Probability and Statistics for Everyone

Bruno Gonçalves

www.data4sci.com/newsletter

graphs4sci.substack.com

https://github.com/DataForScience/Probability-And-Statistics

Question

https://github.com/DataForScience/Probability-And-Statistics

- What's your job title?
    - Data Scientist
    - Statistician
    - Data Engineer
    - Researcher
    - Business Analyst
    - Software Engineer
    - Other

Question

https://github.com/DataForScience/Probability-And-Statistics

- How experienced are you in Python?
    - Beginner (<1 year)
    - Intermediate (1-5 years)
    - Expert (5+ years)

Question

https://github.com/DataForScience/Probability-And-Statistics

- How did you hear about this webinar?
    - O'Reilly Platform
    - Newsletter
    - data4sci.com Website
    - Previous event
    - Other?

Table of Contents

1. Descriptive Statistics

COMPUTERWORLD

WHY BIG DATA IS A BIG DEAL

COVER STORY BI & ANALYTICS

COMMUNICATIONS OF THE ACM

McKinsey Quarterly

nature

THE BITER BIT
Viral infections for viruses

TROPICAL CYCLONES
The strong get stronger

BLACK HOLE PHYSICS
A new window on the Galactic Centre

Science

climate information
research

NUMÉRO SPÉCIAL
POUR LA SCIENCE
BANG

DER SPIEGEL

Pipeline | Volume 10 | Issue 4
Pipeline
RIDING THE BIG DATA
Big Data M
BIG GOVERN

interactions
EXPERIENCE | TECHNOLOGY
Interactions with Big Data

Sci

Nuclear Monarchy · Advanced Leadership
HARVARD MAGAZINE

Computing
SCIENCE & ENGINEERING
BIG

A tale of two Libyas
TIME
Owns a laptop

WINTER 2013
Cornell University School of Electrical and Computer Engineering
ece connections
Data Replica

InformationWeek

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL

Harvard Business Review
GETTING CONTROL OF BIG DA

SPECIAL ISSUE DESIGN SCI-TECH
WIRED

UCSF

cnrs
n° 28
Quarterly
January 2013

significance
statistics making sense
BIG DATA

Also inside: You Want Me in the Office? How 20th Century of You. 16
Cloud Computing's Big Debt to NASA. 10
COMPUTERWORLD

PREMIUM
BIG DATA

May 2013
NACS
PLUS:
Maverik's New Debit Card
Candy's Sweet Spot

Government Executive

THE FUTURIST
Mapping the Future with Big Data

The Economist
The data delu

COSMO
THE SCIENCE OF EVERYTHING

GEOSPATIAL WORLD

Crossroads The ACM Magazine for Students
XRDS

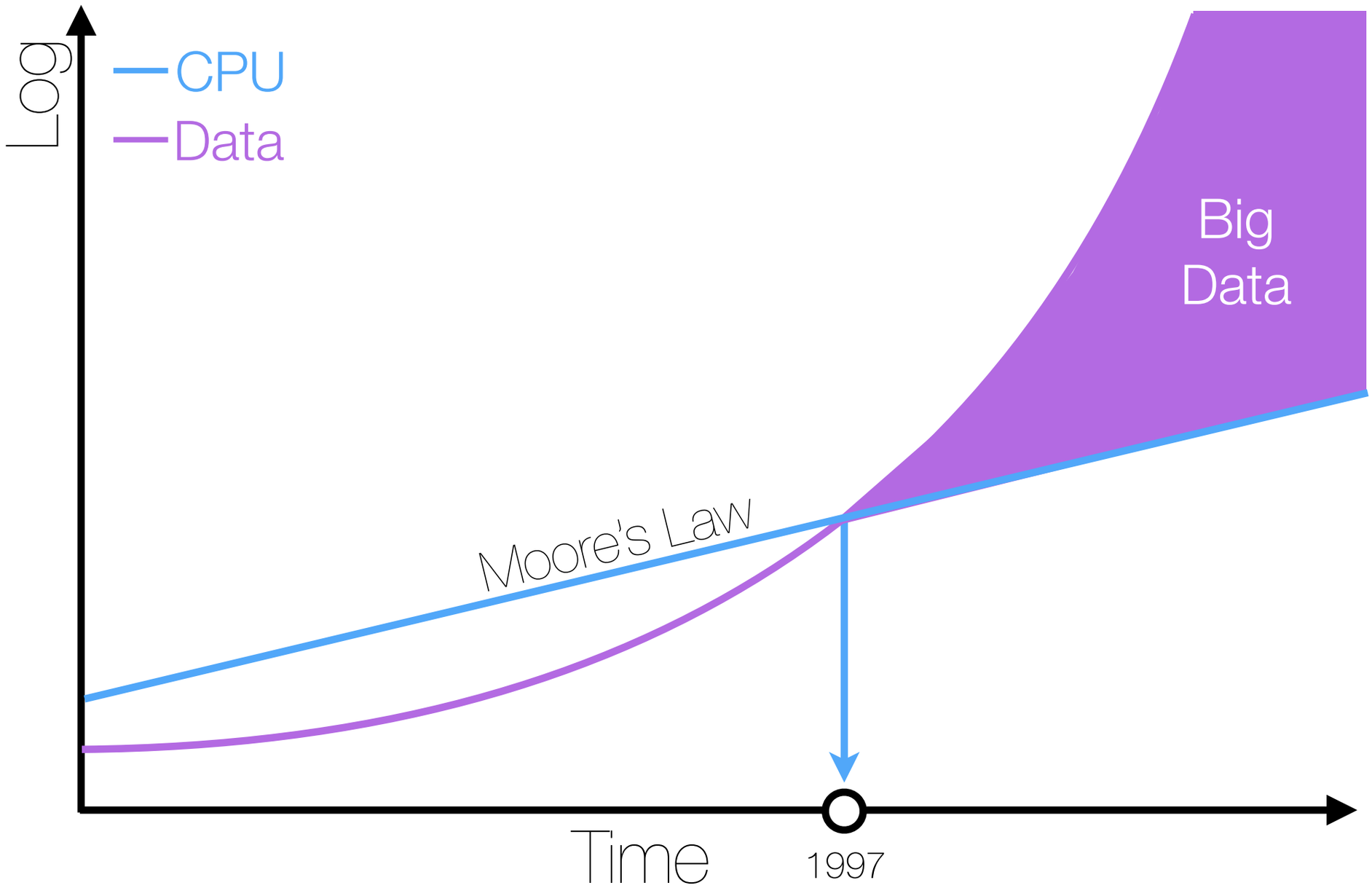
BEST'S REVIEW
Monthly Insurance News Magazine

The Washington and Lee School of Law Magazine
Summer 2014
W&L LAW

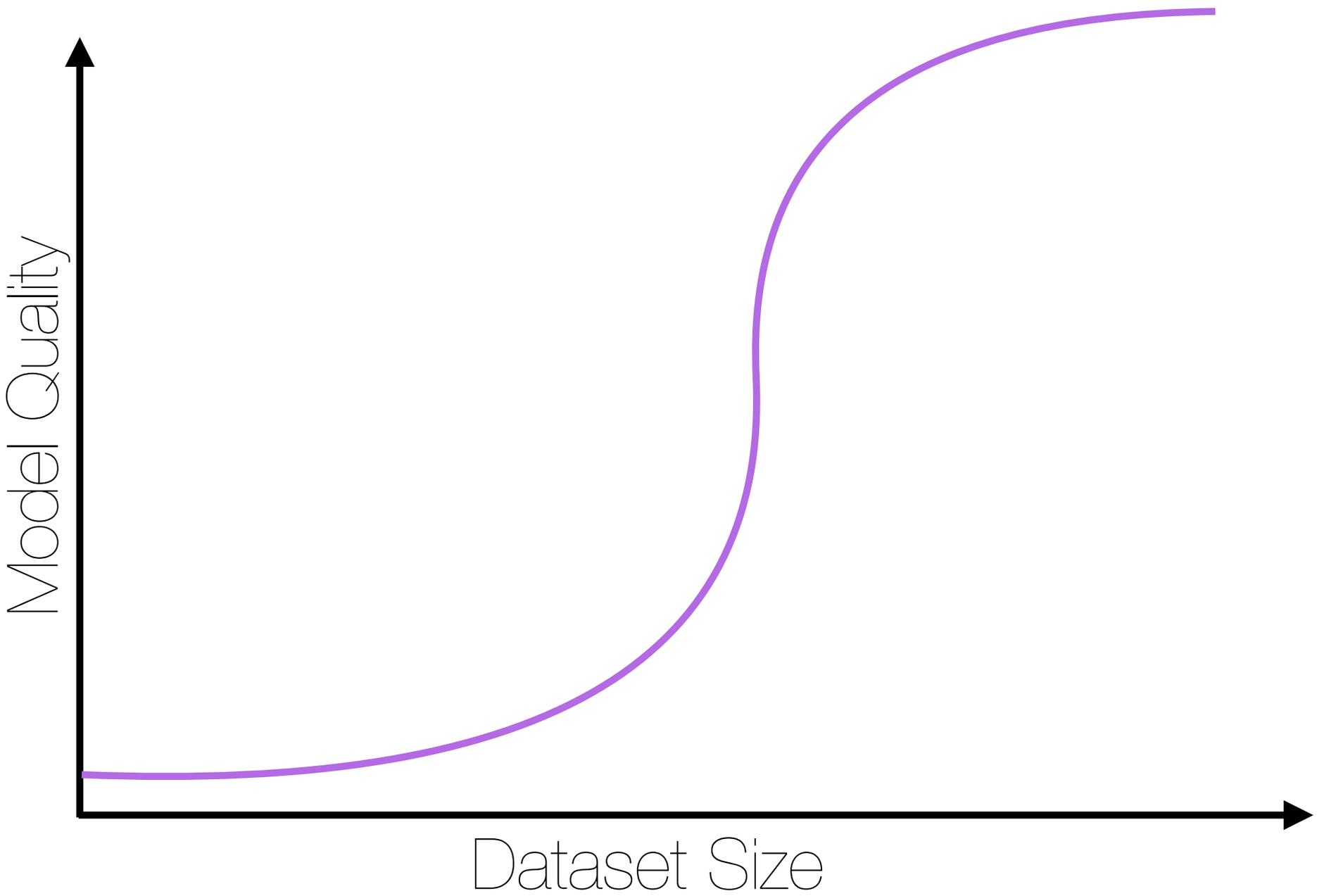
Pipeline | Volume 11 | Issue 3
Pipeline
The Unstructured
GET SMARTER!
NEXT-GEN

A BUSINESS REPORT ON
Big Data Gets Personal

Big Data



Big Data



https://www.wired.com/2008/06/pb-theory/

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

Illustration: Marian Bantjes

Wired Articles

"**All models are wrong**, but some are useful."

https://www.wired.com/2008/06/pb-theory/

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

Illustration: Marian Bantjes

"All models are wrong, but some are useful."

Big Data is a Junkyard

https://bgoncalves.medium.com/big-data-is-a-junkyard-5acaafc9dda6

Big Data is a Junkyard

https://bgoncalves.medium.com/big-data-is-a-junkyard-5acaafc9dda6

- Lots of data (cars)
- Not representative:
    - Biased to the local market (not many American cars in European junkyards and vice versa)
    - Might have a bit of everything, but not in the right proportions
    - Different junkyards (datasets) will have a different mix of the same basic ingredients

Big Data is a Junkyard

http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/

Twitter Demographics
Among internet users, the % who use Twitter

	Internet users
Total	*23%*
Men	25
Women	21
White, Non-Hispanic	20
Black, Non-Hispanic (n=85)	28
Hispanic	28
18-29	32
30-49	29
50-64	13
65+	6
High school grad or less	19
Some college	23
College+	27
Less than $30,000/yr	21
$30,000-$49,999	19
$50,000-$74,999	25
$75,000+	26
Urban	30
Suburban	21
Rural	15

Source: Pew Research Center, March 17-April 12, 2015.

PEW RESEARCH CENTER

LinkedIn Demographics
Among internet users, the % who use LinkedIn

	Internet users
Total	*25%*
Men	26
Women	25
White, Non-Hispanic	26
Black, Non-Hispanic (n=94)	22
Hispanic (n=99)	22
18-29	22
30-49	32
50-64	26
65+	12
High school grad or less	9
Some college	25
College+	46
Less than $30,000/yr	17
$30,000-$49,999	21
$50,000-$74,999	32
$75,000+	41
Employed	32
Not employed*	14
Urban	30
Suburban	26
Rural	12

Source: Pew Research Center, March 17-April 12, 2015.

*Not employed includes those who are retired, not employed for pay, disabled, or students.

PEW RESEARCH CENTER

Pinterest Demographics
Among internet users, the % who use Pinterest

	Internet users
Total	*31%*
Men	16
Women	44
White, Non-Hispanic	32
Black, Non-Hispanic (n=85)	23
Hispanic	32
18-29	37
30-49	36
50-64	24
65+	16
High school grad or less	25
Some college	37
College+	31
Less than $30,000/yr	24
$30,000-$49,999	37
$50,000-$74,999	41
$75,000+	30
Urban	26
Suburban	34
Rural	31

Source: Pew Research Center, March 17-April 12, 2015.

PEW RESEARCH CENTER

Big Data is a Junkyard

http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/

Twitter Demographics
Among internet users, the % who use Twitter

	Internet users
Total	*23%*
Men	25
Women	21
White, Non-Hispanic	20
Black, Non-Hispanic (n=85)	28
Hispanic	28
18-29	32
30-49	
50-64	
65+	
High school grad or less	
Some college	
College+	27
Less than $30,000/yr	21
$30,000-$49,999	19
$50,000-$74,999	25
$75,000+	26
Urban	30
Suburban	21
Rural	15

Source: Pew Research Center, March 17-April 12, 2015.

PEW RESEARCH CENTER

LinkedIn Demographics
Among internet users, the % who use LinkedIn

	Internet users
Total	*25%*
Men	26
Women	25
White, Non-Hispanic	26
Black, Non-Hispanic (n=94)	22
Hispanic (n=99)	22
18-29	22
College+	46
Less than $30,000/yr	17
$30,000-$49,999	21
$50,000-$74,999	32
$75,000+	41
Employed	32
Not employed*	14
Urban	30
Suburban	26
Rural	12

Source: Pew Research Center, March 17-April 12, 2015.

*Not employed includes those who are retired, not employed for pay, disabled, or students.

PEW RESEARCH CENTER

Pinterest Demographics
Among internet users, the % who use Pinterest

	Internet users
Total	*31%*
Men	16
Women	44
White, Non-Hispanic	32
Black, Non-Hispanic (n=85)	23
Hispanic	32
18-29	37
30-49	36
50-64	24
65+	16
High school grad or less	25
Some college	37
College+	31
Less than $30,000/yr	24
$30,000-$49,999	37
$50,000-$74,999	41
$75,000+	30
Urban	26
Suburban	34
Rural	31

Source: Pew Research Center, March 17-April 12, 2015.

PEW RESEARCH CENTER

This is why we need statistics!

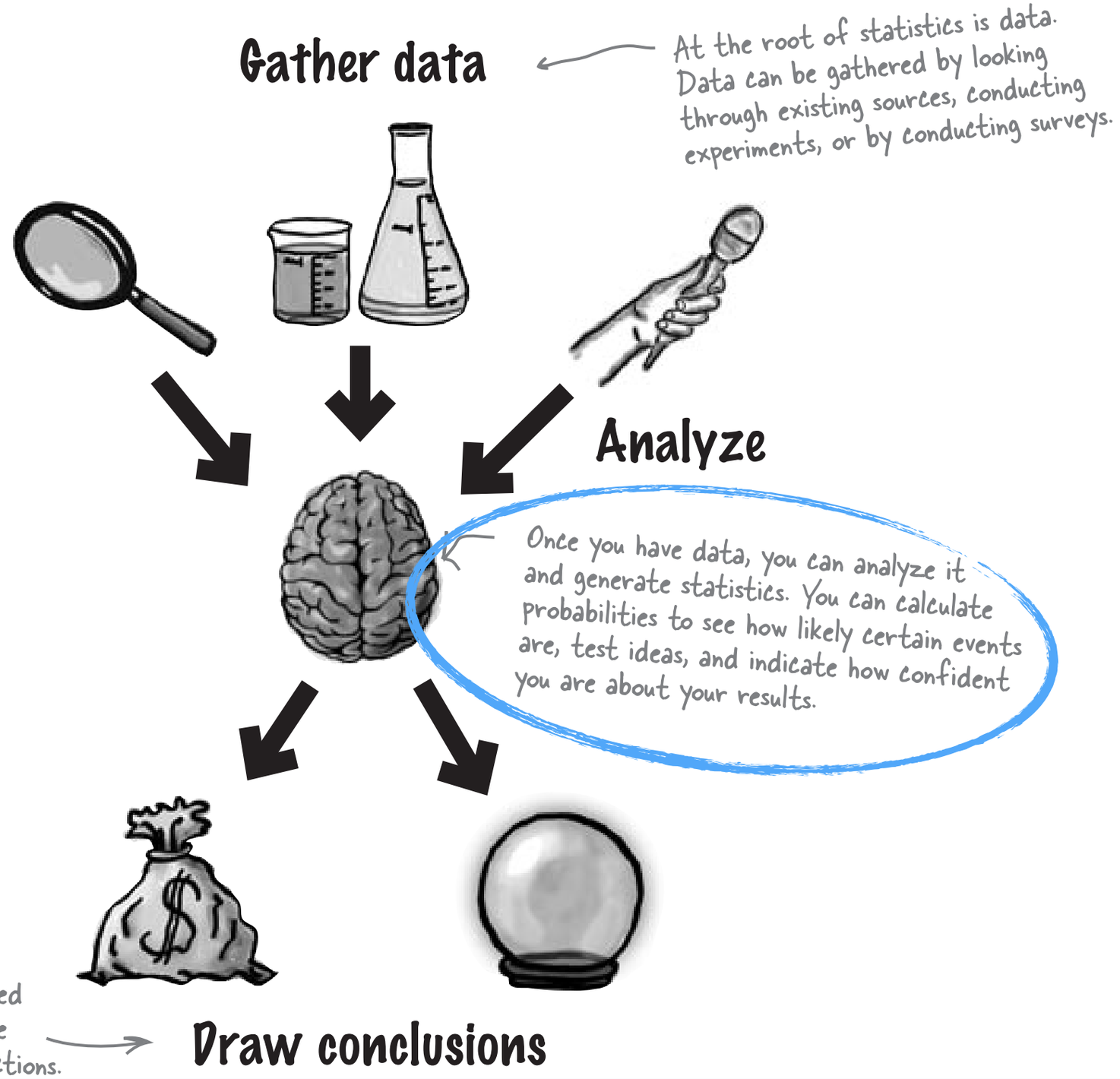
EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

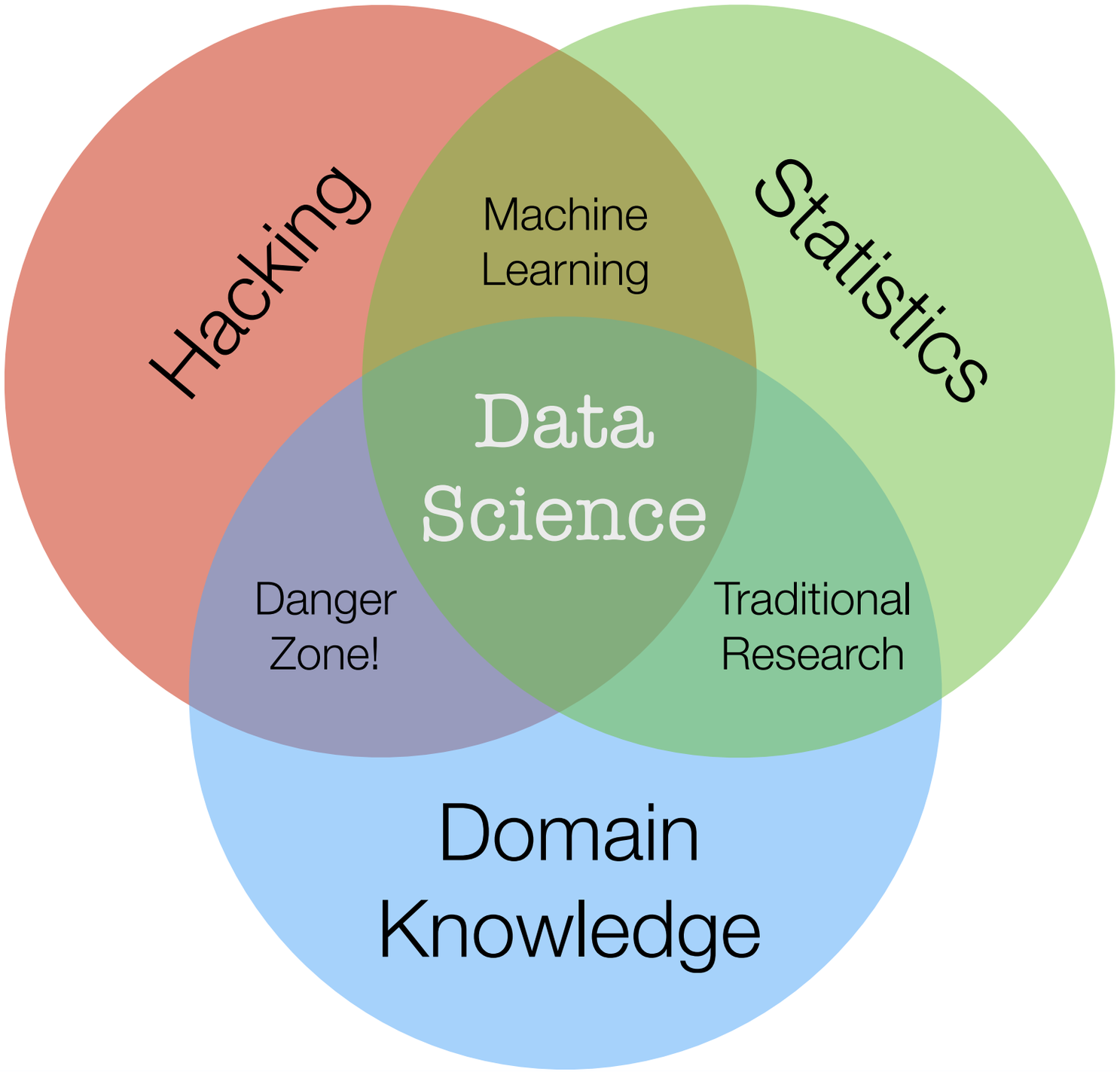
The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

From Data To Information



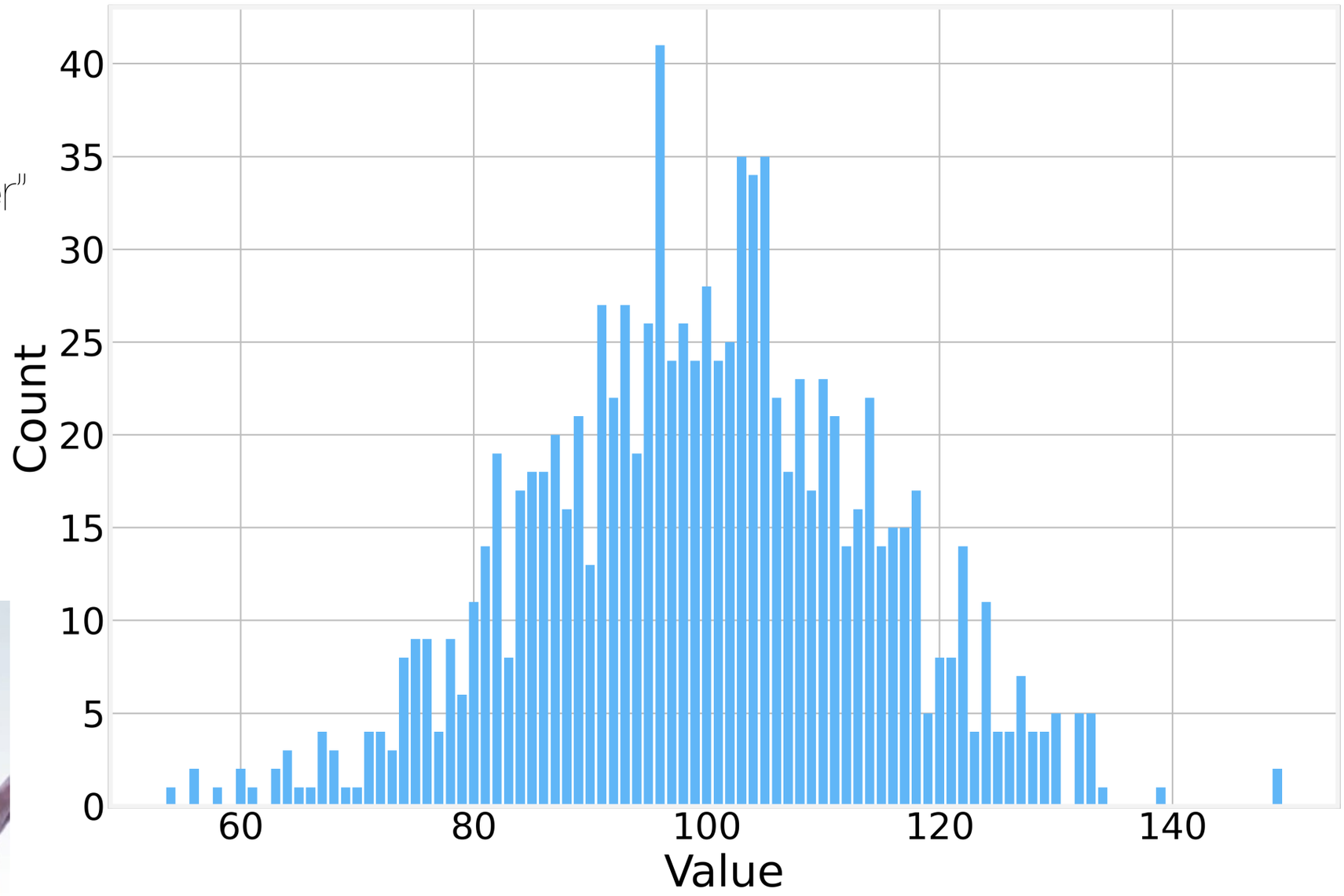
Data Science



Count!

