue of Liberty greeted millions of immigrants as they came to the Americas by ship in the late 19th and early 20th centuries[28] and is a world symbol of the United States and its ideals of liberty and peace.[29] Manhattan became a borough during the consolidation of New York City in 1898.

New York County is the United States' second-smallest county by land area (larger only than Kalawao County, Hawaii), and is also the

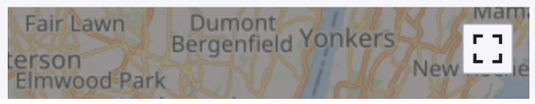
Manhattan
New York County, New York

Borough and county

Midtown Manhattan facing south toward Lower Manhattan

Flag

Etymology: Lenape: **Manaháhtaan** (the place where we get bows)
Nickname(s): *The City*[1]



Markov Chain

- Random walks are simple examples of the more general class of Markov Chain processes

- Memoryless - "Future depends only on present state and not on past history"

$$\vec{\pi}(t+1) = \mathcal{G}\vec{\pi}(t)$$

- Has a stationary state $\vec{\pi} = \mathcal{G}\vec{\pi}$ if the transition matrix $\mathcal{G}$ is:

- Stochastic - Column vectors normalized to 1.

- Irreducible - All nodes are accessible (graph is connected)

- Aperiodic - Return to node i does not occur periodically

PageRank

- PageRank is just a random walk on the web graph

- The underlying intuitions that "A page is important if it is pointed to by other important pages"

$$\pi_i = \sum_j a_{ji} \frac{\pi_j}{k_j}$$

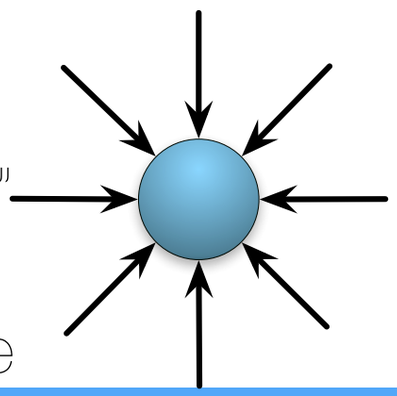
- The PageRank score of a page is the stationary state of a discrete time random walk on the network of pages and links $\vec{\pi}(t+1) = \mathcal{T}\,\vec{\pi}(t)$

- Where $\mathcal{T} = K_0^{-1} A^T$ is the operator that performs one step of the random walk

Problems:

"Dangling nodes"

Slow convergence



Solution:

Connect dangling nodes to every other node

Add damping factor (reset button)

Power Method

- The google transition operator (also called the google matrix) is:

$$\mathcal{G} = (1-\alpha)\,\mathcal{S} + \alpha\mathcal{E}$$

$\mathcal{E}$ - Fully connected matrix

α - damping factor

- The google matrix has a stationary state (converges) if:

$$\vec{\pi} = \mathcal{G}\,\vec{\pi}$$

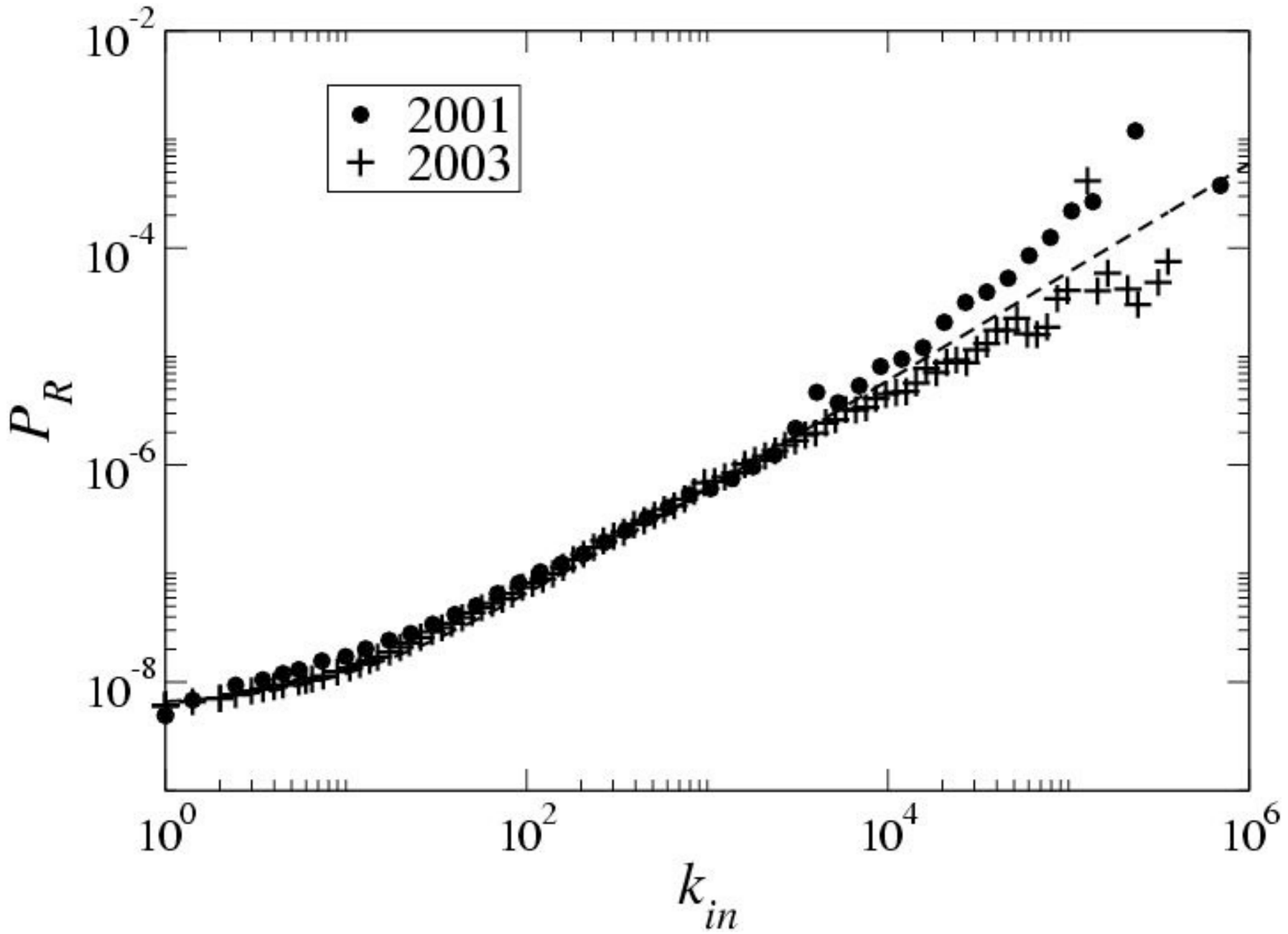
- The Power-Method consists of iterating this process until:

$$\vec{\pi}(t+1) \approx \vec{\pi}(t)$$

- This is equivalent to computing successive powers of

Web Graph

https://link.springer.com/chapter/10.1007%2F978-3-540-78808-9_6



Lesson IV:
A / B Testing

Hypothesis Testing

- The classical question we are trying to answer is: Is my intervention having an actual effect?

- Our hypothesis is that our intervention is effective

- The null-hypothesis is that there is no effect

- The main goal of Hypothesis Testing is to determine under what circumstances we can reject the null-hypothesis with a certain degree of certainty?

- In other words: How sure are we that we're not observing this difference just by chance (due to fluctuations as per the CLT)

- Select an appropriate test statistic to compare the two approaches

Hypothesis Testing

- In the case of binary outcomes, conversions follow a binomial distribution and the test statistic is the Z score:

$$Z = \frac{p_A - p_B}{\sqrt{SE_A^2 + SE_B^2}}$$

- where:

$$SE = \sqrt{\frac{p\left(1-p\right)}{N}}$$

- is the standard error for each instance.

- Under common assumptions, Z follows a Gaussian (normal) distribution centered at zero and with width one.

$$\mathcal{N}\left(0,1\right)$$

- Let's consider a practical example to clarify things

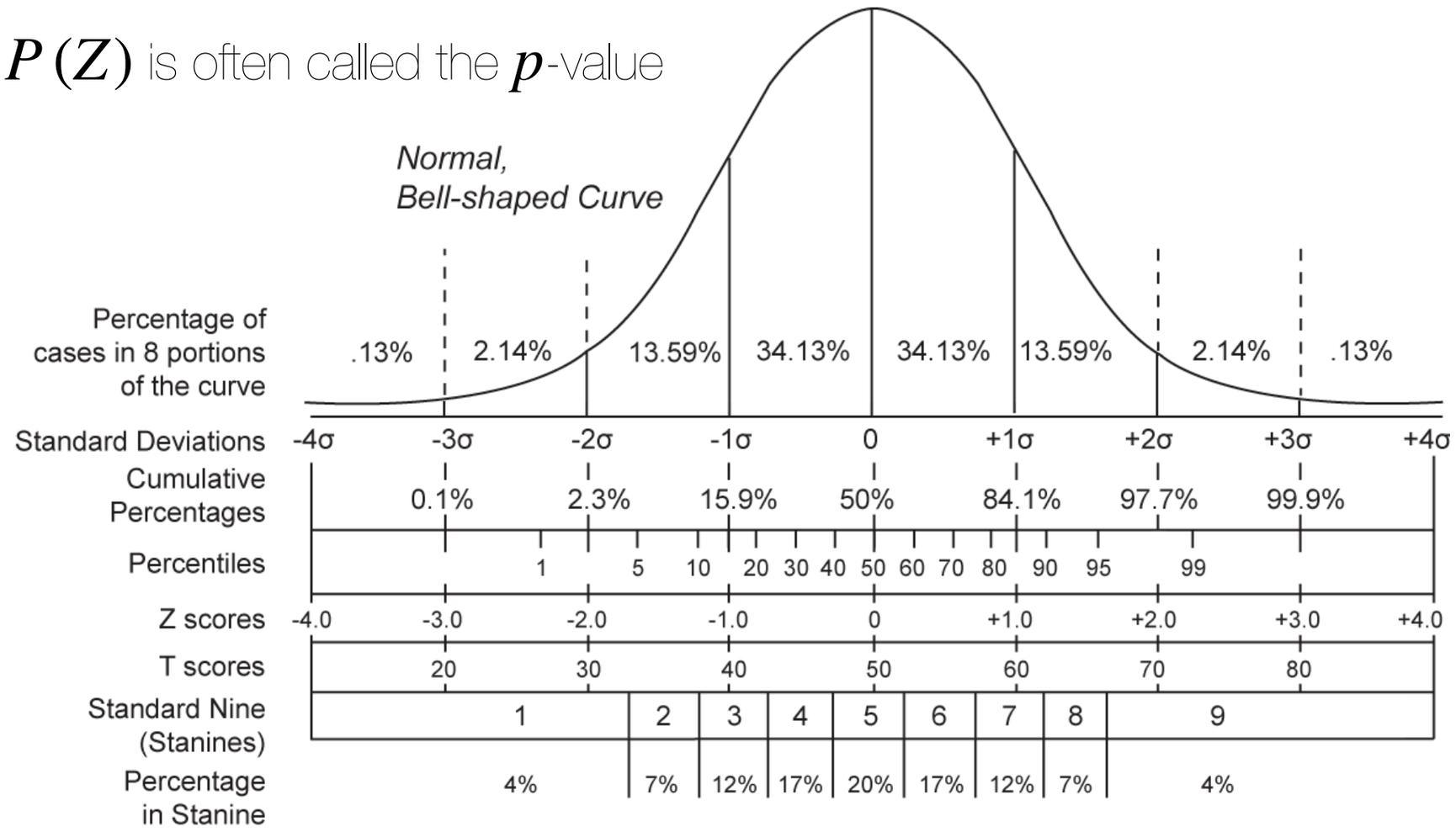
A/B Testing

- Which version of a headline results in more clicks?
- Divide users into two groups A and B and show each of them just one version
- Measure the click probability in each group, p_A and p_B
- The null hypothesis is that $p_A = p_B$. Can we reject it?

A/B Testing

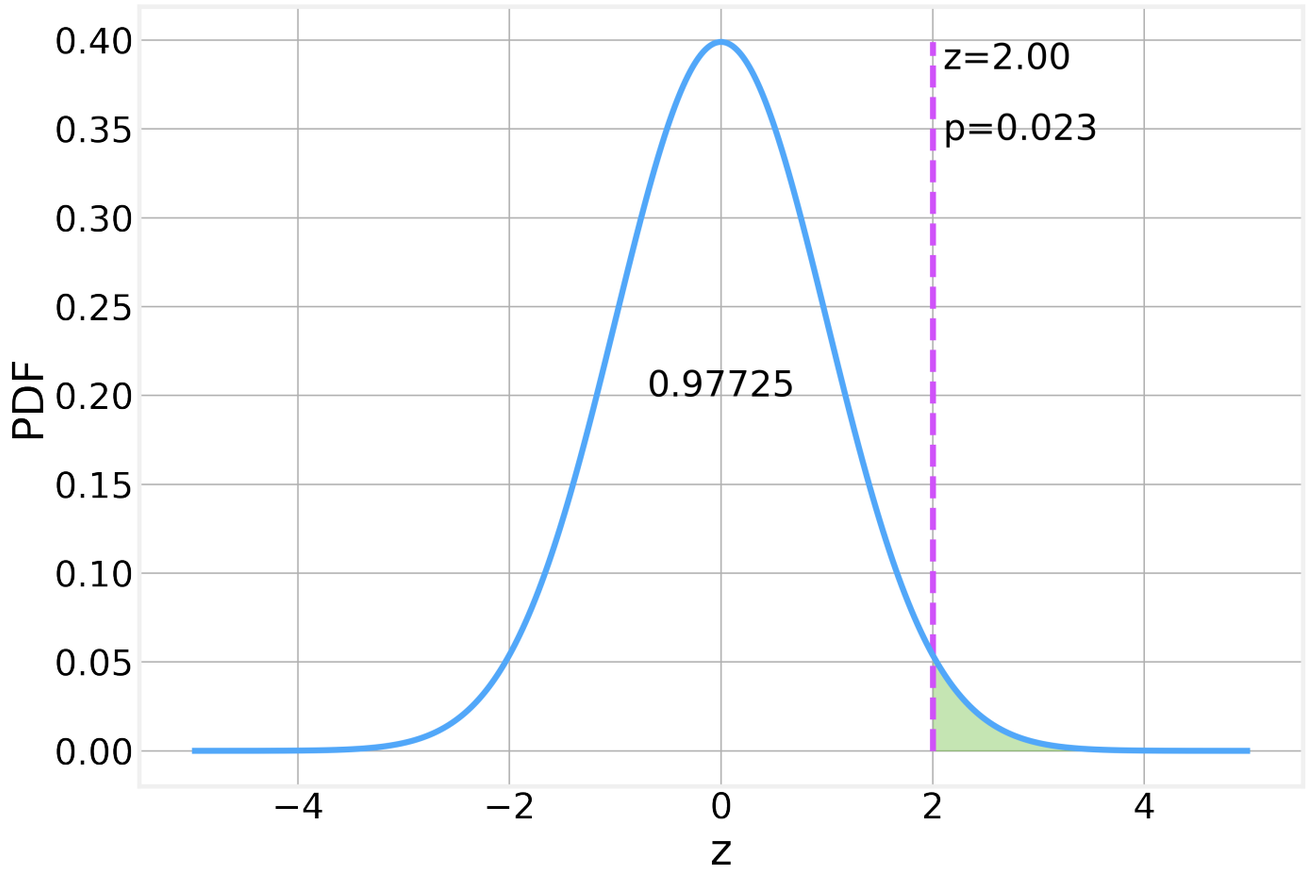
$$Z = \frac{p_A - p_B}{\sqrt{SE_A^2 + SE_B^2}}$$

- The value of $P(Z)$ effectively tells us how likely we are to observe this difference between p_A and p_B just due to sampling effects
- $P(Z)$ is often called the p-value



p-value

- Calculate the probability, p, of an event more extreme that the observation under the "null hypothesis"



- $p < 0.05$ Moderate
- $p < 0.01$ Strong
- $p < 0.001$ Very strong

evidence against the null-hypothesis

- The smaller the p-value the better.

