CSCI 145: Probability for Computing and Data Analysis

Lecture 0: Overview & Course Details

Instructor: Alessio Mazzetto & Eli Upfal

Brown University Computer Science

Fun Quiz

- **ENIAC** (**Electronic Numerical Integrator and Computer**) was the first programmable, electronic, general-purpose digital computer, completed in 1945.

- What was ENIAC mainly used for?
  1. Breaking codes
  2. Image processing
  3. Monte Carlo (probabilistic) simulations
  4. Watching movies

Fun Quiz

- **ENIAC Electronic Numerical Integrator and Computer**) was the first programmable, electronic, general-purpose digital computer, completed in 1945.

- What was ENIAC mainly used for?
  1. Breaking codes
  2. Image processing
  3. Monte Carlo (probabilistic) simulations
  4. Watching movies

Von Neumann, Nicholas Metropolis and others programmed the ENIAC computer to perform the first fully automated Monte Carlo calculations, https://en.wikipedia.org/wiki/ENIAC

Fun Quiz 2

- I flip 3 fair (½, ½) coins. I show you that two of the coins came up Head. What is the probability that the third coin is also Head?

1. 1/2
2. 1/3
3. 1/4
4. 1/8

Fun Quiz 2

- I flip 3 fair (½, ½) coins. I show you that two of the coins came up Head. What is the probability that the third coin is also Head?

1. 1/2
2. 1/3
3. 1/4
4. 1/8

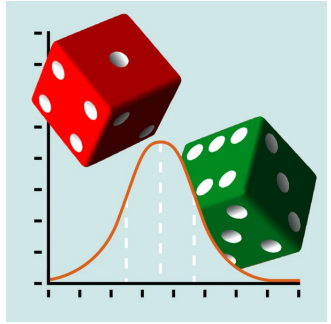
Outcomes of flipping 3 coins:

TTT HTT THT TTH - non of these outcomes

HHH THH HTH HHT - 1 out of these 4 outcomes

CS145: Lecture 0 Outline

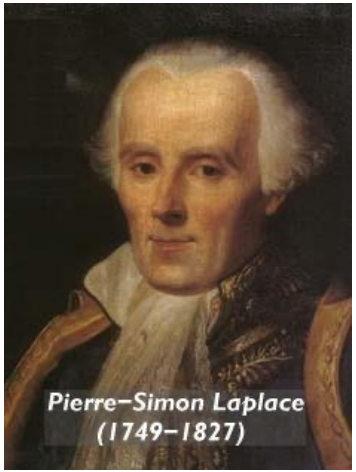
- Why probability and statistics in CS?
- Course overview: Probability and statistics - key concepts & applications in CS
- Course prerequisites
- Course work and evaluation
- *Who should take this class?*
- *Registration, administration, tech details*



Why Probability?

``It is remarkable that this science, which originated in the consideration of games and chances, should have become the most important object of human knowledge... The most important questions of life are, for the most part, really only problems of probability"

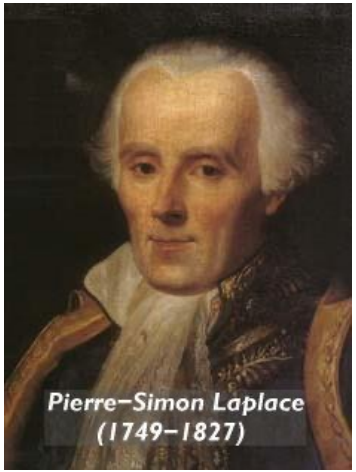
Pierre Simons, Marquis de Laplace (1749-1827).



Why Probability?

``It is remarkable that this science, which originated in the consideration of games and chances, should have become the most important object of human knowledge... The most important questions of life are, for the most part, really only problems of probability"

Pierre Simons, Marquis de Laplace (1749-1827).



Laplace didn't know about:

- Genomics and DNS recombination
- Quantum mechanic (and computing)
- Stochastic finance
- Machine learning and AI
- Statistics

Why Probability?

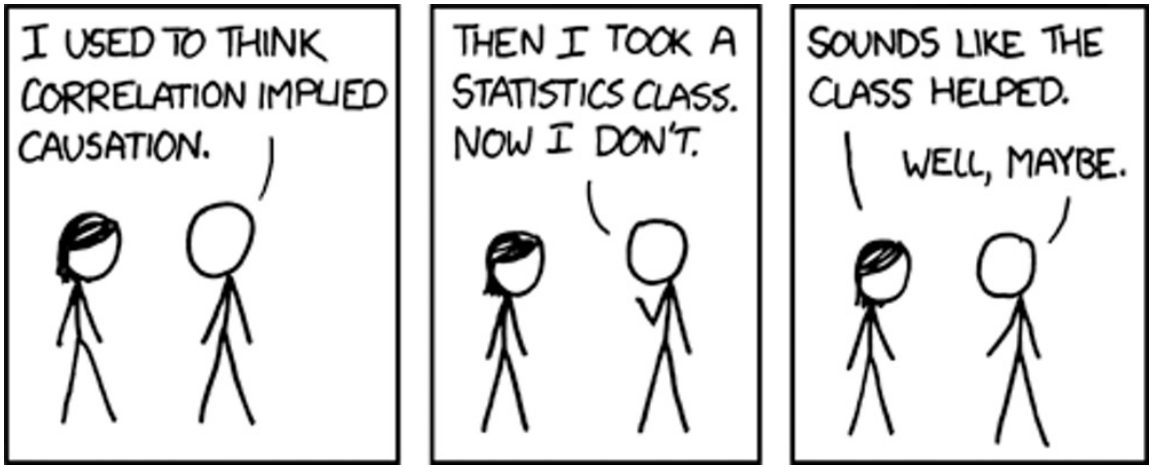
Most advanced computer applications involve randomization:

- Secured web connections are *probabilistically secured*
- Web search engines apply *statistical inference*
- Computer games would be boring without *randomization*
- Spam filters, recommendation systems, web advertising and face recognition use *(statistical) machine learning*

- Efficient data structures are often *randomized* (e.g., hashing)
- Computational finance, computational biology, climate and weather forecast ….

Do I need to understand probability?

- Do I really need to study probability and statistics?
- Isn't it a lot of theory?
- Cannot I just use common sense?



Probability Is Not Intuitive

- I flip 3 fair (½, ½) coins. I show you that two of the coins came up Head. What is the probability that the third coin is also Head?

1. 1/2
2. 1/3
3. 1/4
4. 1/8

Probability Is Not Intuitive

- I flip 3 fair (½, ½) coins. I show you that two of the coins came up Head. What is the probability that the third coin is also Head?

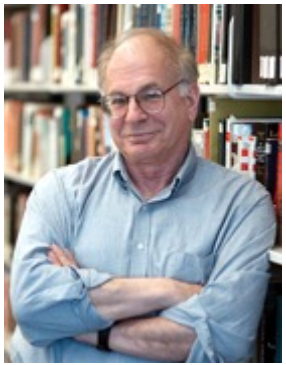
1. 1/2
2. 1/3
3. 1/4
4. 1/8

Outcomes of flipping 3 coins:

TTT HTT THT TTH - non of these outcomes

HHH THH HTH HHT - 1 out of these 4 outcomes

Probability is Often Counterintuitive



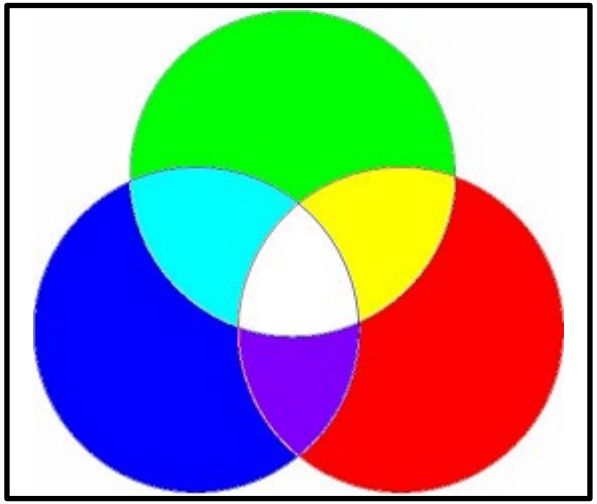
Daniel Kahneman was awarded the 2002 Nobel Prize in Economic Sciences for "…. challenging the assumption of human rationality prevailing in decision-making under uncertainty".

CS145: Lecture 0 Outline

- Why probability and statistics in CS?
- Course overview: Probability and statistics - key concepts & applications
- Course prerequisites:
- Course evaluation:
- *Who should take this class?*
- *Registration, administration, tech details*

Events and Probabilities

Venn diagram for all combinations of 3 binary (true/false) events.