- How many items do we have?

"Zero is the most natural number"
(E. W. Dijkstra)



Descriptive Statistics

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

Standard Deviation $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$

Min

Max

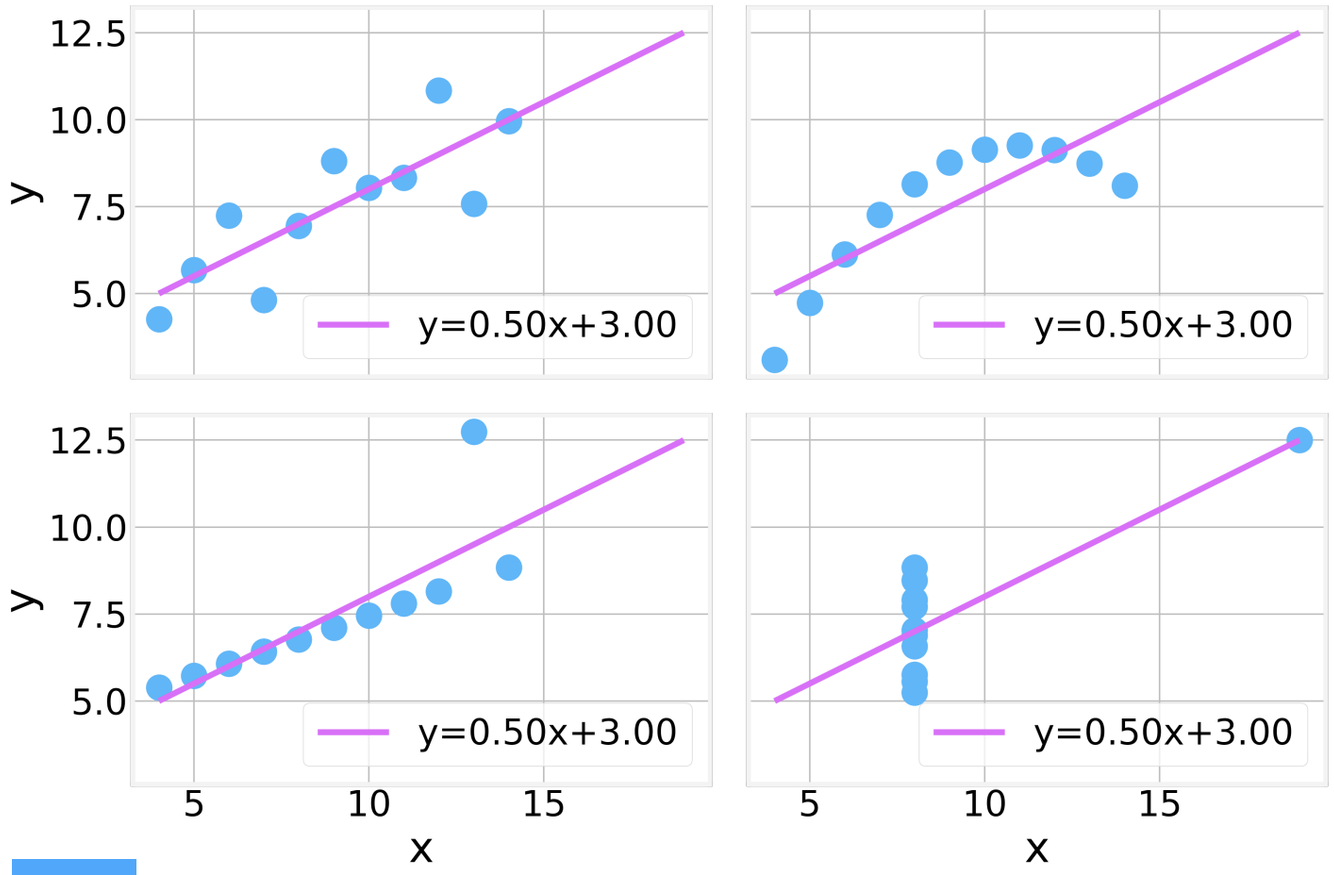
Count

Value

Anscombe's Quartet

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89



μ_x	9
σ_x	11
μ_y	7.5
σ_y	~4.125
ρ	0.816
fit	y=3+0.5

Outliers

- "Bill Gates walks into a bar and on average every patron is a millionaire…"

> **Median** - "the value that separates the lower 50% of the distribution from the higher 50%"

- …but the median remains the same"

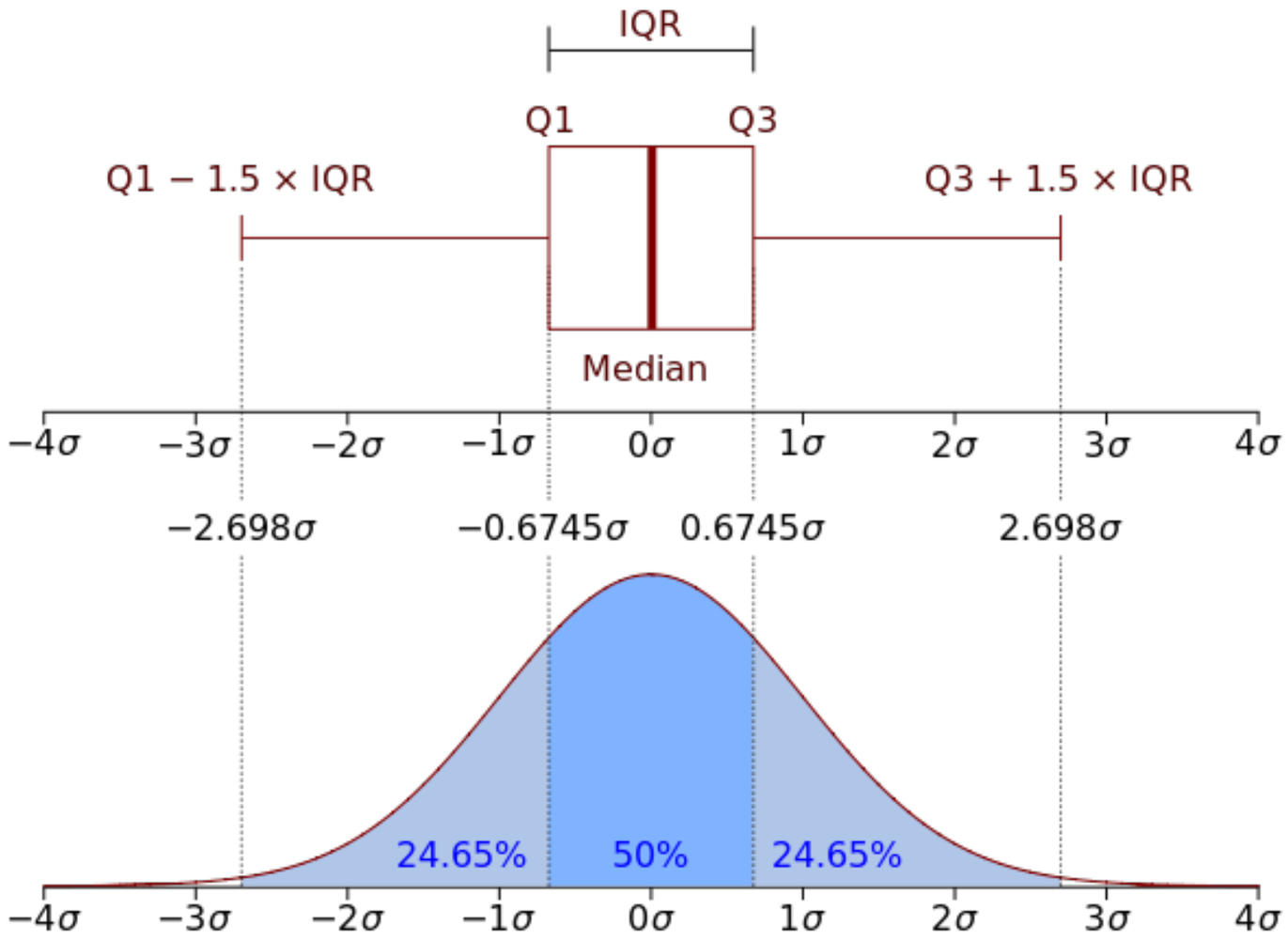
1 1 1 **2** 2 2 1000

- Mean = 144.14
  Median = 2

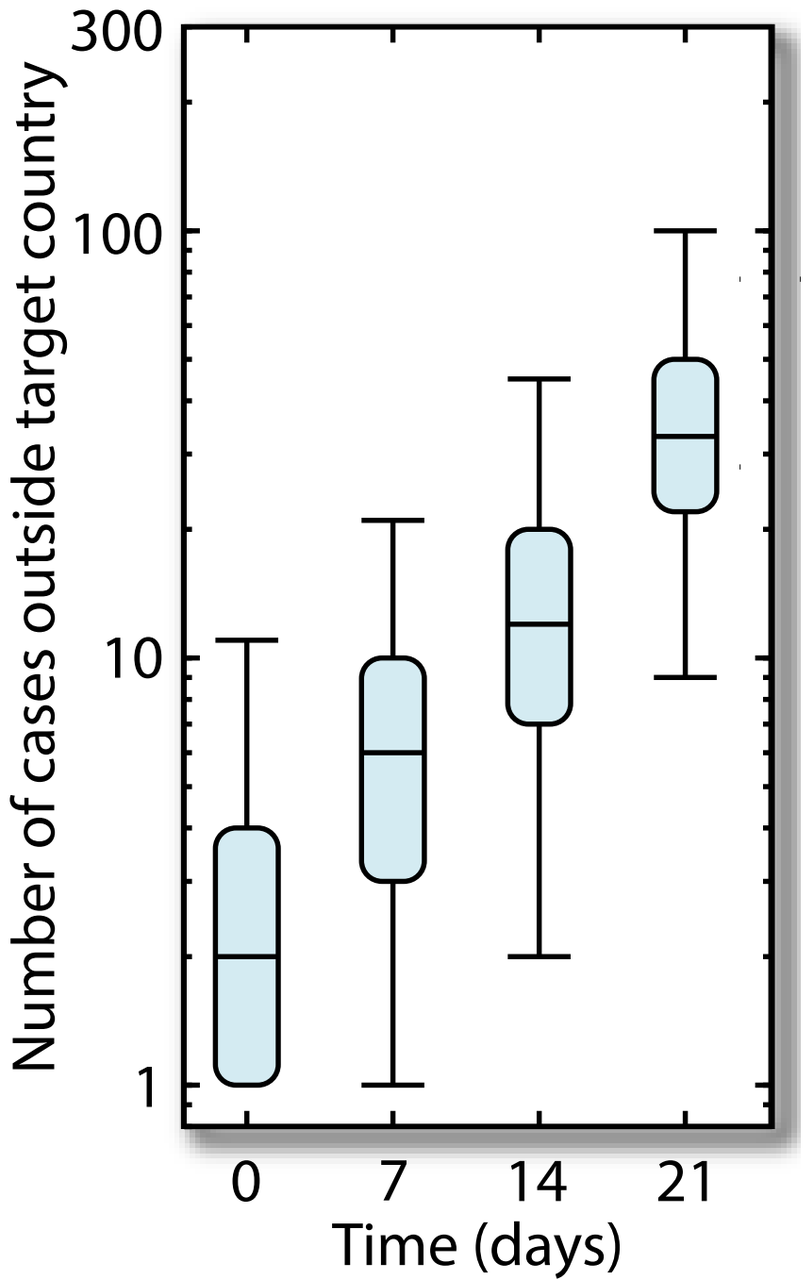
Quantiles

- Quantiles - Points taken at regular intervals of the cumulative distribution function

- Quartiles - Ranked set of points that divide the range in 4 equal intervals (25%, 50%, 75% quantiles)



Box and Whisker Plots



- Show the variation of the data for each bin.
- More informative than just averages or medians.
- Useful to summarize experimental measurements, simulation results, natural variations, etc… when fluctuations are important

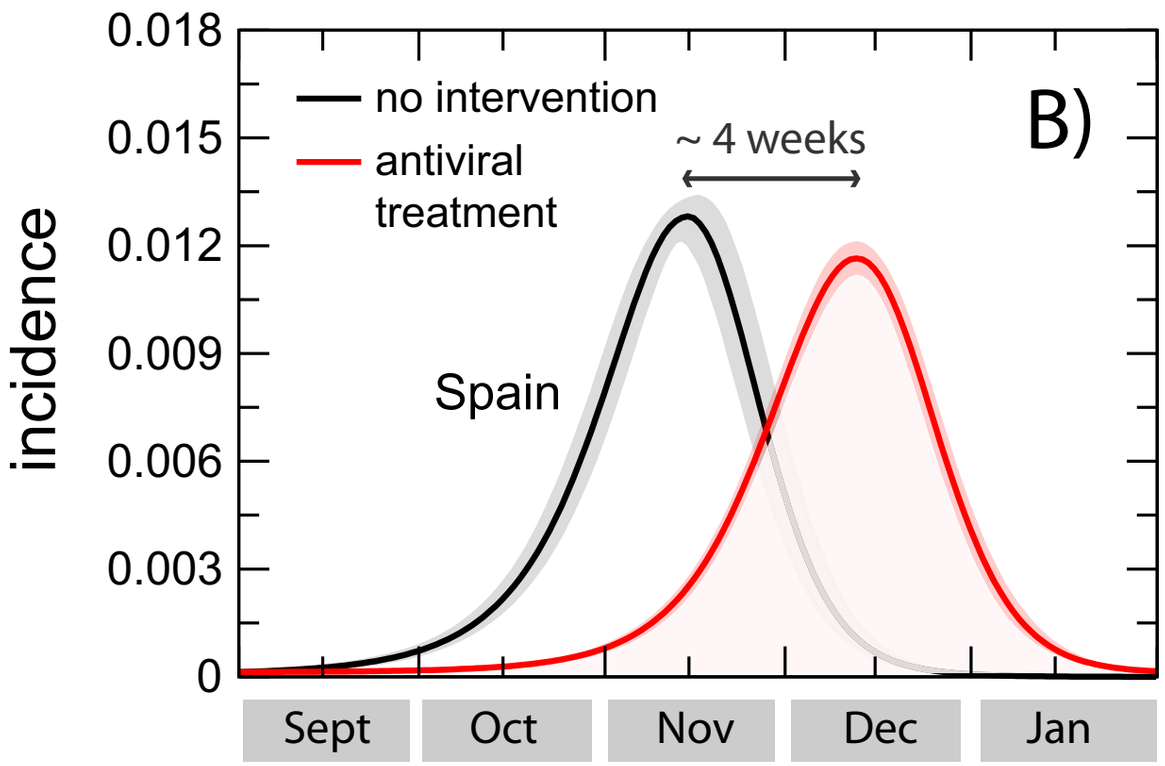
Reference Range

97.5% Percentile

95% RR { — Median

2.5% Percentile

- Useful for continuous curves
- Indicates level of certainty
  "95% of the cases are in this range"



Tools for Statistical Analysis

Name	Advantages	Disadvantages	Open Source
R	Library support and Visualization	Steep learning curve	Yes
Matlab	Native matrix support, Visualization	Expensive, incomplete statistics support	No
Scientific Python	Ease and Simplicity	Heavy development	Yes
Excel	Easy, Visual, Flexible	Large datasets	No
SAS	Large Datasets	Expensive, outdated programming language	No
Stata	Easy Statistical Analysis		No
SPSS	Like Stata but more expensive and less flexible		

Tools for Statistical Analysis

Name	Advantages	Disadvantages	Open Source
R	Library support and Visualization	Steep learning curve	Yes
Matlab	Native matrix support, Visualization	Expensive, incomplete statistics support	No
Scientific Python	Ease and Simplicity	Heavy development	Yes
Excel	Easy, Visual, Flexible	Large datasets	No
SAS	Large Datasets	Expensive, outdated programming language	No
Stata	Easy Statistical Analysis		No
SPSS	Like Stata but more expensive and less flexible		

Tools for Statistical Analysis

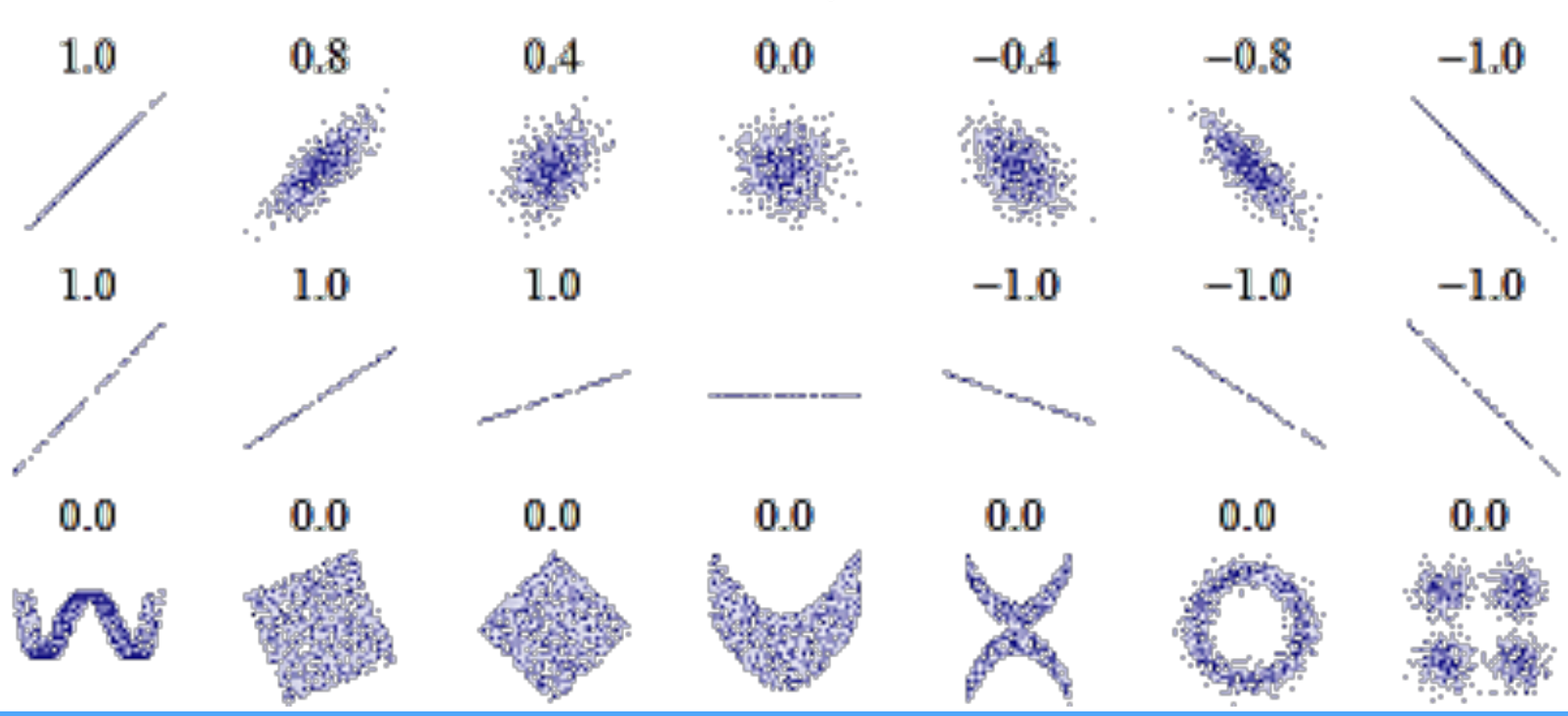
Name	Advantages	Disadvantages	Open Source
R	Library support and Visualization	Steep learning curve	Yes
Matlab	Native matrix support, Visualization	Expensive, incomplete statistics support	No
Scientific Python	Ease and Simplicity	Heavy development	Yes
Excel	Easy, Visual, Flexible	Large datasets	No
SAS	Large Datasets	Expensive, outdated programming language	No
Stata	Easy Statistical Analysis		No
SPSS	Like Stata but more expensive and less flexible		

Correlation

- Many correlation measures have been proposed over the years
- The most well known one is the Pearson Correlation

$$\rho\left(x,y\right) = \sum_{i=1}^{N} \frac{\left(x_i - \mu_x\right)\left(y_i - \mu_y\right)}{\sigma_x \sigma_y}$$

- Assumes a linear relationship between x and y.

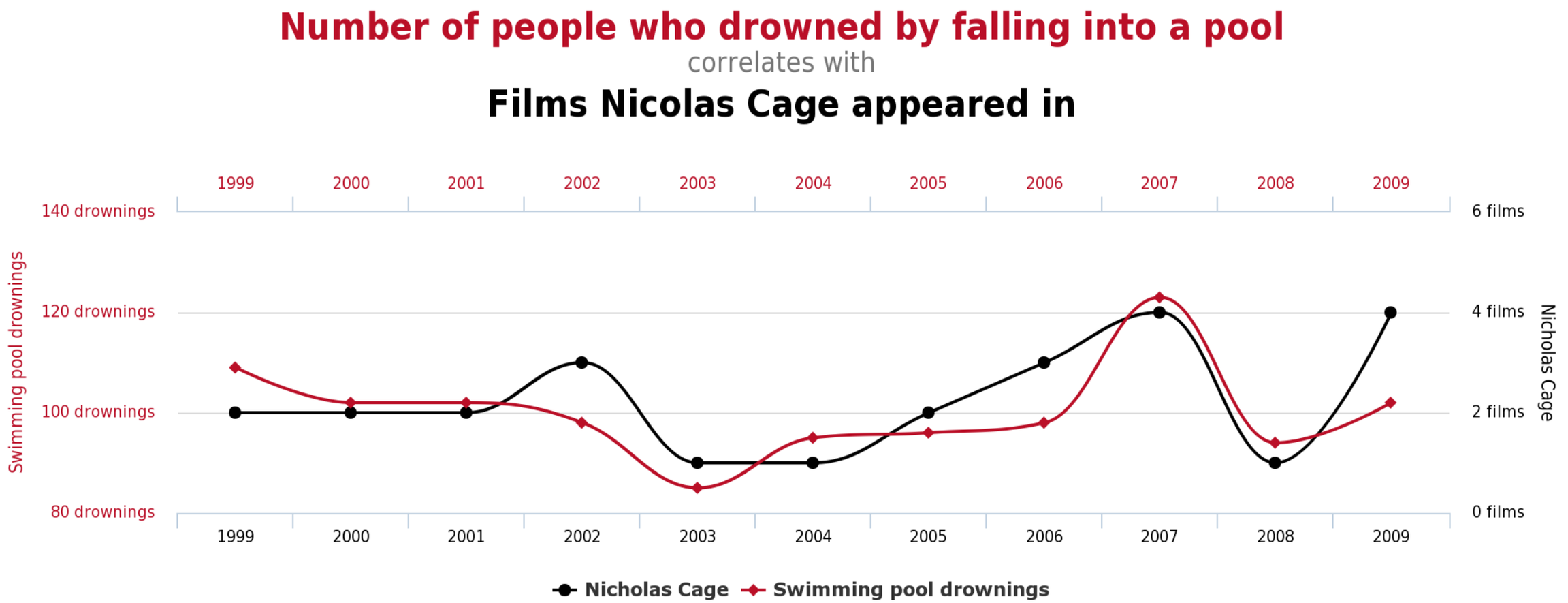


Correlation

- The correlation of two datasets gives you an indication of how similar their behavior is
- Two completely unrelated time series (say, two sequences of random numbers) will have a Pearson correlation coefficient of 0

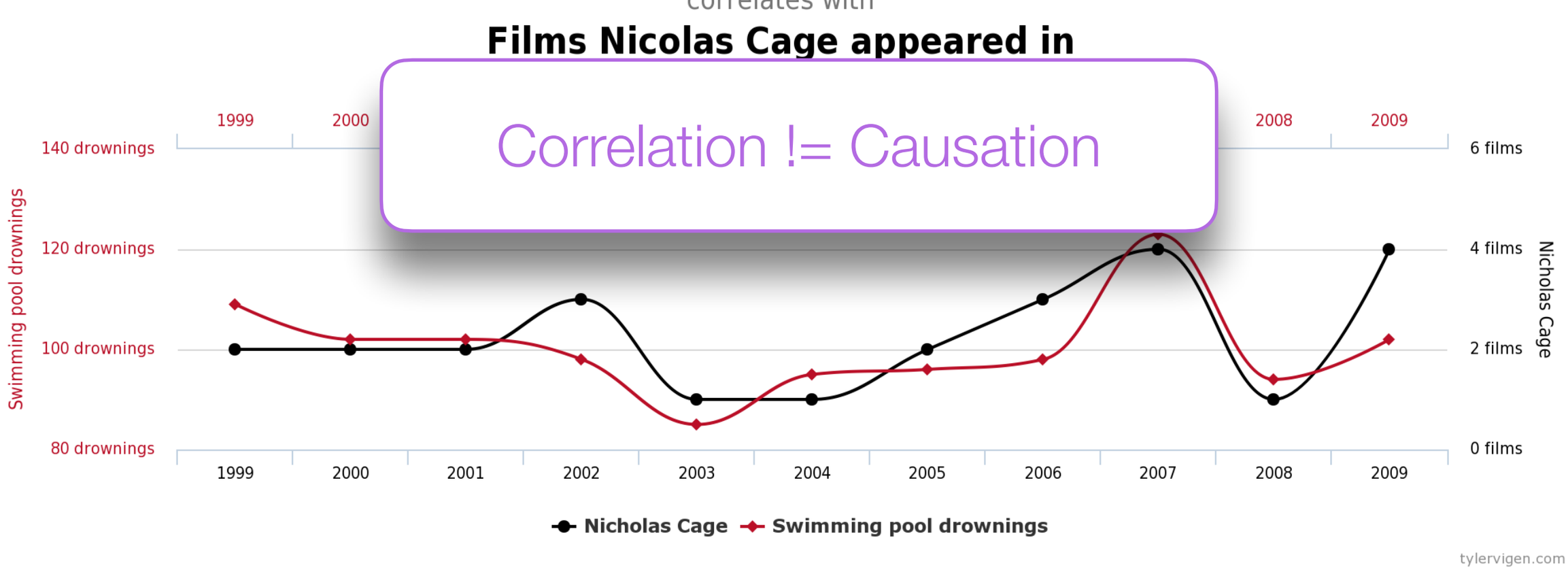
Correlation

- The correlation of two datasets gives you an indication of how similar their behavior is
- Two completely unrelated time series (say, two sequences of random numbers) will have a Pearson correlation coefficient of 0



Correlation

- The correlation of two datasets gives you an indication of how similar their behavior is
- Two completely unrelated time series (say, two sequences of random numbers) will have a Pearson correlation coefficient of 0



R^2

- The square of the Pearson correlation between the data and the fit.

