

# Introduction to Modern Statistics

Wenbin Guo  
Bioinformatics, UCLA  
[wbguo@ucla.edu](mailto:wbguo@ucla.edu)  
2025 Winter

# Notation of the slides

- Code or Pseudo-Code chunk starts with "➤", e.g.  
➤ print("Hello world!")
- Link is underlined
- Important terminology is in **bold** font
- Practice comes with



# Workshop goals

- Recognize the **uncertainty** in life and put chance to work

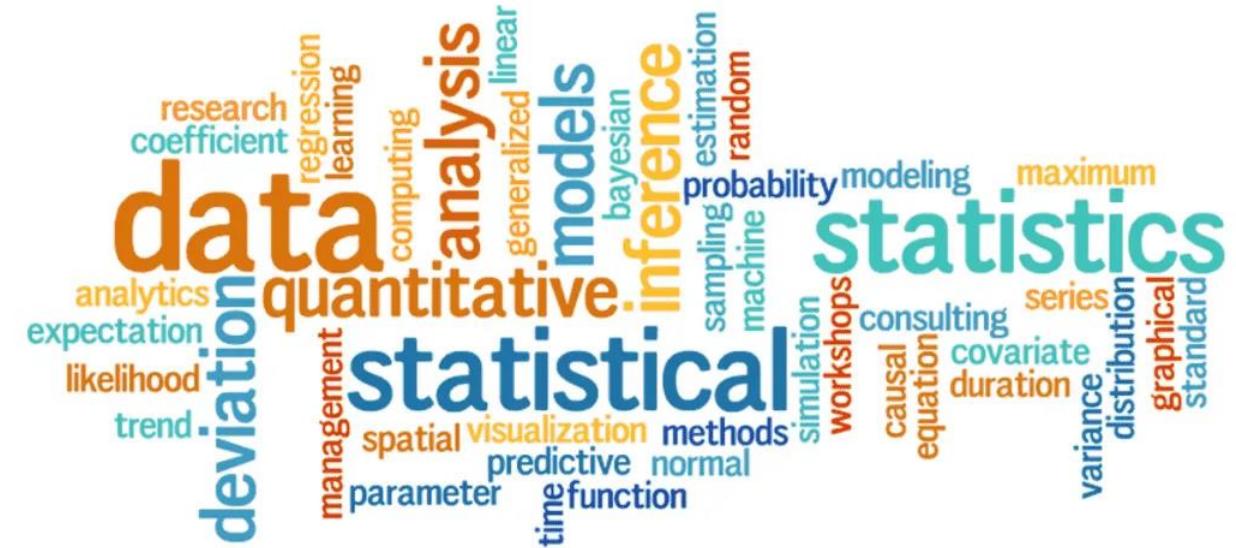
- Understand the nuts and bolts of **statistics**
- Use R to perform **exploratory data analysis** and solve practical problems



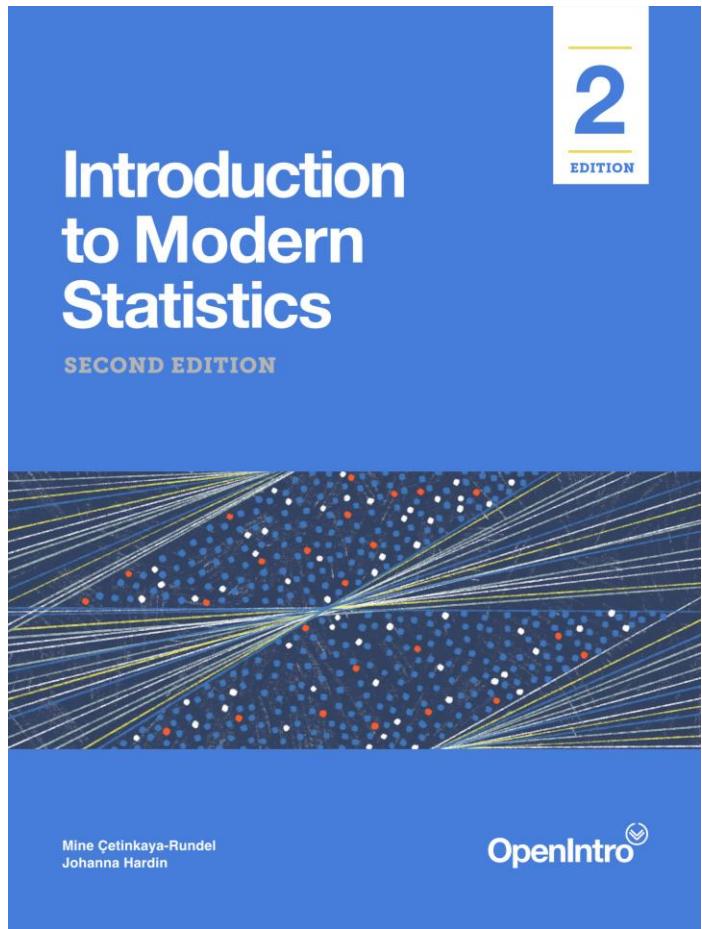
Cross Validated

# Agenda

- Day 1: Probability and Statistics basics
  - Uncertainty; Probability; Distribution
  - Descriptive statistics
- Day 2: Inference
  - Hypothesis testing and  $p$ -values
  - Permutation test and bootstrap
  - False discovery rate control
- Day 3: Modeling
  - Regression techniques
  - Model selection

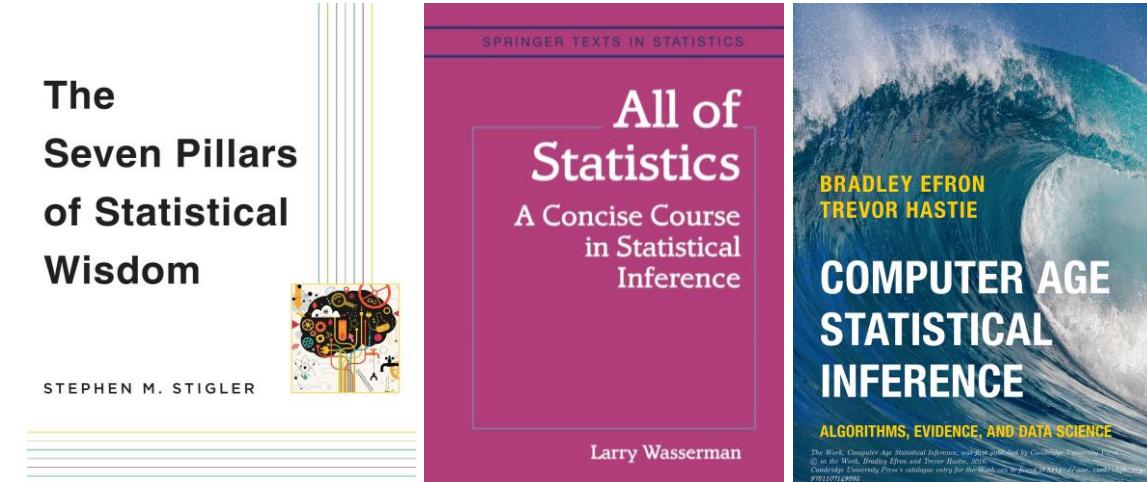


# Reference



[link](#)

- Other useful references



[link](#)

[link](#)

[link](#)

# Working environment

Language



Software



# Environment setup

- Go to the official download [website](#)



## 1: Install R

RStudio requires R 3.6.0+. Choose a version of R that matches your computer's operating system.

*R is not a Posit product. By clicking on the link below to download and install R, you are leaving the Posit website. Posit disclaims any obligations and all liability with respect to R and the R website.*

[DOWNLOAD AND INSTALL R](#)

## 2: Install RStudio

[DOWNLOAD RSTUDIO DESKTOP FOR WINDOWS](#)

Size: 265.27 MB | [SHA-256: 5EFCD188](#) | Version: 2024.12.0+467 | Released: 2024-12-16

# Environment setup

- Go to the official download [website](#)
- Install R and RStudio desktop based on your operating system

The Comprehensive R Archive Network

**Download and Install R**

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux \(Debian, Fedora/Redhat, Ubuntu\)](#)
- [Download R for macOS](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

---

**Source Code for all Platforms**

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2024-10-31, Pile of Leaves) [R-4.4.2.tar.gz](#), read [what's new](#) in the latest version.
- The CRAN directory [src/base-prerelease](#) contains R alpha, beta, and rc releases as daily snapshots in time periods before a planned release.
- Between releases, the same directory [src/base-prerelease](#) contains snapshots of current patched and development versions.  
Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Alternatively, daily snapshots are [available here](#).
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#).

---

**Questions About R**

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

# Environment setup

- Go to the official download [website](#)
- Install R and RStudio desktop based on your operating system
- Install the necessary package(s) in RStudio Console
  - `install.packages("tidyverse")`

# Workshop materials

- git clone <https://github.com/wbvguo/qcbio-Intro2ModernStats.git>



# Day 1: Probability and Statistics basics

Wenbin Guo

Bioinformatics IDP, UCLA

[wbguo@ucla.edu](mailto:wbguo@ucla.edu)

2025 Winter

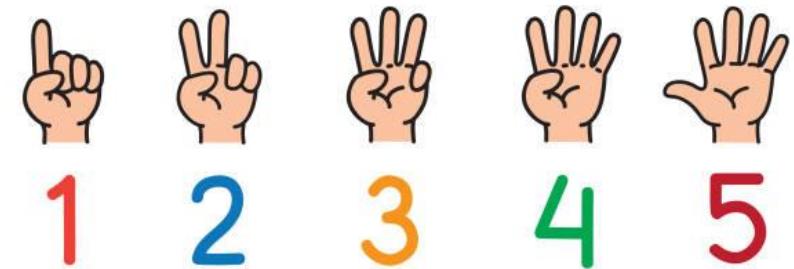
# Overview

## Time

- 2 hours workshop (45min + 45min + practice/Q&A)

## Topics

- Introduction to Statistics
- Uncertainty
- Probability
- Distribution
- Descriptive statistics



Counting as a fundamental part of statistics

# What is statistics?

Statistics is the discipline that concerns the **collection**, **organization**, **analysis**, **interpretation**, and **presentation** of **data**

- experiment design, sampling
- descriptive summary
- hypothesis testing, significance
- visualization
- ...



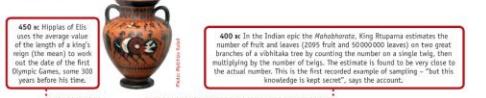
**Data:** samples drawn from a population

Two main statistical methods:

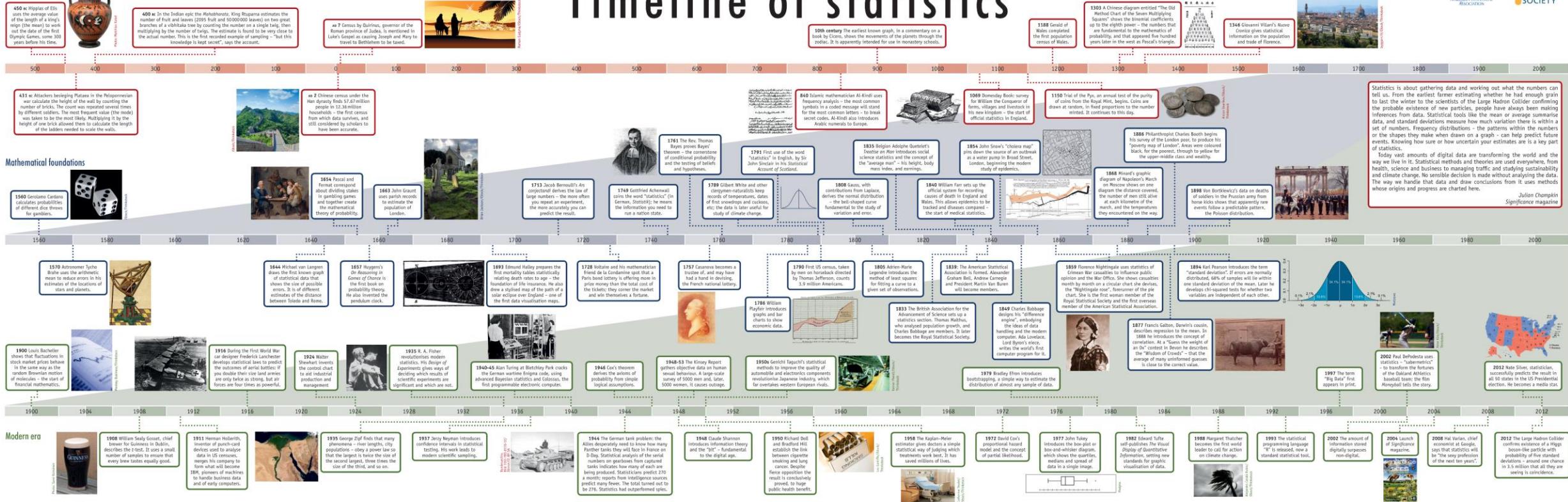
- Descriptive** statistics: summarize the data (describe properties of sample distribution, such as central tendency, dispersion, etc.)
- Inferential** statistics: draw conclusions from the data that are subject to random variation (use probabilistic framework to analyze random phenomena)

# History of Statistics

## Early beginnings



## Timeline of statistics



ASA

ROYAL  
STATISTICAL  
SOCIETY

# Ancient times

450 BC Hippasus of Elis uses the average value of the length of a king's reign (the mean) to work out the date of the first Olympic Games, some 300 years before his time.



Photo: Matthias Kabel



431 BC Attackers besieging Plataea in the Peloponnesian war calculate the height of the wall by counting the number of bricks. The count was repeated several times by different soldiers. The most frequent value (the mode) was taken to be the most likely. Multiplying it by the height of one brick allowed them to calculate the length of the ladders needed to scale the walls.



iStock/Thinkstock

400 BC In the Indian epic the *Mahabharata*, King Ruparna estimates the number of fruit and leaves (2095 fruit and 50 000 000 leaves) on two great branches of a vibhitaka tree by counting the number on a single twig, then multiplying by the number of twigs. The estimate is found to be very close to the actual number. This is the first recorded example of sampling – “but this knowledge is kept secret”, says the account.

AD 7 Census by Quirinus, governor of the Roman province of Judea, is mentioned in Luke's Gospel as causing Joseph and Mary to travel to Bethlehem to be taxed.



AD 2 Chinese census under the Han dynasty finds 57.67 million people in 12.36 million households – the first census from which data survives, and still considered by scholars to have been accurate.



1654 Pascal and Fermat correspond

## Mathematical foundations



by different soldiers. The most frequent value (the mode) was taken to be the most likely. Multiplying it by the height of one brick allowed them to calculate the length of the ladders needed to scale the walls.



households – the first census from which data survives, and still considered by scholars to have been accurate.