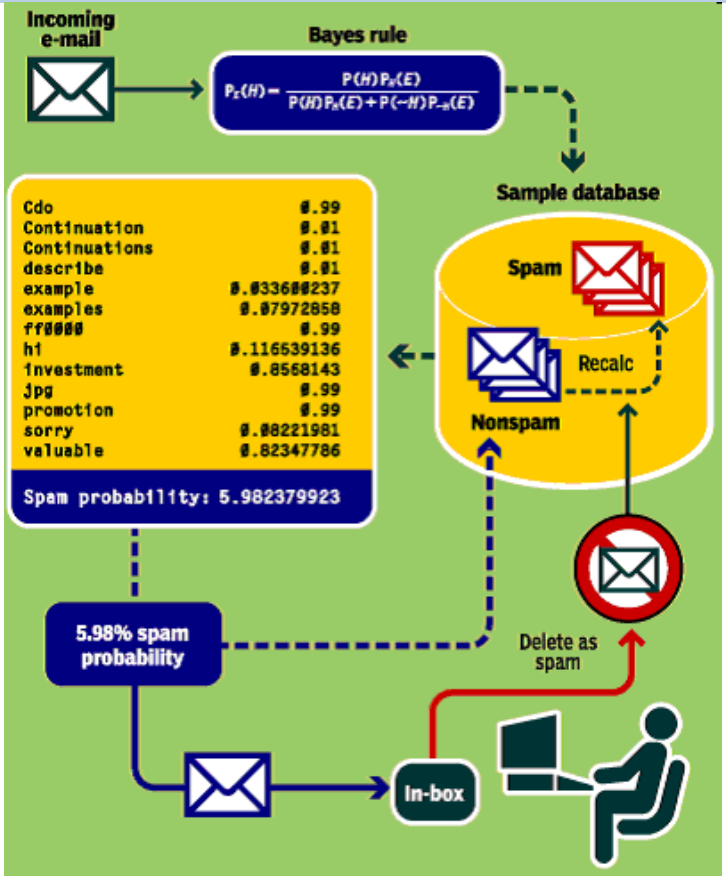
Weather example:

- *Raining or not*
- *Sunny or not*
- *Hot or not*

- **Sample spaces:** When a random event happens, what is the set of all possible outcomes? May be discrete or continuous.
- **Conditioning:** Suppose I observe some data. How does my probability model change?
- **Independence:** Is there any relationship between pairs of variables in my model? Would data provide knowledge?

Bayesian Spam Filtering

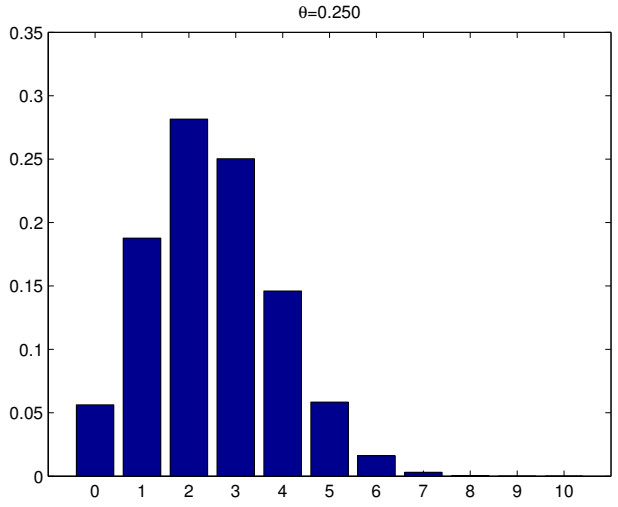
- Binary classification: Is this e-mail useful (ham) or spam?
- Training data: Messages previously marked as spam
- Estimate: Probability that certain words are used in spam and non-spam emails
- Classify: Conditional probability that a mail is spam given the words in the mail.

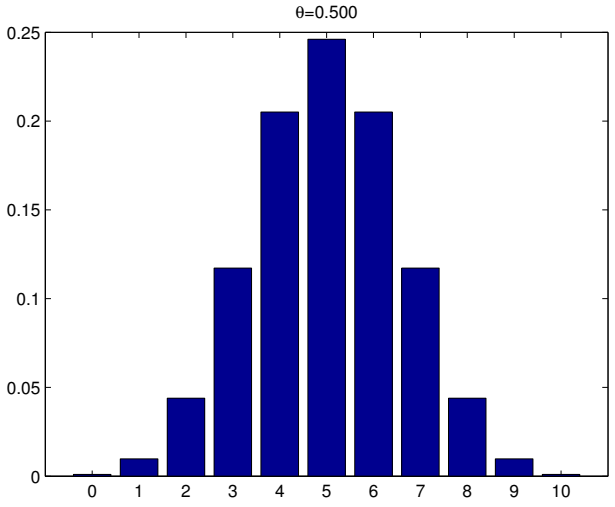


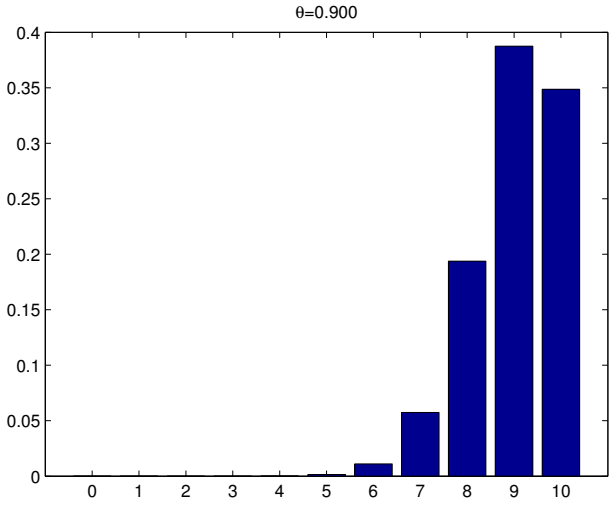
Spam Filter Express: *http://www.spam-filter-express.com/*

Discrete Random Variables

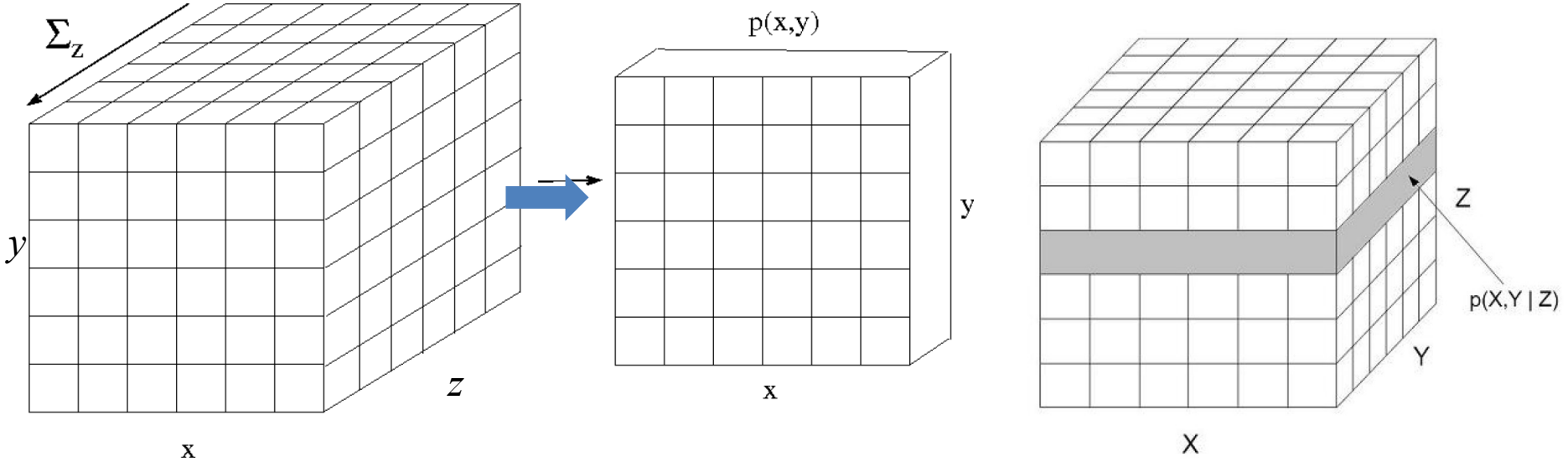
- Suppose I toss a coin 10 times.
- The number of tosses that come up heads, rather than tails, is an example of a *discrete random variable*.
- A *probability mass function* gives the (non-negative) probability of each possible outcome. These probabilities sum to one.