- The amount of variance in the data that is explained by the "model".

Spearman Rank Correlation

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N\left(N^2 - 1\right)}$$

- Equivalent to the Pearson Correlation Coefficient of the ranked variables

- d_i^2 squared difference in ranks

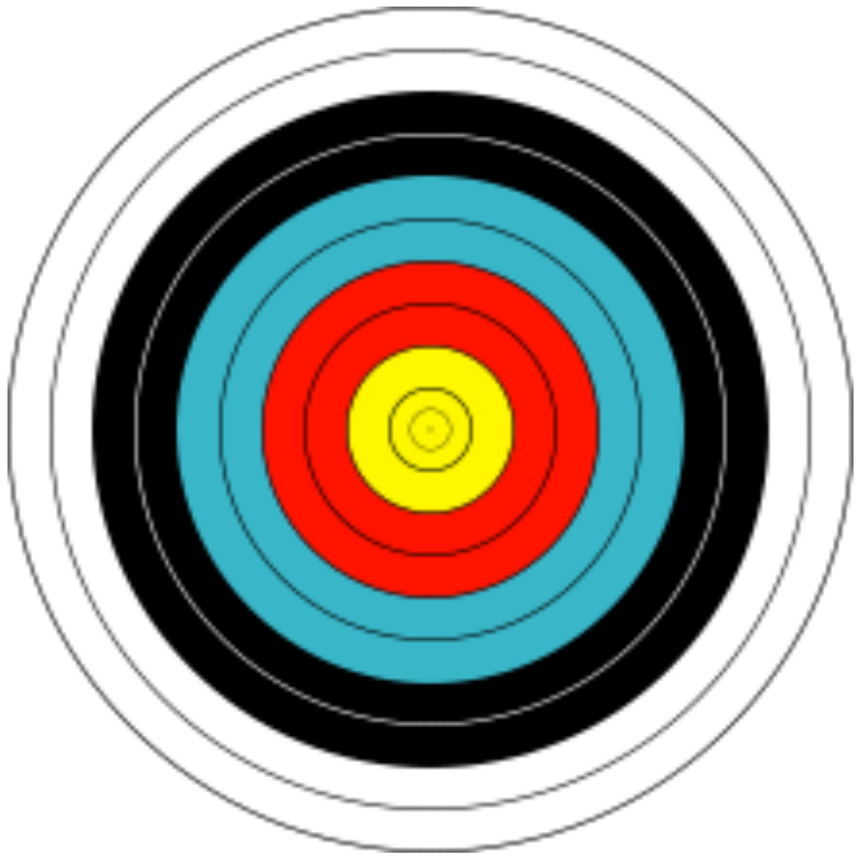
- less sensitive to outliers as values are limited by rank

Code - Descriptive Statistics

https://github.com/DataForScience/Probability-And-Statistics

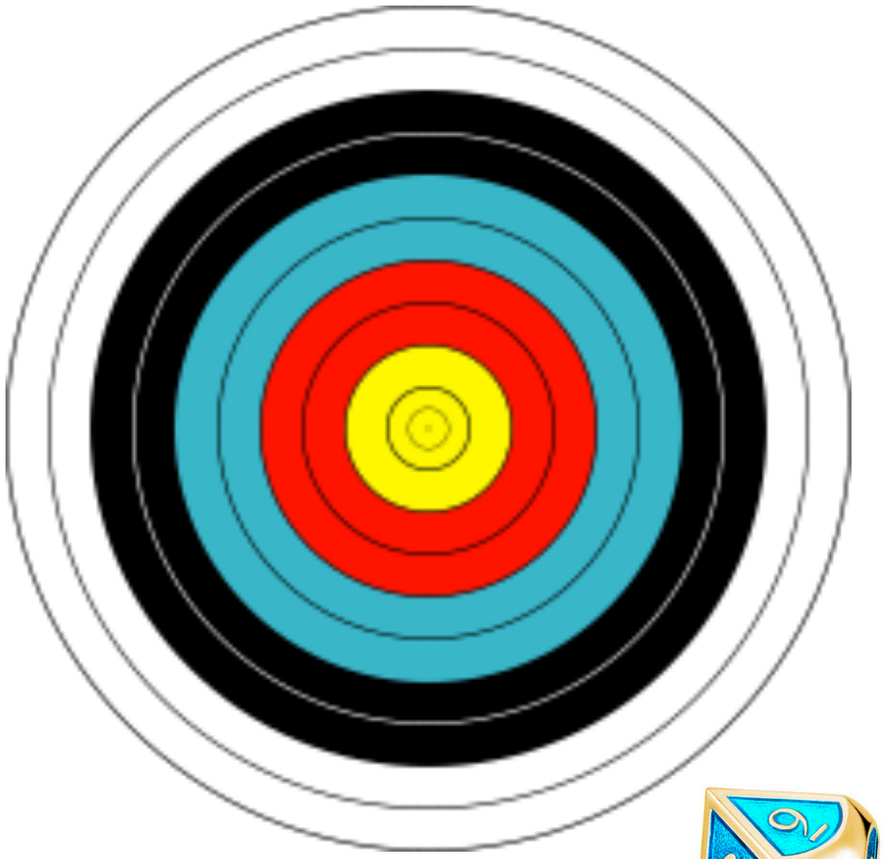
2. Fundamentals of Probability

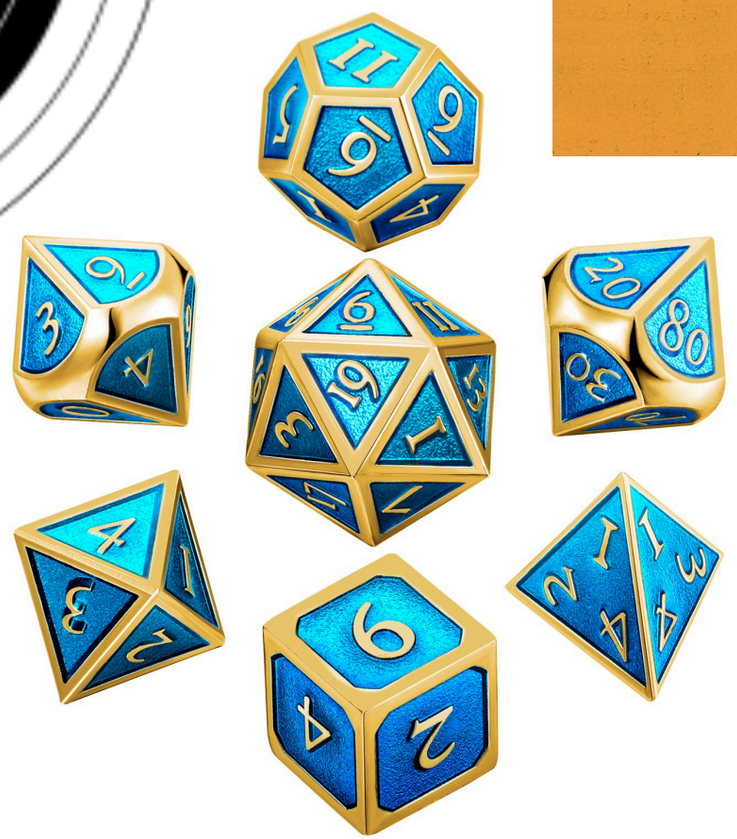
Randomness



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.

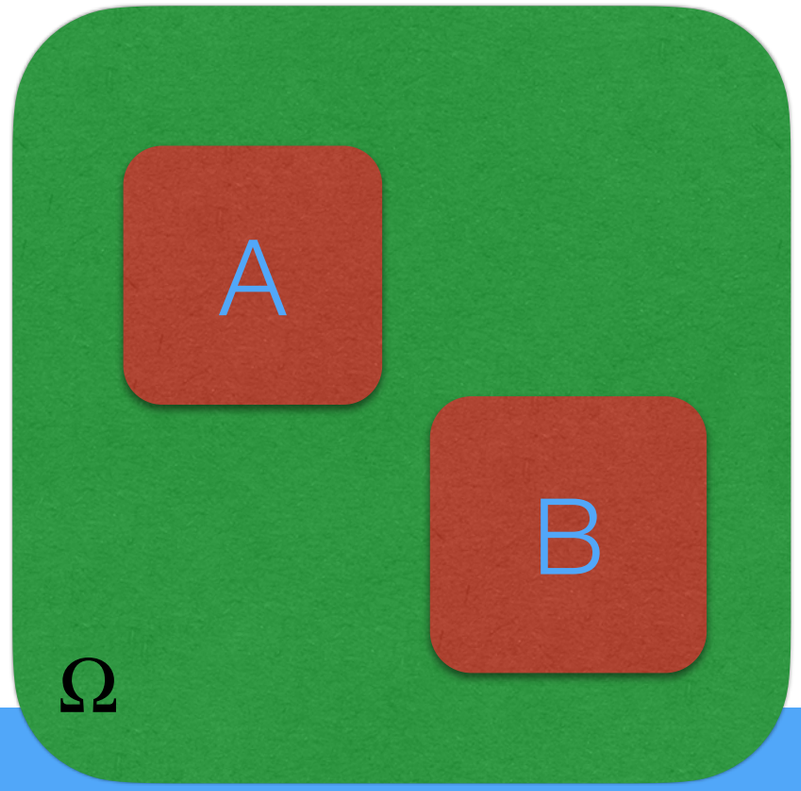
Probability = Area

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.

Probability = Area



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

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

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

Kolmogorov's Probability Axioms

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

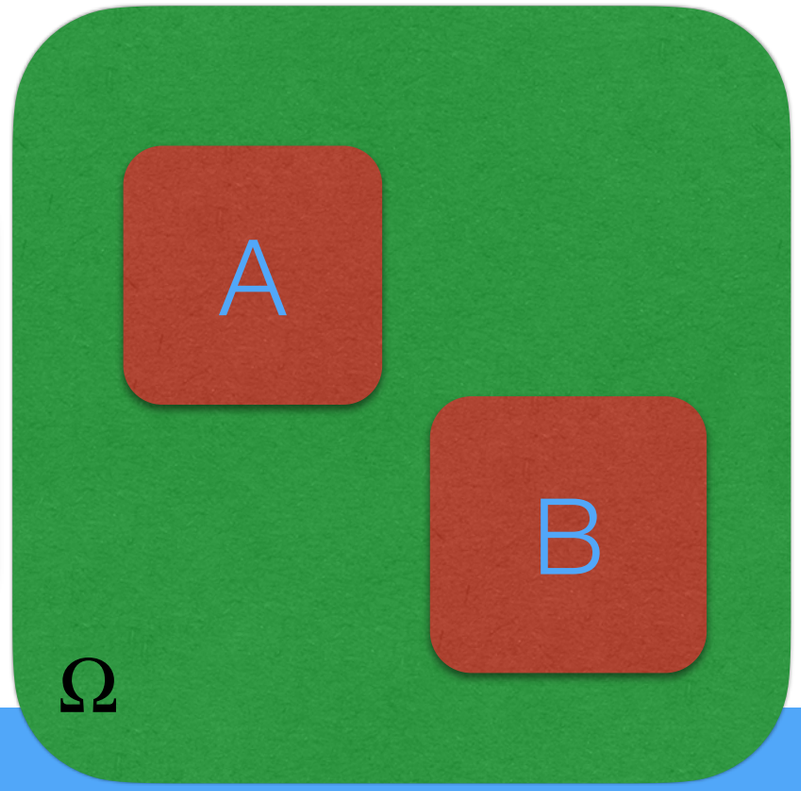
- Axiom 1: Probability is a real number greater or equal to 0.
- Axiom 2: Total probability is equal to 1.
- Axiom 3: Probability of mutually exclusive events is the sum of the probabilities.