## Mathematical foundations



**1560** Gerolamo Cardano calculates probabilities of different dice throws for [gamblers](#).



**1654** Pascal and Fermat correspond about dividing stakes in gambling games and together create the mathematical theory of probability.

1560

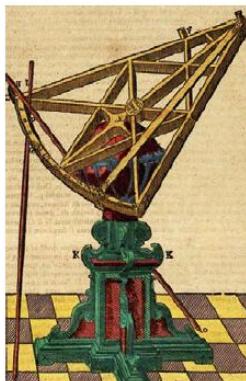
1580

1600

1620

1640

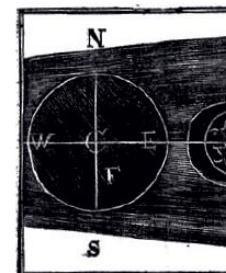
1660



**1570** Astronomer Tycho Brahe uses the arithmetic mean to reduce errors in his estimates of the locations of stars and planets.

**1644** Michael van Langren draws the first known graph of statistical data that shows the size of possible errors. It is of different estimates of the distance between Toledo and Rome.

**1657** Huygens's *On Reasoning in Games of Chance* is the first book on probability theory. He also invented the pendulum clock.





**1713** Jacob Bernoulli's *Ars conjectandi* derives the law of large numbers – the more often you repeat an experiment, the more accurately you can predict the result.

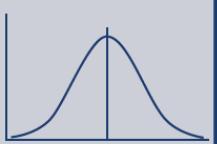
**1761** The Rev. Thomas Bayes proves Bayes' theorem – the cornerstone of conditional probability and the testing of beliefs and hypotheses.

**1791** First use of the word "statistics" in English, by Sir John Sinclair in his *Statistical Account of Scotland*.

**1835** Belgian Adolphe Quetelet's *Treatise on Man* introduces social science statistics and the concept of the "average man" – his height, body mass index, and earnings.

**1749** Gottfried Achenwall coins the word "statistics" (in German, *Statistik*); he means the information you need to run a nation state.

**1789** Gilbert White and other clergymen-naturalists keep records of temperatures, dates of first snowdrops and cuckoos, etc; the data is later useful for study of climate change.



**1808** Gauss, with contributions from Laplace, derives the normal distribution – the bell-shaped curve fundamental to the study of variation and error.

**1840** William Farr's official system for causes of death in Wales. This allows events tracked and diseases at the start of medical

1700

1720

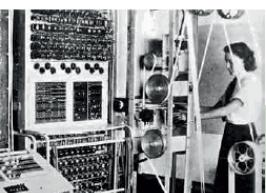
1740

1760

1780

1800

1820

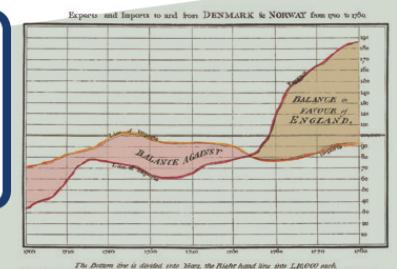


Edmund Halley prepares the mortality tables statistically by death rates to age – the basis of life insurance. He also produced a stylised map of the path of a solar eclipse over England – one of the first data visualisation maps.

**1728** Voltaire and his mathematician friend de la Condamine spot that a Paris bond lottery is offering more in prize money than the total cost of the tickets; they corner the market and win themselves a fortune.



**1757** Casanova becomes a trustee of, and may have had a hand in devising, the French national lottery.

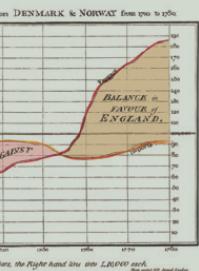


**1786** William Playfair introduces graphs and bar charts to show economic data.

**1833** The British Association for the Advancement of Science sets up a statistics section. Thomas Malthus and Charles Babbage are members. It becomes the Royal Statistical Society.

**1839:** The Association of Graham and Prentiss will

census, taken  
back directed  
person, counts  
Americans.



**1835** Belgian Adolphe Quetelet's *Treatise on Man* introduces social science statistics and the concept of the "average man" – his height, body mass index, and earnings.

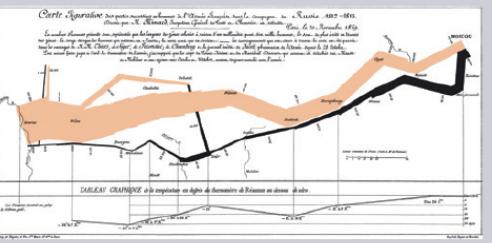
**1808** Gauss, with contributions from Laplace, derives the **normal distribution** – the bell-shaped curve fundamental to the study of variation and error.

**1854** John Snow's "cholera map" pins down the source of an outbreak as a water pump in Broad Street, London, beginning the modern study of epidemics.



**1886** Philanthropist Charles Booth begins his survey of the London poor, to produce his "poverty map of London". Areas were coloured black, for the poorest, through to yellow for the upper-middle class and wealthy.

**1840** William Farr sets up the official system for recording causes of death in England and Wales. This allows epidemics to be tracked and diseases compared – the start of medical statistics.



**1868** Minard's graphic diagram of Napoleon's March on Moscow shows on one diagram the distance covered, the number of men still alive at each kilometre of the march, and the temperatures they encountered on the way.

**1898** Von Bortkiewicz's data on deaths of soldiers in the Prussian army from horse kicks shows that apparently rare events follow a predictable pattern, the **Poisson distribution**.

1800

1820

1840

1860

1880

1900

1920

**1805** Adrien-Marie Legendre introduces the method of **least squares** for fitting a curve to a given set of observations.

**1839**: The American Statistical Association is formed. Alexander Graham Bell, Andrew Carnegie and President Martin Van Buren will become members.

**1833** The British Association for the Advancement of Science sets up a statistics section. Thomas Malthus, who analysed population growth, and Charles Babbage are members. It later becomes the Royal Statistical Society.

**1849** Charles Babbage designs his "difference engine", embodying the ideas of data handling and the modern computer. Ada Lovelace, Lord Byron's niece, writes the world's first

**1859** Florence Nightingale uses statistics of Crimean War casualties to influence public opinion and the War Office. She shows casualties month by month on a circular chart she devises, the "Nightingale rose", forerunner of the pie chart. She is the first woman member of the Royal Statistical Society and the first overseas member of the American Statistical Association.



**1877** Francis Galton, Darwin's cousin, describes **regression to the mean**. In 1888 he introduces the concept of correlation. At a "Guess the weight of an Ox" contest in Devon he describes



errors. It is of different estimates of the distance between Toledo and Rome.

probability theory. He also invented the pendulum clock.

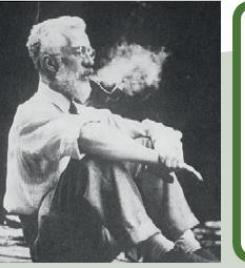


drew a stylised map of the path of a solar eclipse over England – one of the first data visualisation maps.

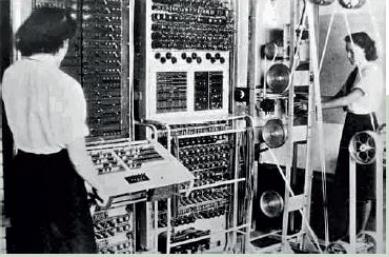
the tickets; they corner the market and win themselves a fortune.

**1916** During the First World War car designer Frederick Lanchester develops statistical laws to predict the outcomes of aerial battles: if you double their size land armies are only twice as strong, but air forces are four times as powerful.

**1924** Walter Shewhart invents the control chart to aid industrial production and management



**1935** R. A. Fisher revolutionises modern statistics. His *Design of Experiments* gives ways of deciding which results of scientific experiments are significant and which are not.



**1940-45** Alan Turing at Bletchley Park cracks the German wartime Enigma code, using advanced Bayesian statistics and Colossus, the first programmable electronic computer.

**1946** Cox's theorem derives the axioms of probability from simple logical assumptions.

1916

1920

1924

1928

1932

1936

1940

1944



**1935** George Zipf finds that many phenomena – river lengths, city populations – obey a power law so that the largest is twice the size of the second largest, three times the size of the third, and so on.



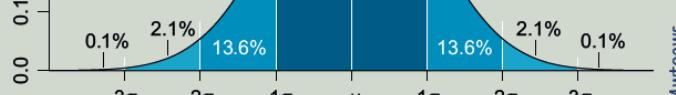
Bundesarchiv, Bild 101I-783-0110-12/ Dömer/CC-BY-SA

**1937** Jerzy Neyman introduces confidence intervals in statistical testing. His work leads to modern scientific sampling.

**1944** The German tank problem: the Allies desperately need to know how many Panther tanks they will face in France on D-Day. Statistical analysis of the serial numbers on gearboxes from captured tanks indicates how many of each are being produced. Statisticians predict 27 a month; reports from intelligence sources predict many fewer. The total turned out to be 276. Statistics had outperformed spies.

of the pie  
er of the  
st overseas  
assocation.

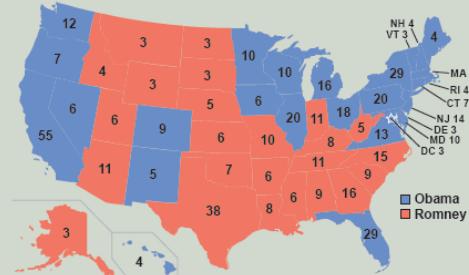
develops chi-squared tests for whether two variables are independent of each other.



877 Francis Galton, Darwin's cousin, describes regression to the mean. In 1888 he introduces the concept of correlation. At a "Guess the weight of an Ox" contest in Devon he describes the "Wisdom of Crowds" – that the average of many uninformed guesses is close to the correct value.



Peter Kim/Stock/  
Thinkstock



2012 Nate Silver, statistician, successfully predicts the result in all 50 states in the US Presidential election. He becomes a media star.

1997 The term "Big Data" first appears in print.

2002 Paul DePodesta uses statistics – "sabermetrics" – to transform the fortunes of the Oakland Athletics baseball team; the film *Moneyball* tells the story.

1984

1988

1992

1996

2000

2004

2008

2012

1982 Edward Tufte publishes *The Visual Display of Quantitative Information*, setting new standards for graphic visualisation of data.

1988 Margaret Thatcher becomes the first world leader to call for action on climate change.



Alejandro Catalá Rubio/  
iStock/Thinkstock

1993 The statistical programming language "R" is released, now a standard statistical tool.

2002 The amount of information stored digitally surpasses non-digital.

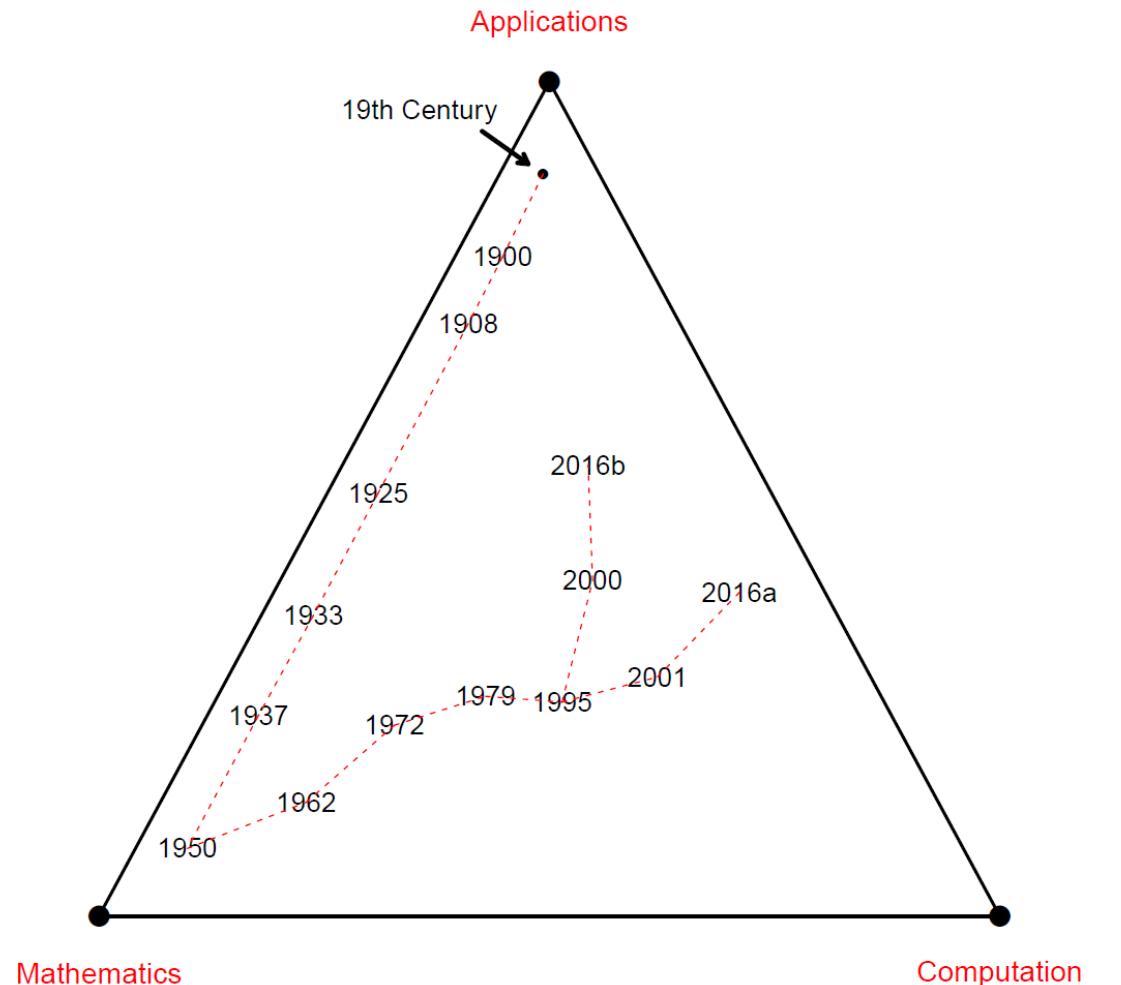


2008 Hal Varian, chief economist at Google, says that statistics will be "the sexy profession of the next ten years".

2012 The Large Hadron Collider confirms existence of a Higgs boson-like particle with probability of five standard deviations – around one chance in 3.5 million that all they are seeing is coincidence.