Joint, Marginal, & Conditional Distributions

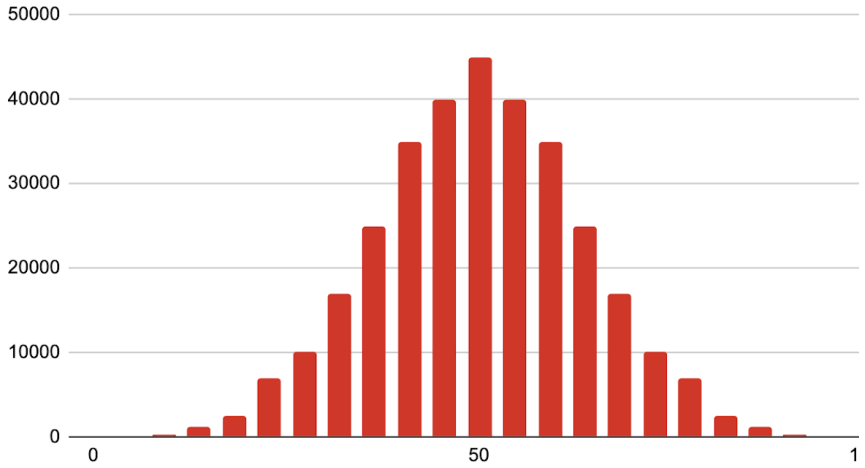


Example: (age, height, weight)

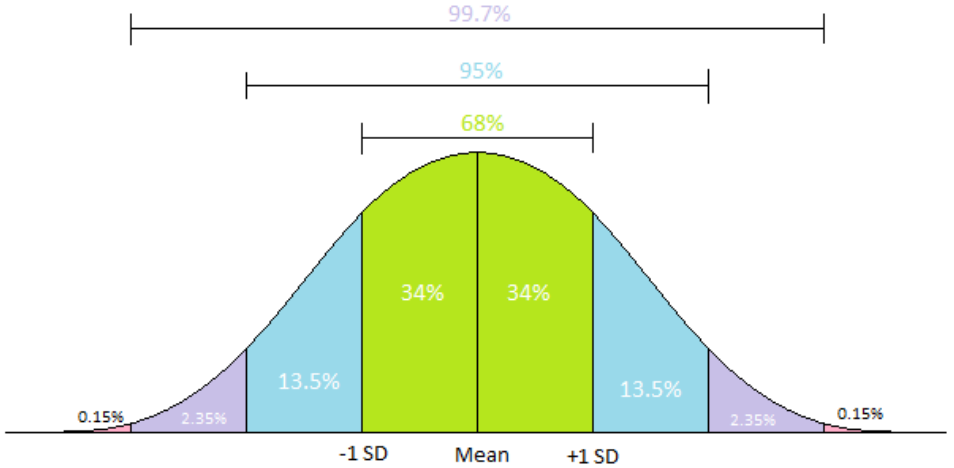
- **Joint Distribution:** Probability of each possible outcome.
- **Marginal Distribution:** If some variables are not observed and not relevant, how do I remove them from the model?
- **Conditional Distribution:** What if I observe some data?

Expectation, Variance, & Standard Deviation

What can you **expect,** and how **confident** can you be with this expectation?



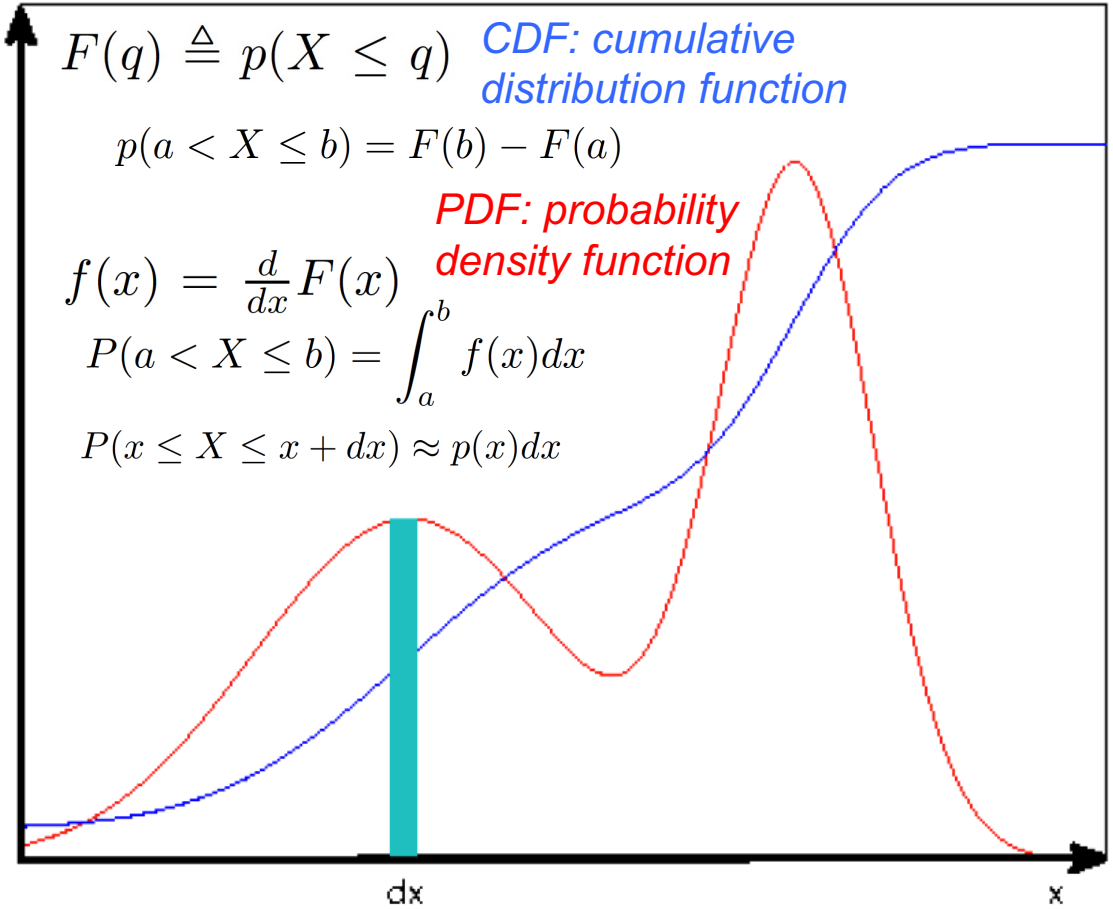
Number of Heads in 100 Coin Flips 200,000 runs.



https://cetking.com/solving-standard-deviations-sums-without-formula/

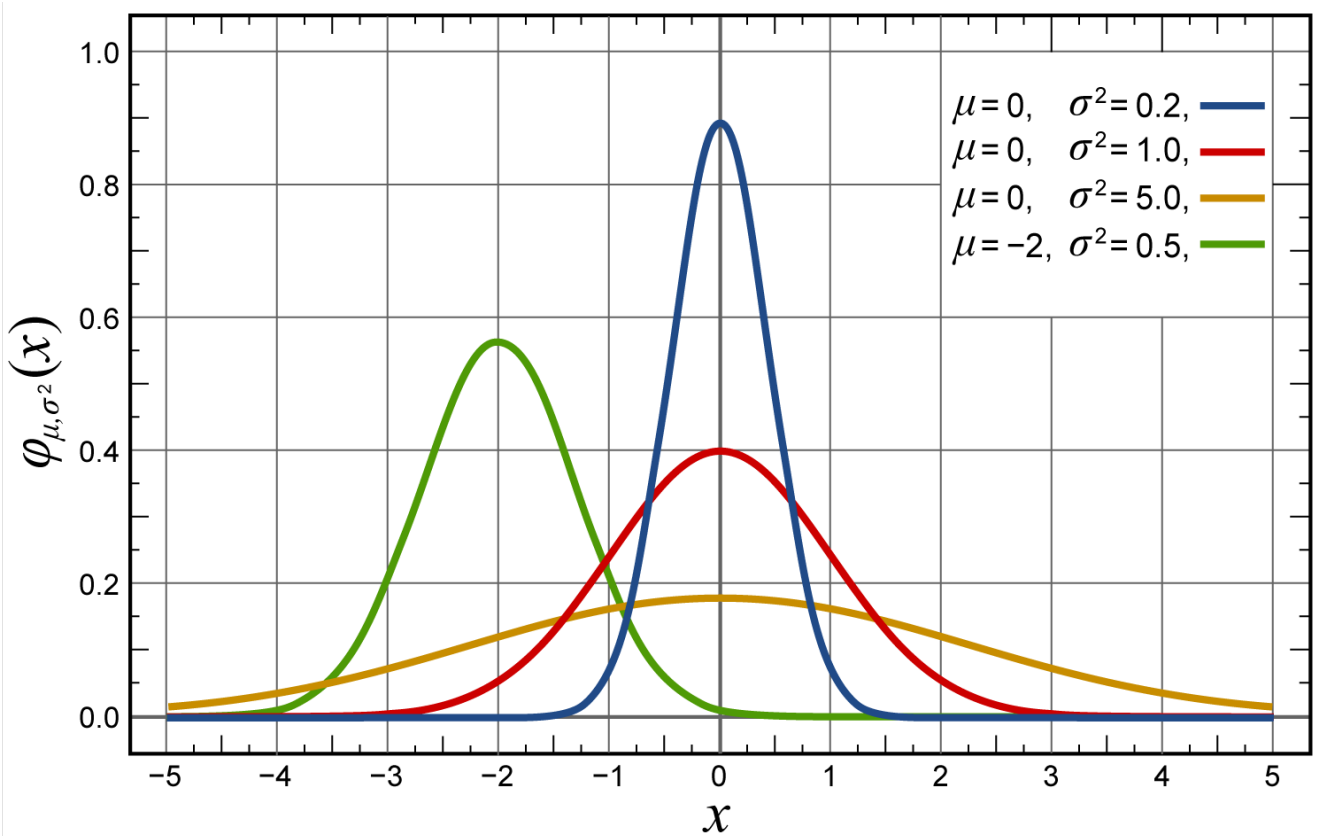
https://www.inchcalculator.com/binomial-distribution-calculator/