Probability = Area

Prob(A) = Area A

Total Area = 1

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

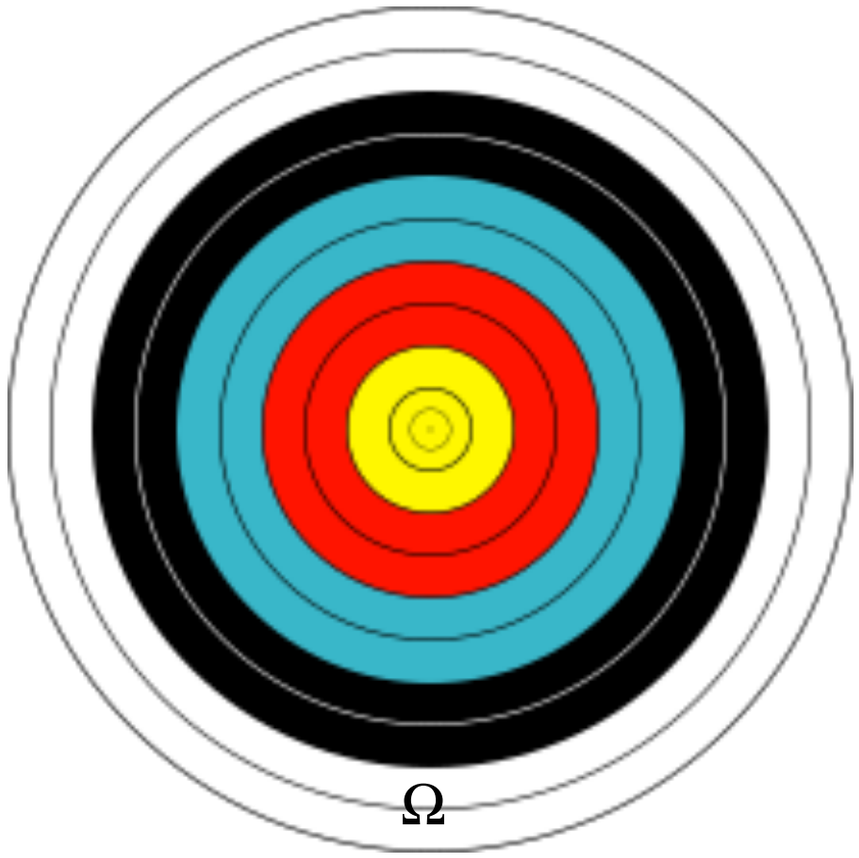


$$0 \leq P(A) \leq 1$$

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

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

Probability = Frequency



Rolling Dice

- 6 sided die

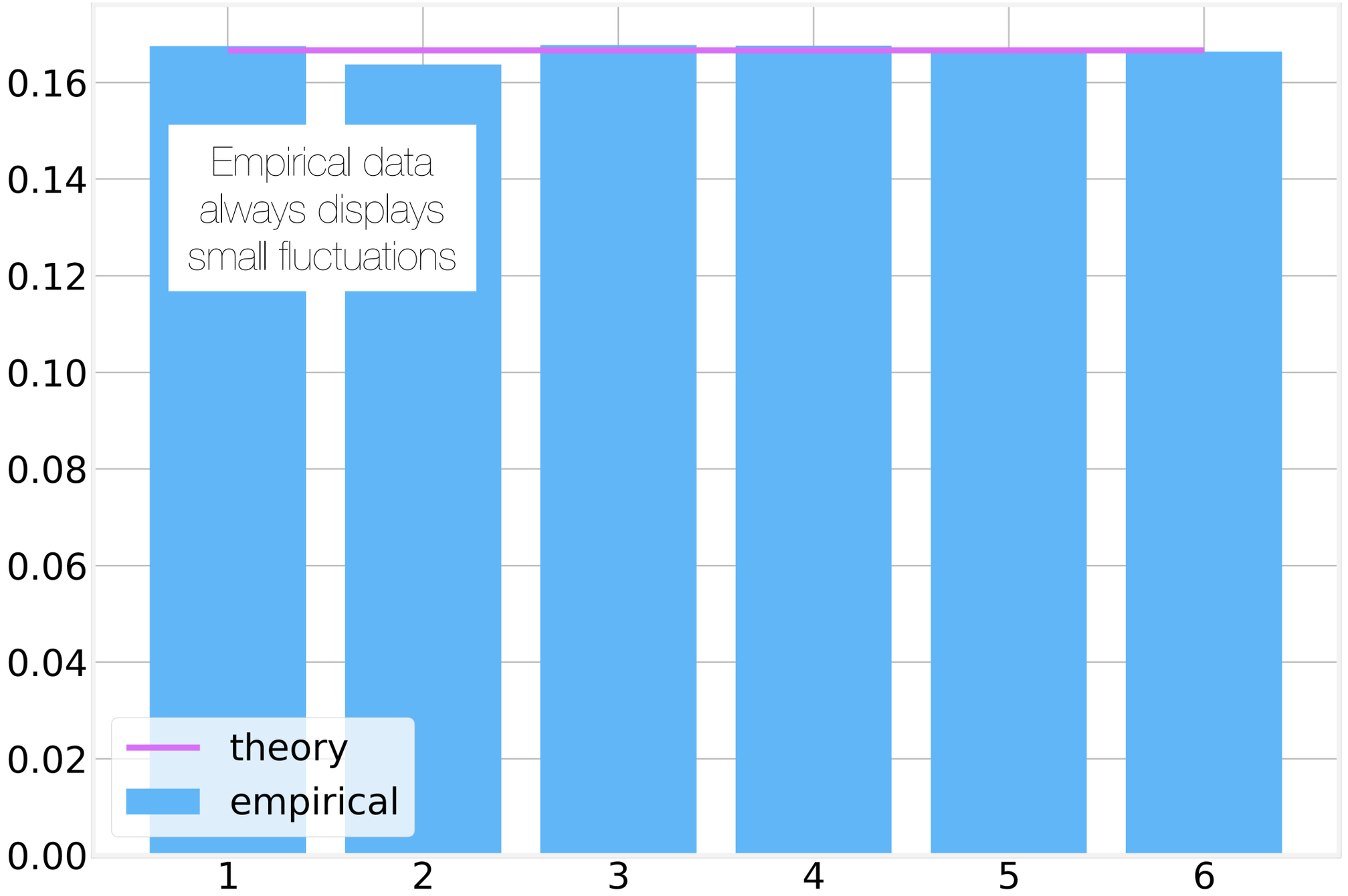
- 6 possibilities

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

1	2	3
4	5	6

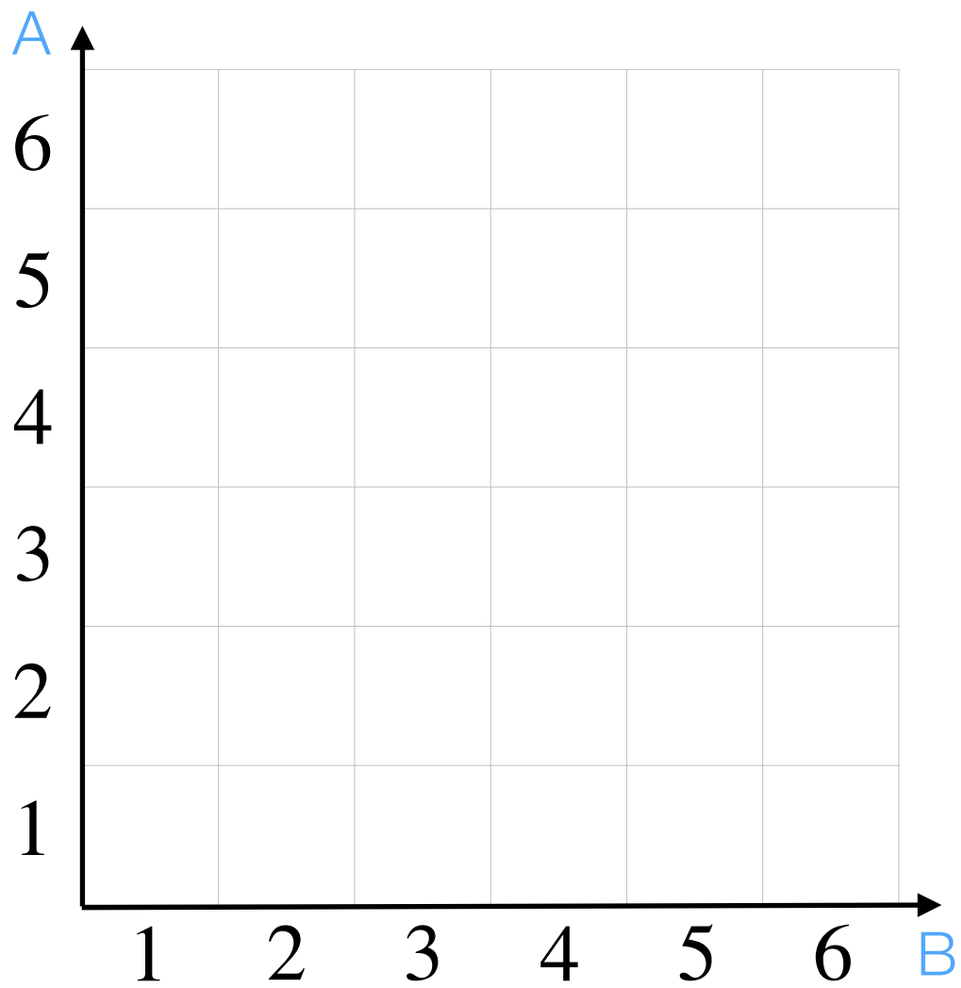
Ω

Rolling Dice



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

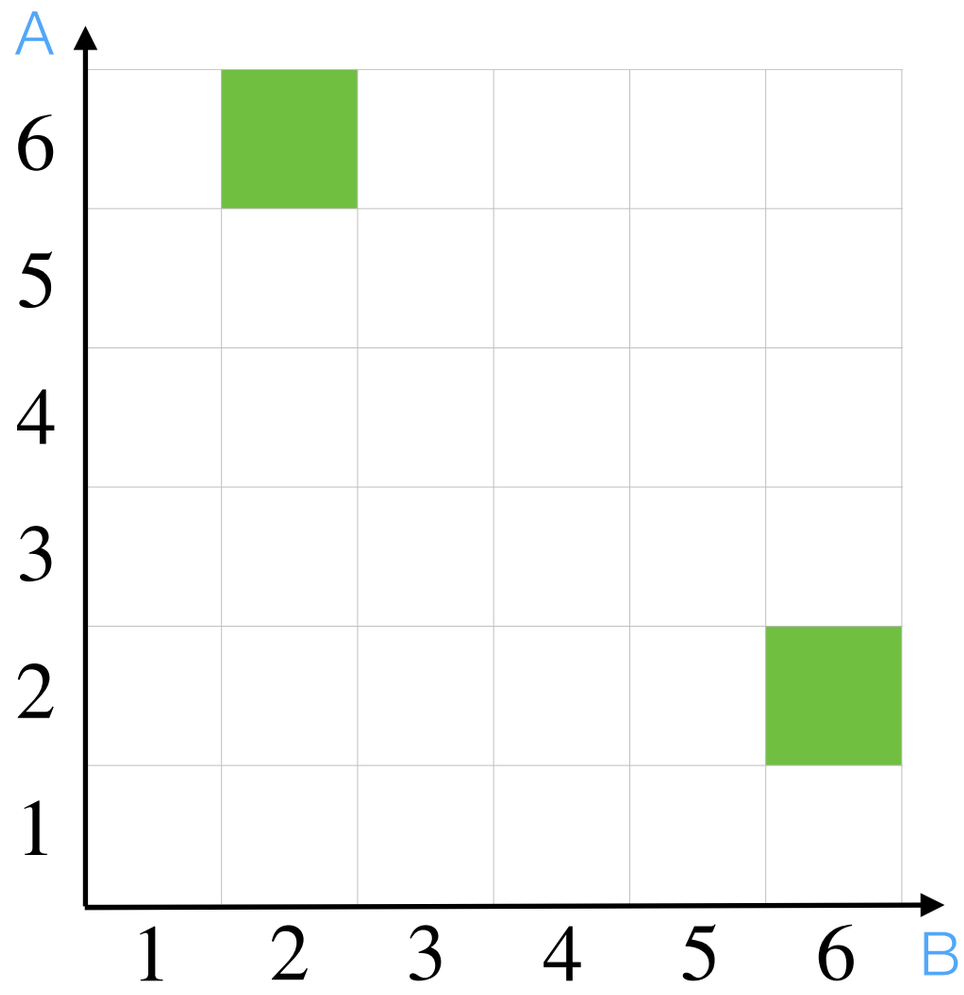


Multiply

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



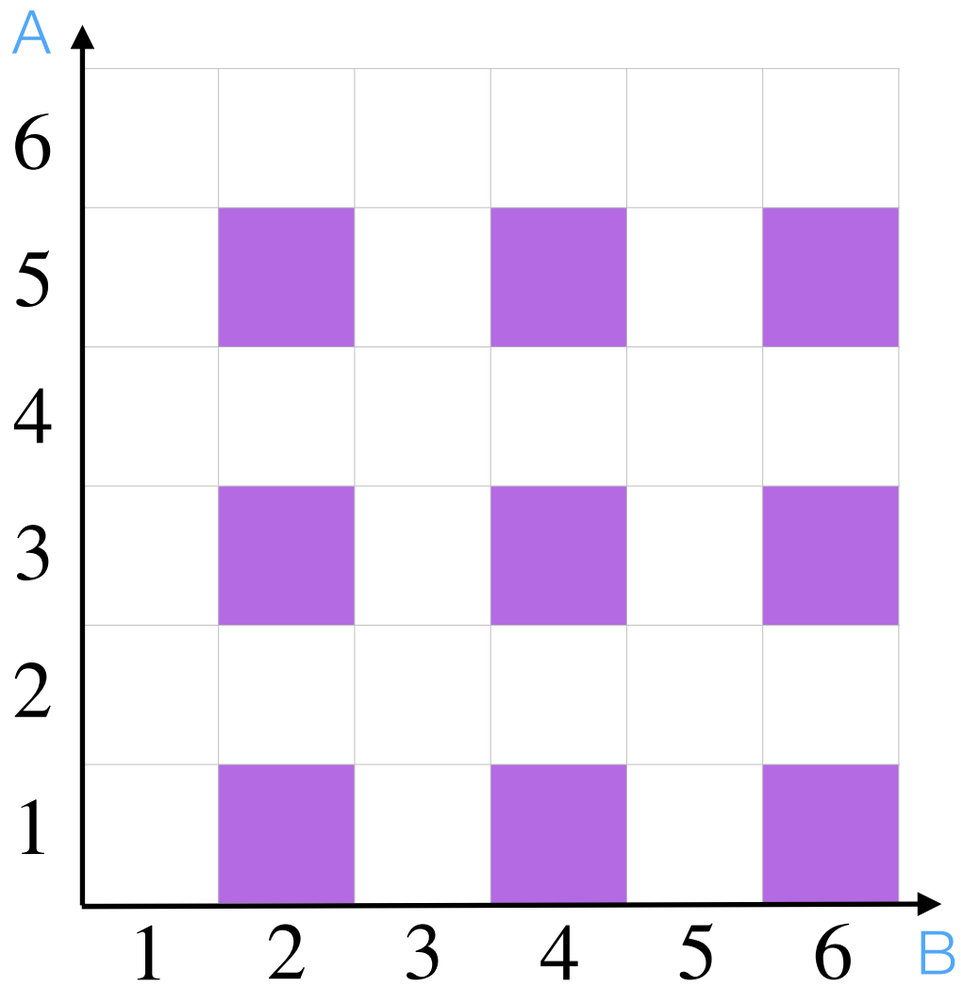
$$P\left(2,6\right) = \frac{2}{36}$$

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



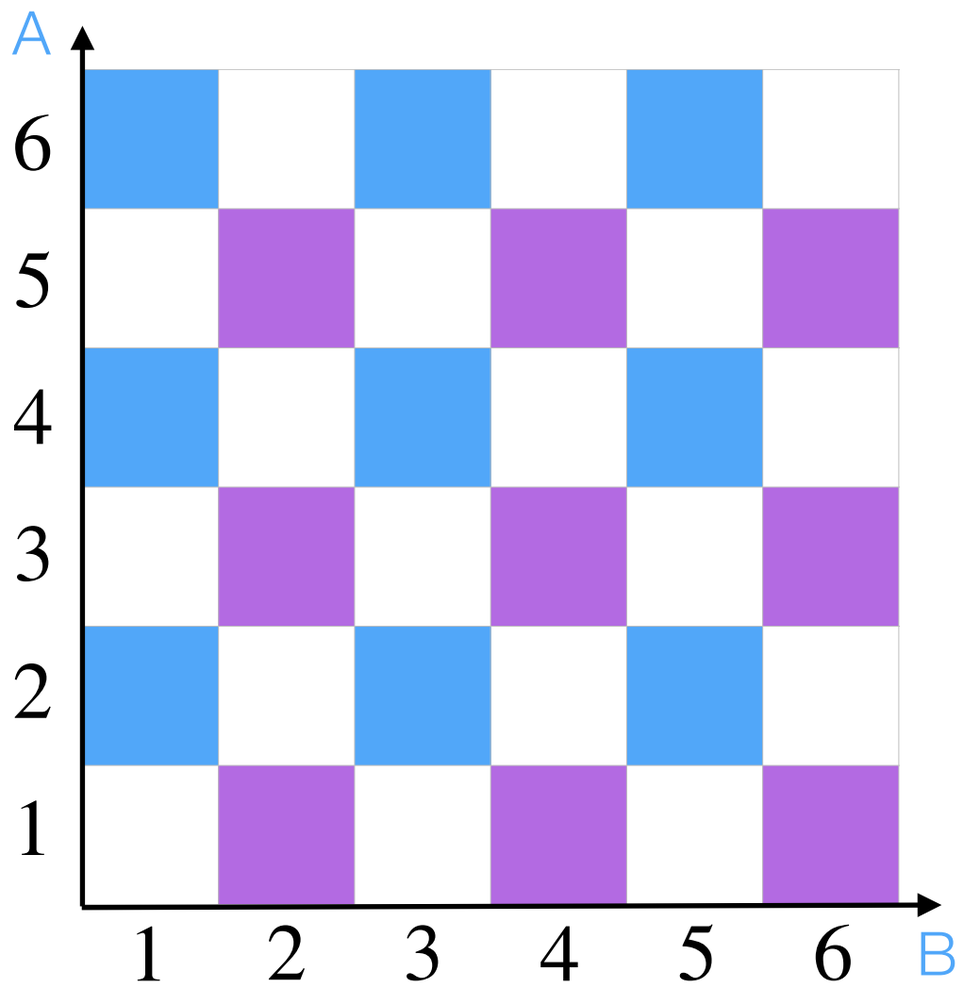
$$P\left(odd, even\right) = \frac{9}{36}$$

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:

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

Shuffling Cards

- 52 cards. The probability of picking any card is:

$$P\left(\text{any card}\right) = \frac{1}{52}$$

- 4 suits of 13 cards each. The probability of picking a Diamond:

$$P\left(\text{diamond}\right) = 13 \cdot \frac{1}{52} = \frac{1}{4}$$

- 3 figure cards per suit (K, Q and J). The probability of picking a figure card of hearts:

$$P\left(\text{figure of hearts}\right) = \frac{3}{52}$$

- The probability of picking any figure card:

$$P\left(\text{figure}\right) = 4 \cdot 3 \cdot \frac{1}{52} = \frac{3}{13}$$

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

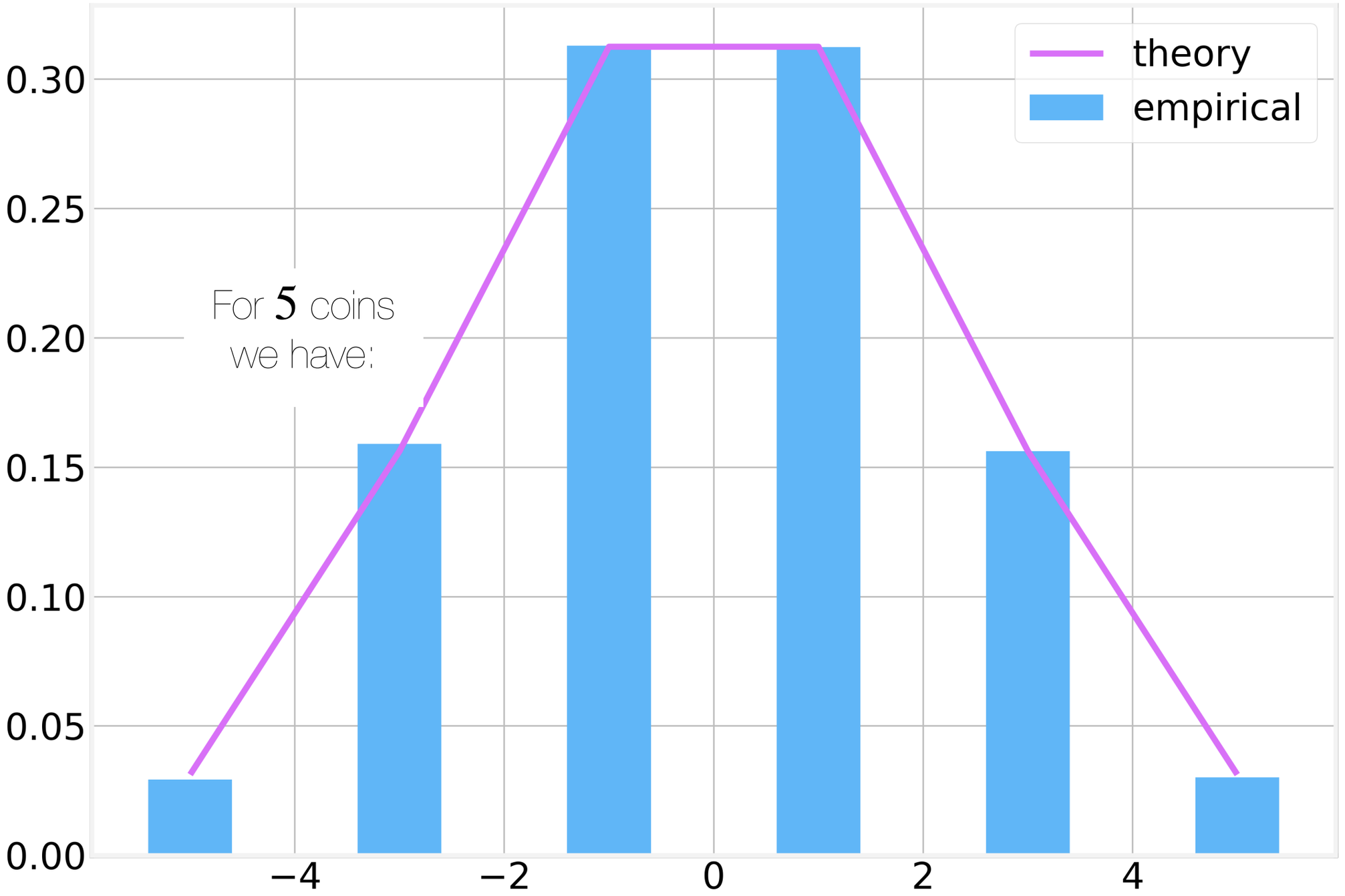
- From combinatorics we also know that the number of ways to get N_h heads out of total N flips is:

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

- 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



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

Code - Probability

https://github.com/DataForScience/Probability-And-Statistics

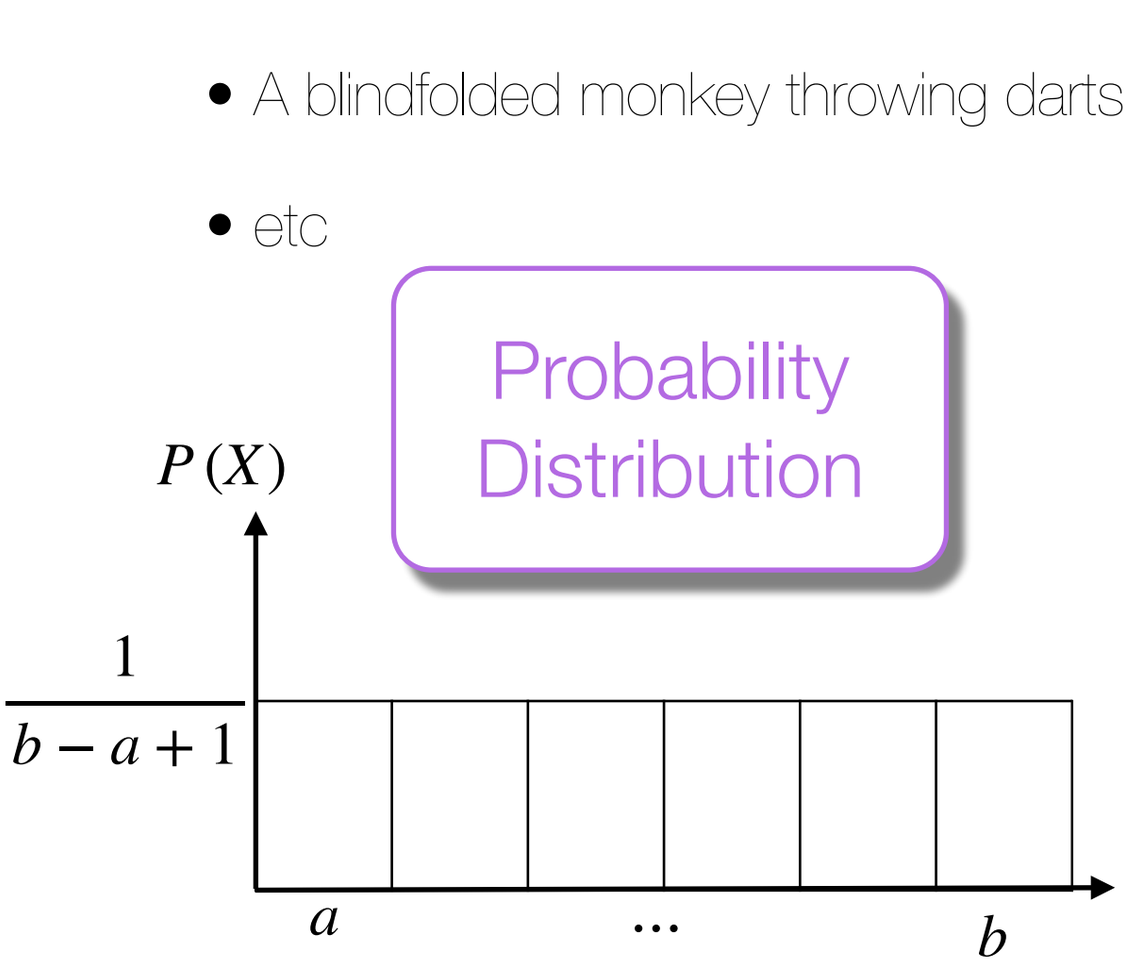
3. Probability Distributions

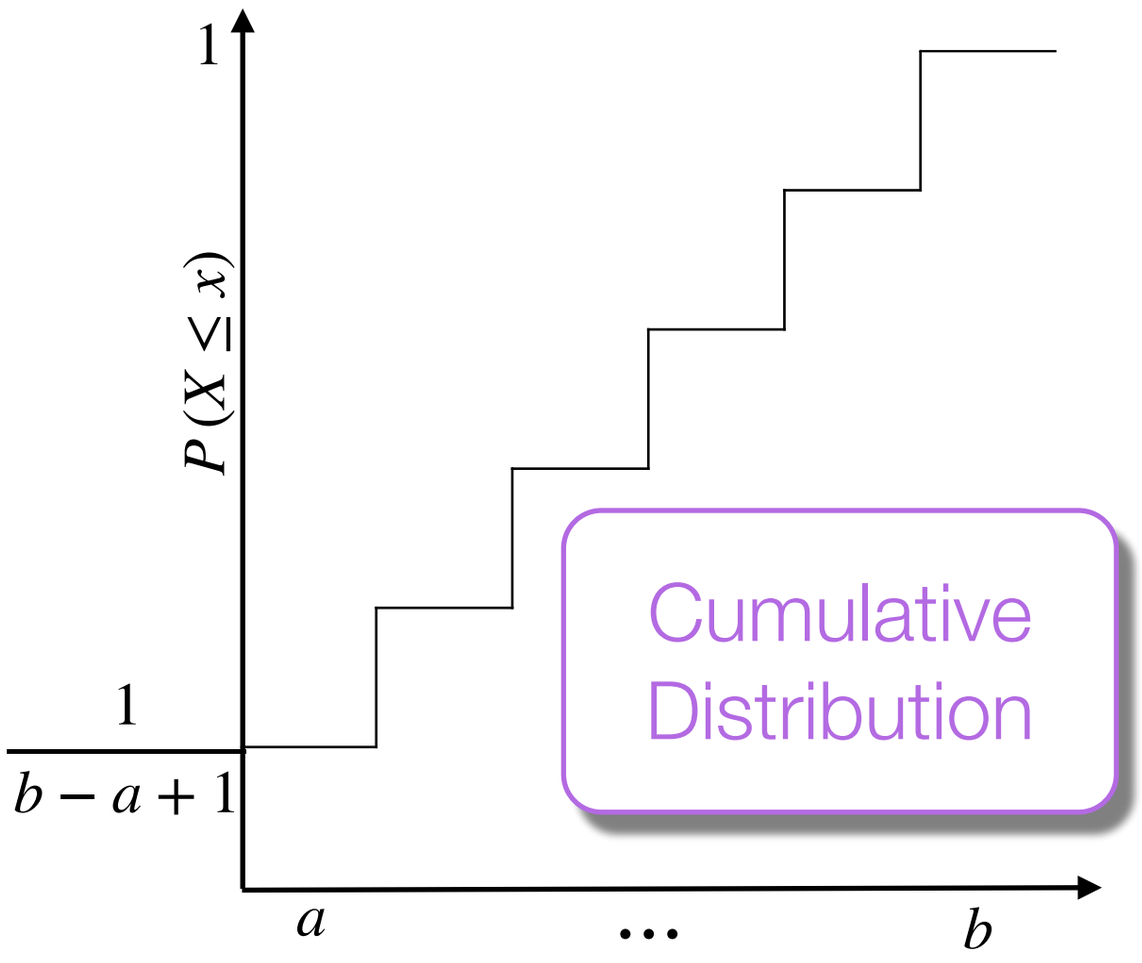
Uniform Distribution

- All N outcomes have the same probability, $1/N$
    - A fair die
    - A deck of cards
    - etc

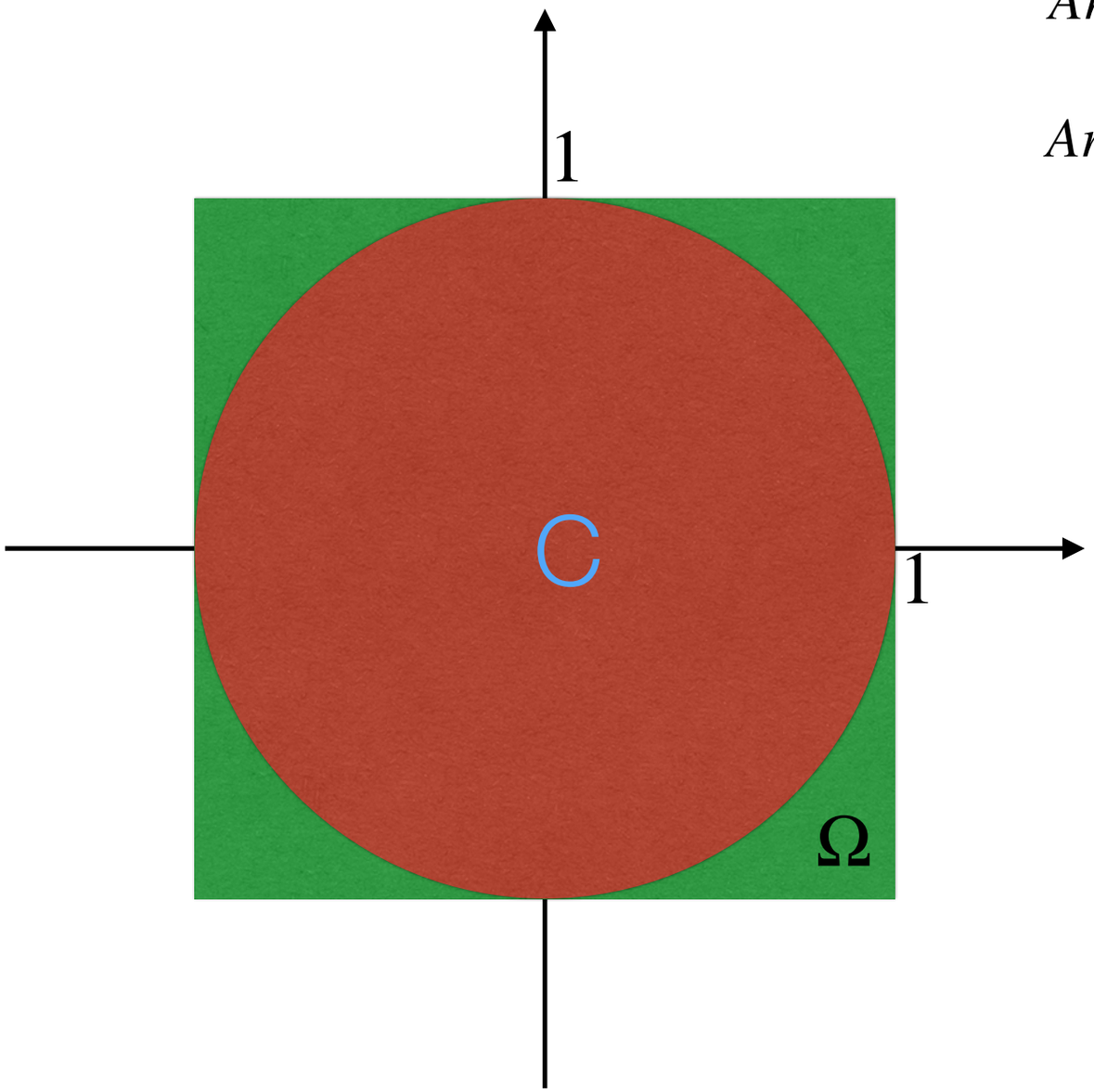
Uniform Distribution

- All N outcomes have the same probability, $1/N$
  - A fair die
  - A deck of cards
  - A blindfolded monkey throwing darts
  - etc





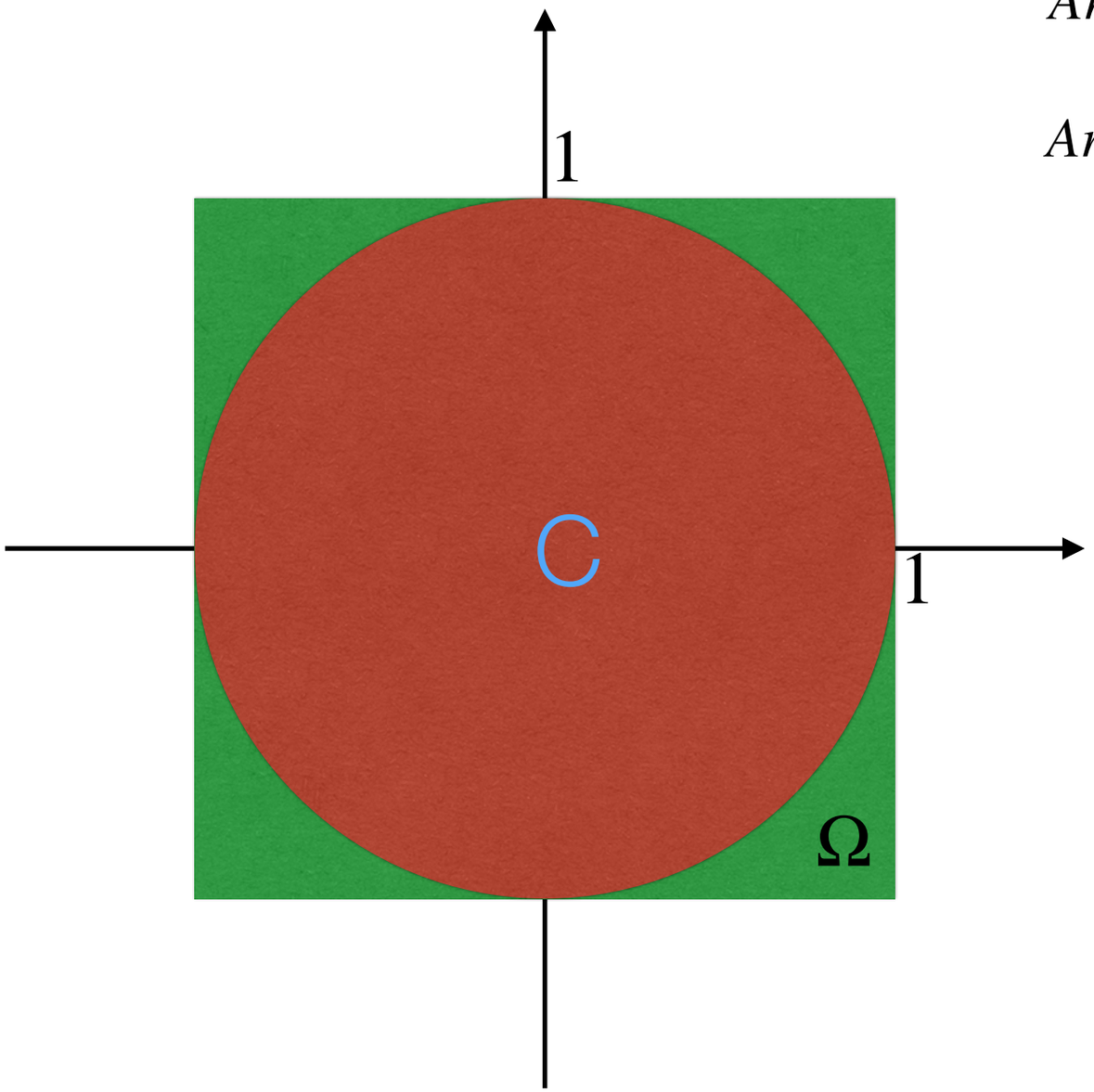
Continuous Version



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

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

Continuous Version

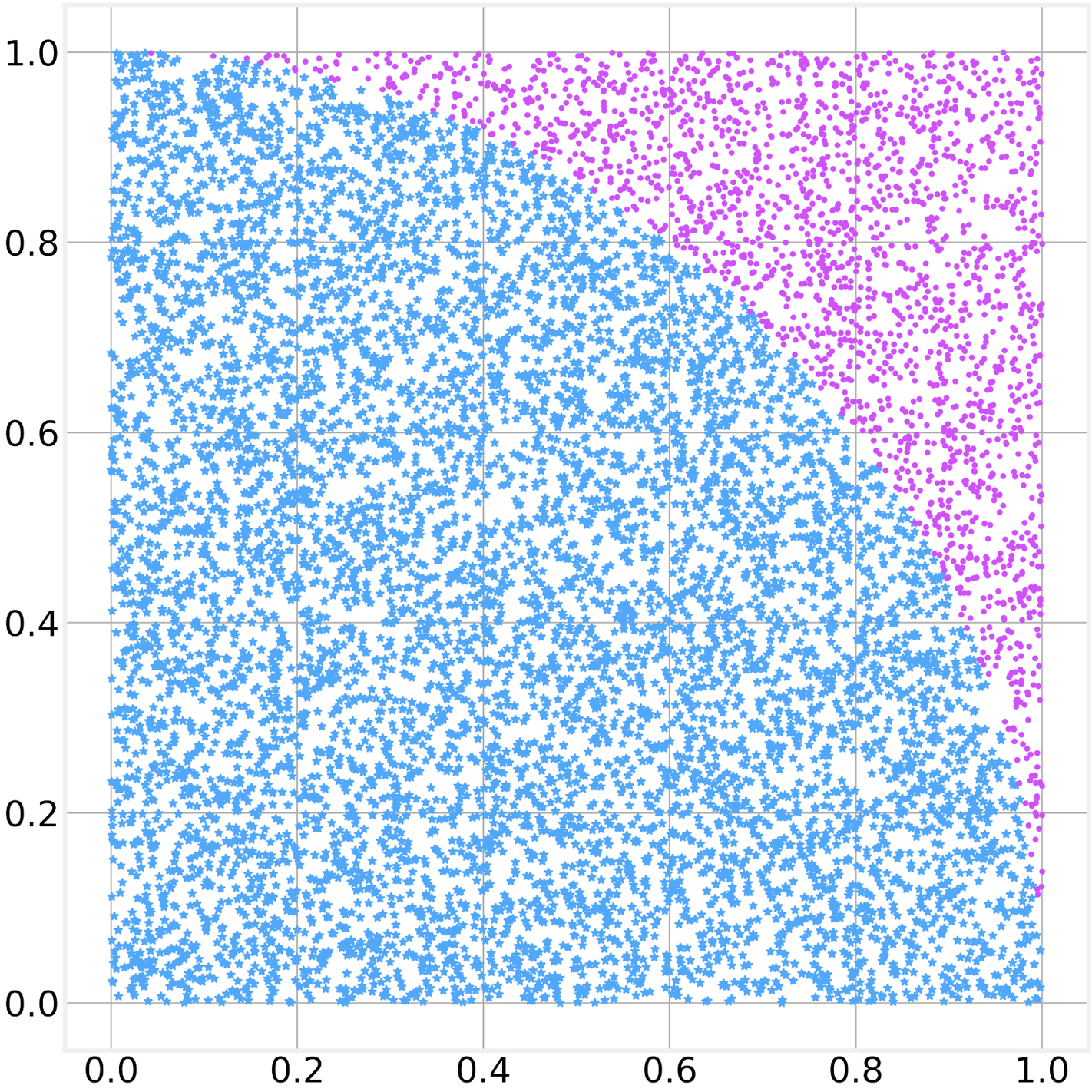


$$Area\left(C\right) = \pi$$

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

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

Continuous Version



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:

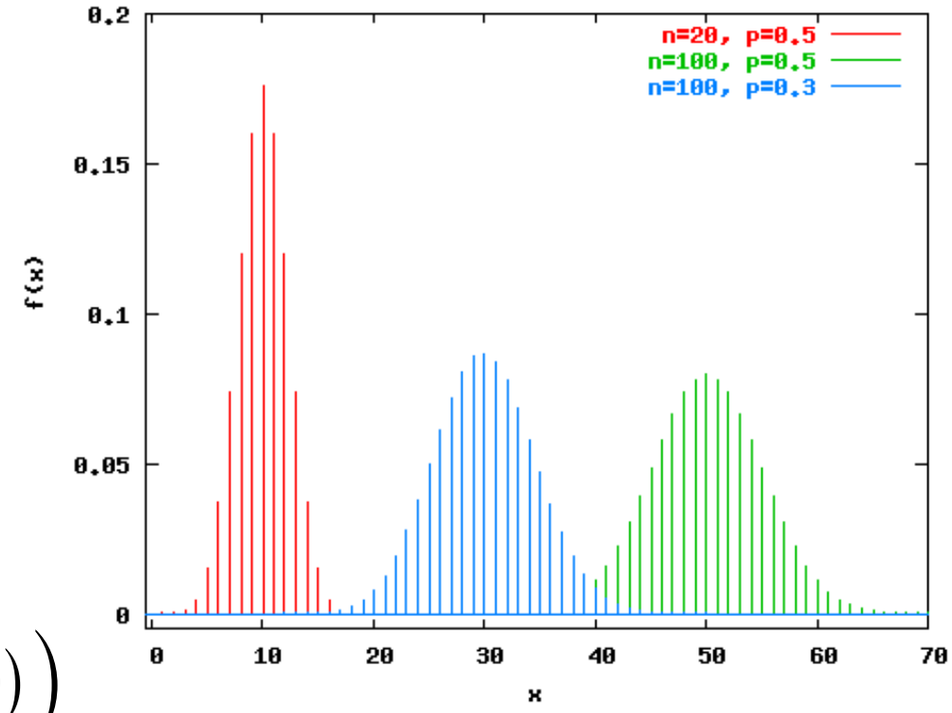
$$\mu = np$$

- and the variance:

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

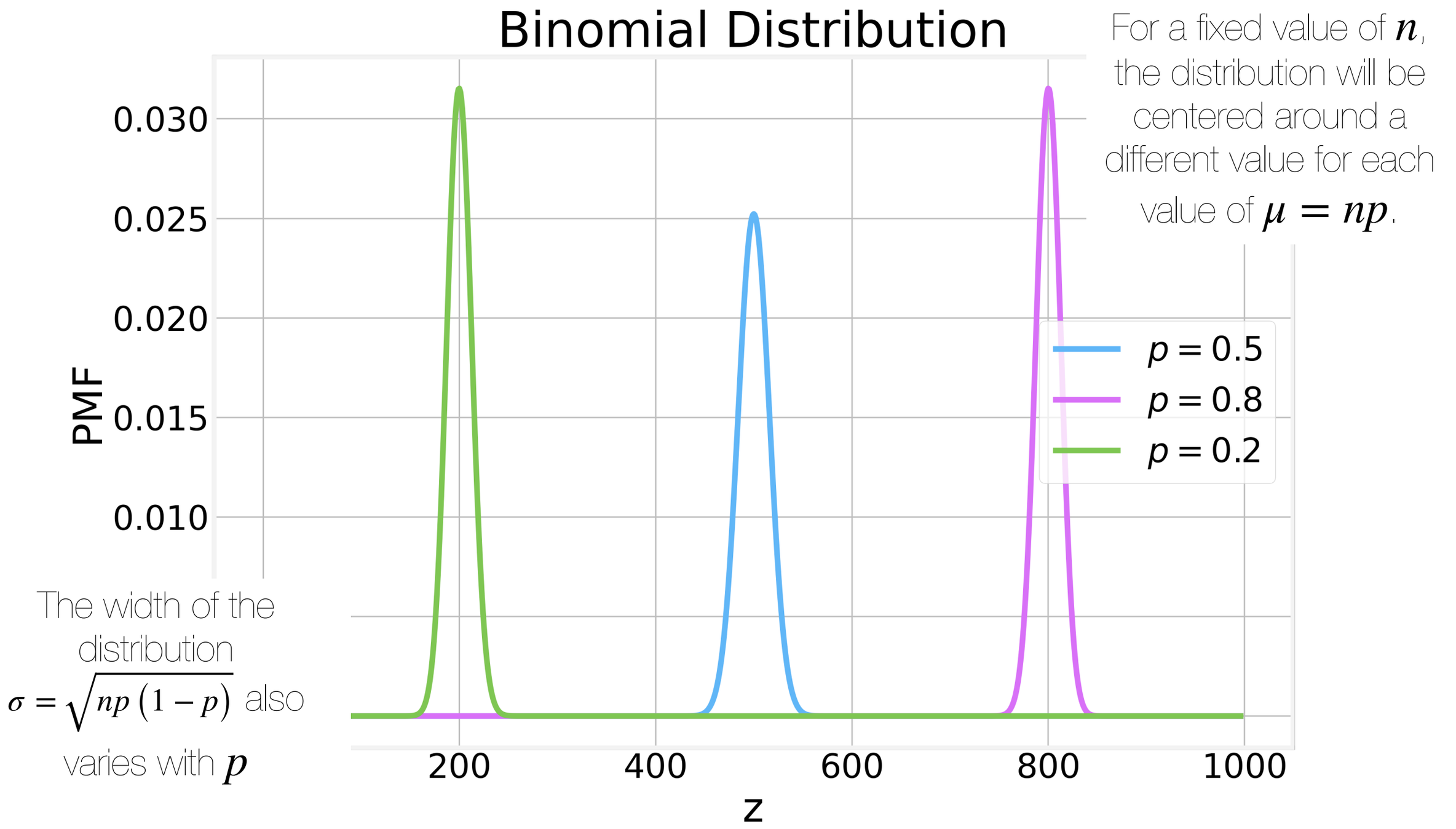
- and for sufficiently large n:

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



- Useful to understand the properties of any process with a probabilistic outcome

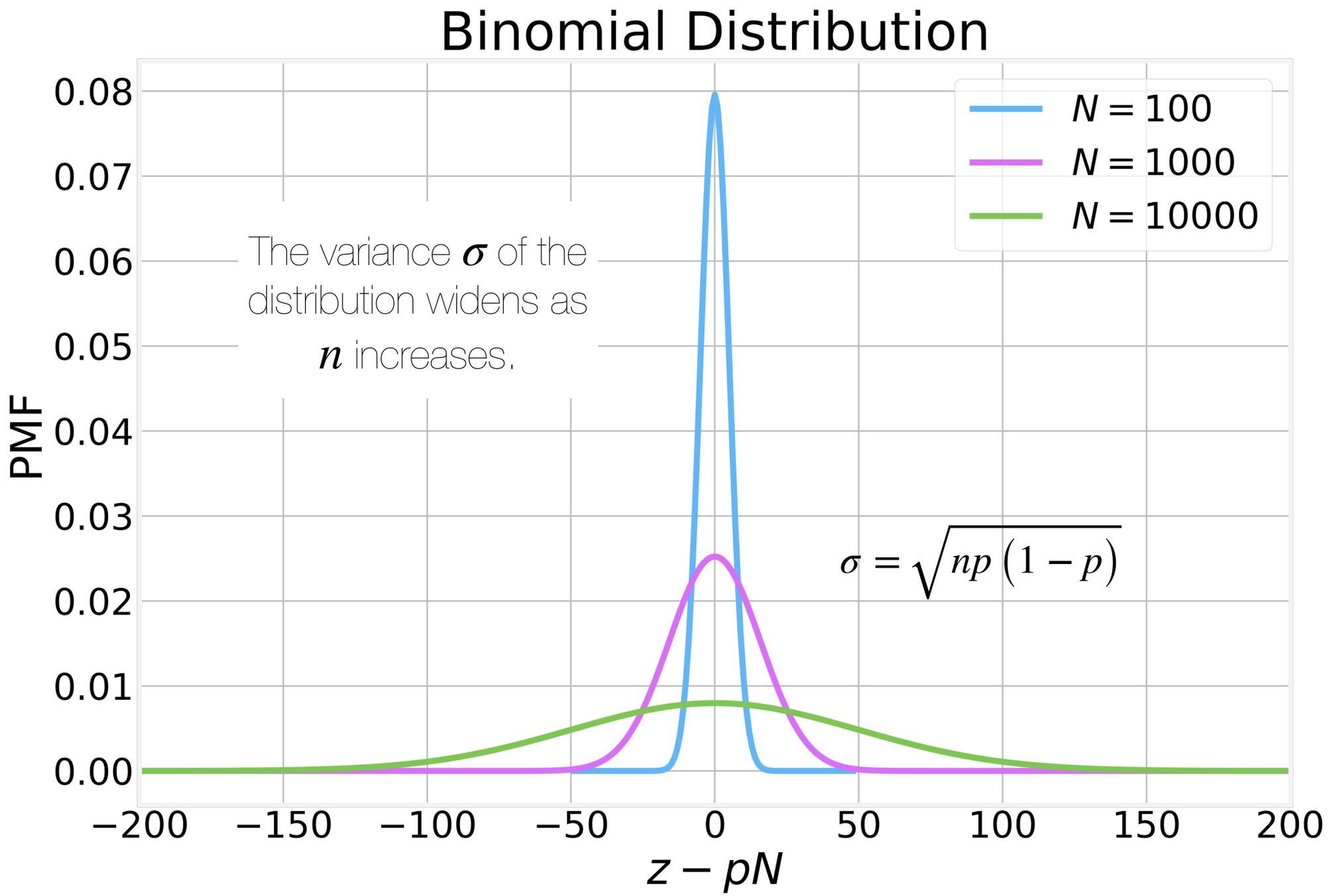
Binomial Distribution



For a fixed value of n, the distribution will be centered around a different value for each value of $\mu = np$.

The width of the distribution $\sigma = \sqrt{np\left(1-p\right)}$ also varies with p

Binomial Distribution



Gaussian/Normal Distribution

- Perhaps the most common distribution, especially in biology and social sciences
    - Human height
    - IQ
    - Birth weight
    - SAT scores
    - etc
- Implies a typical (normal) value
- Occurs whenever we take an average over different measurements

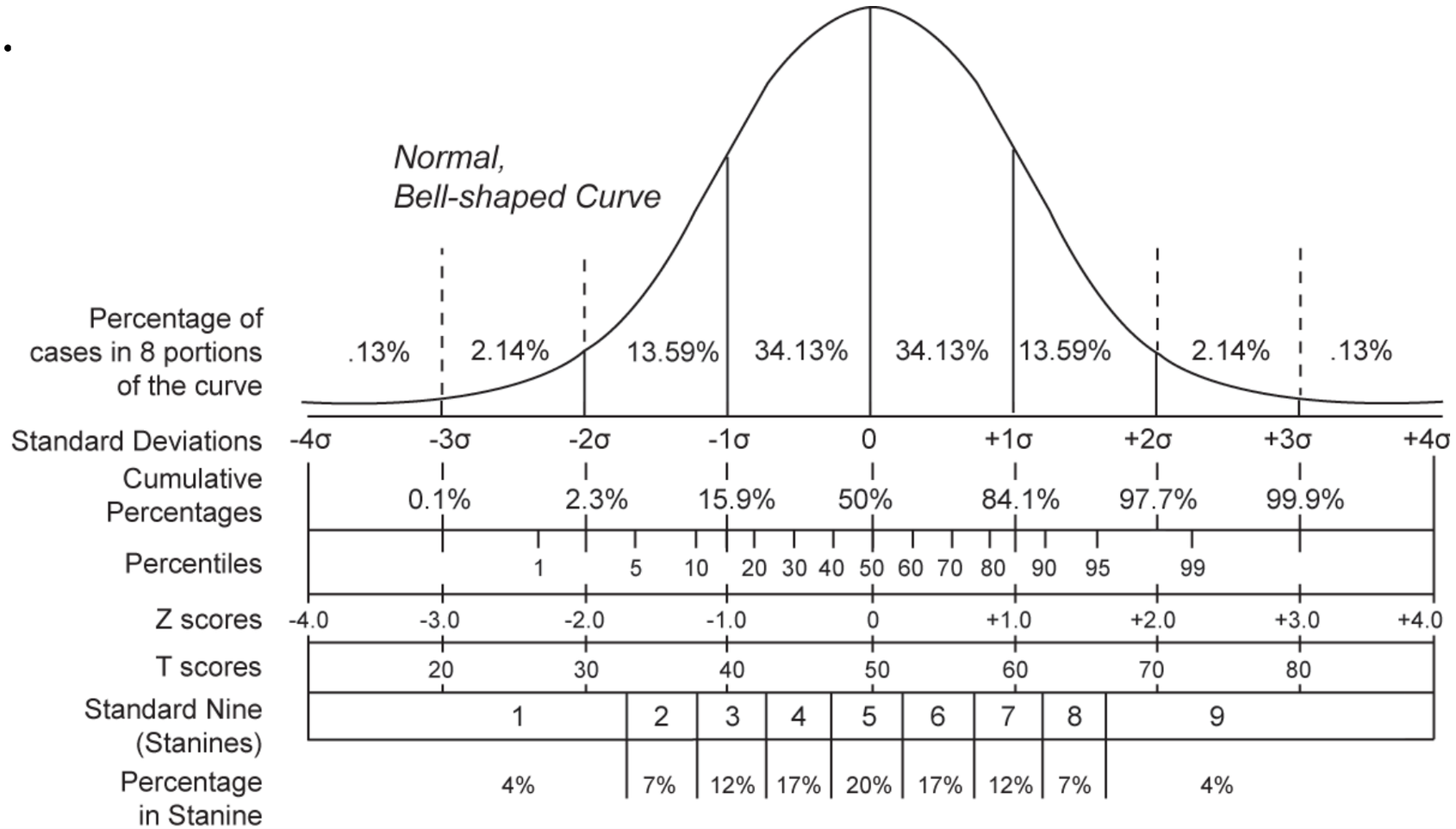
Gaussian/Normal Distribution

- The probability distribution is given by:

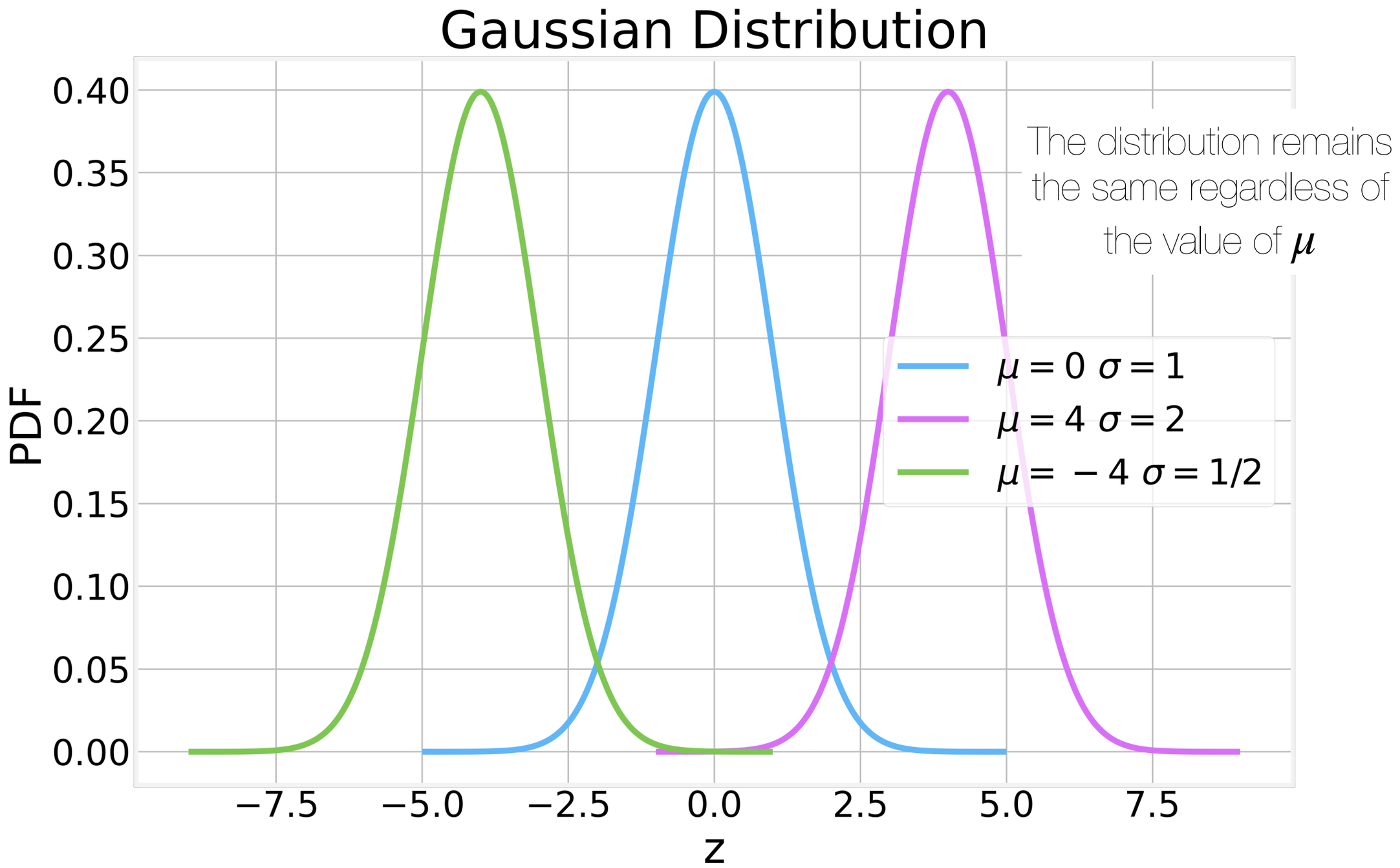
$$P_G\left(x,\mu,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right]$$

- Where the mean value is μ and the variance is σ^2.

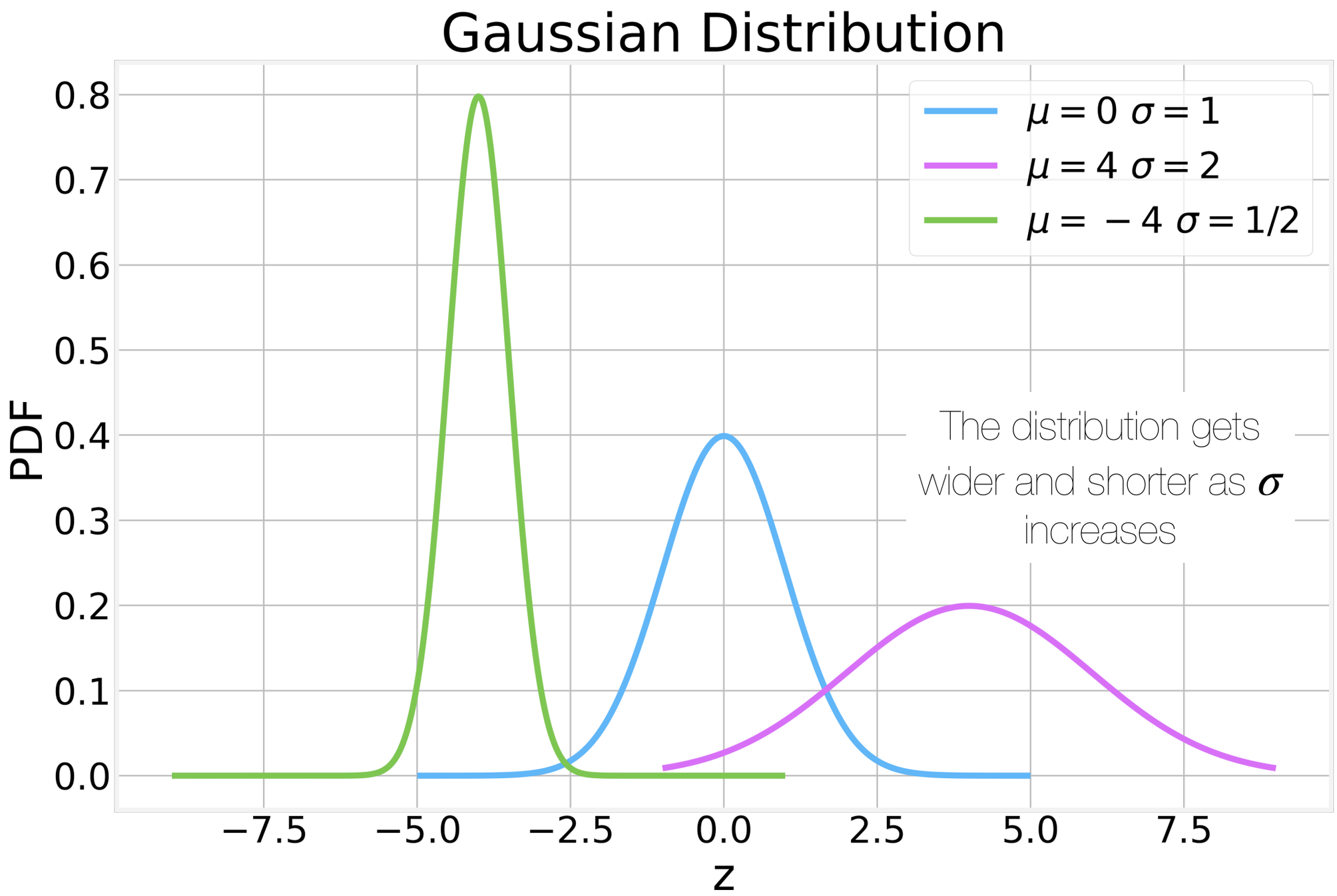
- 99.