# Modern Statistics

- 1900: Chi-squared test
- 1908:  $t$  statistic
- 1925: Maximum likelihood theory, etc.
- 1933: Neyman–Pearson lemma
- 1937: Confidence interval
- 1950: Decision theory
- 1962: Data analysis
- 1972: Cox's proportional hazards ratio
- 1979: Bootstrap; Monte-Carlo Markov Chain (MCMC)
- 1995: False Discovery Rate; Lasso
- 2000: Large scale inference (microarray data)
- 2001: Random Forest, boosting
- 2016a: Data science (even without parametric models or formal inference)
- 2016b: Big data application (Genome-wide association studies)

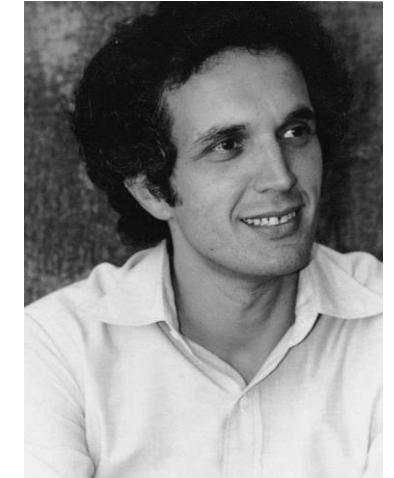
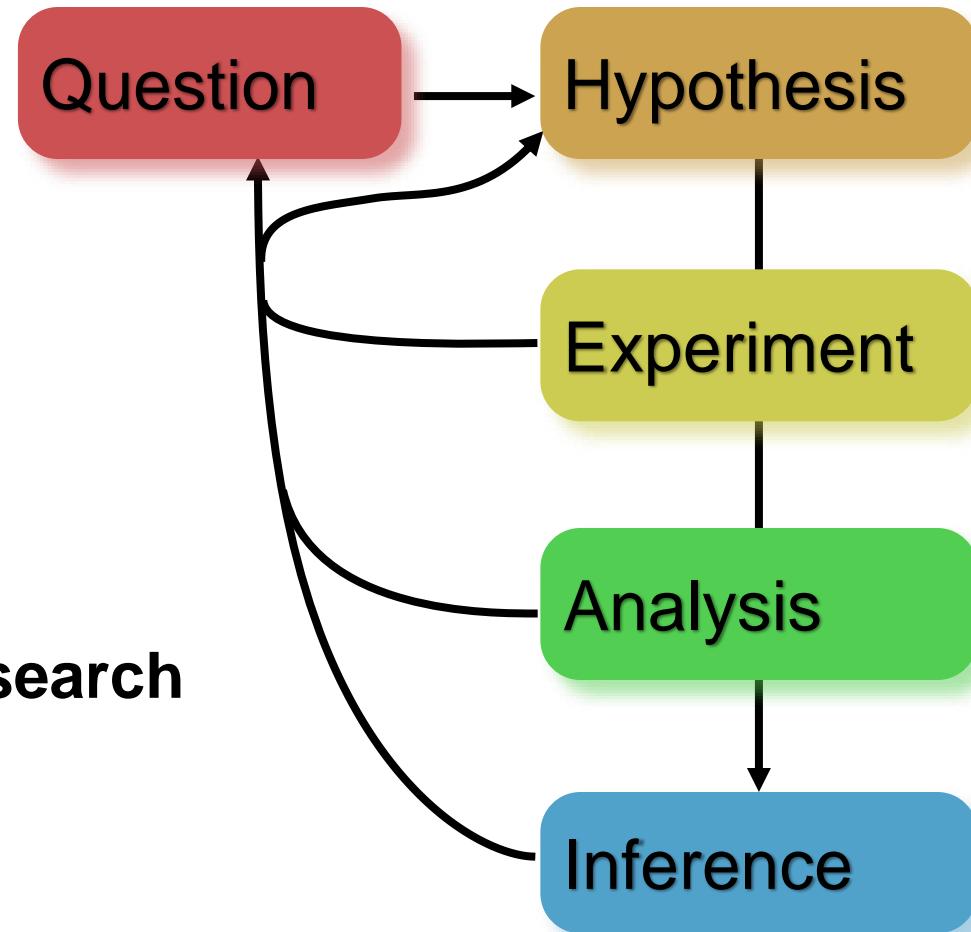


Development of the statistics discipline since the end of 20-th century

# Why statistics?

“Those who ignore statistics are condemned to reinvent it”

**Statistics in scientific research**



**Bradley Efron**  
(1938- )

# Statistics and learning

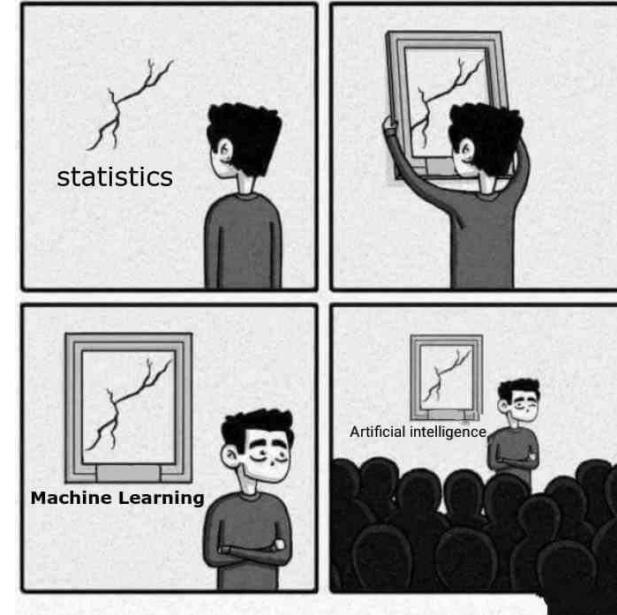
- Training data come from a probability distribution

$$(x_i, y_i, i = 1, \dots, n) \sim p(x, y)$$

- Learn about this probability distribution

$$\max_{\theta} \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(y_i|x_i) \rightarrow p_{\hat{\theta}}(y|x)$$

- Generalize to testing data (Memorization and generalization)

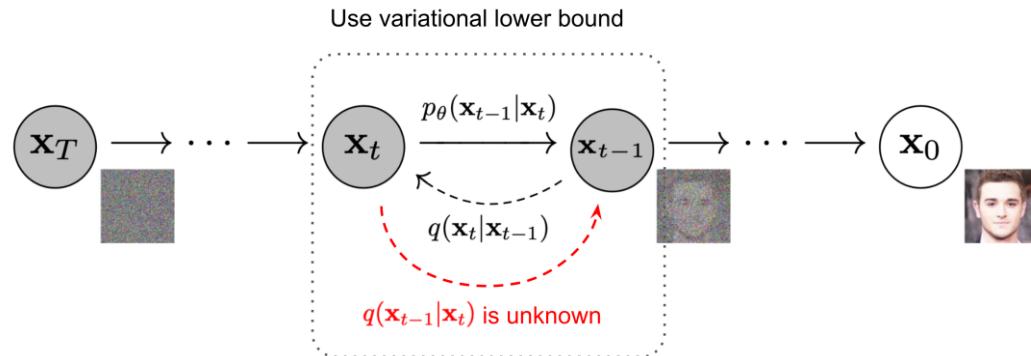


# Statistics and learning

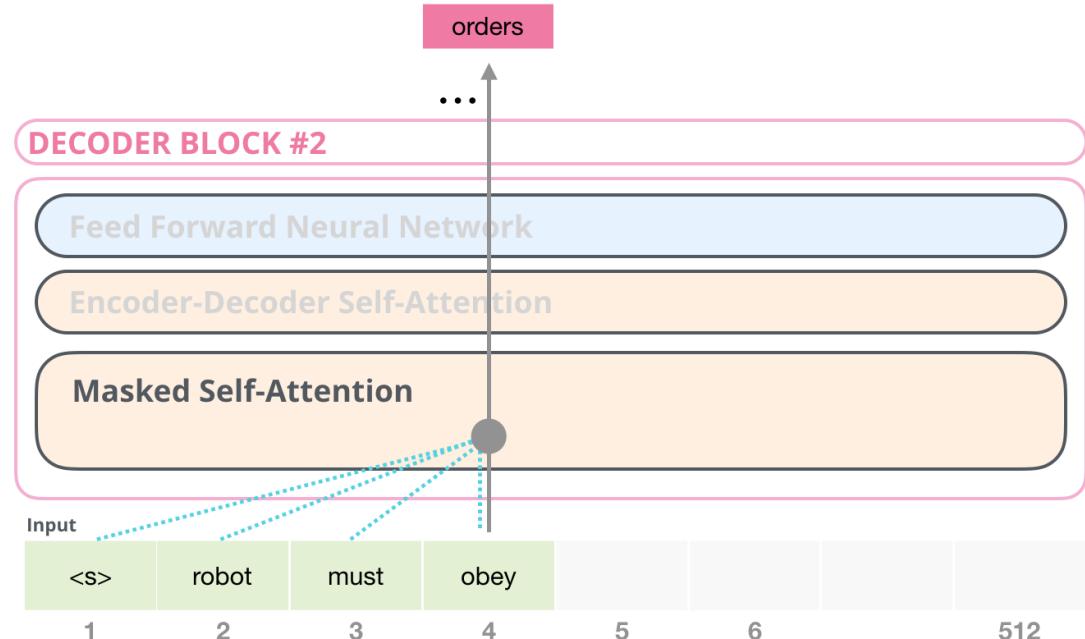
- Natural Language Processing:

**next word prediction**  $\prod_t p_\theta(y_t | y_{<t}, x)$

- Diffusion model:



**Diffusion denoising probability model**  $\prod_t p_\theta(y_{t-1} | y_t, x)$



# Statistics and decision

*Knowledge is what we know  
Also, what we know we do not know.  
We discover what we do not know  
Essentially by what we know.*

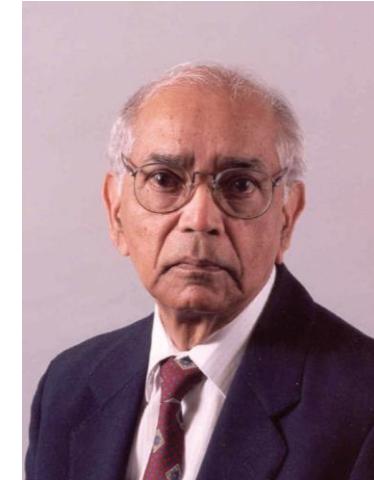
*Thus knowledge expands.*

*With more knowledge we come to know  
More of what we do not know.*

*Thus knowledge expands endlessly.*

\* \* \*

*All knowledge is, in final analysis, history.  
All sciences are, in the abstract, mathematics.  
All judgements are, in their rationale, statistics.*



C.R. Rao  
(1920-2023)

One of the greatest  
statisticians

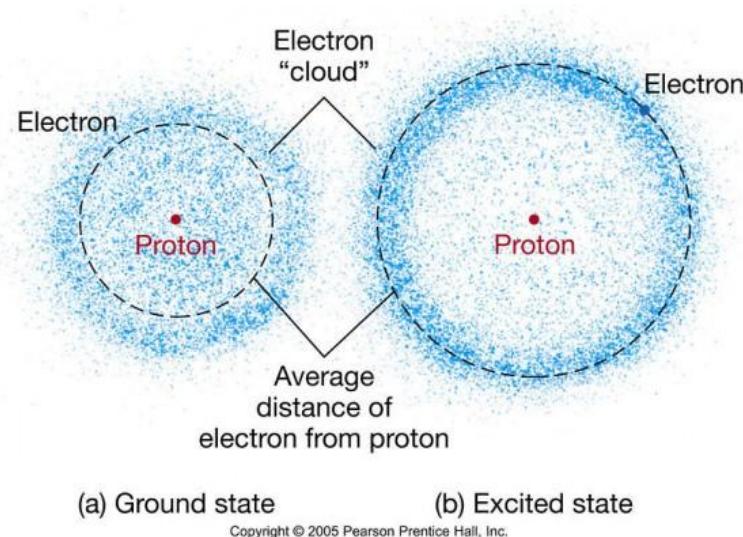
— *Statistics And Truth: Putting Chance To Work*, 1997

# Why statistics?

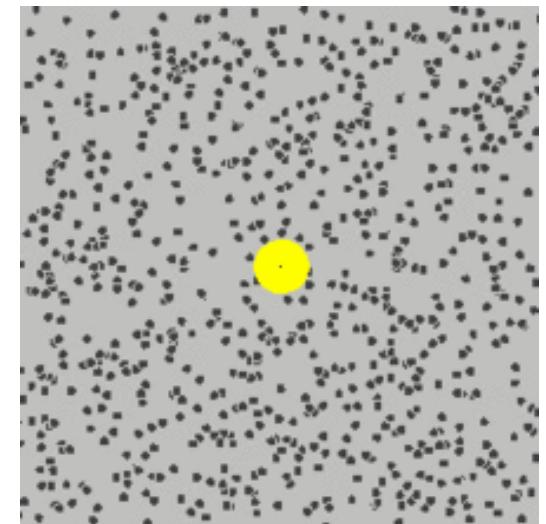
- The real world is full of randomness
  - At the most fundamental level, the physical laws are probabilistic



Double-slit experiment  
Wave function → probability density



Electron cloud



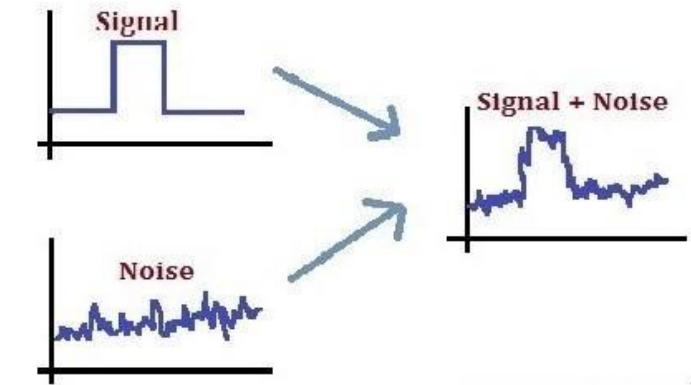
Brownian motion

# Why statistics?

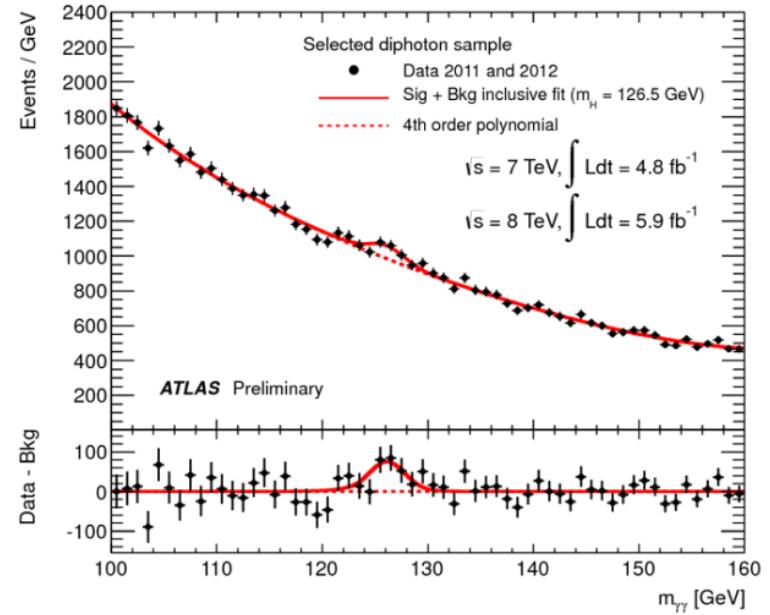
□ The real world is full of randomness

□ Our observations are not perfect

- Limited observations due to resource constraints
- Measurement errors (noise, bias)
- ...