Continuous Random Variables



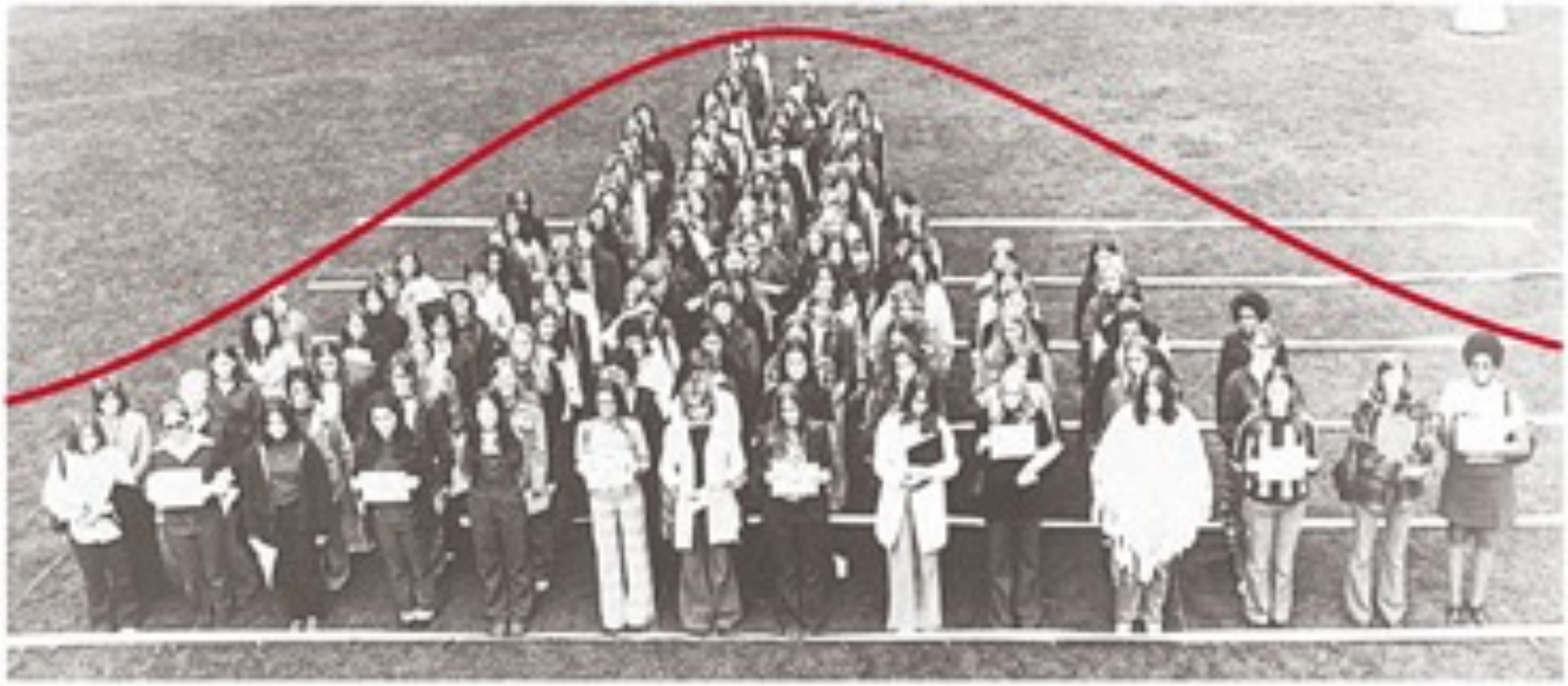
Model processes or data which are encoded as real numbers: *temperature, commodity price, DNA expression level, light on camera sensor, …*

Gaussian (Normal) Distributions



Summaries: *Mean, median, mode, variance, standard deviation, …*

“Nature’s Distribution”



A group of women arranged by height, ScienceBlogs.com Feb. 2009.

- In a large population, how likely is a person to be much taller than average?
- How likely is a request on my web server to be much larger than average?

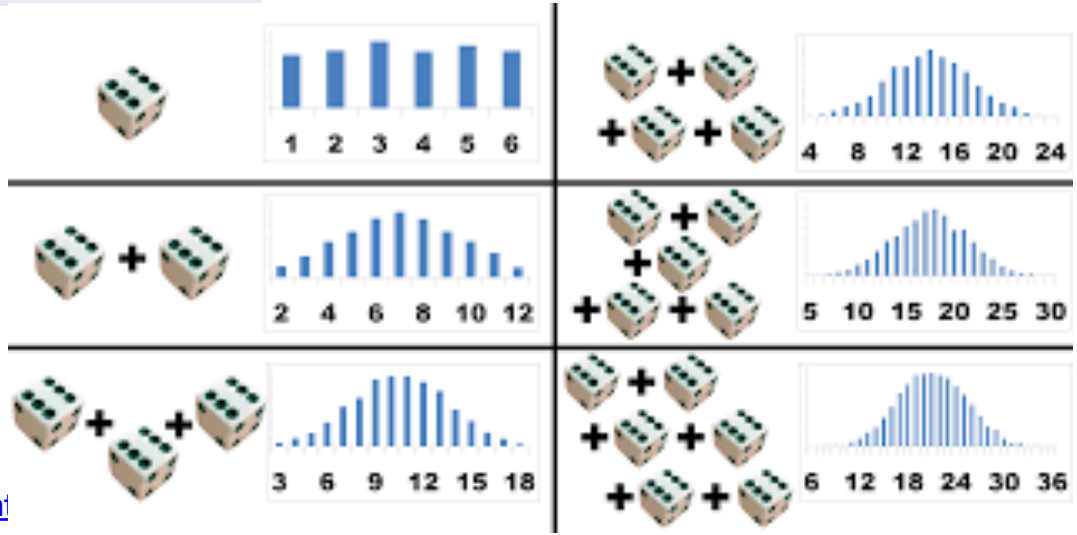
Central Limit Theorem

Theorem (DeMoivre-Laplace-Liapounoff)

Let $x_1, ..., x_n$ be n independent, identically distributed random variables with mean μ and variance σ^2. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, then

$$P(a \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq b) \rightarrow \Phi(b) - \Phi(a)$$