Berkeley Discrimination Case Part I

	Candidates	Acceptance Rate
Men	8442	0.44
Women	4321	0.35

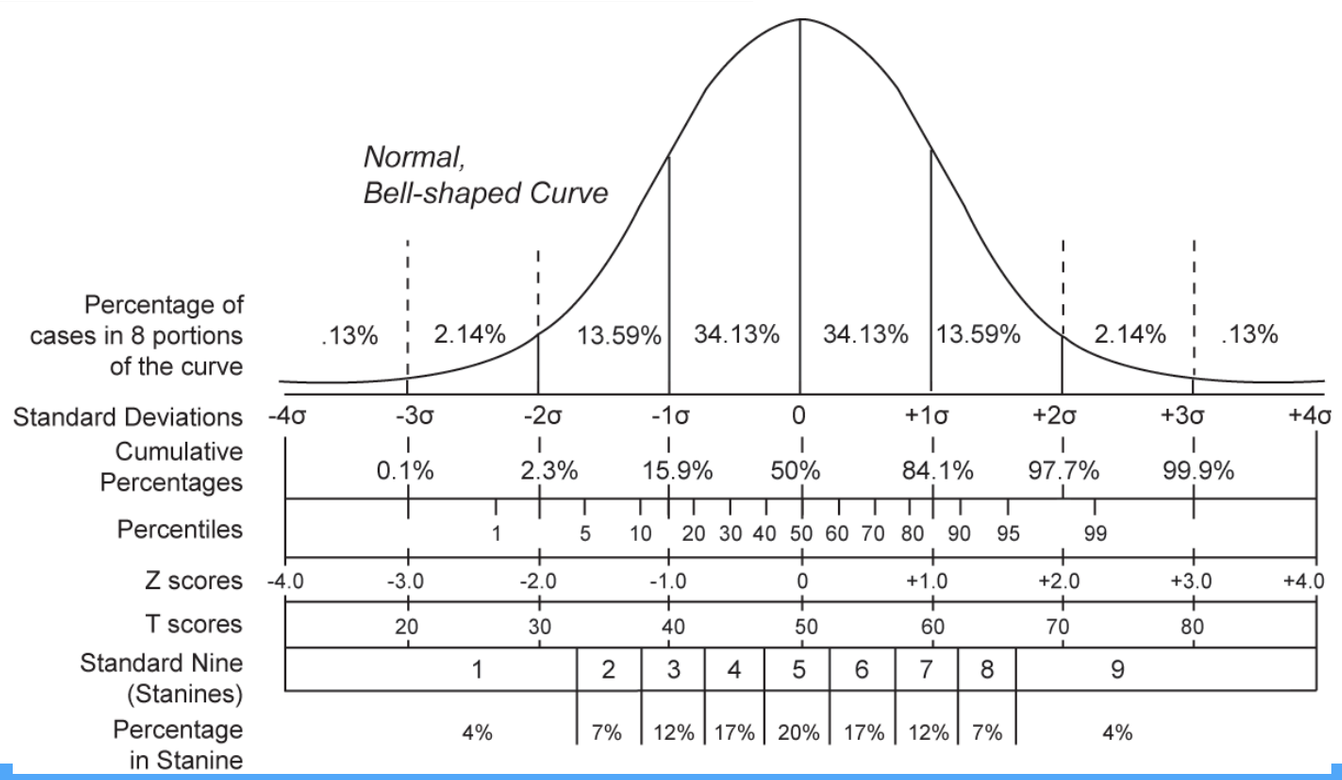
Were women being discriminated against when they applied to Berkley?

Berkeley Discrimination Case Part I

	Candidates	Acceptance Rate	SE
Men	8442	0.44	5.4×10^{-3}
Women	4321	0.35	7.2×10^{-3}

Were women being discriminated against when they applied to Berkley?