2013 Nobel prize in physics  
(Higgs-boson particle)



We need to understand the world, and make decisions under **uncertainties**

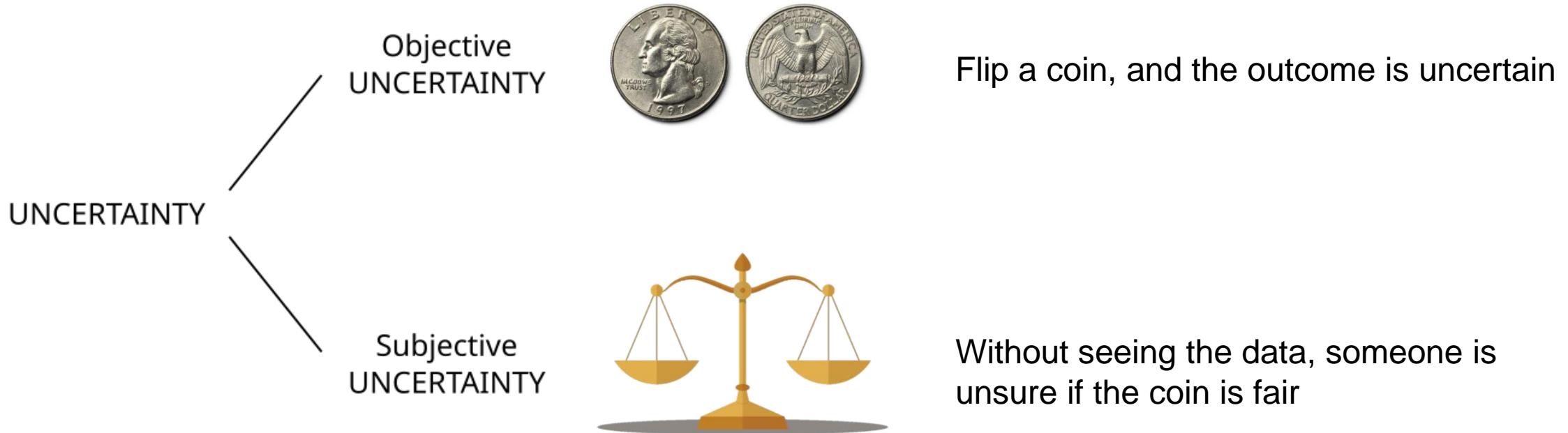
# Uncertainty

“The only certainty in life is uncertainty itself”



# Recognizing uncertainty in life

**Uncertainty:** A state of limited knowledge where the current state or future outcome cannot be precisely determined.



# Source of uncertainty

## □ lack of knowledge/Information

e.g. go to the bus station, and see the bus is there

## Question:

If we have all information, there will be no uncertainty?

No...The Heisenberg uncertainty principle



Measuring the position of a particle with high accuracy inevitably introduces uncertainty in its momentum, and vice versa, meaning that **knowing both with perfect precision at the same time is impossible**.

$$\Delta x \Delta p \geq \frac{h}{4\pi} \quad (h \text{ is plank constant})$$

Uncertainty in position

Uncertainty in momentum

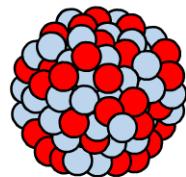
# Source of uncertainty

- lack of knowledge/Information

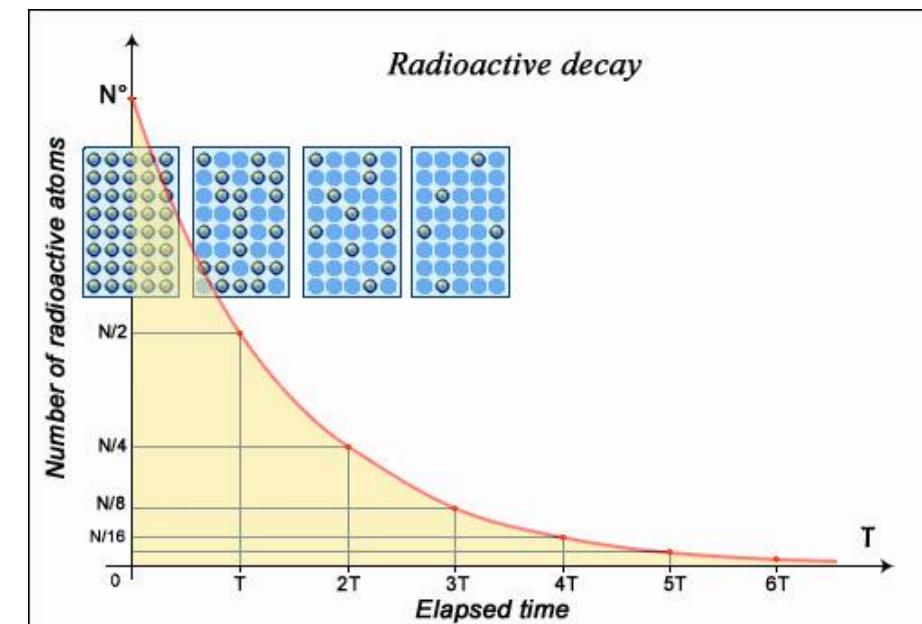
e.g. go to the bus station, and see the bus is there

- fundamental property of nature (true randomness)

e.g. Atom's radioactive decay

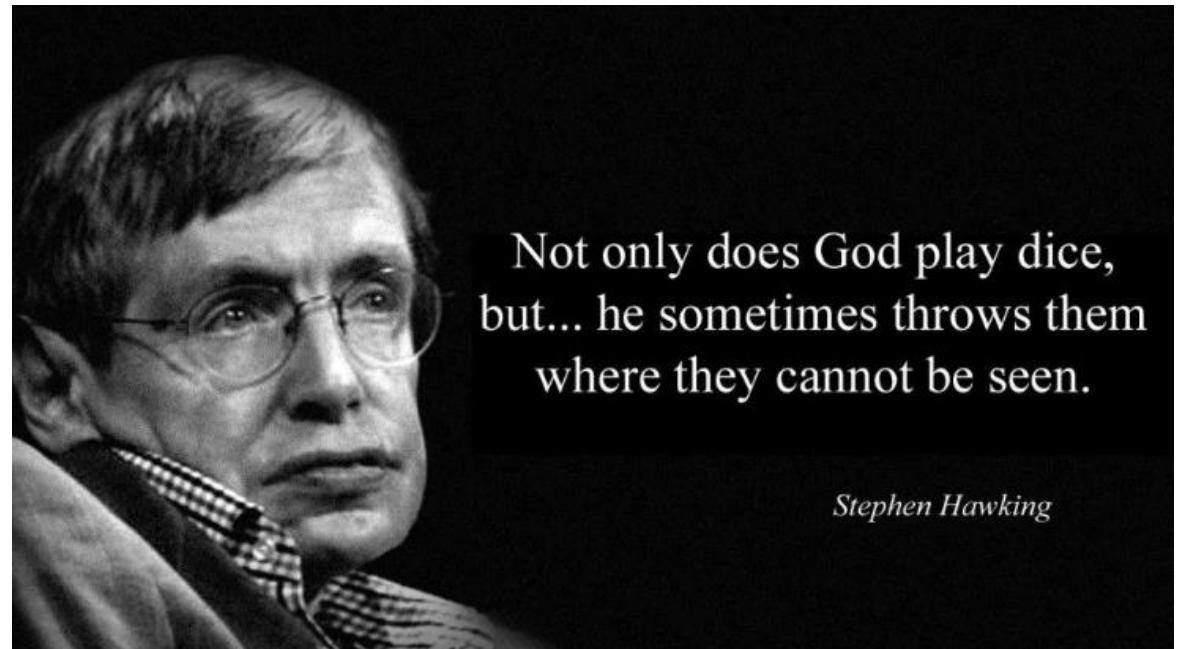
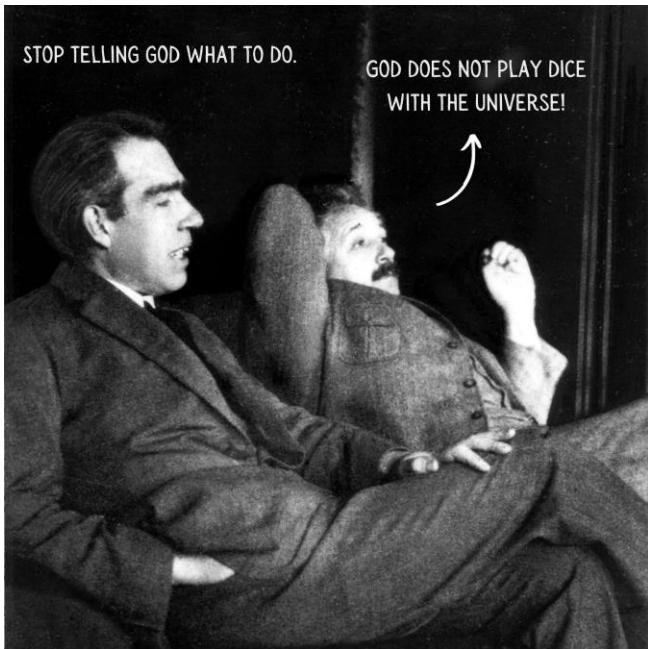


$^{226}_{88}\text{Ra}$



# Deterministic vs Stochastic view of the world

Einstein famously declared, '**He does not play dice**', in response to Bohr's probabilistic interpretation of quantum mechanics, expressing his belief in a deterministic universe.



# More examples of uncertainty in life



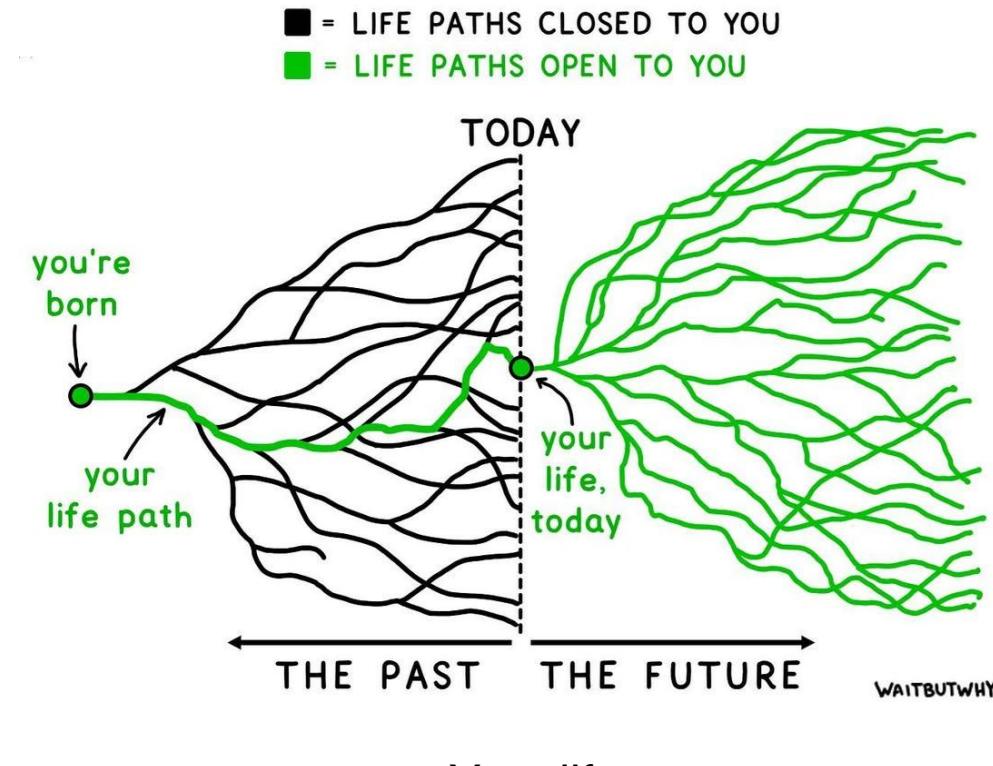
Lottery ball



Earthquake



Weather forecasting



Your life

# How do we feel about uncertainty?

- We **hate** it, we try to avoid it
  - Drive slowly when the vision is not good
  - Invest less if a stock fluctuates too much
  - ...
- We **like** it, we play with it



Texas poker



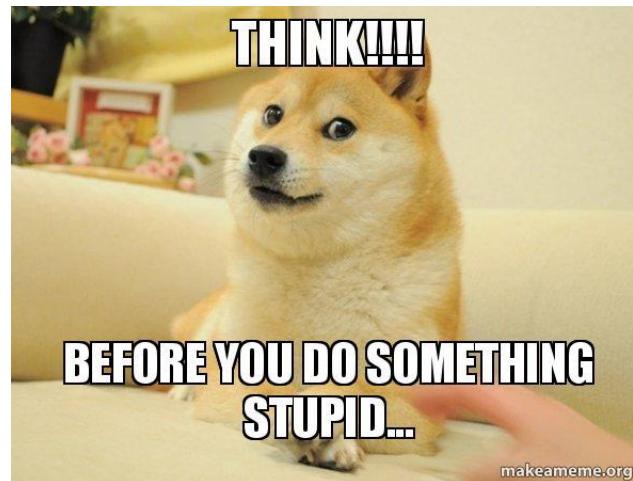
Monopoly



Quant trading

# How do we deal with uncertainty?

We try to make a **guess on the consequence** before we act



Did you notice that, we seems to **do some implicit calculations** in our mind when making decisions with uncertainty?

The chances of winning  
the lottery is 50%. You  
either win or don't



# Probability

*“Nothing but common sense reduced to calculation” – Laplace*

# Not all events are created equal

Consider two events

- a) I flip a coin, and get a head
  - Odds:  $1/2$
- b) I buy a lottery ticket, and win \$1,000,000
  - Odds:  $1/1,219,304$

Both of them are uncertain, which one is more likely?

We need a quantitative measure for the uncertainty!