Normal Distribution



Normal Distribution



Poisson Distribution

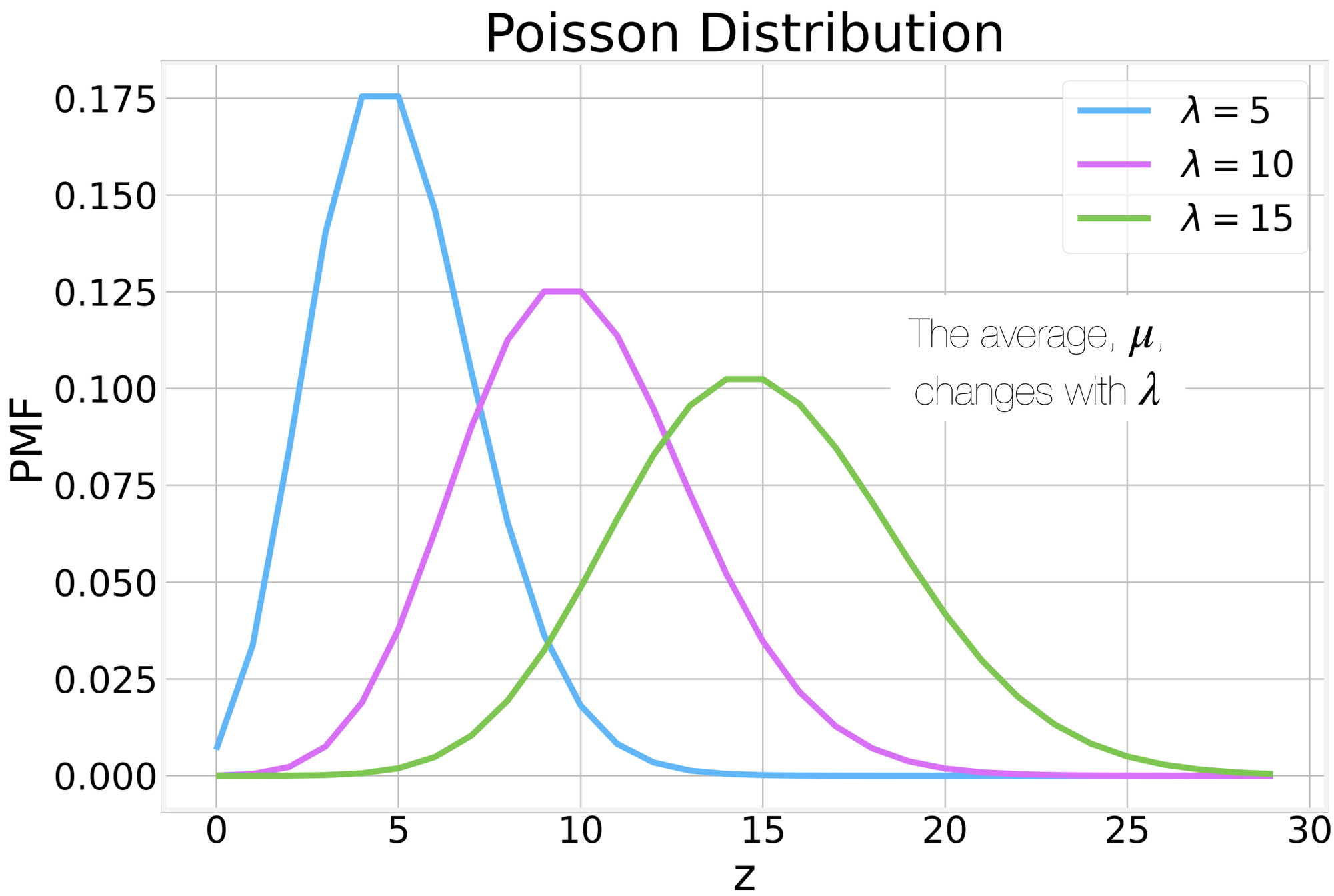
- The Poisson distribution describes the probability of a given number of events, λ, occurring in a fixed interval of time if these events occur with a known constant mean rate, $1/\lambda$:
    - Calls received in a call center per day
    - Number of cars passing an intersection
    - Number of floods per year
    - etc
- The probability of k events occurring in our reference time period is:

$$P_P\left(k,\lambda\right) = \frac{\lambda^k e^{-\lambda}}{k!}$$

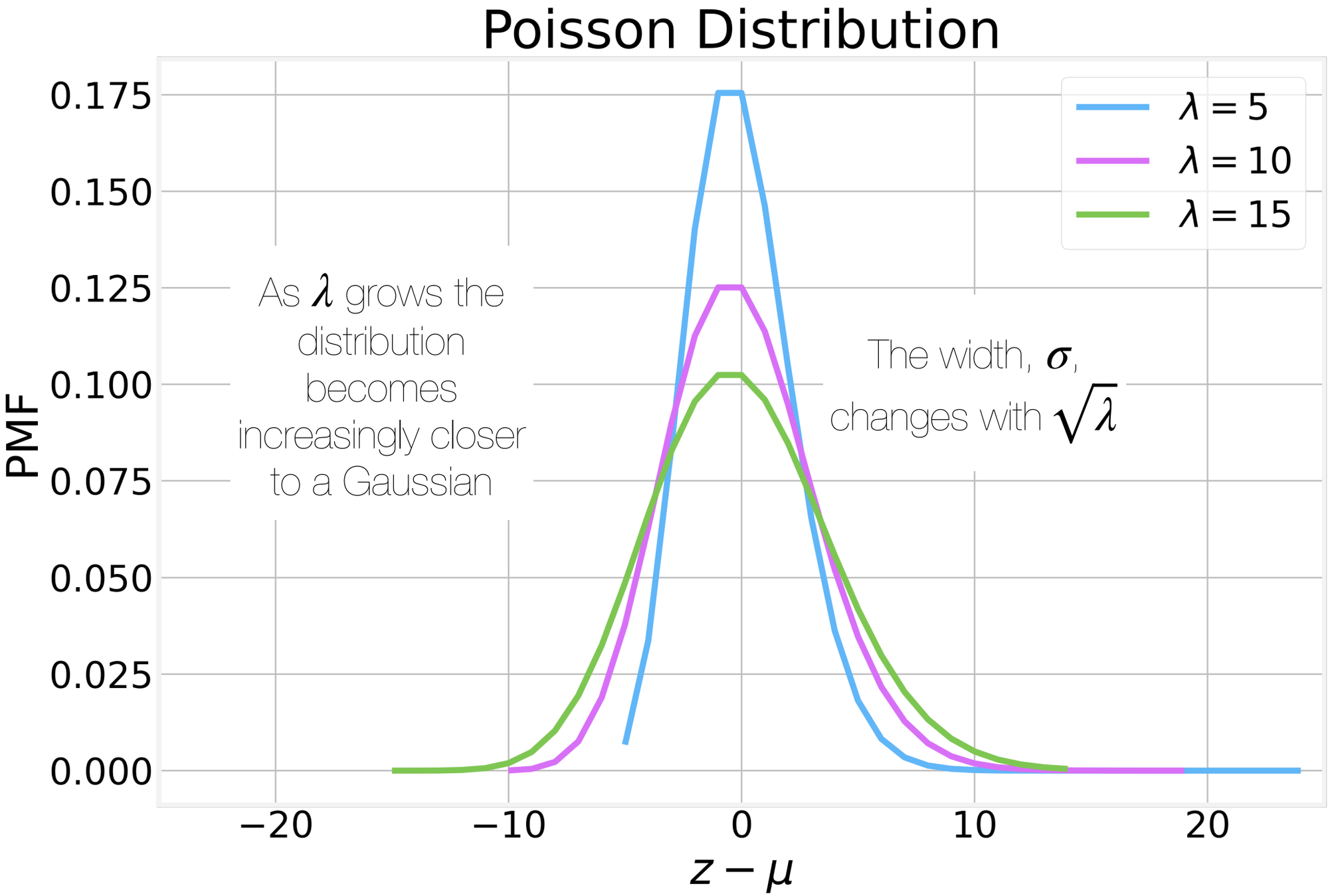
- With a mean $\mu = \lambda$ and standard deviation σ given by:

$$\sigma = \sqrt{\lambda}$$

Poisson Distribution



Poisson Distribution



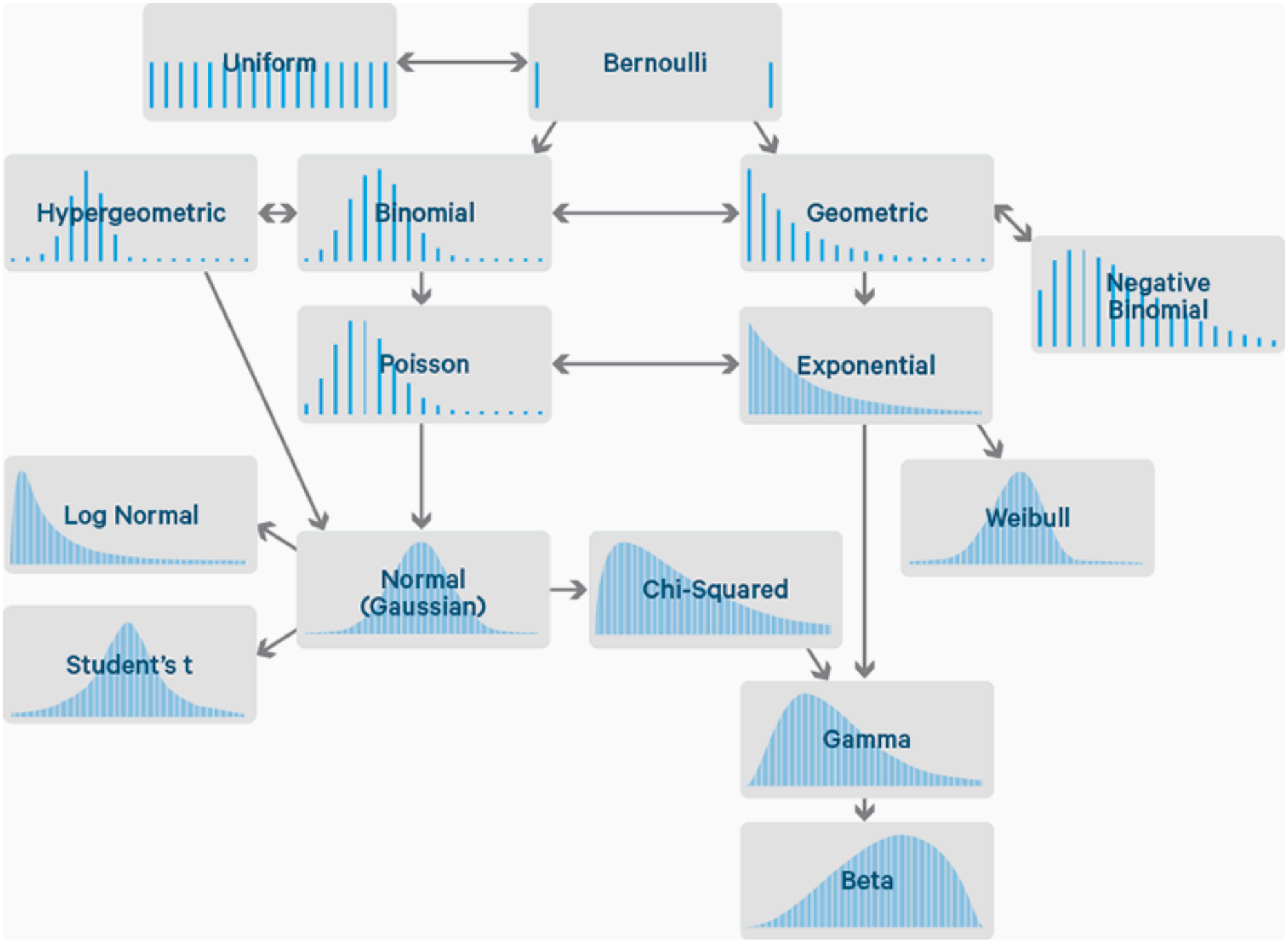
SciPy

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
  - scipy.stats.normal() - Normal 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

Probability distributions

https://math.stackexchange.com/questions/3050352/relationship-between-probability-distributions



Probability distributions

- 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:

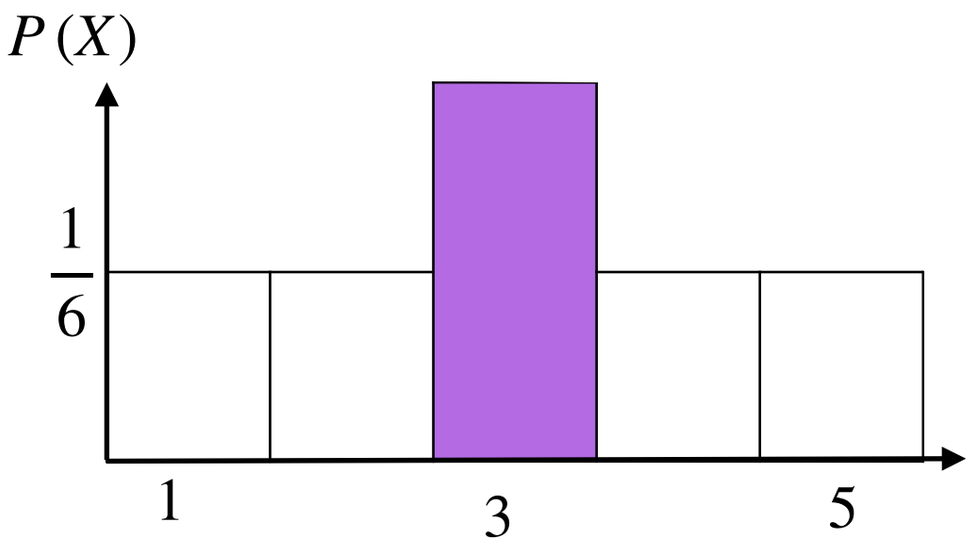
$$P(X \leq x)$$

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

Non-Uniform Distribution

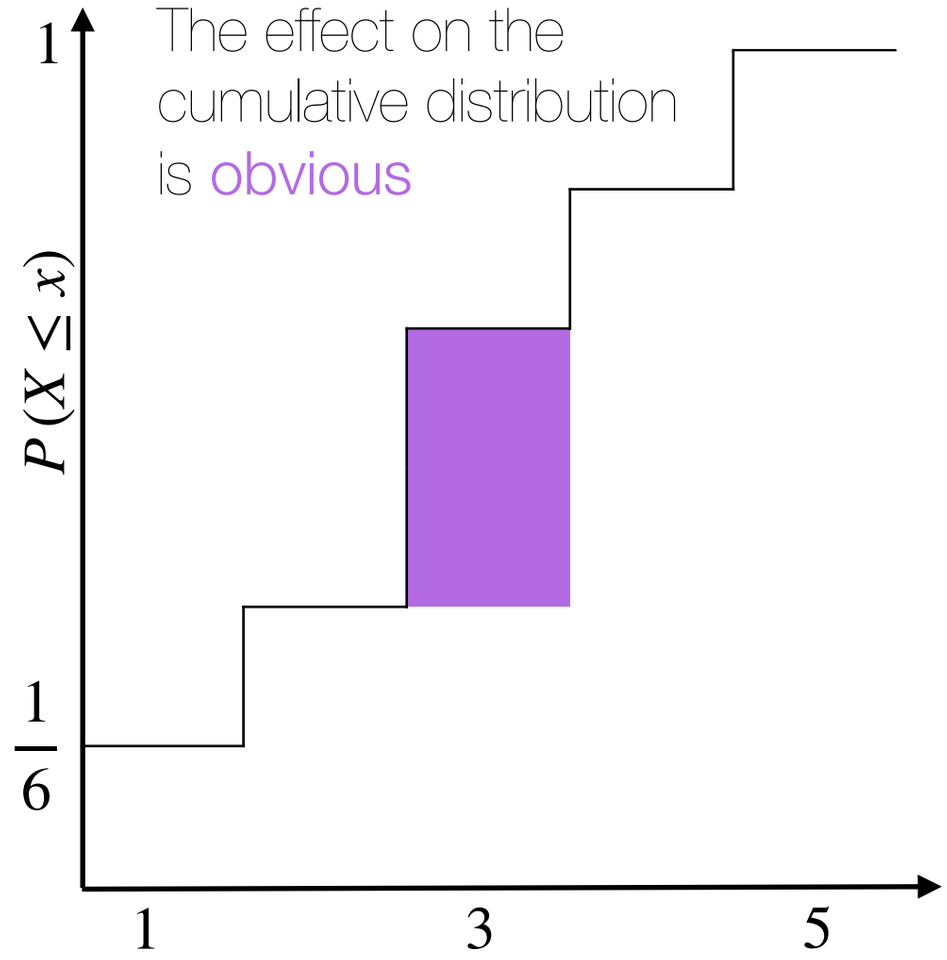
Probability Distribution

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



Cumulative Distribution

The effect on the cumulative distribution is obvious



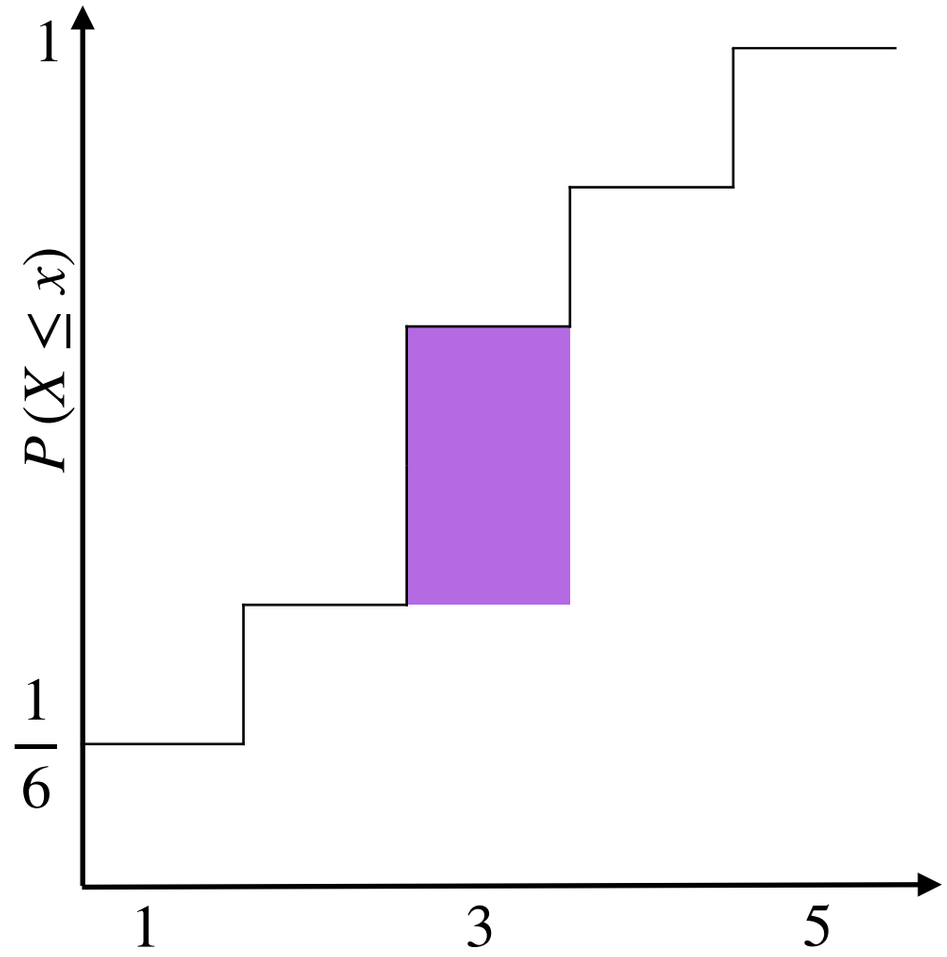
Non-Uniform Distribution

- 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



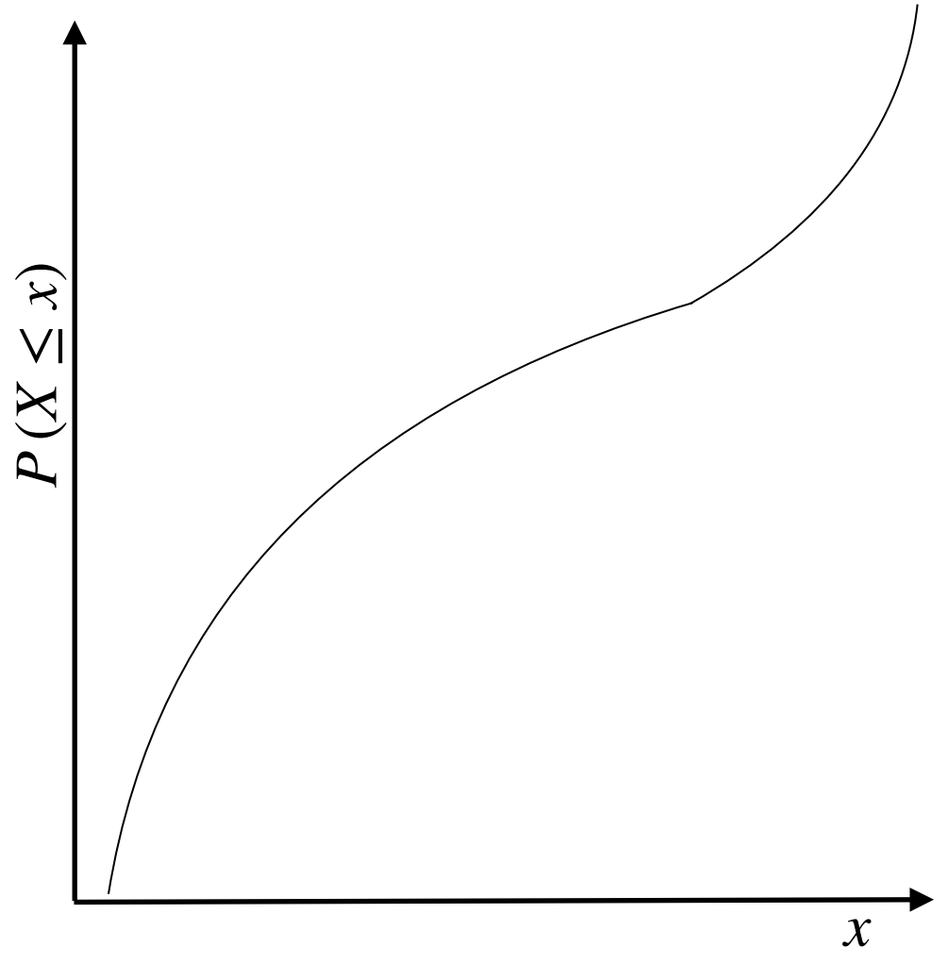
Generating arbitrary distributions

- 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



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?

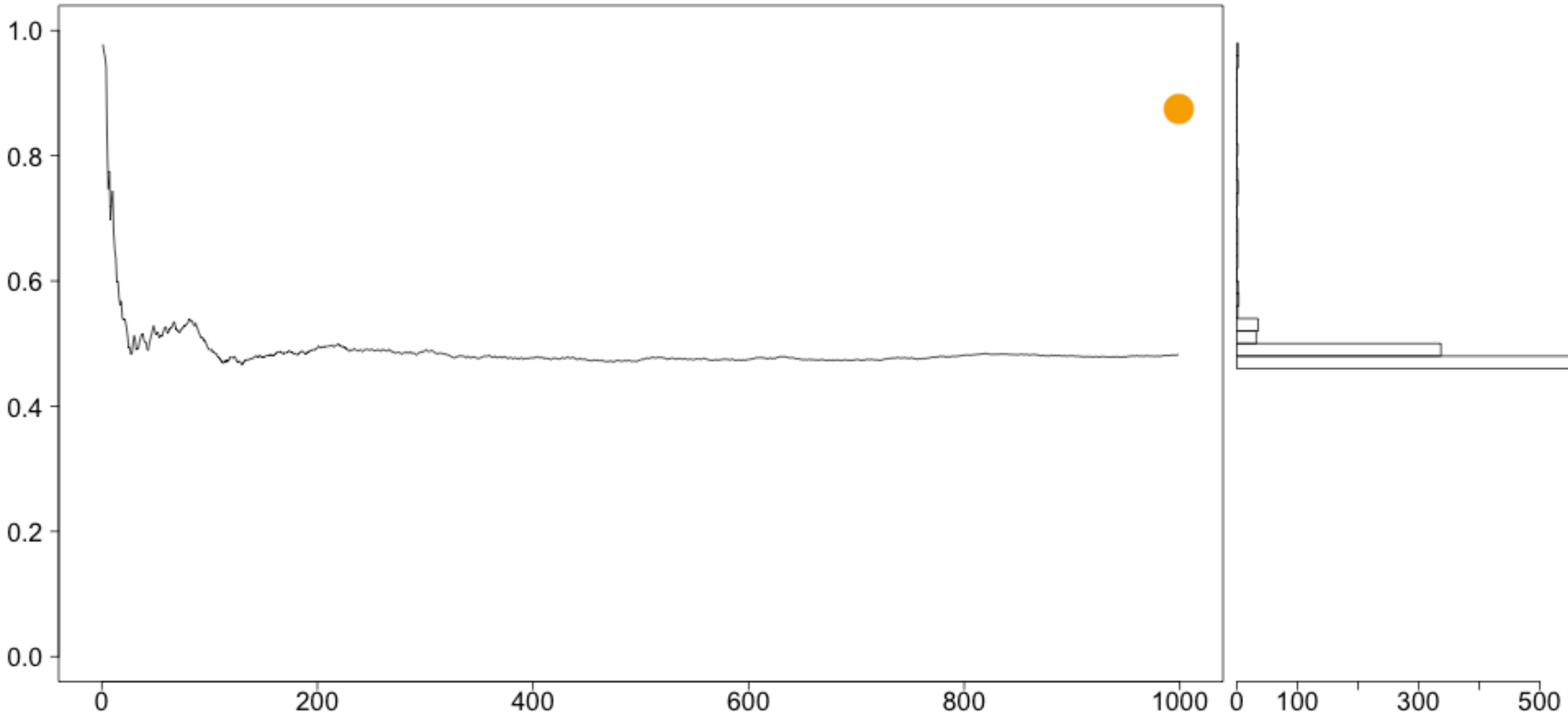
- For any given set of dice rolls, we are only estimating the true value.