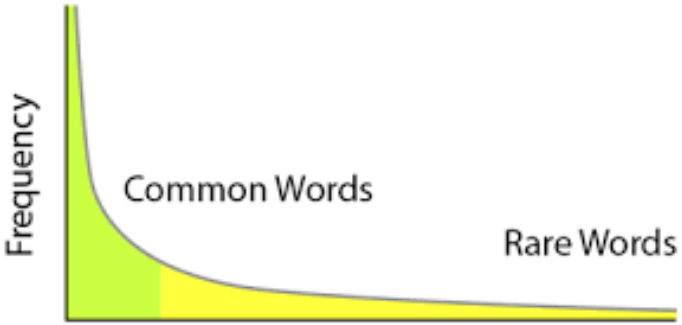
as $n \rightarrow \infty$,

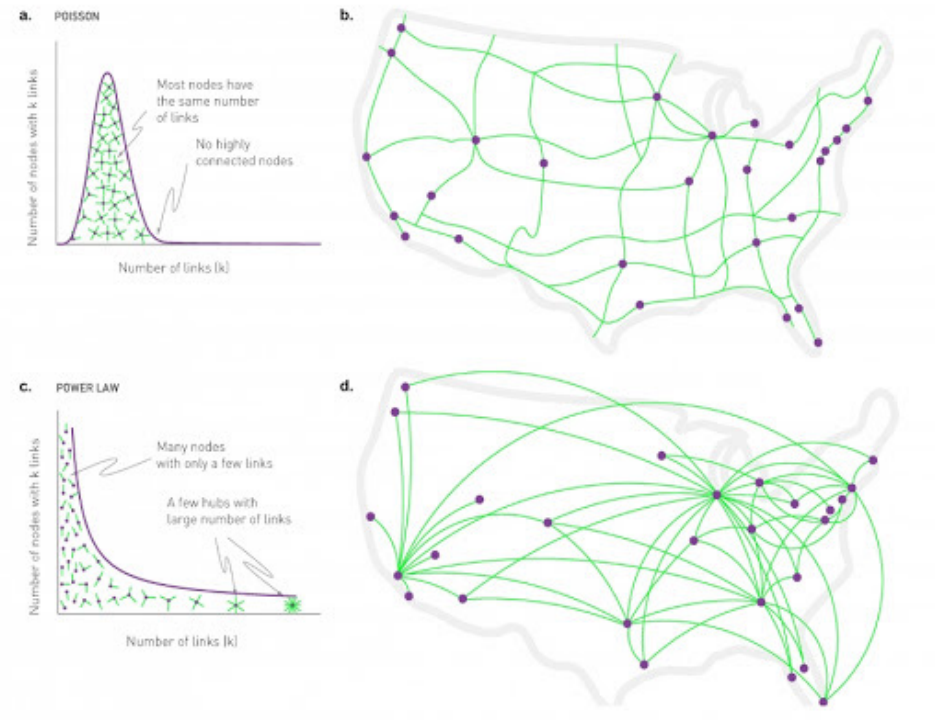


http://www.animatedsoftware.com/statglos/sgcltheo.ht

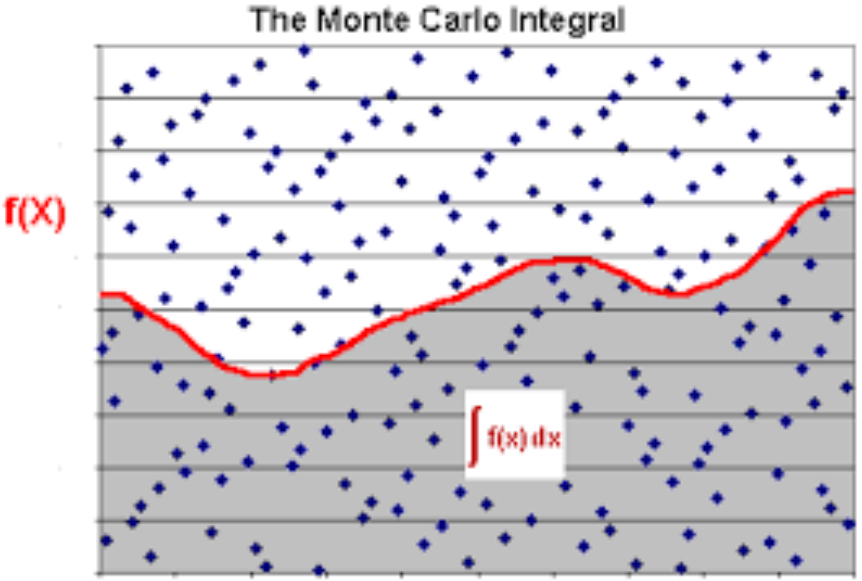
But not all distributions are Normal…

Frequency of using English words

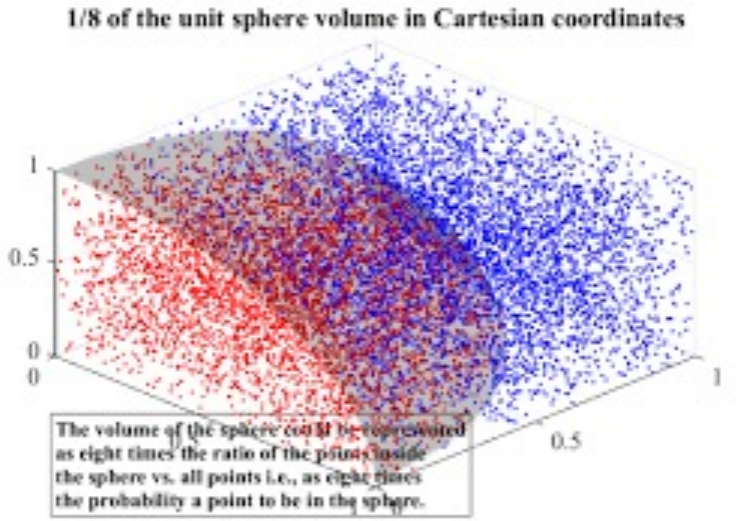




Monte Carlo Methods

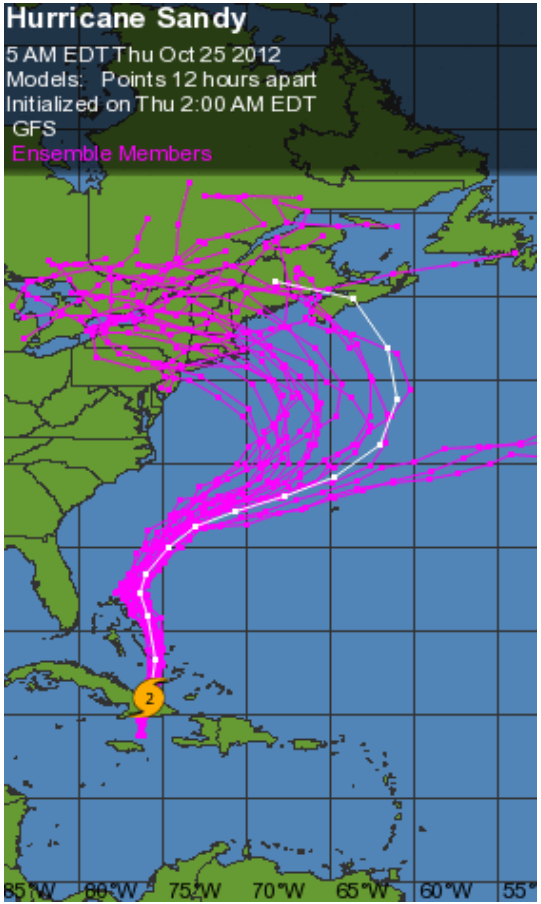


http://marcoagd.usuarios.rdc.puc-rio.br/quasi_mc.html

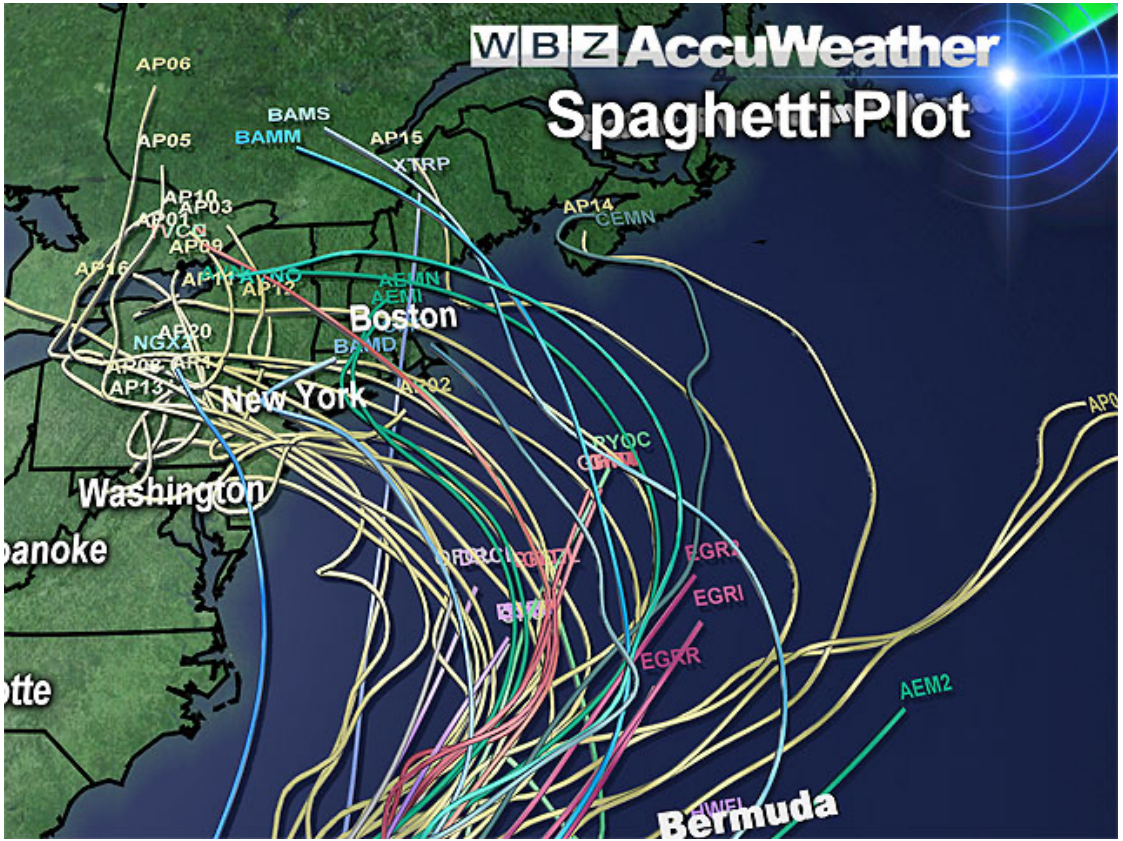


https://www.mathworks.com/matlabcentral/fileexchange/55306-monte-carlo-estimation-examples-with-matlab