P.S. Possible doesn't imply probable



# Probability (take coin toss as an example)

Probability is a mathematical language for quantifying uncertainty

- **Sample space  $\Omega$ :** the set of possible outcomes of an experiment



- **Event A:** a subset of sample space



If each outcome is equally likely, then **probability**  $P(A) = \frac{|A|}{|\Omega|}$

# Properties of probability

$$\text{Probability } P(A) = \frac{|A|}{|\Omega|}$$

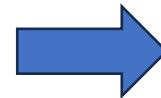
Axiom:

**Axiom 1:**  $P(A) \geq 0$  for every  $A$

**Axiom 2:**  $P(\Omega) = 1$

**Axiom 3:** If  $A_1, A_2, \dots$  are disjoint then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$



$$P(\emptyset) = 0$$

$$A \subset B \implies P(A) \leq P(B)$$

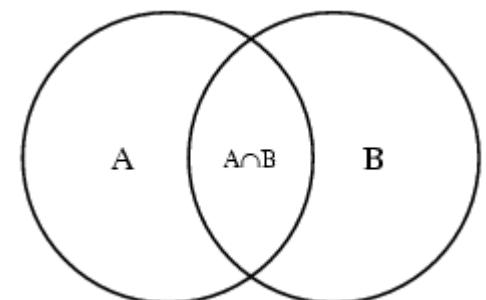
$$0 \leq P(A) \leq 1$$

$$P(A^c) = 1 - P(A)$$

Lemma:

For any events A and B

$$P(A \cup B) = P(A) + P(B) - P(AB)$$



# Let's do a small exercise

- Consider tossing a die twice and summing the values. What is the probability that the sum equals 10?
- Given the number of people in the classroom, assuming birthdays are equally likely, calculate the probability that at least two students share the same birthday (same month and day)
- Code practice: estimate  $\pi$



# Independent events

Two events are **independent** if

$$P(AB) = P(A) P(B)$$

e.g. Flip the coin twice, what's the probability of getting both heads?

Solution 1:

- $\Omega = \{HH, HT, TH, TT\}$ ;  $A = \{HH\}$ ;
- $P(A) = \frac{|A|}{|\Omega|} = \frac{1}{4}$

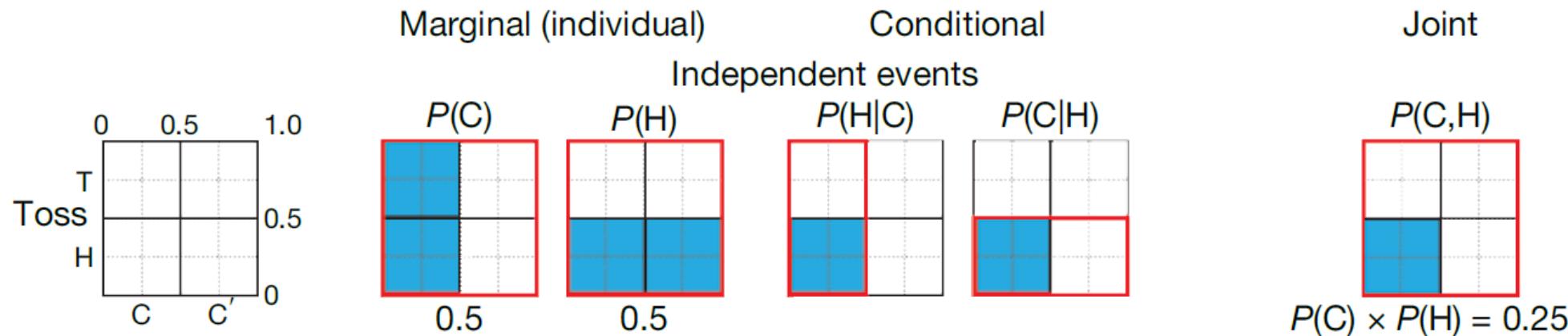
Solution 2:

$$P(A_1=H) = \frac{1}{2}; P(A_2=H) = \frac{1}{2}; P(A = A_1 A_2 = HH) = \frac{1}{4}$$

# Conditional Probability

If  $P(B) > 0$ , the conditional probability of  $A$  given  $B$  is given by

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



Apparently,  $A$  and  $B$  are independent events, if and only if

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

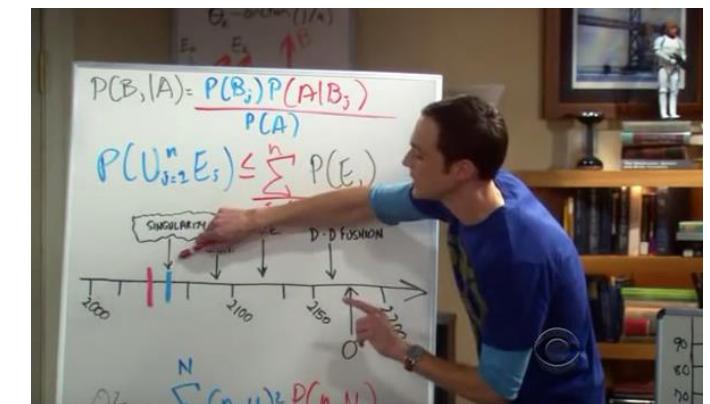
# Bayes theorem

Given that  $P(AB) = P(A | B) * P(B) = P(B | A) * P(A)$ , we derive

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

Prior

Posterior



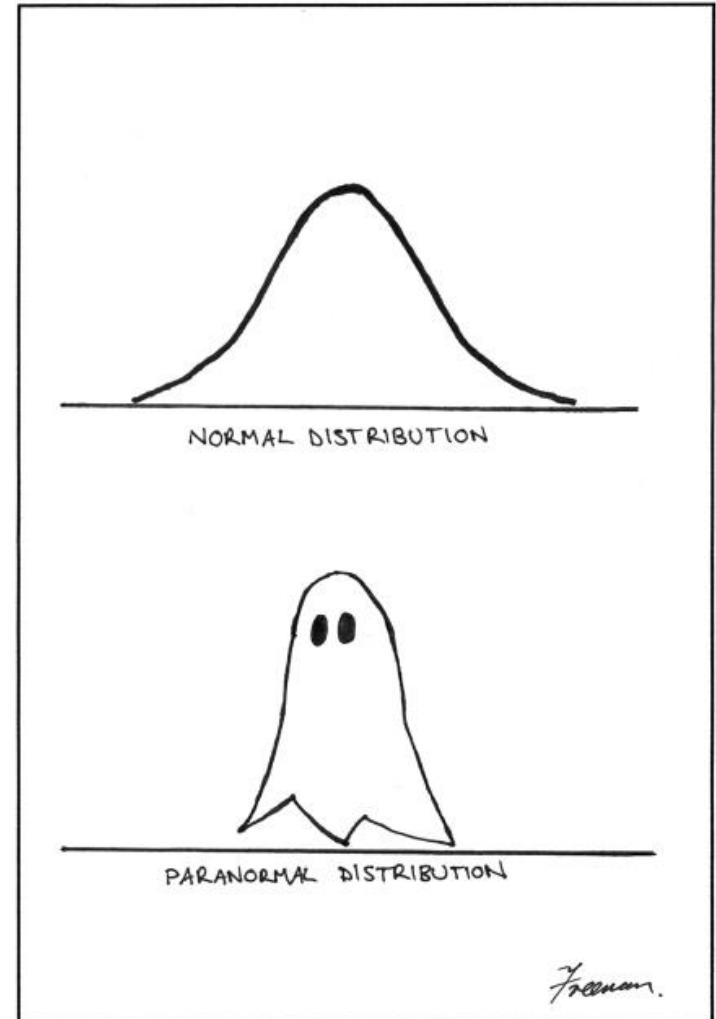
# Let's do a small exercise!



- A particular heart disease has a prevalence of 1/1000 people.
- The test used to detect this disease has a 5% false positive rate, meaning that the probability of a positive result given that a person does not have the disease ( $P(B | \text{not } A)$ ) is 0.05.
- Assume the test correctly identifies every individual who does have the disease.
- What is the chance that a randomly selected person found to have a positive result actually has the disease?

# Distribution

A map of possibilities.



# Probability function

For a **random variable**  $X$ , the **cumulative distribution function** (CDF)  $F_X: R \rightarrow [0,1]$  is defined as

$$F_X(x) = P(X \leq x)$$

□ If  $X$  is discrete, define **probability mass function** (PMF) as

$$f_X = P(X = x)$$

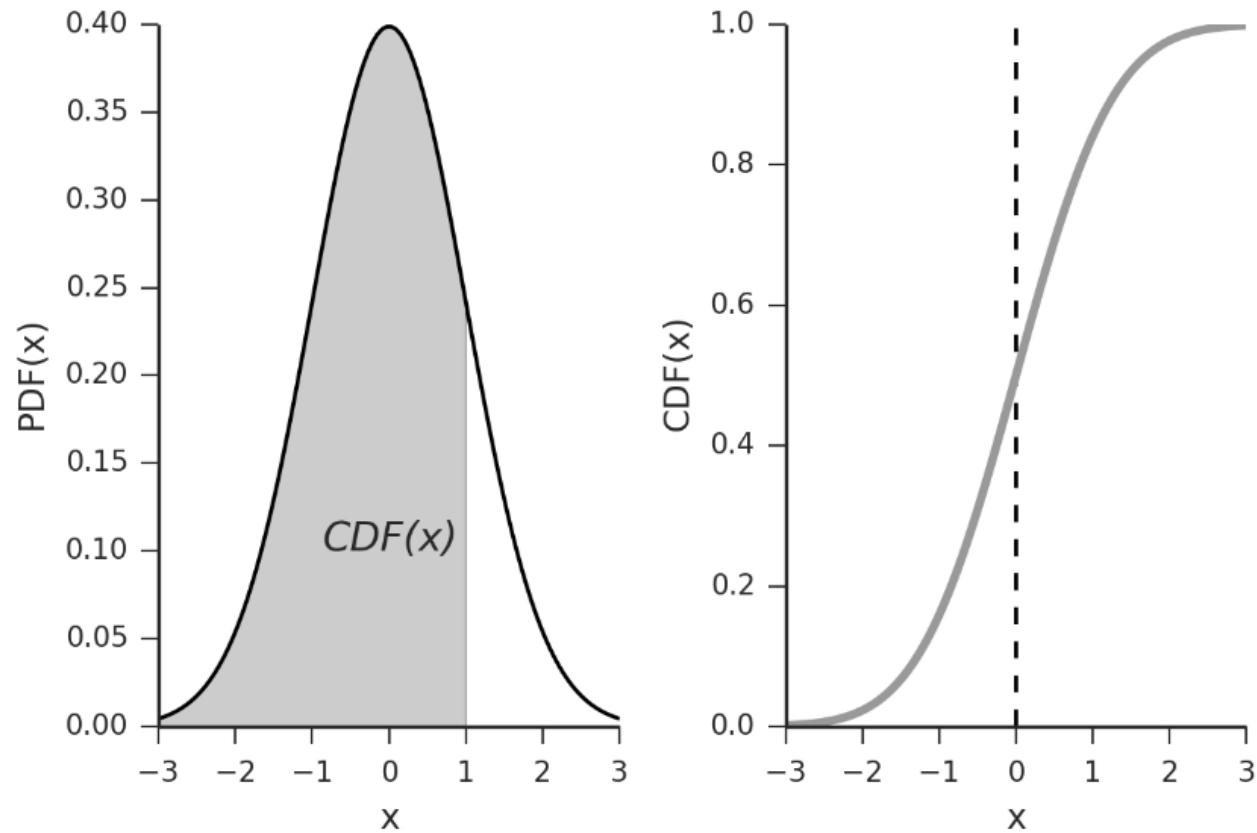
□ If  $X$  is continuous, define **probability density function** (PDF) as

$$f_X = F'_X(x)$$

That is

$$F_X(x) = \int_{-\infty}^x f_X(t)dt$$

# Comparison between PDF and CDF



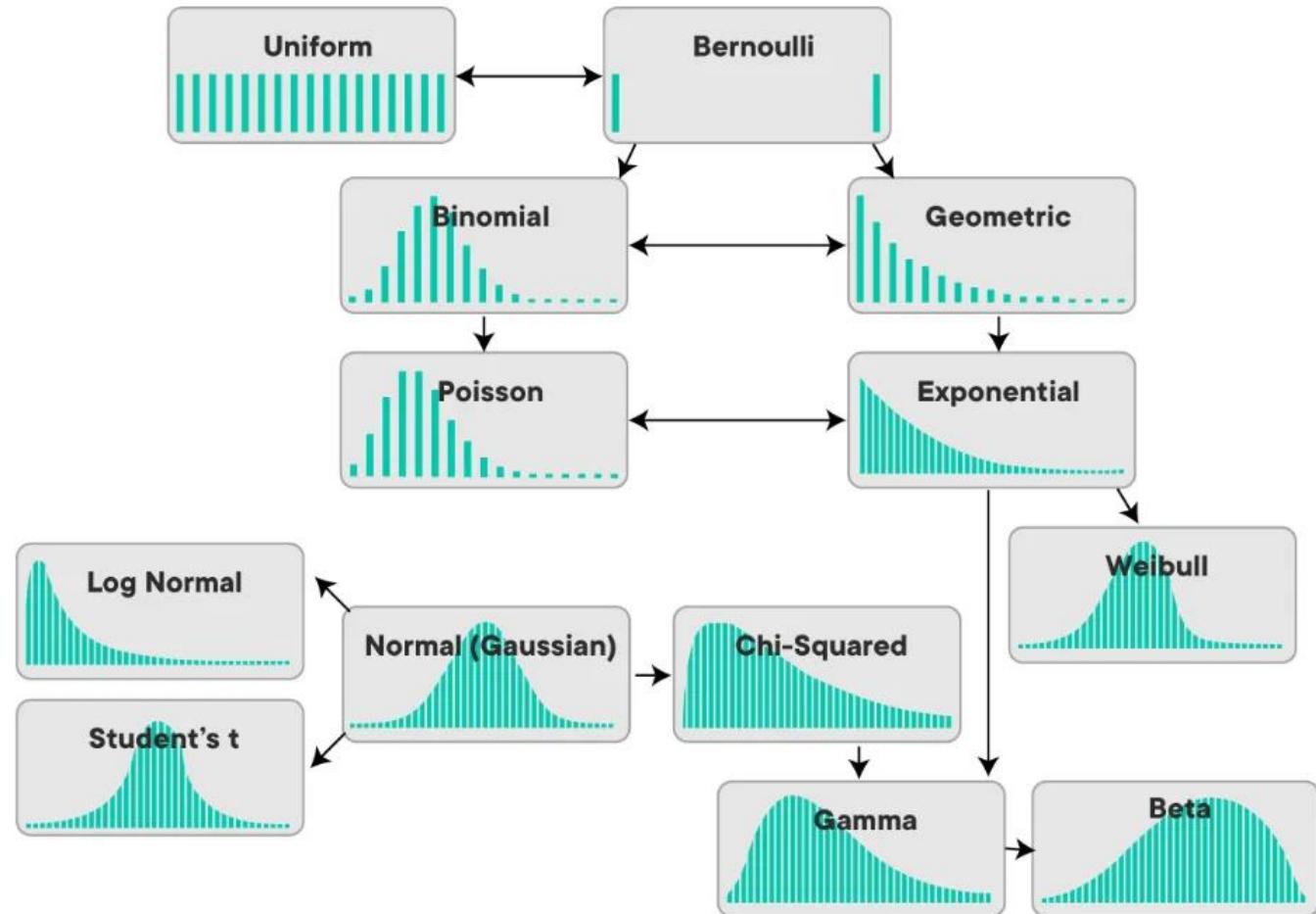
# Distributions

## Discrete

- Uniform
- Bernoulli
- Binomial
- Negative Binomial
- Geometric
- Hypergeometric
- Poisson