$$Z = \frac{p_A - p_B}{\sqrt{SE_A^2 + SE_B^2}}$$



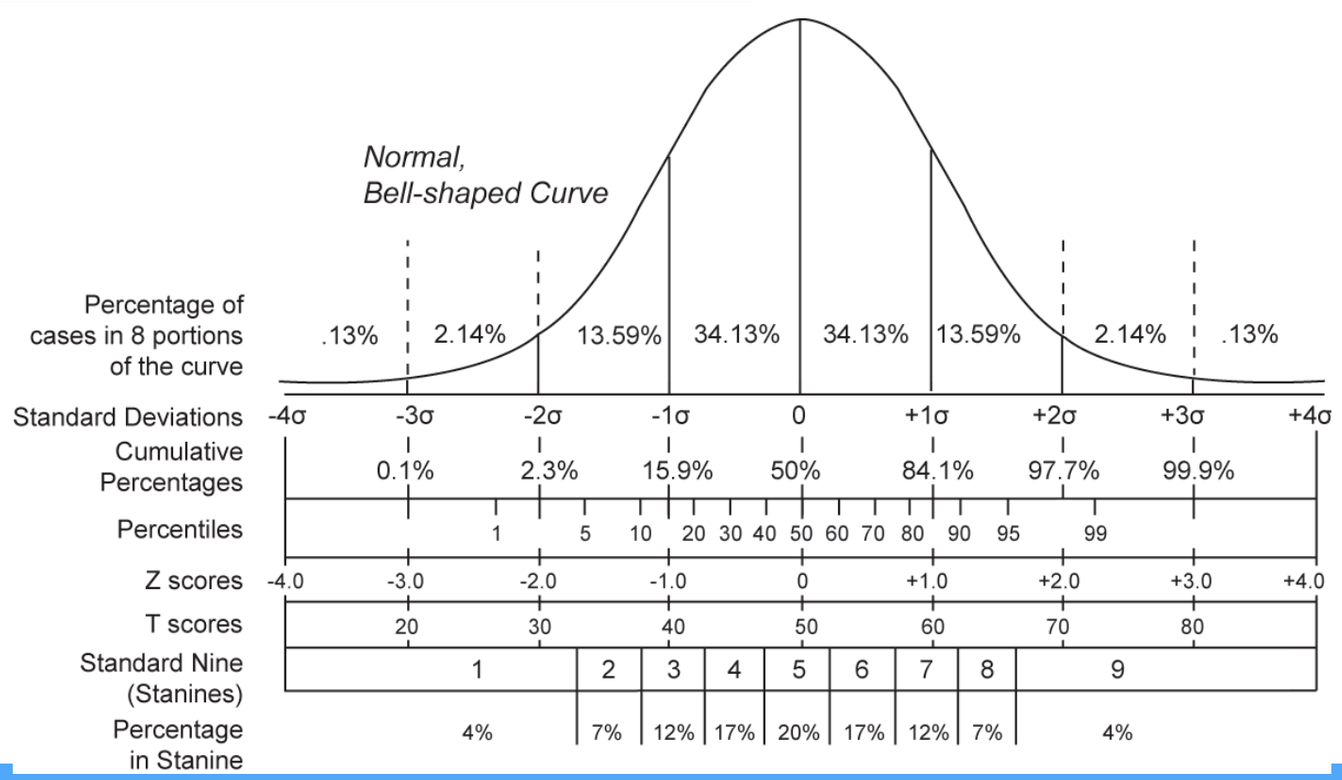
Berkeley Discrimination Case Part I

	Candidates	Acceptance Rate	SE
Men	8442	0.44	5.4×10^{-3}
Women	4321	0.35	7.2×10^{-3}

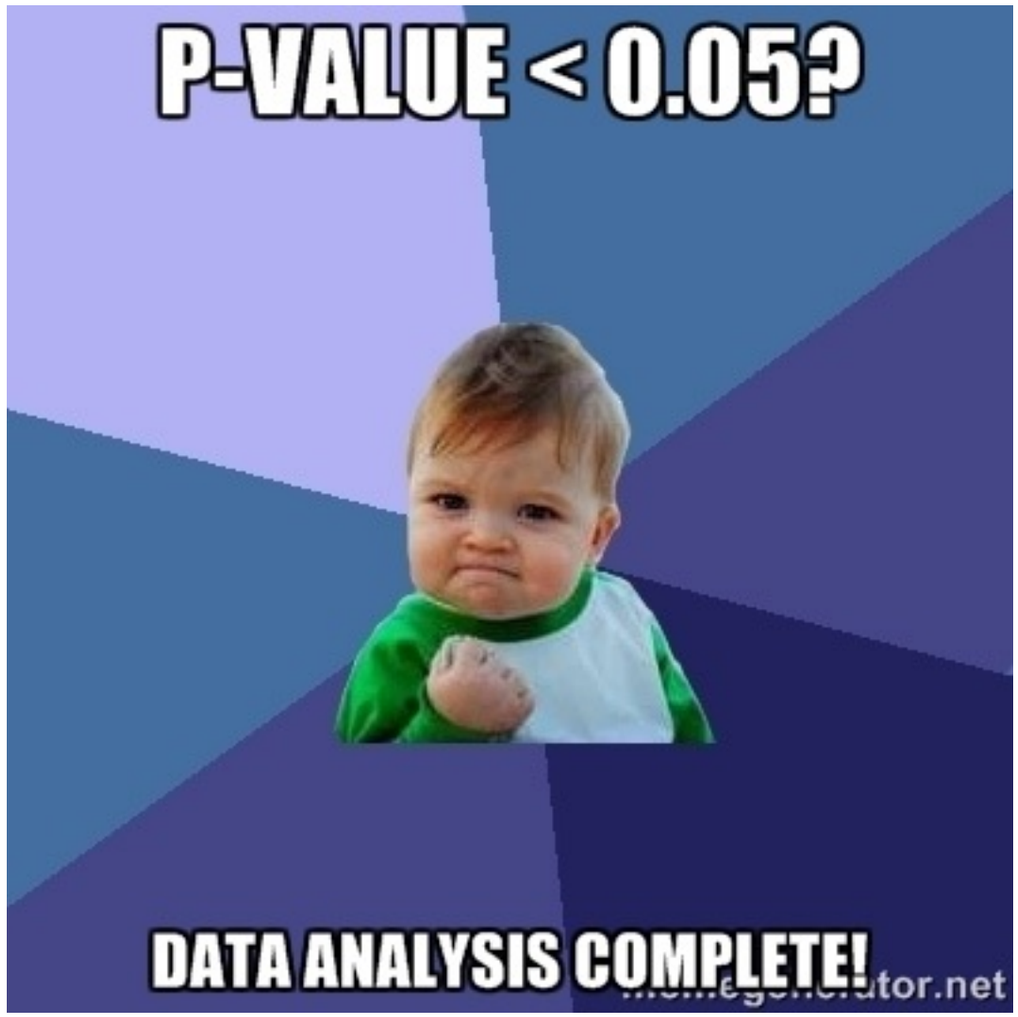
Were women being discriminated against when they applied to Berkley?