Monte Carlo Methods

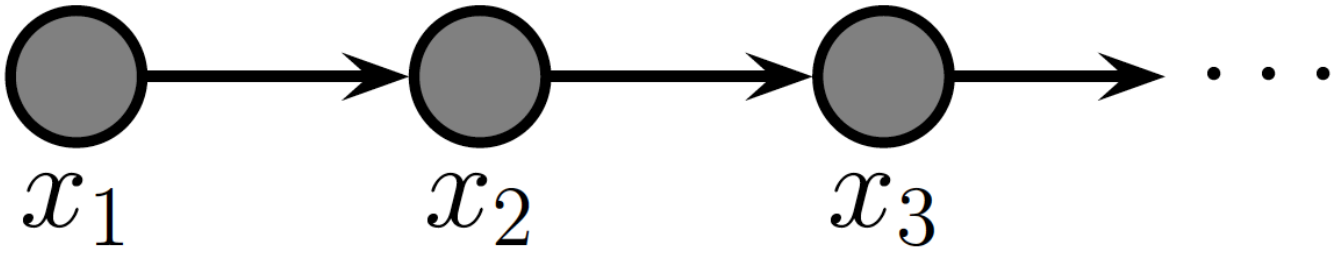


Weather Wisdom, Boston.com



Hurricane Sandy made landfall in New Jersey on October 29, 2012.

Markov Chains



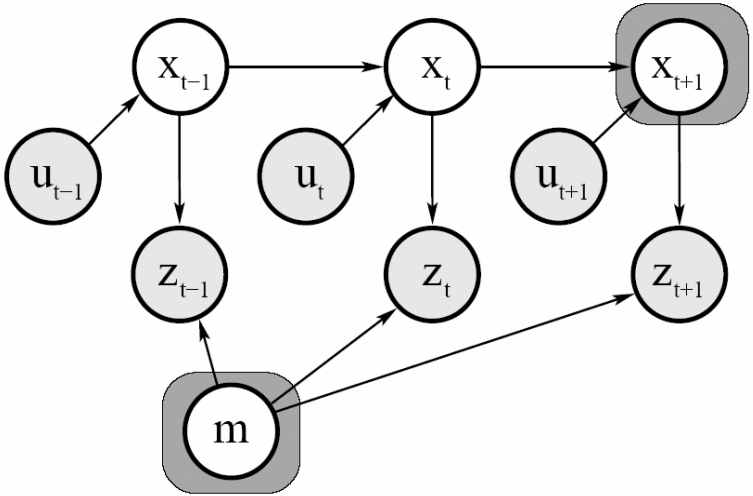
$$p(x) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)p(x_4 \mid x_3)\cdots = p(x_1)\prod_{t=2}^{T} p(x_t \mid x_{t-1})$$

Markov Property: *Conditioned on the present, past & future are independent*

- Building block for modeling random processes that evolve and change over time.
- What is the long-term behavior of some process? What is the probability of reaching a good state? A bad state?
- Allows agents to reason about future consequences of actions.

Markov Chains for Robot Navigation

Simultaneous Localization & Mapping (FastSLAM, Montemerlo 2003)



$$p(x_t, m \mid z_{1:t}, u_{1:t})$$

x_t = State of the robot at time t

m = Map of the environment

$z_{1:t}$ = Sensor inputs from time 1 to t

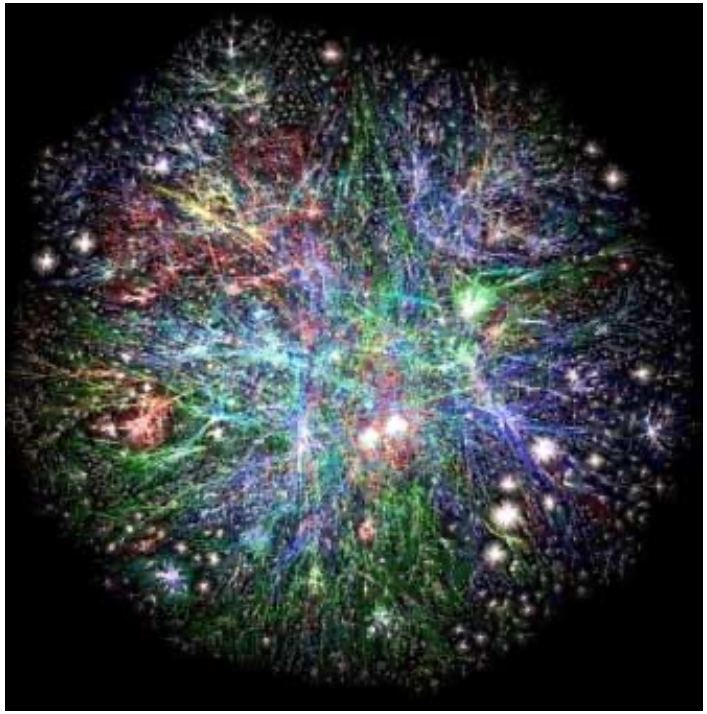
$u_{1:t}$ = Control inputs from time 1 to t

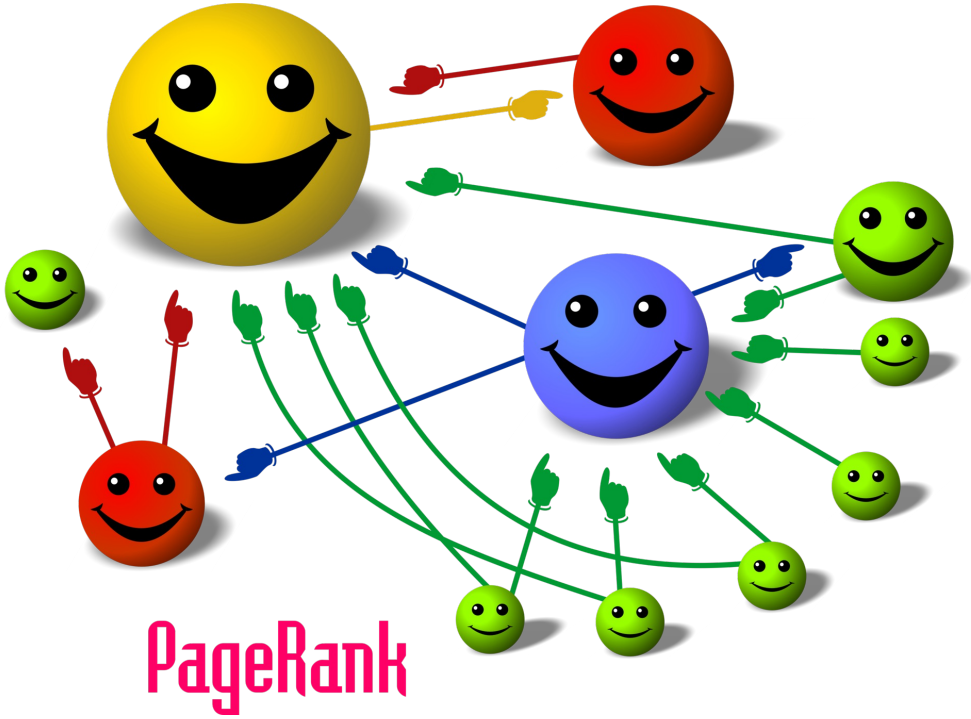
Raw odometry (controls)

True trajectory (GPS)