## Continuous

- Uniform
- Normal
- $t$
- Chi-square
- F
- Exponential
- Beta

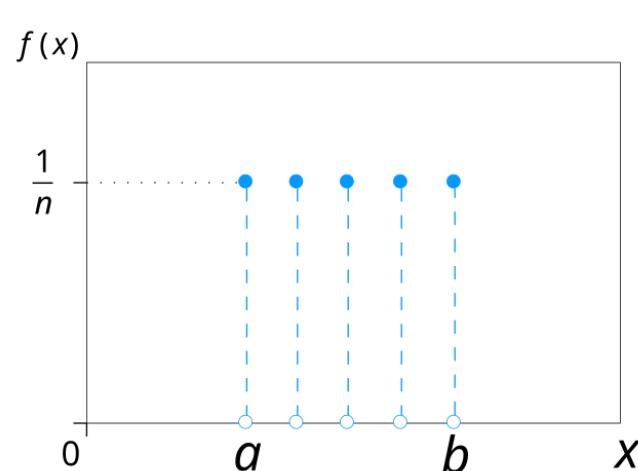


# Uniform (discrete)

**Intuition:** A discrete uniform distribution assigns **equal probability** to every integer between  $a$  and  $b$

$$P(X = x) = \frac{1}{b - a + 1}, \quad x = a, a + 1, \dots, b$$

**Example:** Rolling a fair 6-sided die is a uniform distribution with possible outcomes 1 to 6, each with equal probability 1/6

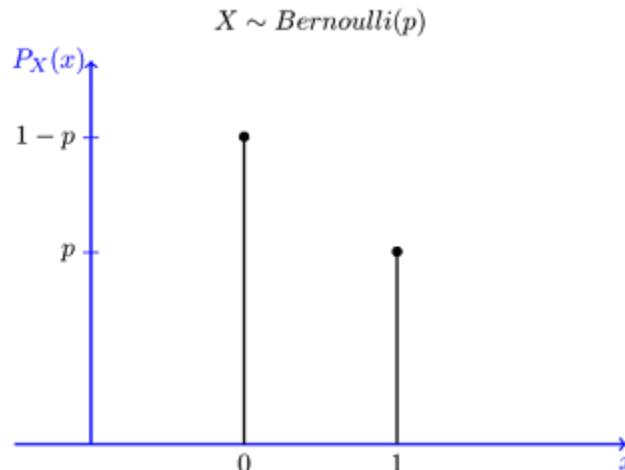


# Bernoulli

**Intuition:** A Bernoulli distribution represents a **single trial** with two possible outcomes: success (1) with probability  $p$ , and failure (0) with probability  $1 - p$ .

$$P(X = x) = p^x(1 - p)^{1-x}, \quad x = 0, 1$$

**Example:** Flipping a coin once — with heads as success and tails as failure — is a Bernoulli trial with  $p = 0.5$ .

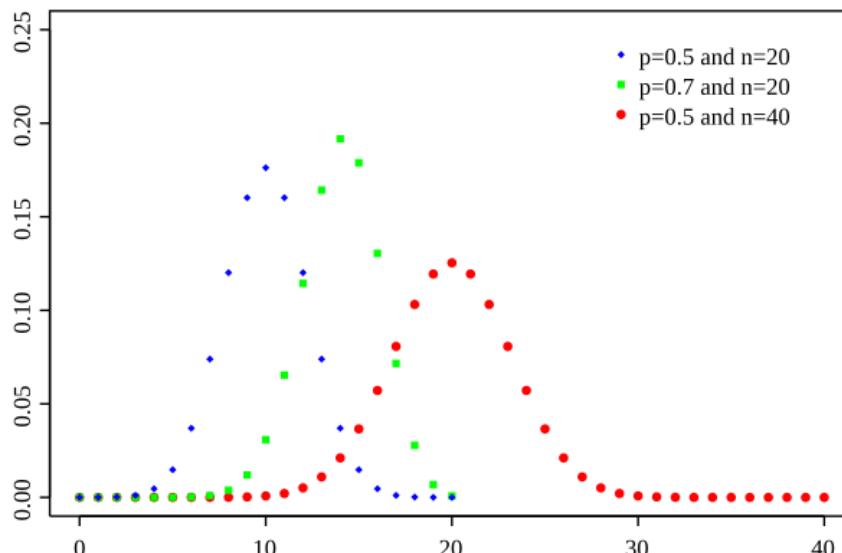


# Binomial

**Intuition:** The binomial distribution represents the **number of successes in  $n$  independent Bernoulli trials**, each with success probability  $p$ .

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n$$

**Example:** Flipping a coin  $n$  times and counting the number of heads is a binomial experiment.

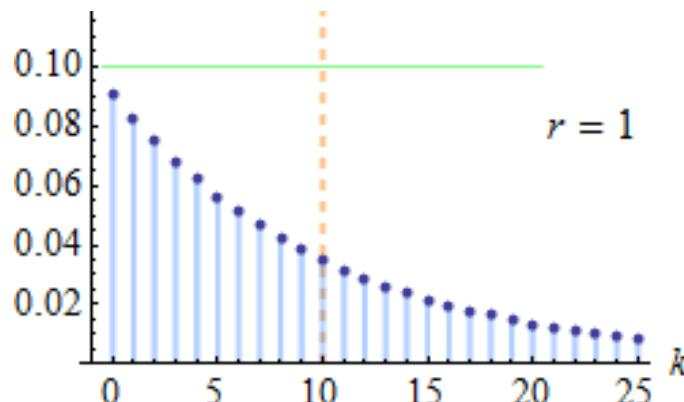


# Negative binomial

**Intuition:** The negative binomial distribution represents the **number of failures before achieving  $r$  successes** in a series of independent Bernoulli trials with success probability  $p$ .

$$P(X = k) = \binom{k + r - 1}{k} p^r (1 - p)^k, \quad k = 0, 1, 2,$$

**Example:** The number of times you need to roll a die before getting a 6 three times is a negative binomial experiment.

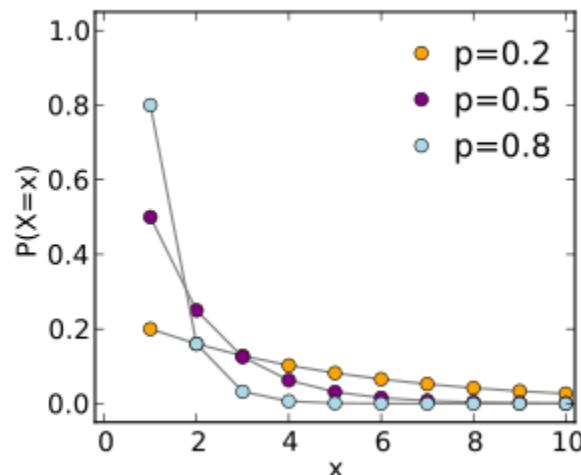


# Geometric

**Intuition:** The geometric distribution represents the **number of trials needed to get the first success** in a series of Bernoulli trials with success probability  $p$ .

$$P(X = k) = (1 - p)^{k-1}p, \quad k = 1, 2, 3,$$

**Example:** The number of coin flips needed to get the first heads is a geometric distribution.



# Hypergeometric

**Intuition:** The hypergeometric distribution models **the number of successes in a sample of size  $n$  drawn without replacement** from a population of size  $N$  containing  $K$  successes. This test is usually used in enrichment analysis.

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}, \quad \max(0, n - N + K) \leq k \leq \min(n, K)$$

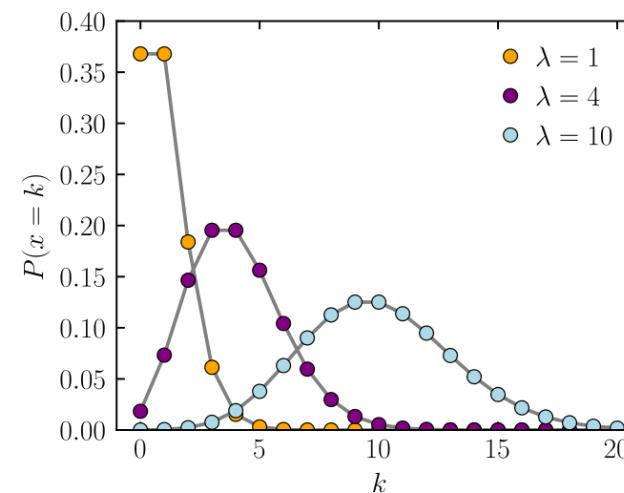
**Example:** Drawing cards from a deck — the number of red cards in a hand of 5 cards is a hypergeometric distribution.

# Poisson

**Intuition:** The Poisson distribution models the **number of events occurring in a fixed interval of time or space**, assuming the events occur randomly and independently at a constant average rate  $\lambda$ .

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2,$$

**Example:** The number of phone calls received at a call center in an hour follows a Poisson distribution.

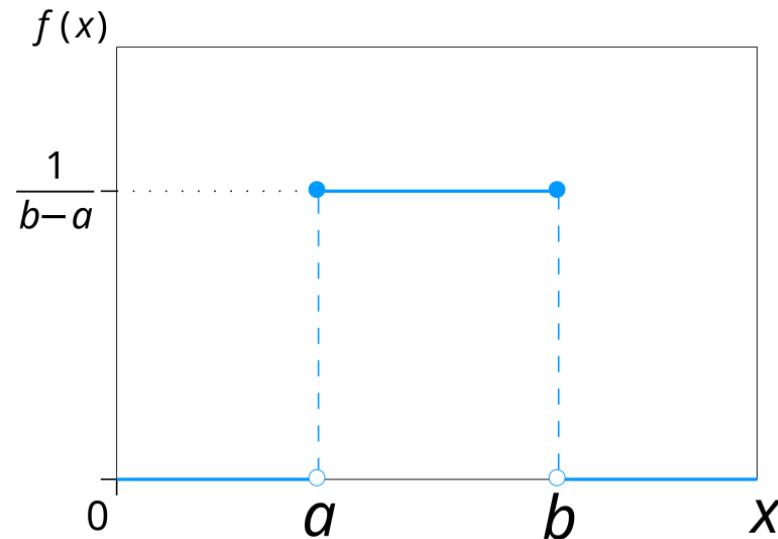


# Uniform (Continuous)

**Intuition:** Every value in the interval  $[a,b]$  is **equally likely**.

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

**Example:** The waiting time for a bus that arrives at a random time within a known 10-minute window is uniformly distributed.

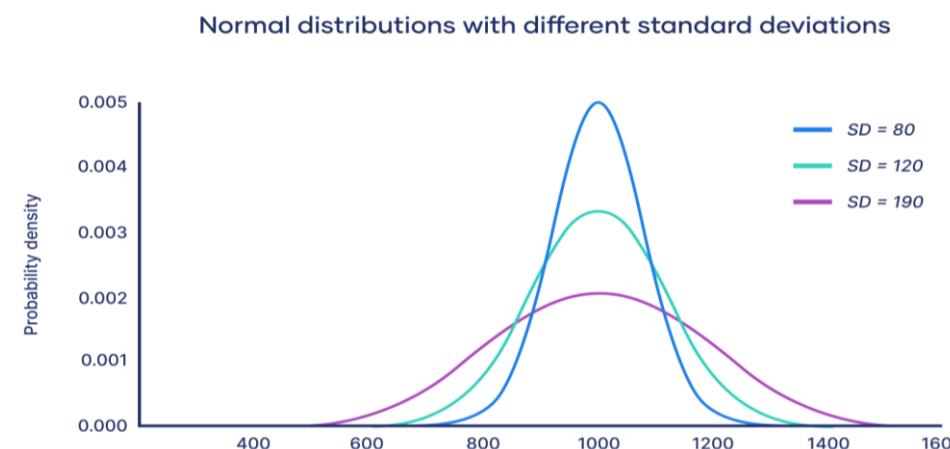
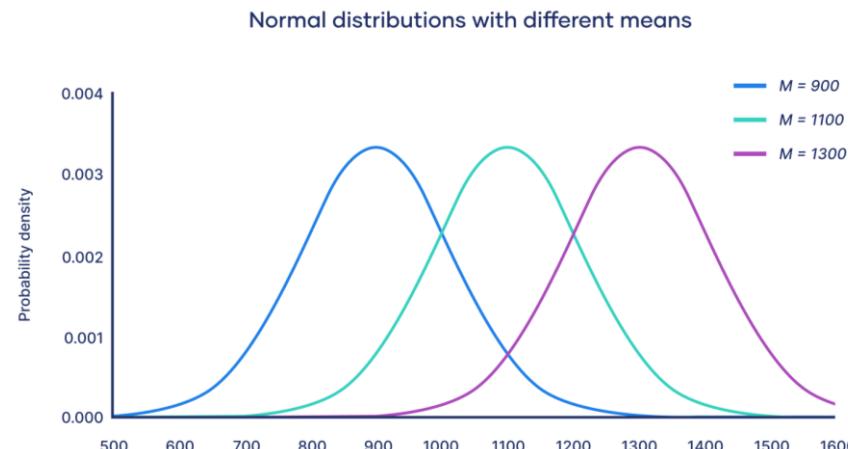