$$Z = \frac{p_A - p_B}{\sqrt{SE_A^2 + SE_B^2}}$$

$$p \approx 10^{-23}$$

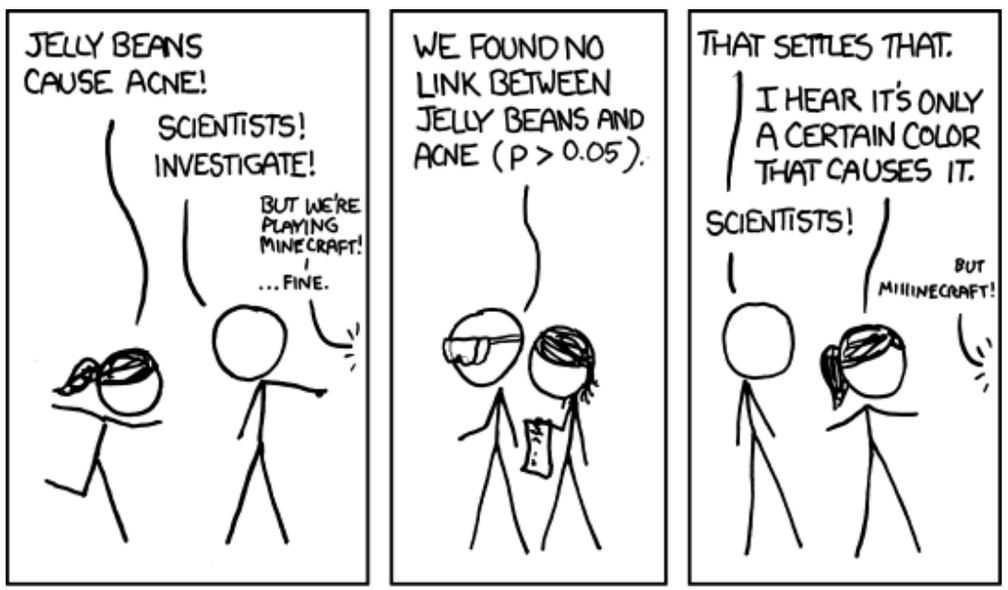


p-value

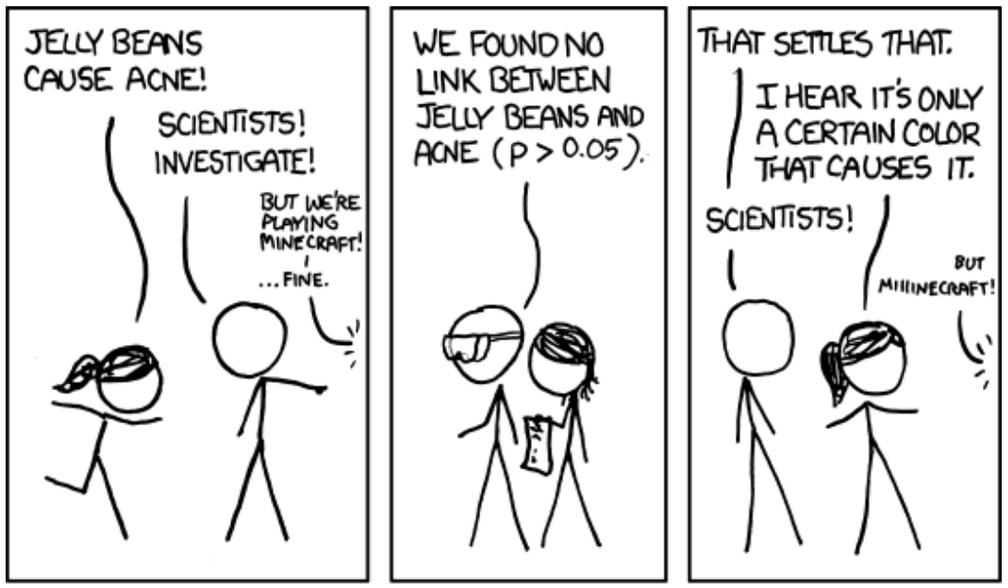


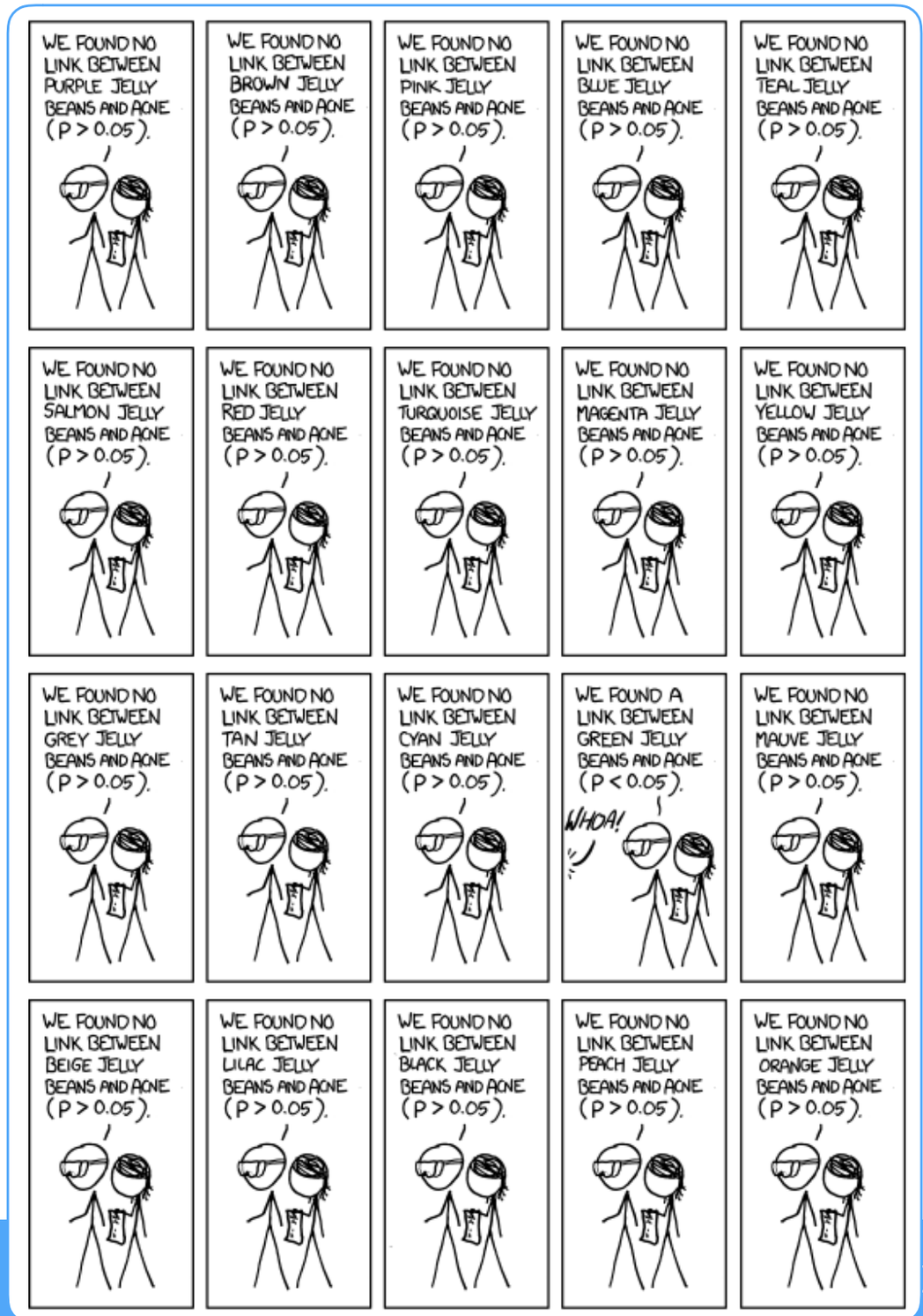
"Statistical significance does not imply scientific significance"