Inferred trajectory & landmarks

Markov Chains for Web Search: PageRank





Wikipedia

Randomized Algorithm

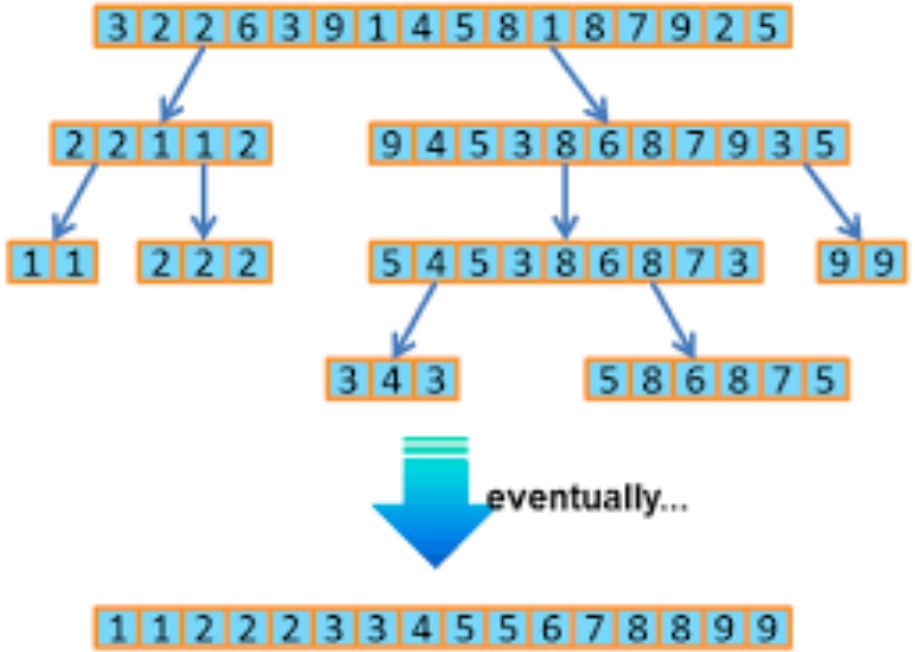
Randomized algorithms makes random choices:

Their run-time, and even correctness are random variables

Expected time to quicksort n elements:

$$
\begin{aligned}
\mathrm{E}[X] &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2}{j-i+1} \\
&= \sum_{i=1}^{n-1} \sum_{k=1}^{n-i} \frac{2}{k+1} \\
&< \sum_{i=1}^{n-1} \sum_{k=1}^{n} \frac{2}{k} \\
&= \sum_{i=1}^{n-1} O(\lg n) \\
&= O(n \lg n) .
\end{aligned}
$$

Quicksort Execution:



https://www.hpcwire.com/

From Probability to Statistics

- In *probability theory* we compute the probability that 20 independent flips of a fair (unbiased) coin give the sequence

  HTTHTHTHHTTHTHTHHTTT

- In *statistics* we ask: Given that we observed the sequence

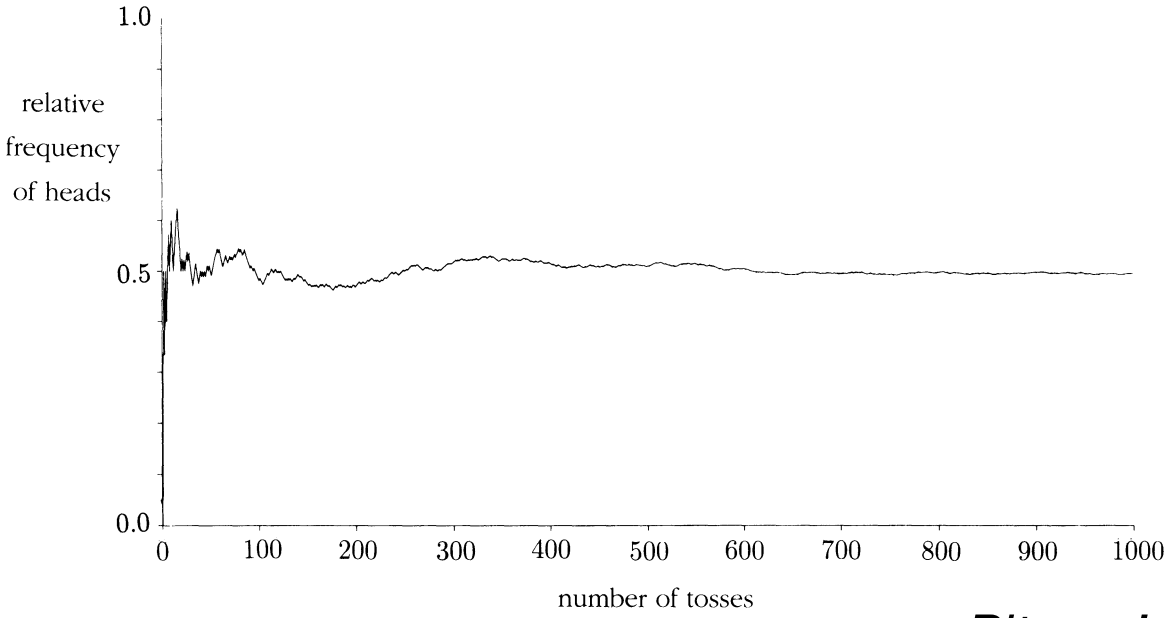
  HTTHTHTHHTTHTHTHHTTT

  what is the likelihood that the coin is fair (unbiased)?

https://hattonsoflondon.co.uk/

From Probability to Statistics

The Frequentist Model: The probability of an outcome in a trial is the frequency of that outcome in a long sequence of identical and independent such trials (limiting frequency).



Pitman's Probability, 1999

From Probability to Statistics

The Frequentist Model: The probability of an outcome in a trial is the frequency of that outcome in a long sequence of identical and independent such trials (limiting frequency).

But then what's the meaning of:
Candidate X has 60% probability of winning the election?
The election happens only once – no frequencies.

From Probability to Statistics

The Bayesian Belief Model: Based on all the information we have seen so far, the probability is our best estimate for the chance of a particular outcome.

I offer a bet that X wins the election at q to 1 odds: you lose $1 if you're wrong, and earn $q if you're right. A rational person would take this bet only if they believe the odds of winning are at least 1/(1+q).

Probabilities and odds express the same tradeoff.

From Probability to Statistics

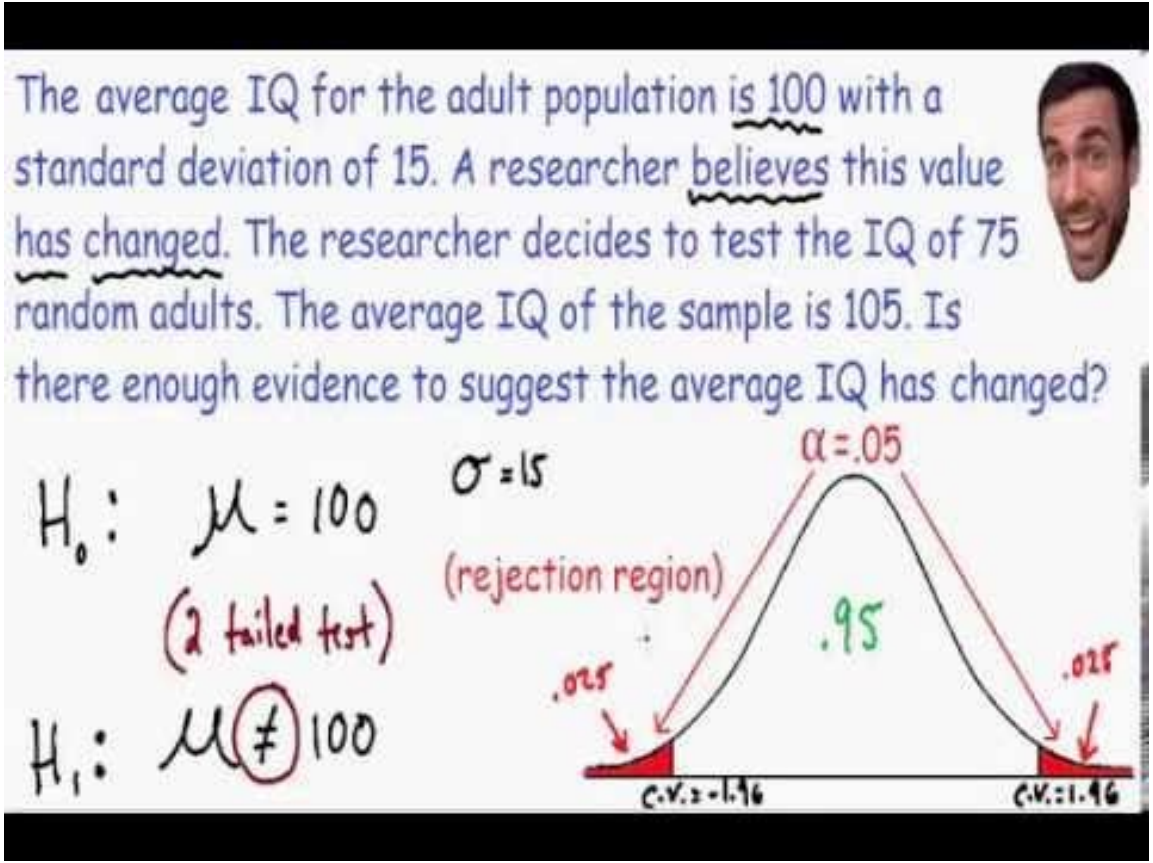
The Frequentist Model: The probability of an outcome in a trial is the frequency of that outcome in a long sequence of identical and independent such trials (limiting frequency).