# Normal

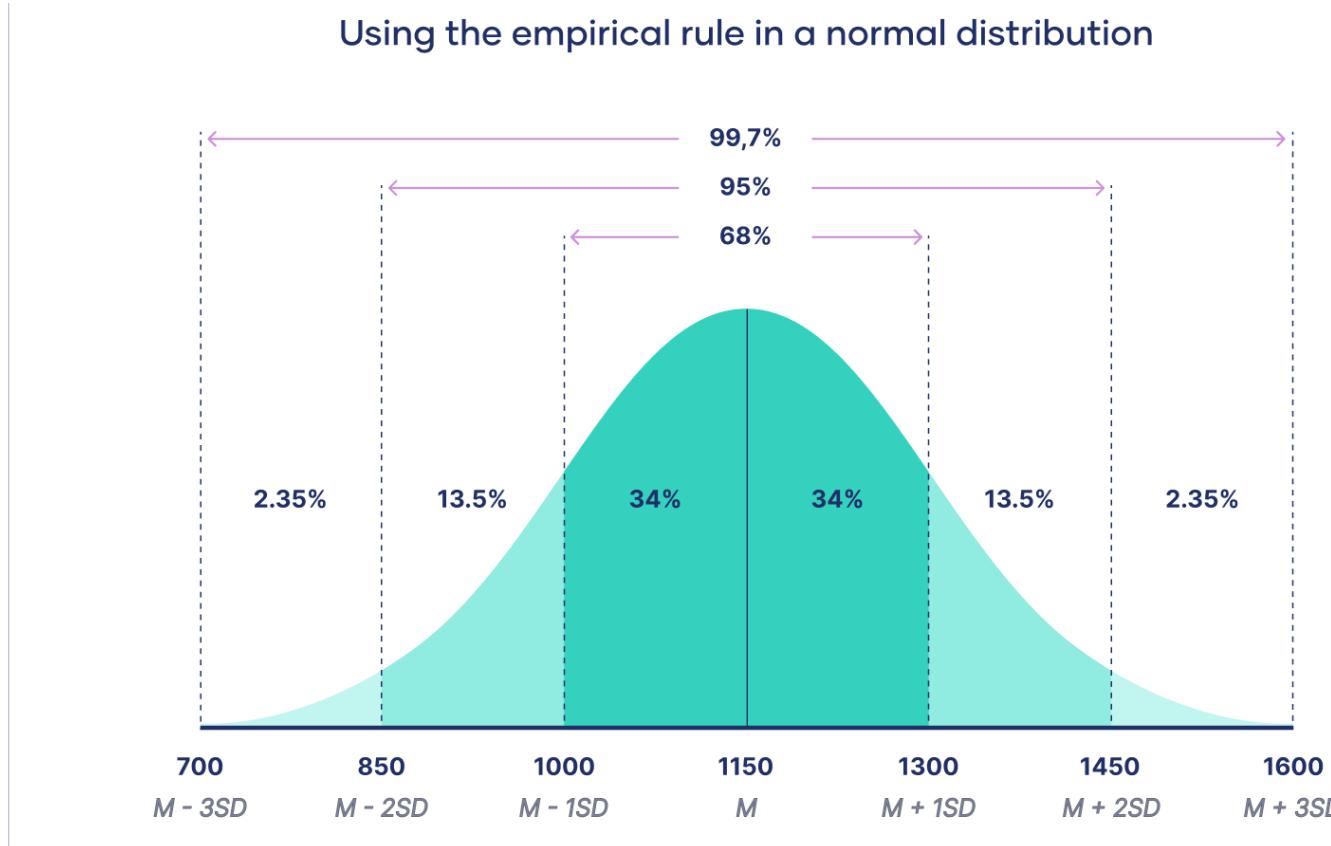
**Intuition:** The normal distribution (bell curve) models many natural phenomena like heights, weights, and measurement errors. It is **symmetric around the mean  $\mu$ , with standard deviation  $\sigma$ .**

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

**Example:** The distribution of heights in a large population



# Normal distribution and three-sigma rule of thumb



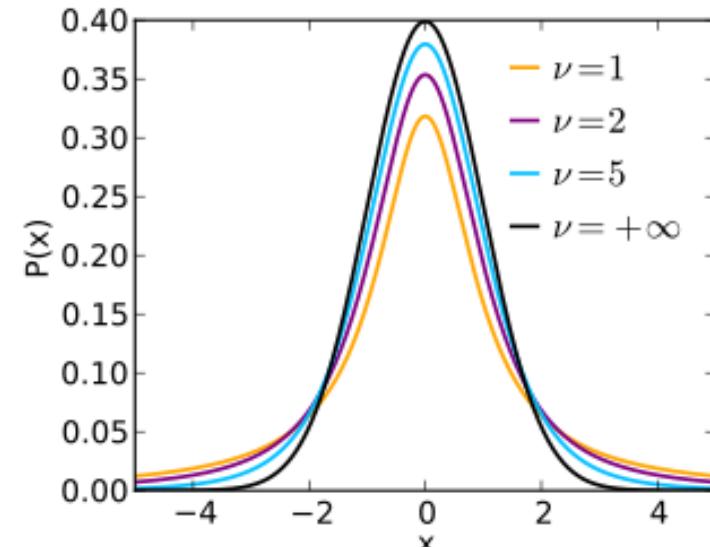
nearly all values are taken to lie within three standard deviations of the mean

# *t*-distribution

**Intuition:** The *t*-distribution is similar to the normal distribution but has **heavier tails**. It is used in hypothesis testing when the sample size is small.

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Where  $\nu$  is the degrees of freedom.



# t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Nonlinear dimensionality reduction

- PCA uses the global covariance matrix
  - t-SNE focus more on the local structure

- Core algorithm

- In high dimensional space

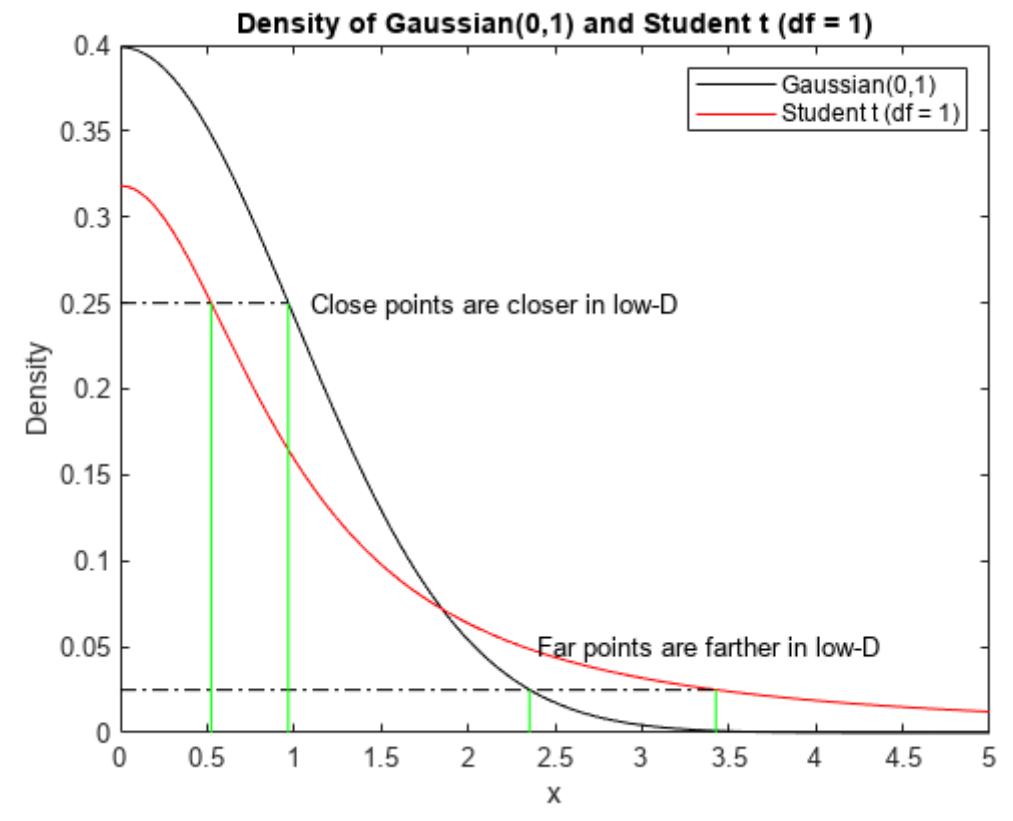
$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq l} \exp(-\|x_k - x_l\|^2 / 2\sigma_i^2)}$$

- In lower dimensional space

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

- Try to minimize the difference of 2 distributions

$$C = D_{\text{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{P(x)}{Q(x)} \right)$$

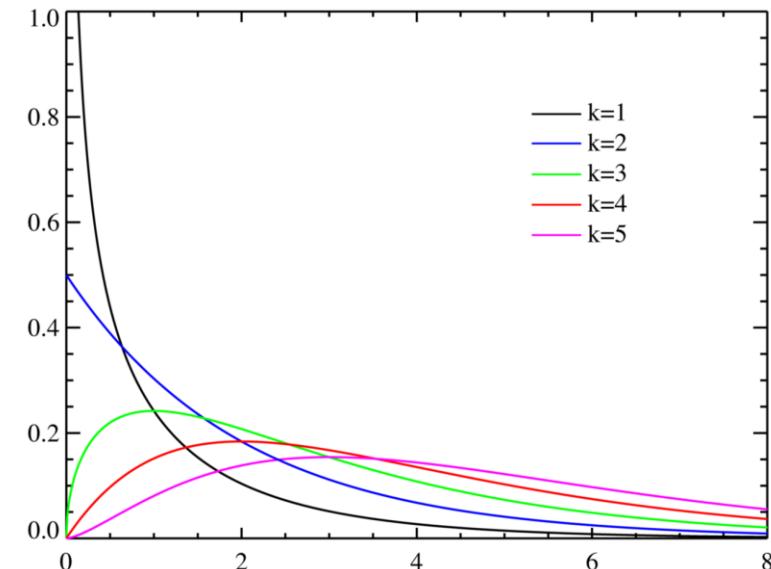


# Chi-square

**Intuition:** The chi-square distribution models the **sum of the squares of  $k$  independent standard normal variables**. It's commonly used in tests of independence and goodness-of-fit.

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \quad x \geq 0$$

where  $k$  is the degrees of freedom.

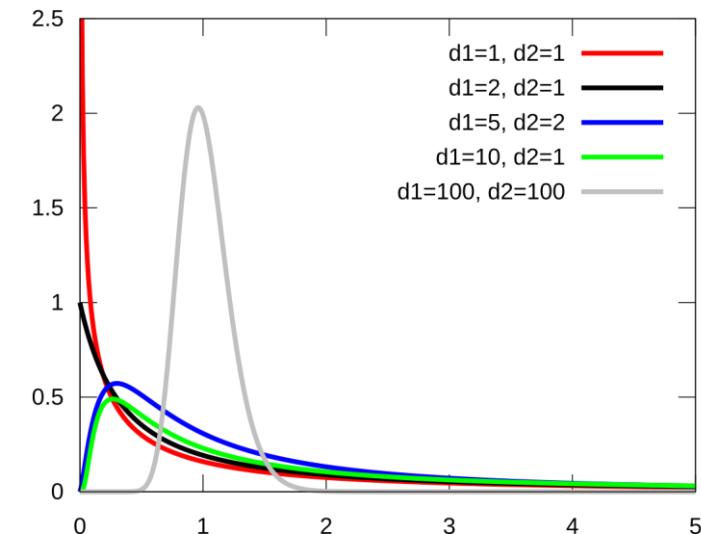


# F

**Intuition:** The F-distribution is used to **compare variances** between two groups. It arises in ANOVA and regression analysis.

$$f(x) = \frac{\left(\frac{d_1}{d_2}\right)^{d_1/2} x^{d_1/2-1}}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right) \left(1 + \frac{d_1}{d_2}x\right)^{(d_1+d_2)/2}}, \quad x \geq 0$$

where  $d_1$  and  $d_2$  are degrees of freedom.

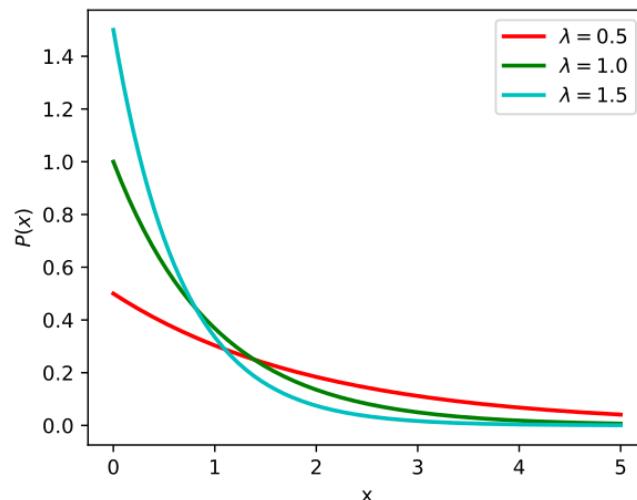


# Exponential

**Intuition:** The exponential distribution models the **time between events** in a Poisson process.

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

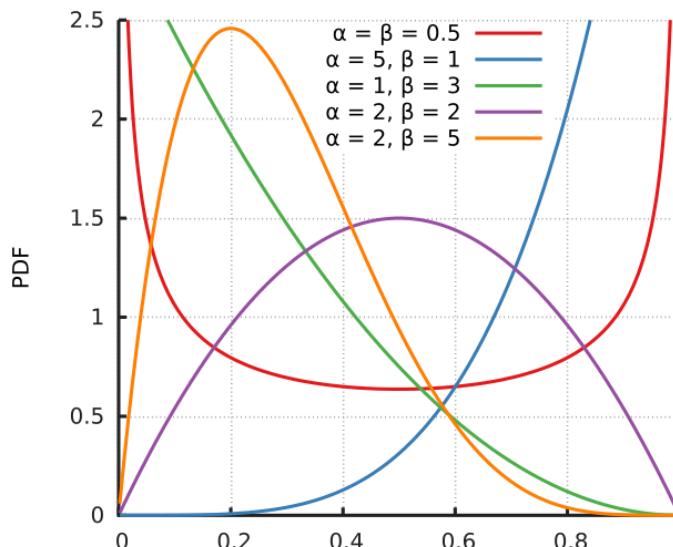
**Example:** The time between arrivals of customers at a service desk follows an exponential distribution.