Bonferoni Correction

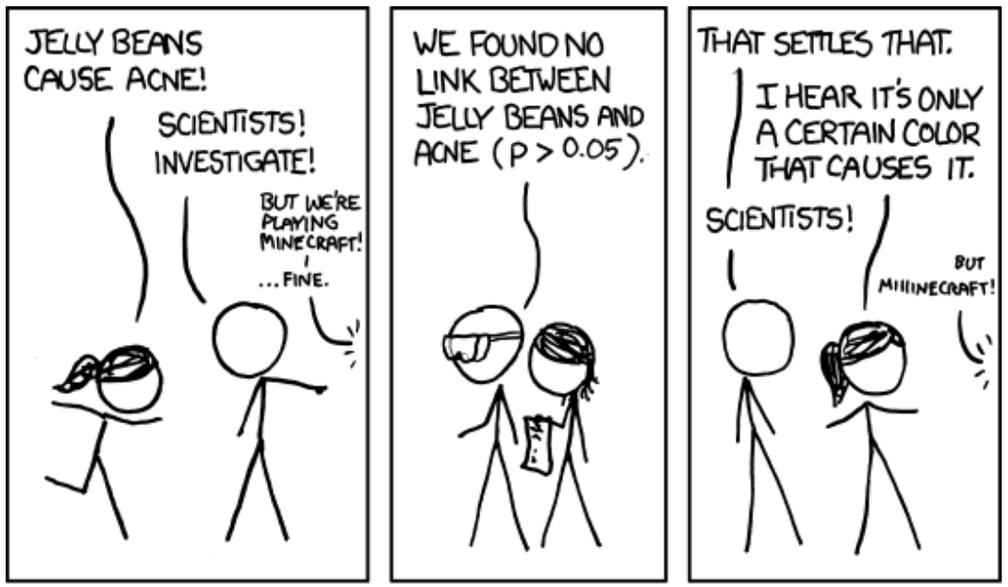


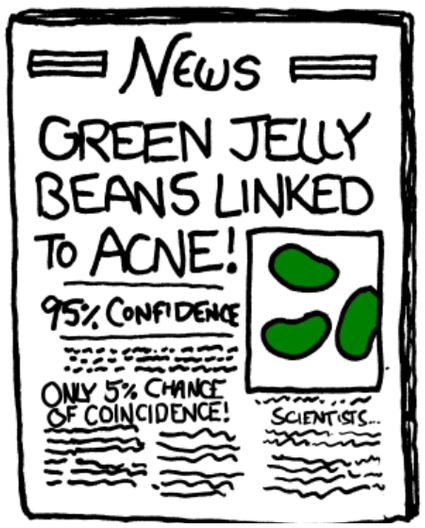
Bonferoni Correction

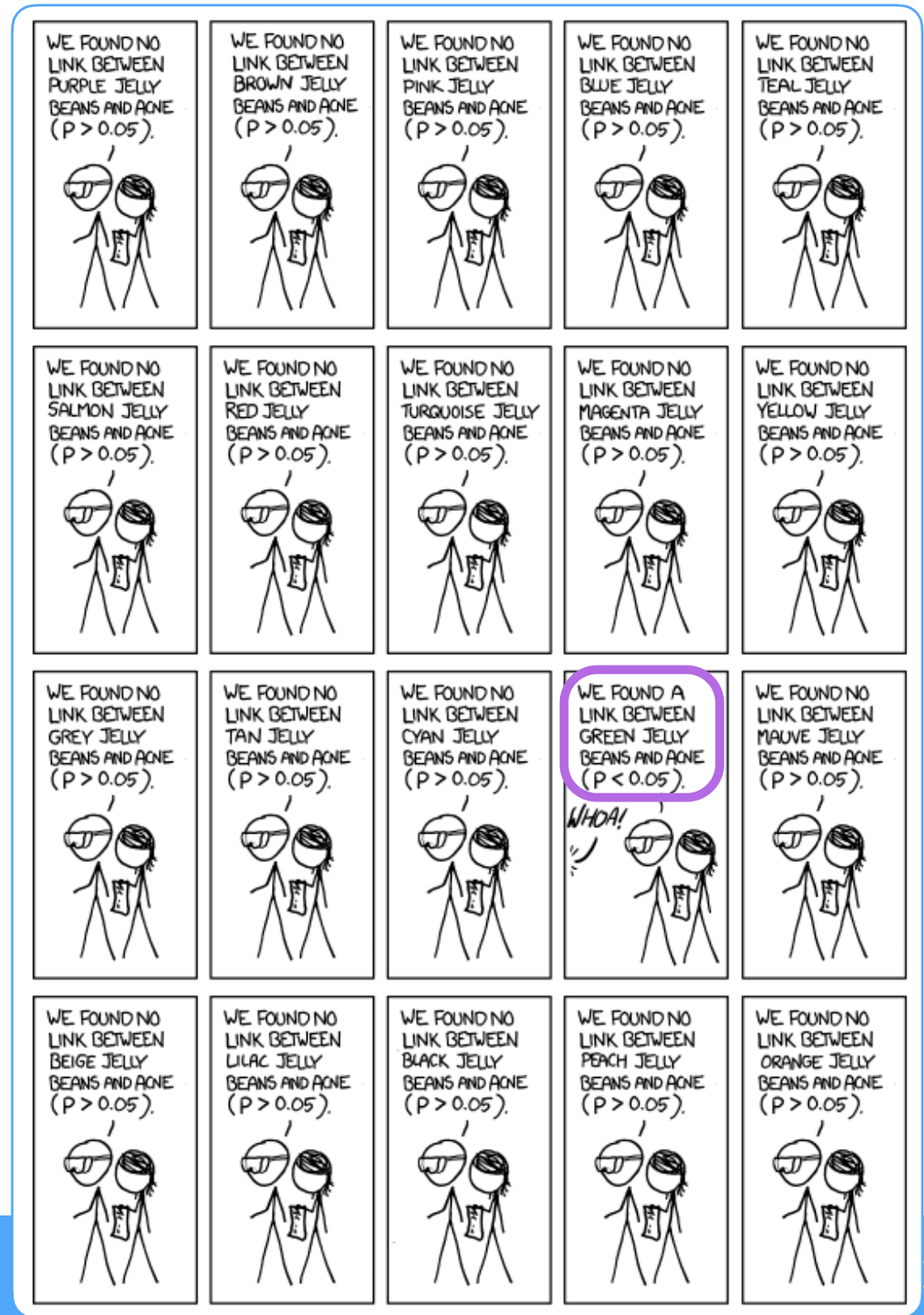




Bonferoni Correction







Bonferoni Correction

- You can think of the p-value as the probability of observing a result as extreme by chance. With n comparisons, this probability becomes:

$$p_n = 1 - \left(1 - p\right)^n$$

which quickly goes to 1 as n increases.

- However, if we replacing p by $\frac{p}{n}$ for each individual comparison, we obtain:

$$p_n = 1 - \left(1 - \frac{p}{n}\right)^n$$

- and for sufficiently large n:

$$p_n \approx 1 - e^{-p} \approx p$$

- allowing us to keep the probability of false results arbitrarily low even with arbitrarily large numbers of comparisons.

Simpsons Paradox

Science 187, 398 (1975)

	Candidates	Acceptance Rate
Men	8442	0.44
Women	4321	0.35

Berkeley Discrimination Case Part II:
The statisticians strike back.

Dept	Men		Women	
	Candidates	Acceptance	Candidates	Acceptance