Classic hypothesis testing; confidence interval, etc.

The Bayesian Belief Model: Based on all the information we have seen so far, the probability is our best estimate for the chance of a particular outcome.

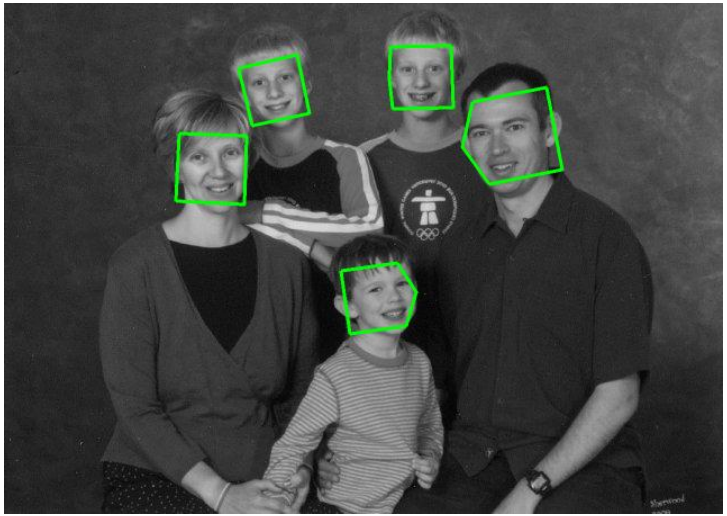
Machine learning, Bayesian analysis

Hypothesis Testing



Hypothesis Testing - Proportion Example

Bayesian Method: Face Detection



K. Murphy & Family

Based on classifiers trained from tens of thousands of example faces (Viola & Jones, 2004)

Digit & Hand Gesture Recognition

Athitsos et al., CVPR 2004 & PAMI 2008

Summary of Course Topics

I. Probability Models
II. Discrete Random Variables
III. Continuous Random Variables
IV. Normal Distributions
V. Limit Theorems
VI. Markov Chains
VII. Randomized Algorithms
VIII. Monte Carlo Methods
IX. Bayesian Statistical Inference
X. Frequentist Statistical Inference

CS145: Lecture 0 Outline

- Probability and statistics: key concepts & applications
- Course Details: People
- Course prerequisites:
  calculus, programming
- Course work and evaluation:
  homework, midterm exam, final exam
- *Who should take this class?*
- *Registration, administration, tech details*

Course Prerequisites

Not formally enforced, but we will assume comfort with:

Calculus

- AP Calculus BC exam, or Brown MATH 0100/0170
- Topics: *limits, basic derivatives & chain rule, basic integrals & fundamental theorem of calculus, sequences & series, …*

Course Prerequisites

Not formally enforced, but we will assume comfort with:

Calculus

- AP Calculus BC exam, or Brown MATH 0100/0170
- Topics: *limits, basic derivatives & chain rule, basic integrals & fundamental theorem of calculus, sequences & series, …*

Programming (optional)

- Any single-semester programming course: CS4, CS15, CS17, CS19, etc.
- Or, other experience that gives comfort with writing simple functions

CS145: Lecture 0 Outline

- Probability and statistics: key concepts & applications
- Course Details: People
- Course prerequisites:
  calculus, programming
- Course work and evaluation:
  homework, midterm exam, final exam
- *Who should take this class?*
- *Registration, administration, tech details*

Course Work and Evaluation

Homework 30%, Midterm 30%, Final 40% of course grade

- Weekly, equally weighted homework assignments
  - Can work in a group but write your own submission.
- Probabilistic derivations, calculations, and reasoning (i.e., math)
- Usually, some math questions can be substitute by an easy implementation assignment.
- Submitted electronically, out for one week
- Read the syllabus and grading policies document on the website
- In class midterm and final

- We are very flexible – when warranted

CS145: Lecture 0 Outline

- Probability and statistics: key concepts & applications
- Course Details: People
- Course prerequisites:
  calculus, programming
- Course work and evaluation:
- *Why should take this class?*
- *Registration, administration, tech details*

Who Should Take CS 1450?

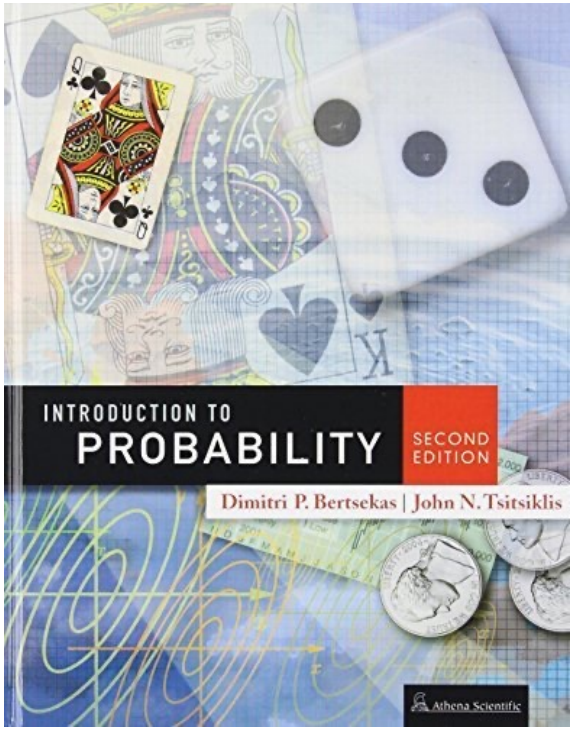
- CS concentrators can satisfy the CS requirement with APMA 1650/1655
- DSI master's students can satisfy the DSI requirement with APMA 1690
- CS 1450 covers probability/statistics as in APMA 1655

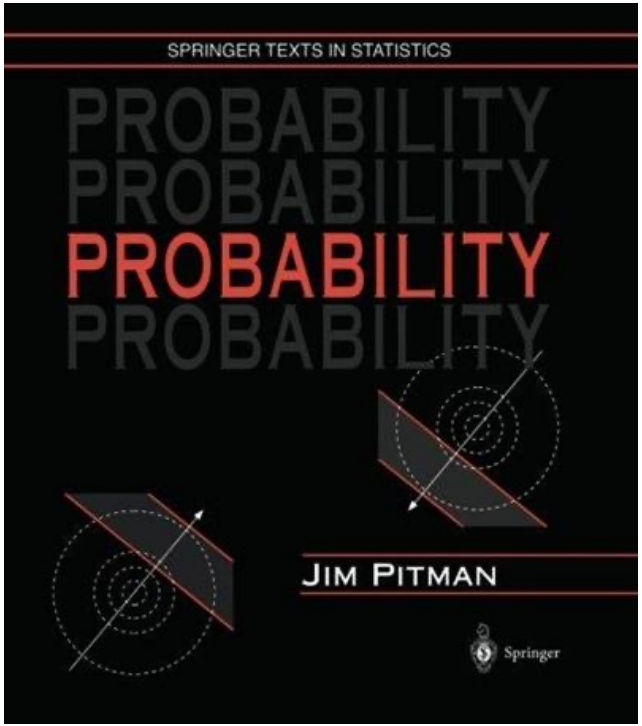
  plus, algorithmic/data science applications.
- If you plan to study machine learning, data science, or theory, consider taking CS 1450 (even if it's harder).

CS145: Lecture 0 Outline

- Probability and statistics: key concepts & applications
- Course Details: People
- Course prerequisites
  calculus, programming
- Course work and evaluation:
- *Why you shouldn't take this class*
- *Registration, administration, tech details*

Course Textbook





- Primary: Bertsekas & Tsitsiklis, *Introduction to Probability*, 2nd ed. (2008)
- Alternative for some topics (no statistics): Pitman, *Probability* (1999)
- Supplemental readings (online) for a few advanced topics

Course Details

- Resources:
  - Course web site: http://cs.brown.edu/courses/csci1450/index.html
    - Slides, homework assignments
  - Ed discussion: https://edstem.org

- Registration:
  - Send a request through CAB

Questions?