- 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:

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

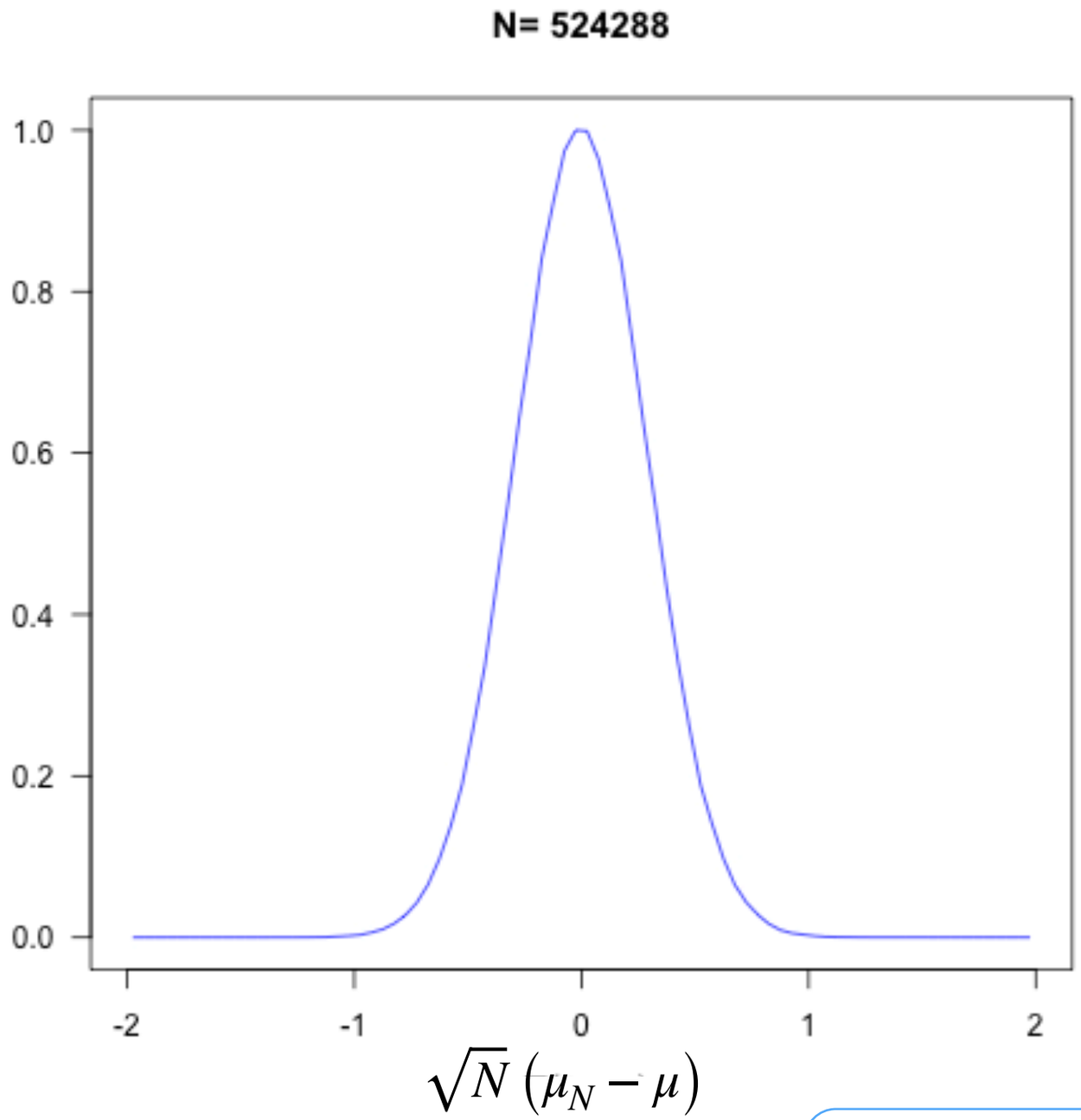
Error Estimate

- 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:

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

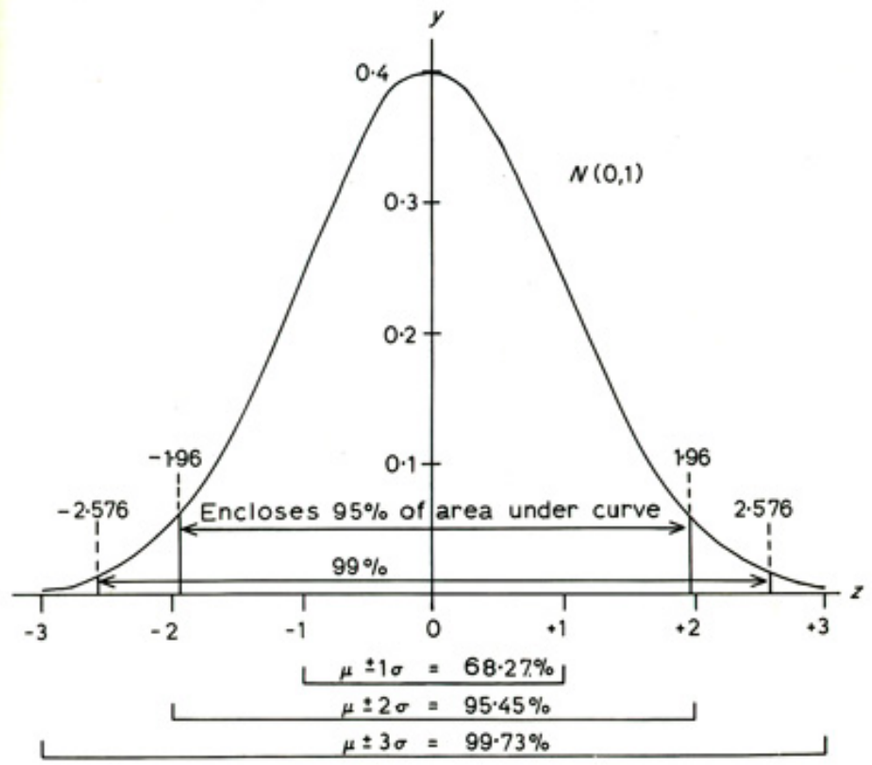
- The mean value is:

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

- 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

Code - Probability Distributions

https://github.com/DataForScience/Probability-And-Statistics

4. Bayesian Statistics

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:

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

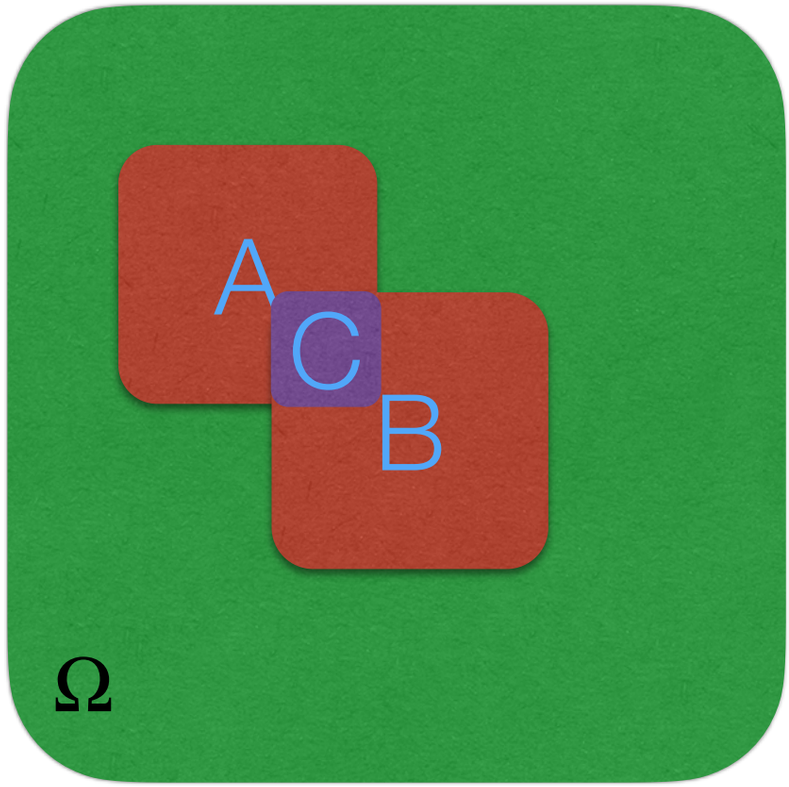
$$P(C) = P(B|A)\,P(A)$$

- 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.

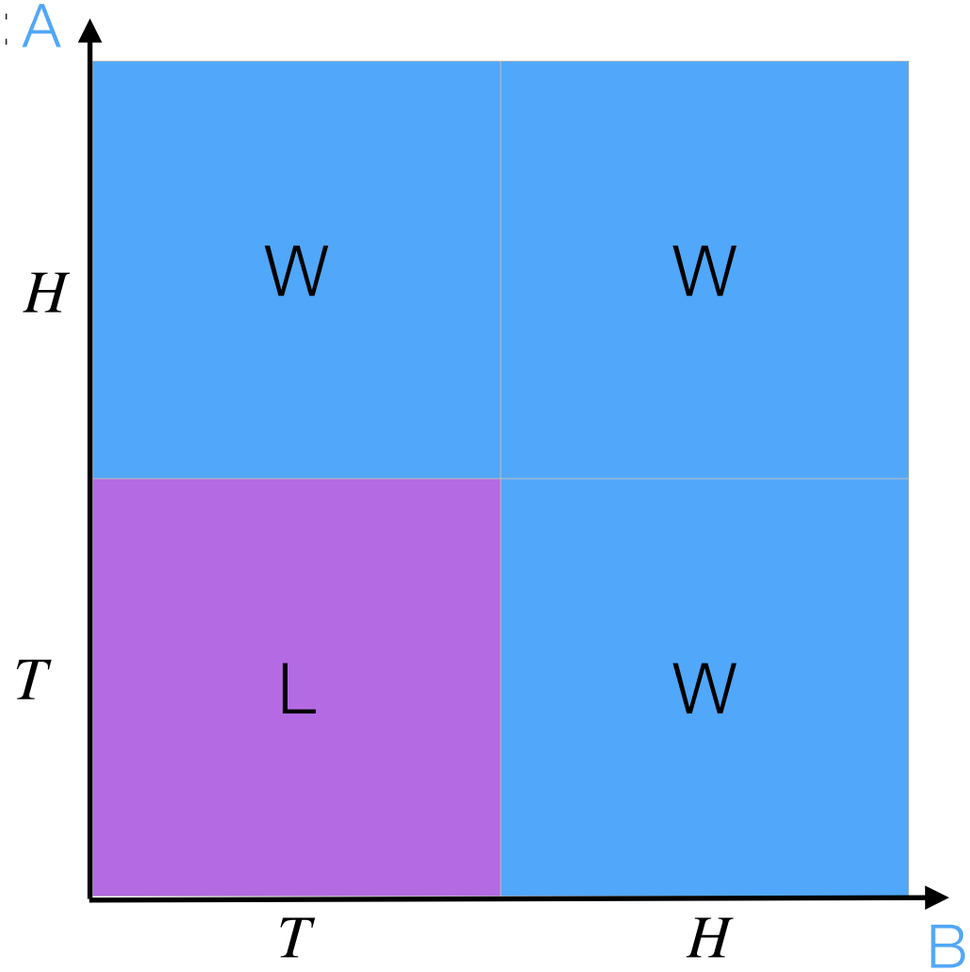
- We also know that

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

- 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

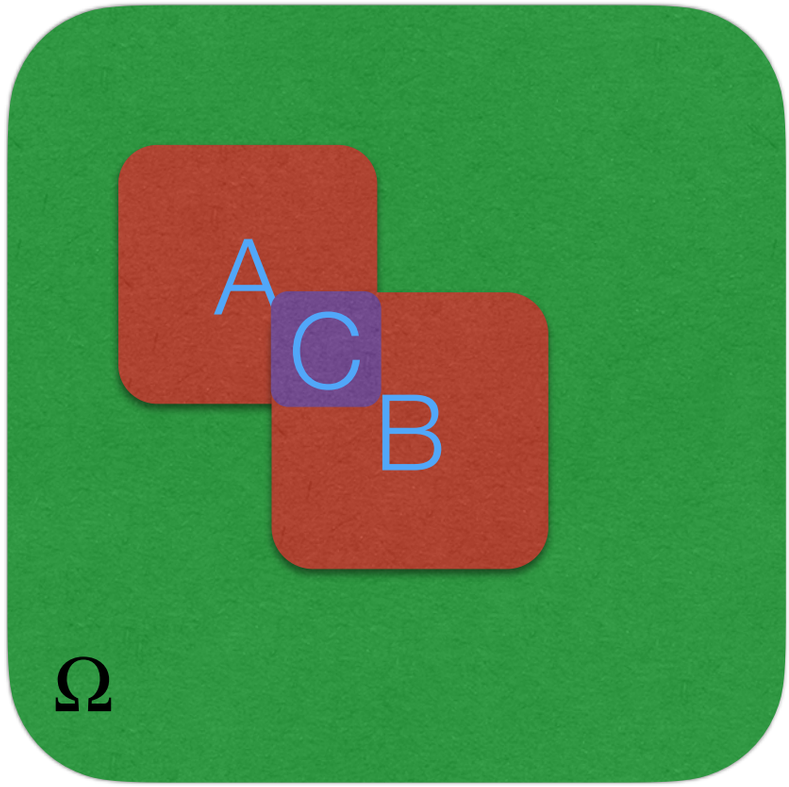
- A simple way of remembering this formula is to remember that:

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

- 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

Likelihood

Posterior

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

Prior

Evidence

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

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

$$P\left(disease \,|\, negative\ test\right) = \frac{TP}{TN + FN} = 0.99999$$

Medical Tests

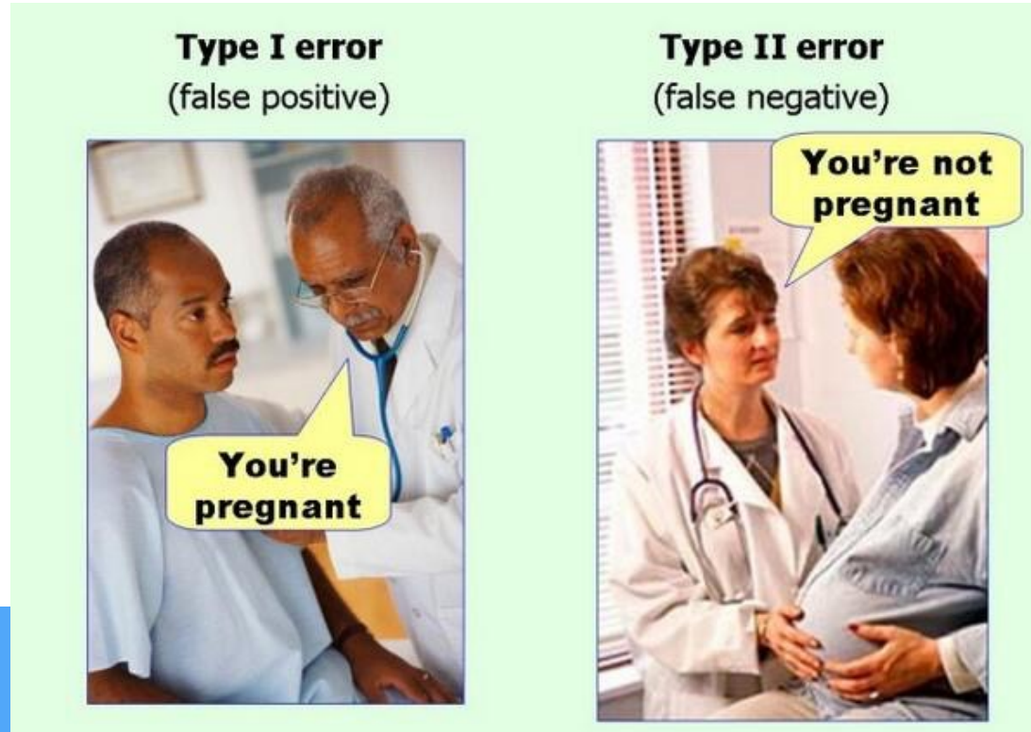
Consider a population of 1,000,000 individuals. The numbers we should expect are:

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:

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

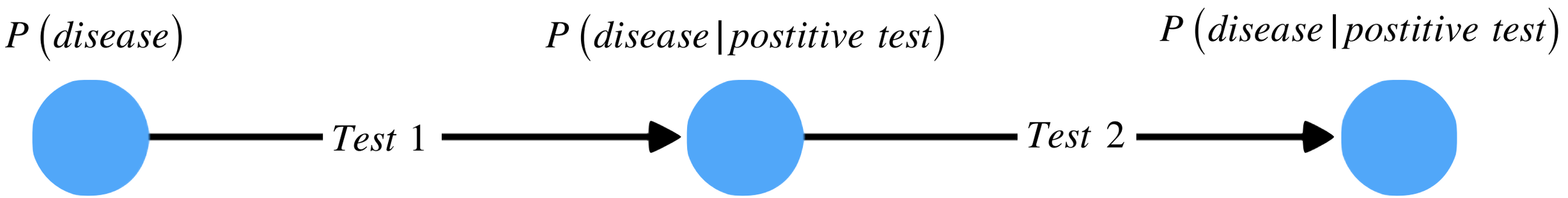
but now the probability that we have the disease has been updated:

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

So this time we find:

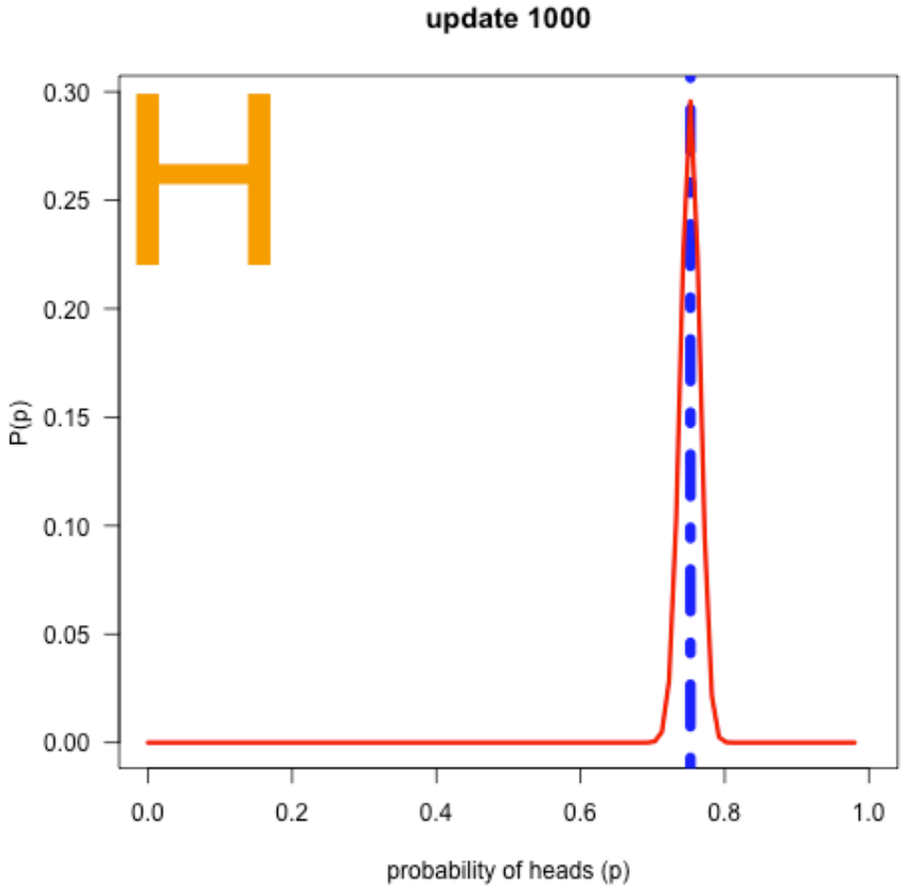
$$P\left(disease|positive\ test\right) = 0.4949$$

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



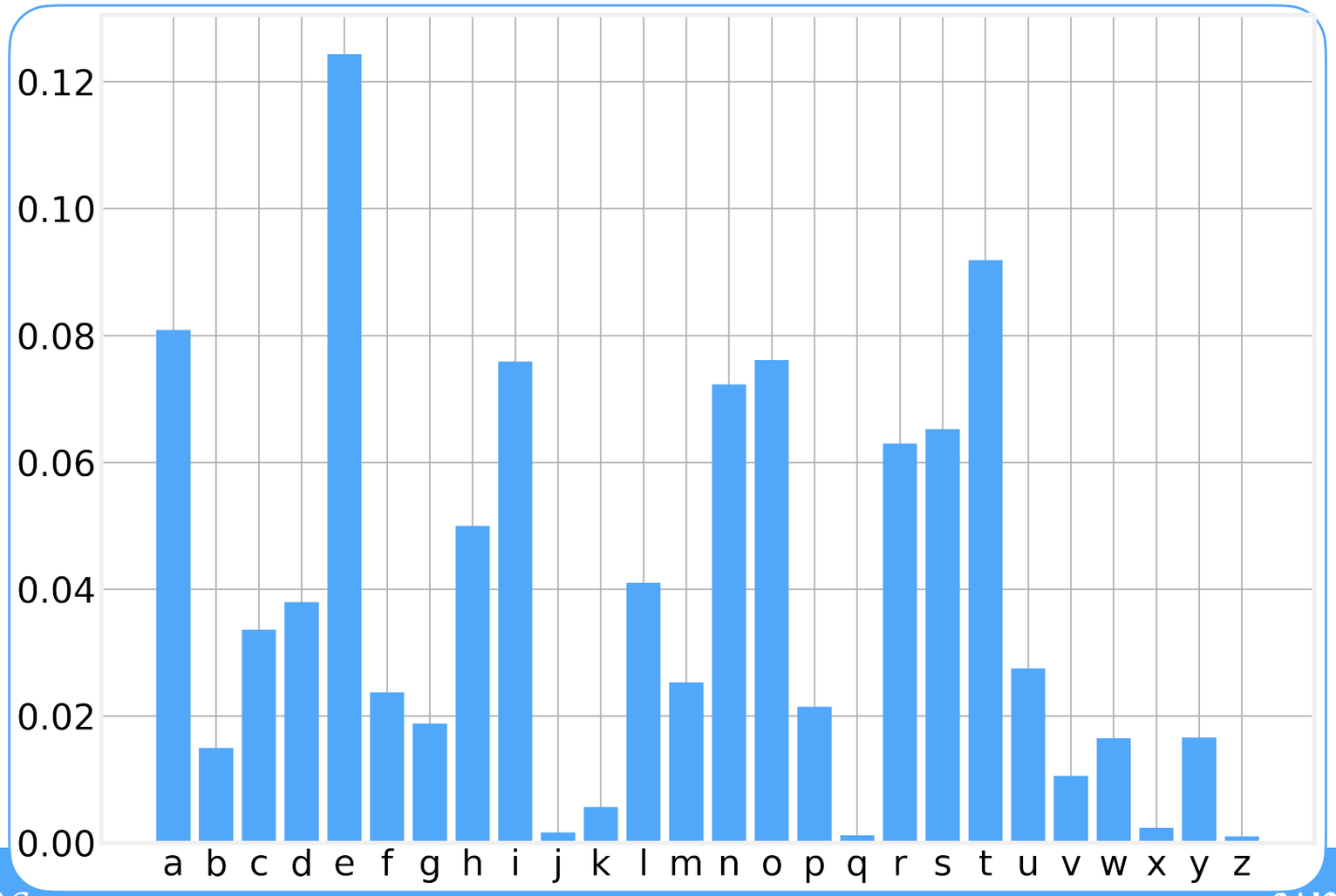
$$P\left(p\,|\,heads\right) = \frac{P\left(heads\,|\,p\right)P\left(p\right)}{P\left(heads\right)}$$

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

http://youtu.be/GTx0D8VY0CY

Language Detection (Naive Bayes Classification)

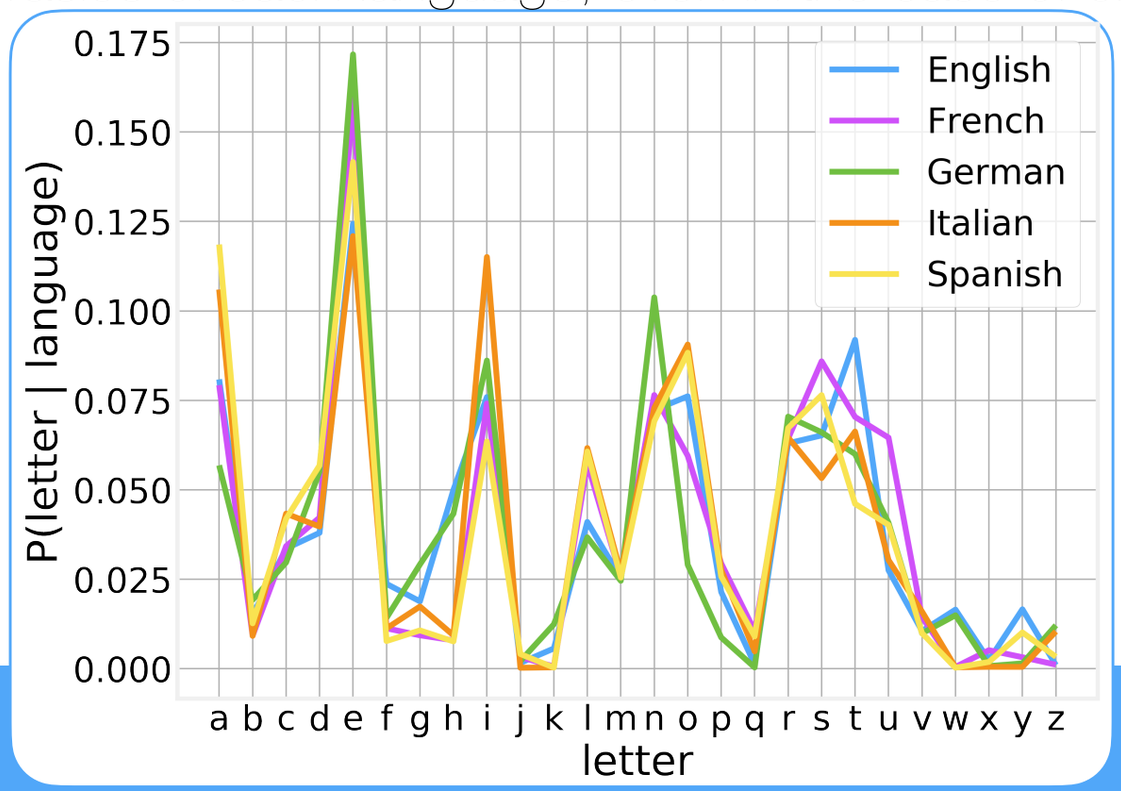
- Language detection allows us to preprocess our corpus so that we can focus on specific languages, sort documents by language, etc.
- Languages can be characterized by their character (letter) distribution.
- The character level distribution for English, obtained using Google Books 1-gram dataset is:



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.

Code - Bayesian Statistics

https://github.com/DataForScience/Probability-And-Statistics

5. A / B Testing

Randomized Controlled Trial

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

- The classical question we are trying to answer is: Does my new treatment have an actual effect?
- Each patient either gets the medication or doesn't. No redos!
- How can we make sure the outcome doesn't depend on the individual patients?

Randomized Controlled Trial

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

- The classical question we are trying to answer is: Does my new treatment have an actual effect?
- Each patient either gets the medication or doesn't. No redos!
- How can we make sure the outcome doesn't depend on the individual patients?