A	825	0.62	108	0.82
B	560	0.63	25	0.68
C	325	0.37	594	0.34
D	417	0.33	375	0.35
E	191	0.28	393	0.24
F	272	0.06	341	0.07
	2590	**0.46**	**1835**	**0.30**

Simpsons Paradox

Science 187, 398 (1975)

	Candidates	Acceptance Rate
Men	8442	0.44
Women	4321	0.35

"aggregated data can appear
to reverse important trends
in the numbers being combined"
WSJ, Dec 2, 2009

Dept	Men		Women	
	Candidates	Acceptance	Candidates	Acceptance
A	825	0.62	108	**0.82**
B	560	0.63	25	**0.68**
C	325	0.37	594	0.34
D	417	0.33	375	0.35
E	191	**0.28**	393	0.24
F	272	**0.06**	341	0.07
	2590	**0.46**	**1835**	**0.30**

Thank You!

data4sci.com/newsletter

Natural Language Processing (NLP) from Scratch
Nov 11, 2019 - 7am-11am (PST)

Graphs and Network Algorithms from Scratch
Nov 18, 2019 - 5am-9am (PST)

Data Visualization with matplotlib and seaborn
Dec 4, 2019 - 5am-9am (PST)

Deep Learning From Scratch
Dec 11, 2019 - 5am-9am (PST)

Natural Language Processing (NLP) from Scratch
http://bit.ly/LiveLessonNLP - On Demand