Randomly Assign Patients

- Make sure that no specific individual characteristic determines the group (treatment/placebo) a specific patient is assigned to.
- Compare the outcomes

Hypothesis Testing

- 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

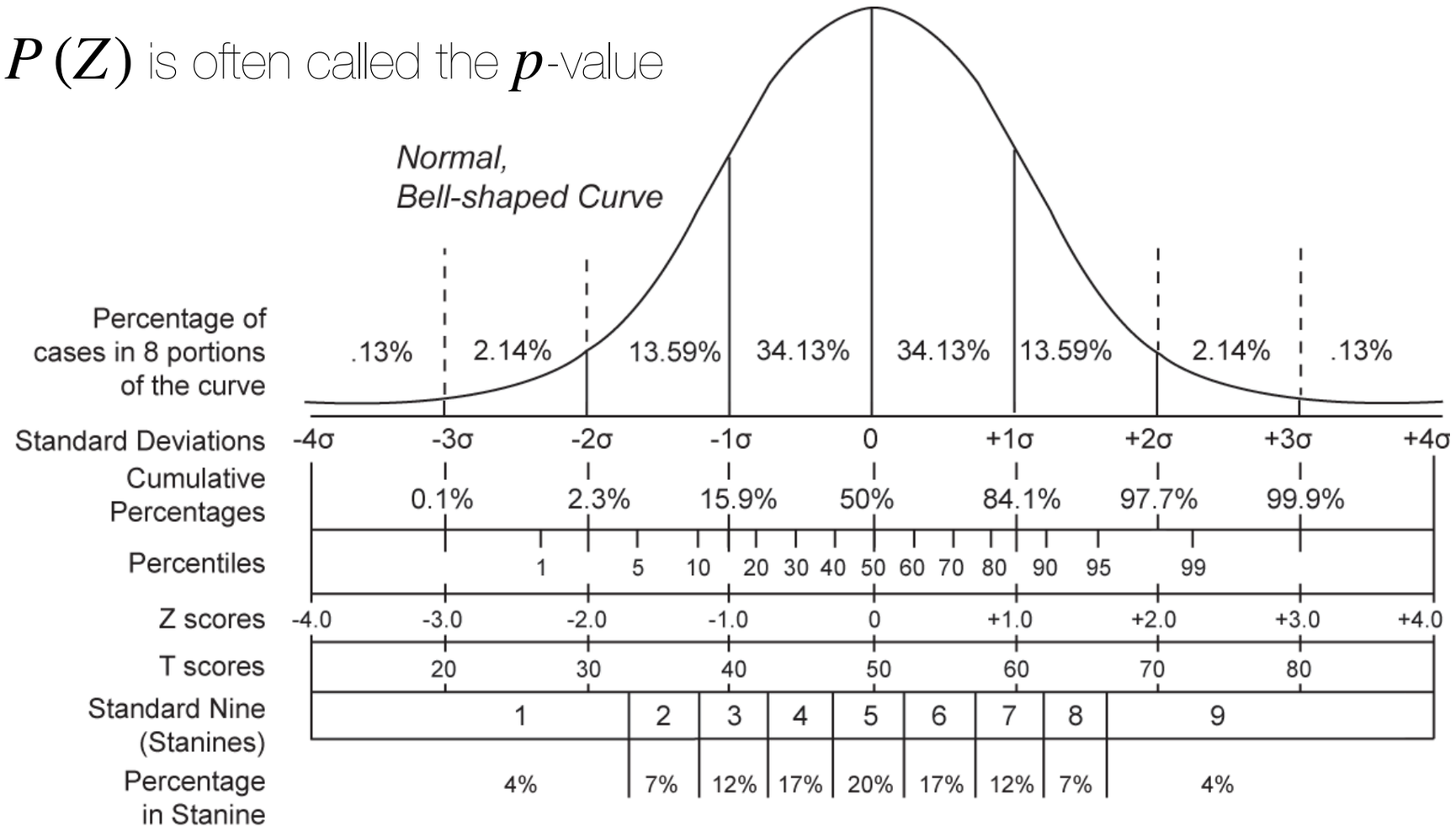
A/B Testing

- 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

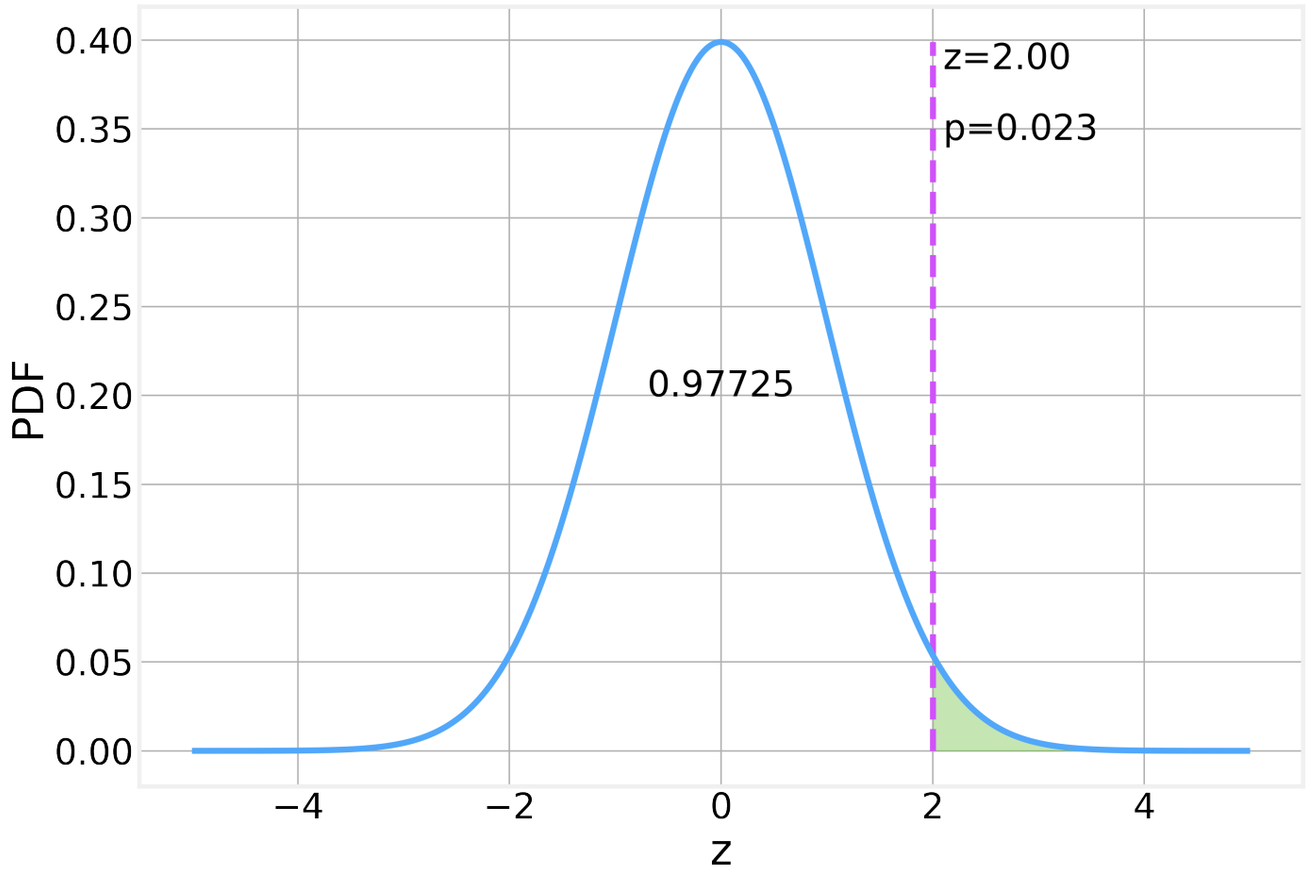
$$Z = \frac{p_A - p_B}{\sqrt{SE_A^2 + SE_B^2}}$$

- 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

	Candidates	Acceptance Rate
Men	8442	0.44
Women	4321	0.35

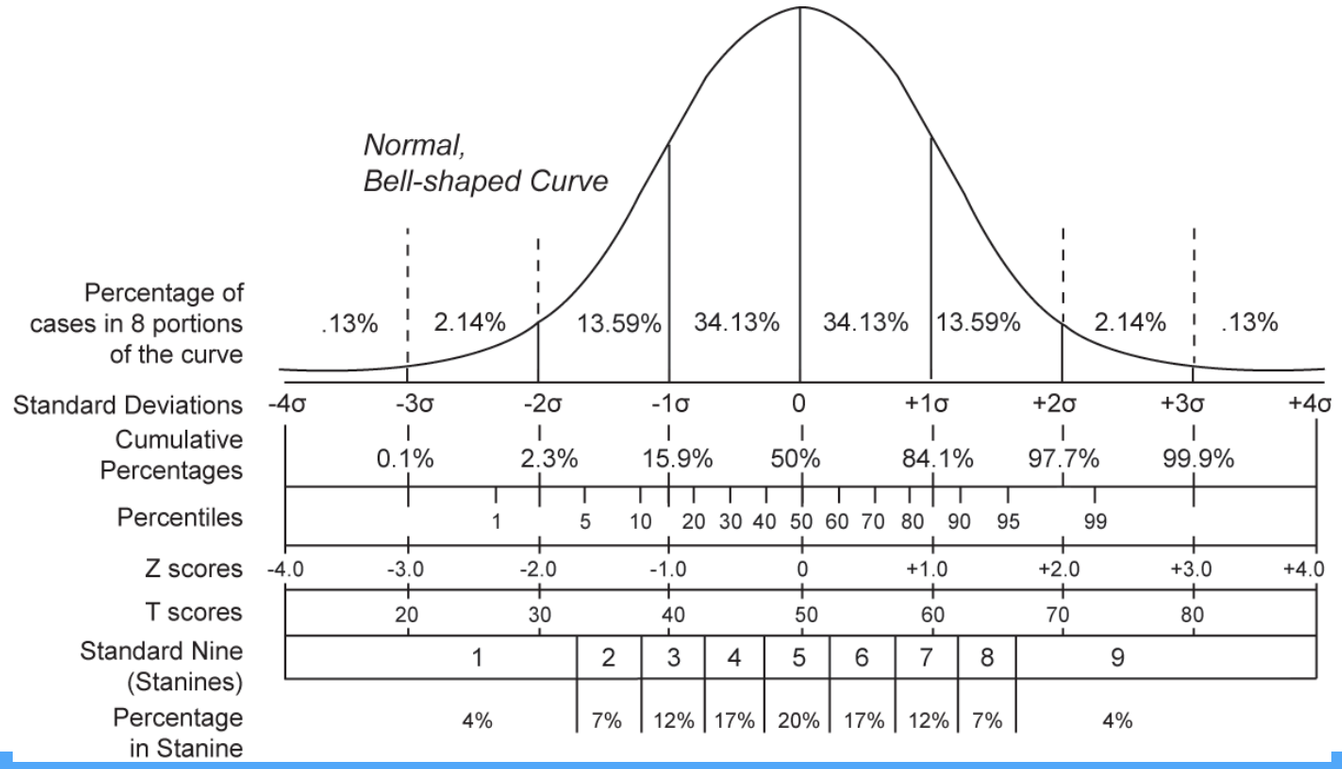
Were women being discriminated against when they applied to Berkley?

Berkeley Discrimination Case

	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?

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



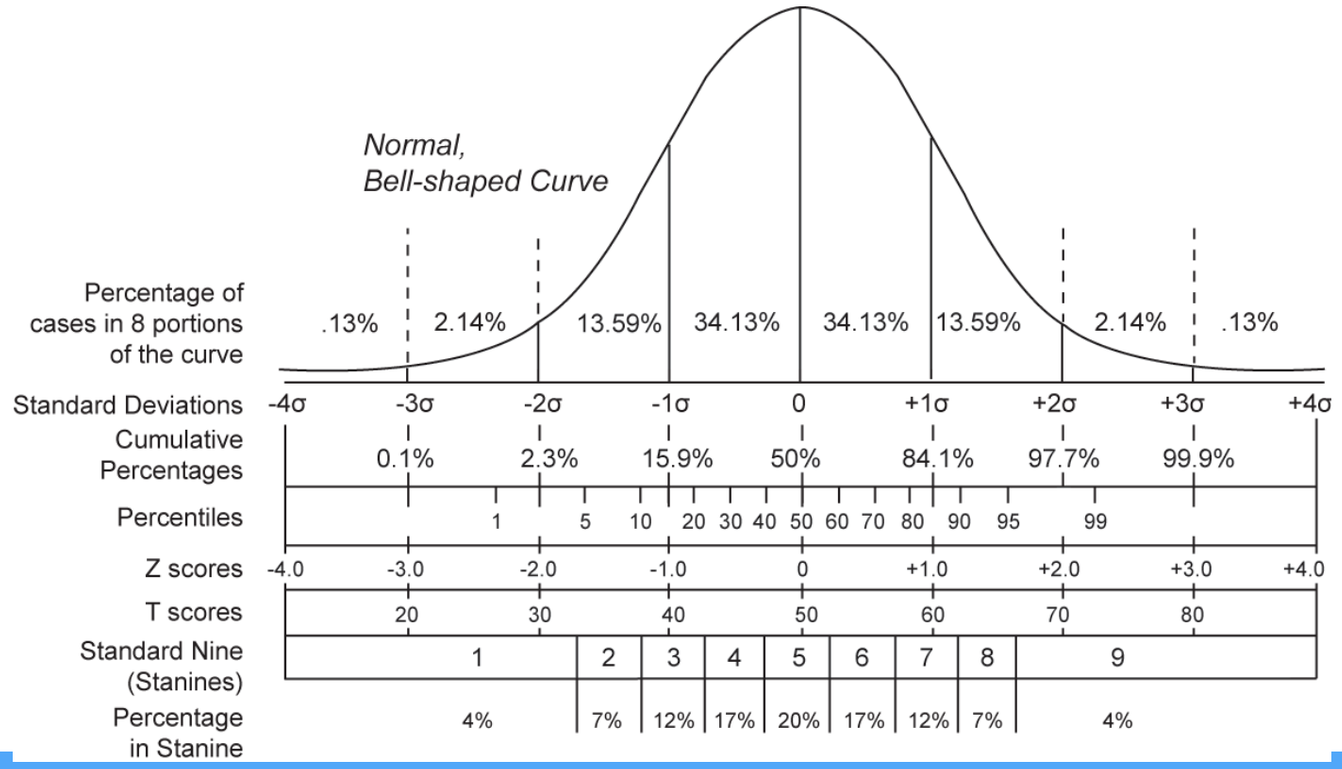
Berkeley Discrimination Case

	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?

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

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



Berkeley Discrimination Case

Science 187, 398 (1975)

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

	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
Total	2590	0.46	1835	0.30

Berkeley Discrimination Case

Science 187, 398 (1975)

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

	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
Total	2590	0.46	1835	0.30

Simpson's Paradox

Science 187, 398 (1975)

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

	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
Total	2590	0.46	1835	0.30

Simpson's Paradox

https://en.wikipedia.org/wiki/Simpson%27s_paradox

"aggregated data can appear to reverse important trends in the numbers being combined"

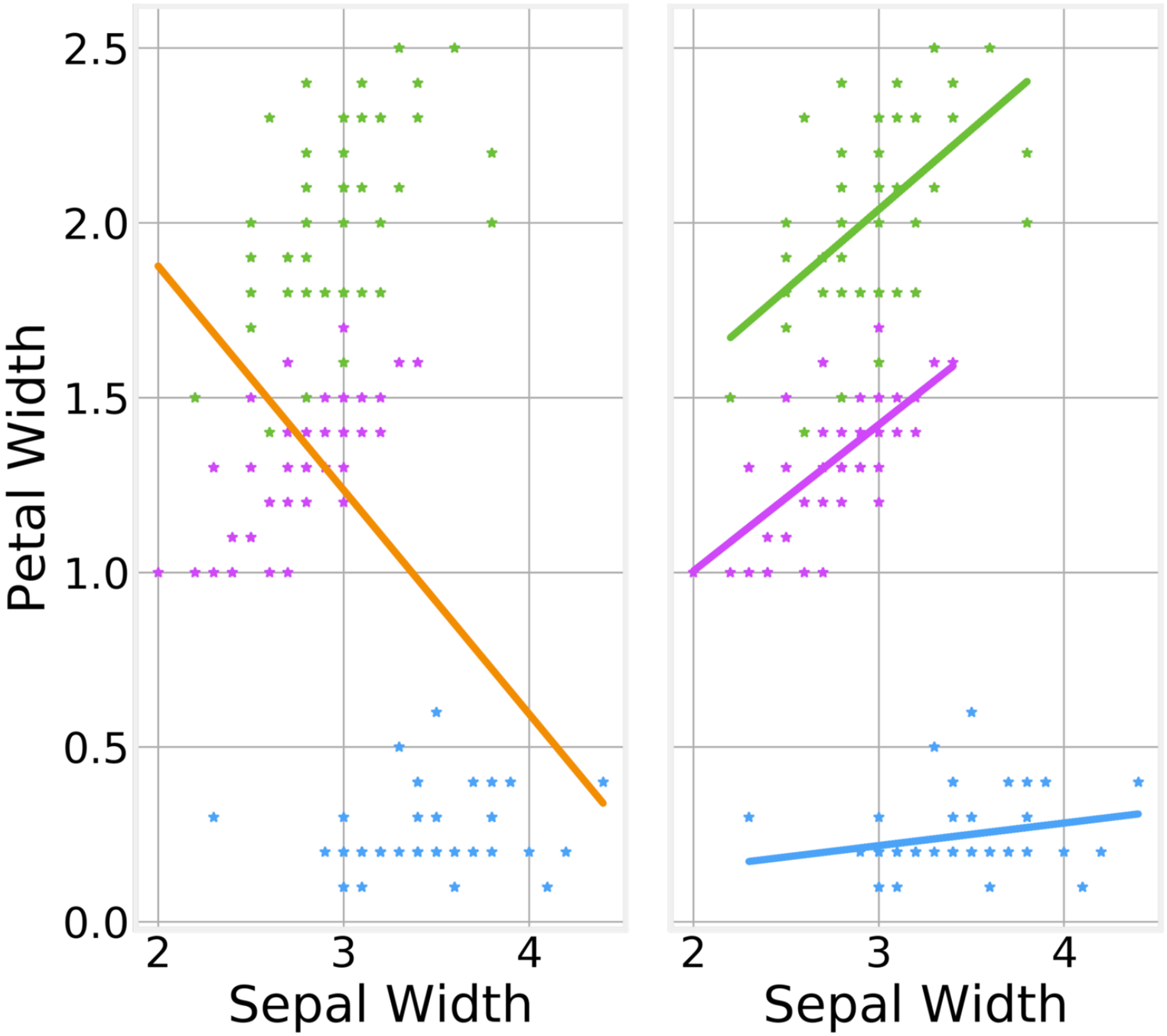
WSJ, Dec 2, 2009

- Simpson's Paradox is likely to appear whenever you have confounding factors.

- In the Berkeley Case, we had two factors:
  - Men and Women prefer different departments
  - Departments have widely varying acceptance rates
- Most women applied to departments with low acceptance rates, while most men applied to departments with high acceptance rates

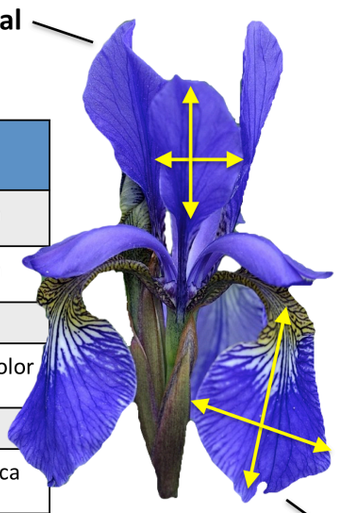
Simpson's Paradox

https://en.wikipedia.org/wiki/Simpson%27s_paradox



Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

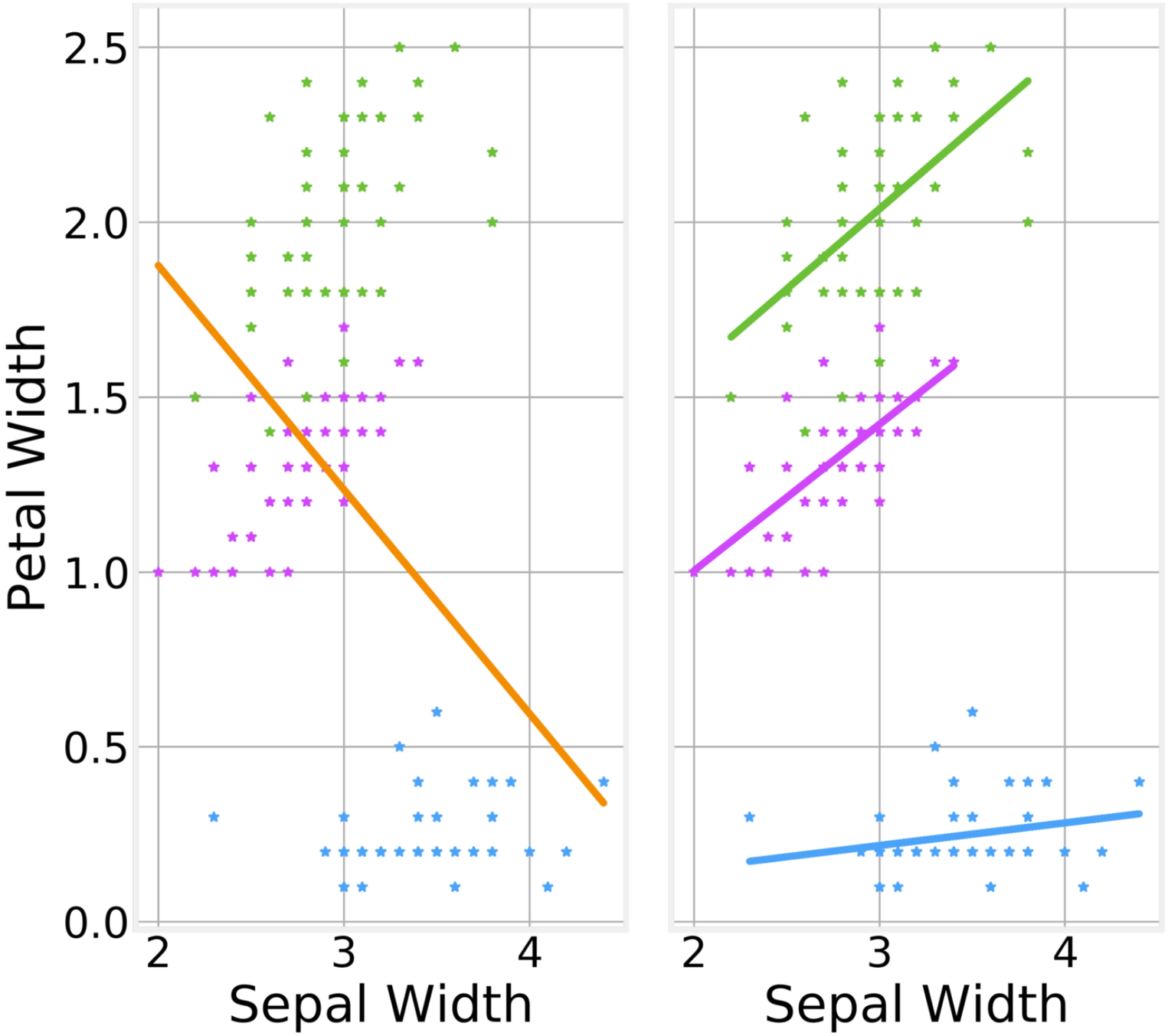


Class labels
(targets)

Features
(attributes, measurements, dimensions)

Simpson's Paradox

https://en.wikipedia.org/wiki/Simpson%27s_paradox



Understanding the data generating process (the history behind the data) is paramount to properly understand causal relationships

Code - A / B Testing

https://github.com/DataForScience/Probability-And-Statistics

Question

- How was the technical level?
    - 1 — Too Low (too many details)
    - 2 — Low
    - 3 — Just Right
    - 4 — High
    - 5 — Too High (not enough details)

Question

- How was the level of Python code/explanations?
    - 1 — Too Low (too many details)
    - 2 — Low
    - 3 — Just Right
    - 4 — High
    - 5 — Too High (not enough details)

Events

graphs4sci.substack.com

Graphs for Data Science

May 10, 2023 - 10am-2pm (PST)

graphs4sci.substack.com

Deep Learning for Everyone

Jun 20, 2023 - 10am-2pm (PST)

VIDEO

Times Series Analysis for Everyone

★★★★★ 1 review

By Bruno Goncalves

https://bit.ly/Timeseries_LL

TIME TO COMPLETE:
6h

TOPICS:
Time Series

PUBLISHED BY:
Pearson

PUBLICATION DATE:
November 2021

Start

6 Hours of Video Instruction

The perfect introduction to time-based analytics

Overview

Times Series Analysis for Everyone LiveLessons covers the fundamental tools and techniques for the analysis of time series data. These lessons introduce you to the basic concepts, ideas, and algorithms necessary to develop your own time series applications in a step-by-step, intuitive fashion. The lessons follow a gradual progression, from the more specific to the more abstract, taking you from the very basics to some of the most recent and sophisticated algorithms by leveraging the statsmodels, arch, and Keras state-of-the-art models.

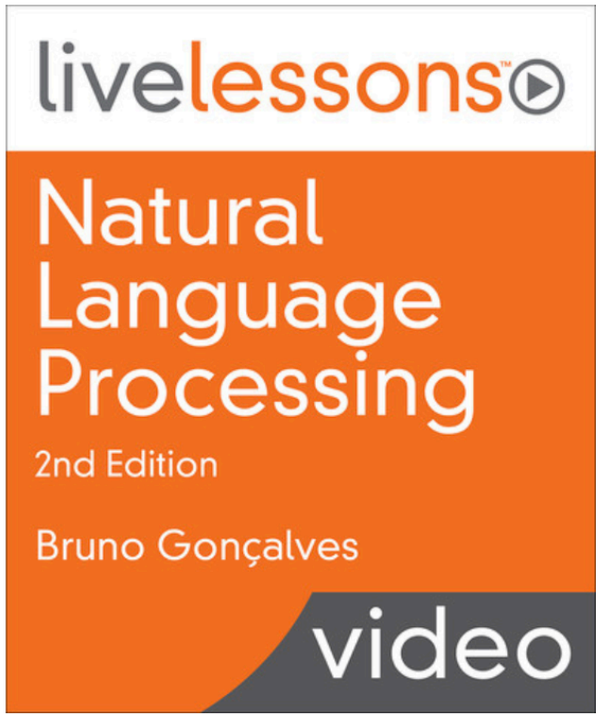
VIDEO

Natural Language Processing, 2nd Edition

Write the first review

By Bruno Goncalves

https://bit.ly/NLP_LL



TIME TO COMPLETE:
5h 23m

TOPICS:
Natural Language Processing

PUBLISHED BY:
Addison-Wesley Professional

PUBLICATION DATE:
October 2021

Start

5 Hours of Video Instruction

Overview

Natural Language Processing LiveLessons covers thefundamentals of Natural Language Processing in a simple and intuitive way,empowering you to add NLP to your toolkit. Using the powerful NLTK package, itgradually moves from the basics of text representation, cleaning, topicdetection, regular expressions, and sentiment analysis before moving on to theKeras deep learning framework to explore more advanced topics such as textclassification and sequence-to-sequence models. After successfully completingthese lessons you'll be equipped with a fundamental and practical understandingof state-of-the-art Natural Language Processing tools and algorithms.

END