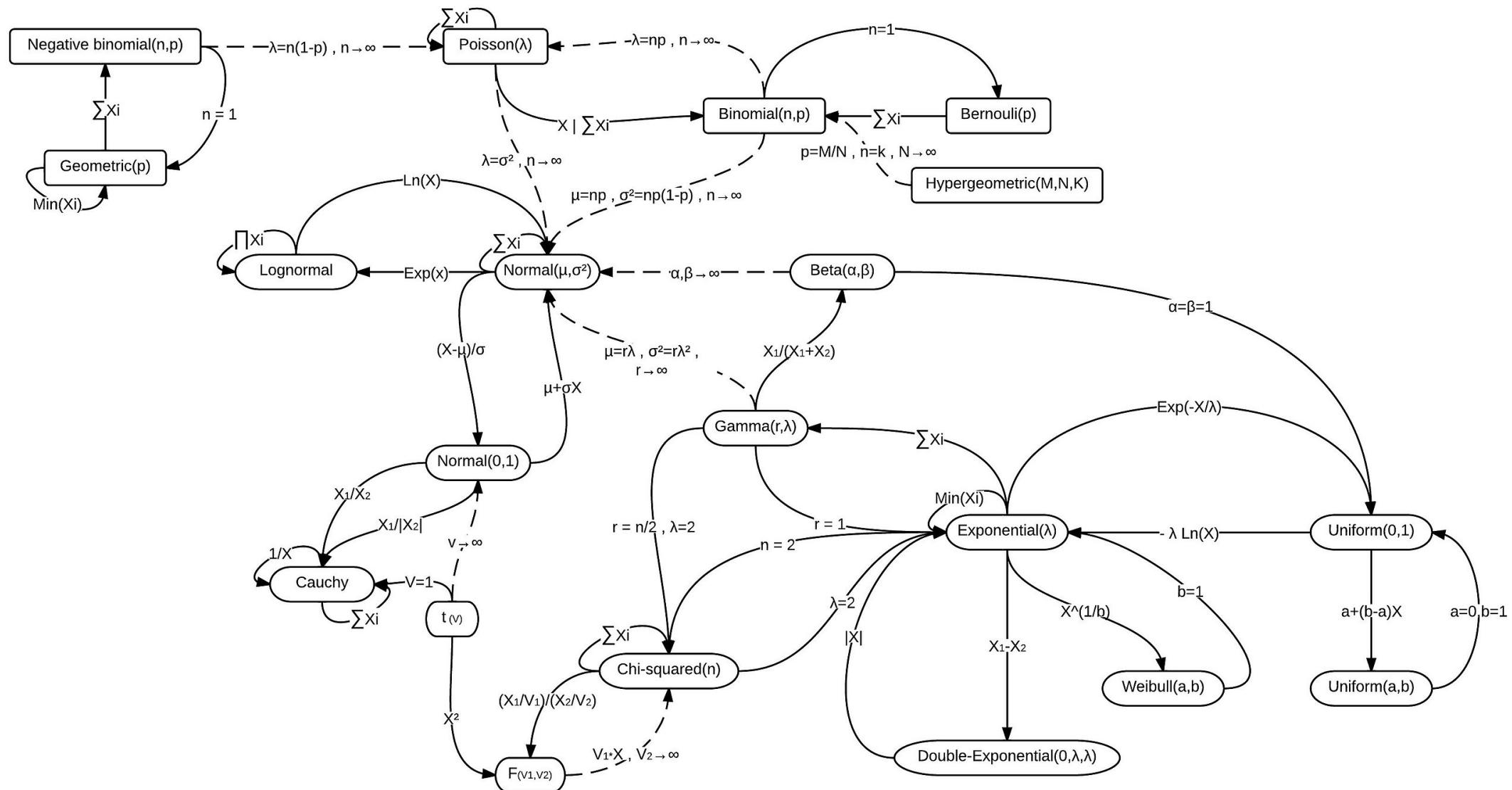
# Beta

**Intuition:** The beta distribution is a continuous distribution on  $[0, 1]$ . It's often used to **model proportions and probabilities**.

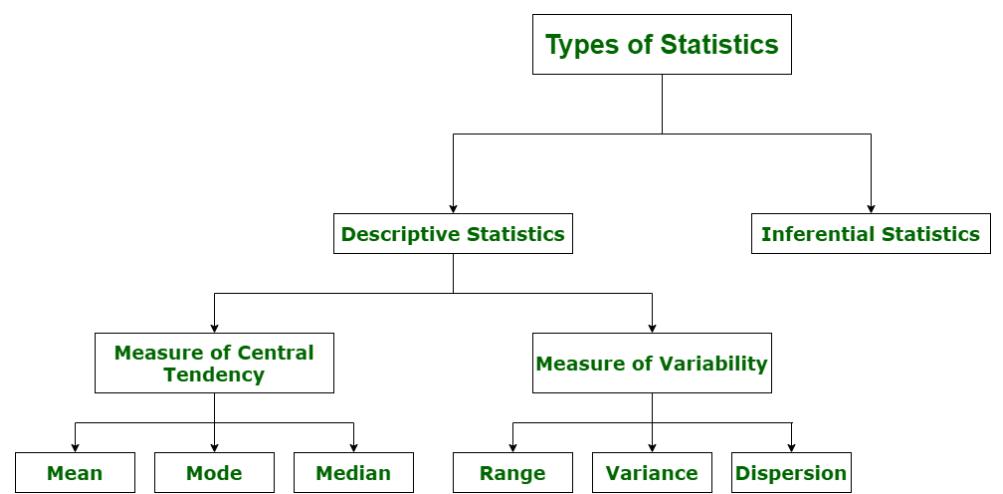
$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 \leq x \leq 1$$



# Relationship between distributions



## Types of Statistics



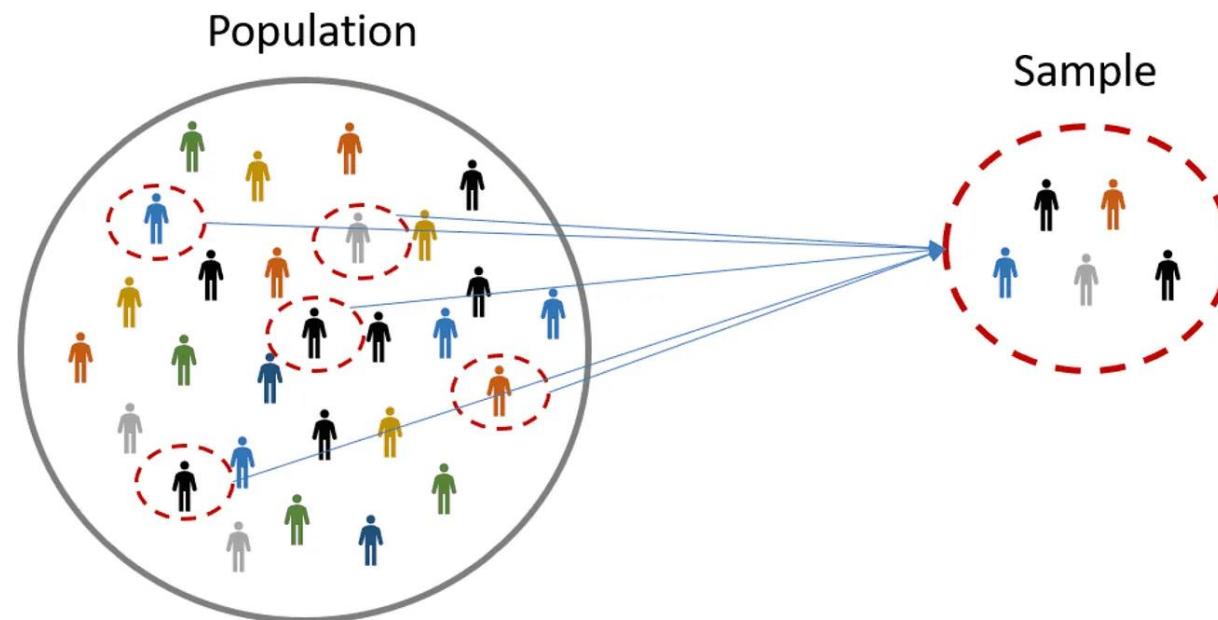
# Descriptive statistics

The first step in understanding is to describe.

# Population and samples

We study a finite set of samples from the population

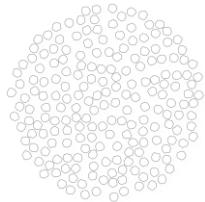
- It's too costly to measure the entire population
- It's impossible to iterate through the entire population
- ...



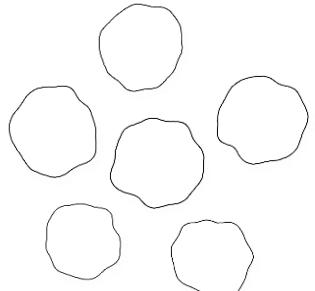
# Common Type of Experiment



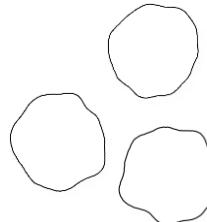
The whole population



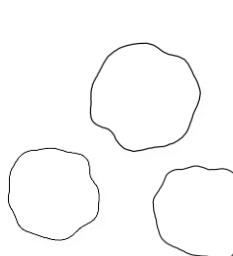
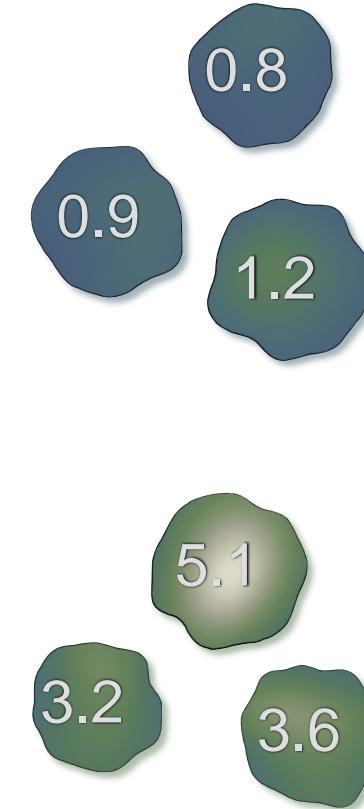
Random sample



Random assign



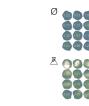
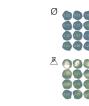
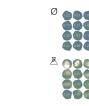
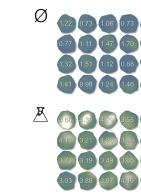
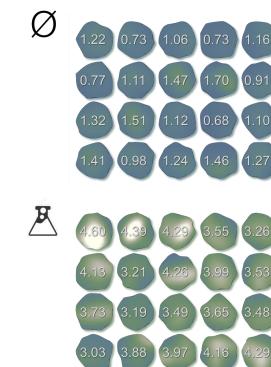
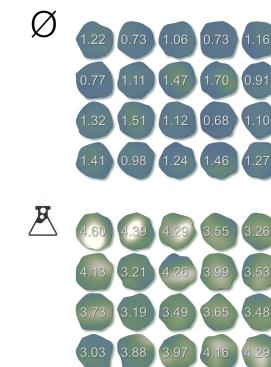
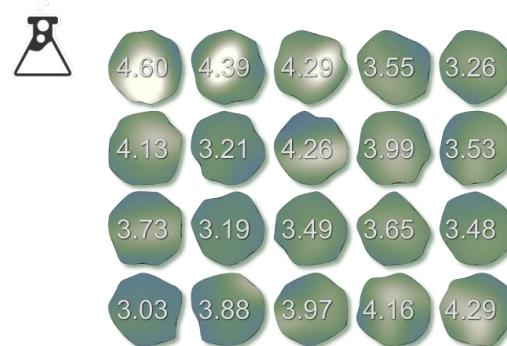
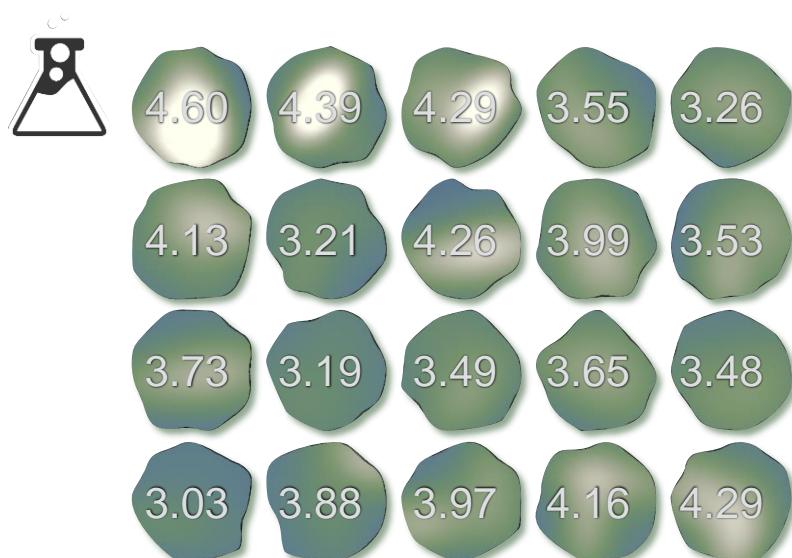
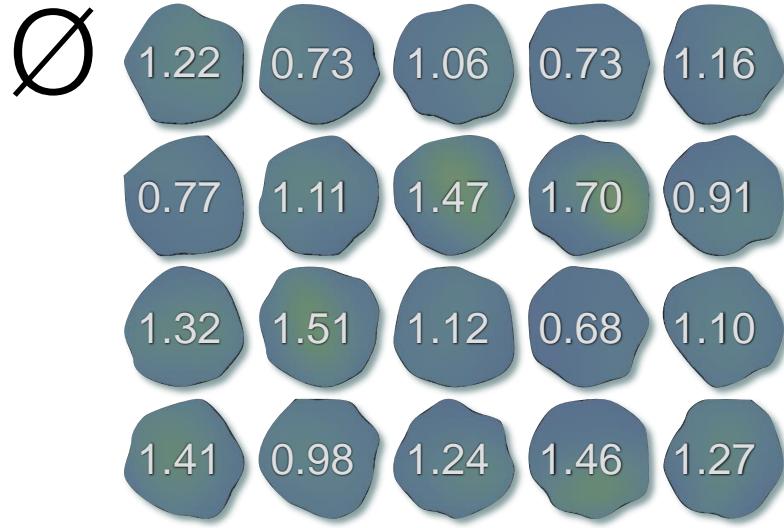
Control



Treated



# Summarizing data



Consider a **data compression** process

**Data → Information**

# Some summary statistics

**Tendency:** characterize the distribution's central or typical value

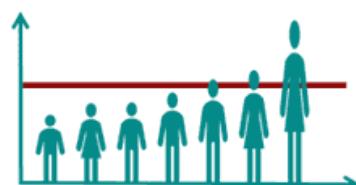
□ **Mean:**  $\frac{\sum_{i=1}^n x_i}{n}$

□ **Median:**

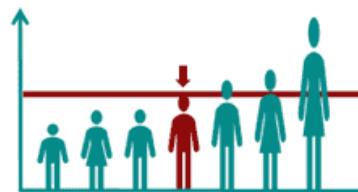
- For odd number observations:  $x_{(\frac{n+1}{2})}$
- For even number observations:  $\frac{x_{(n/2)} + x_{(n/2+1)}}{2}$

□ **Mode:** value or values that appear most frequently in a dataset.

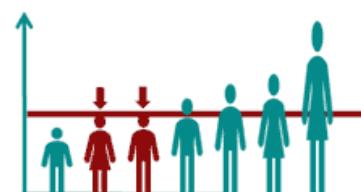
Mean



Median



Mode



# Some summary statistics

**Dispersion:** measures how far distribution members deviate from the center and each other.

## □ Variance/Standard deviation

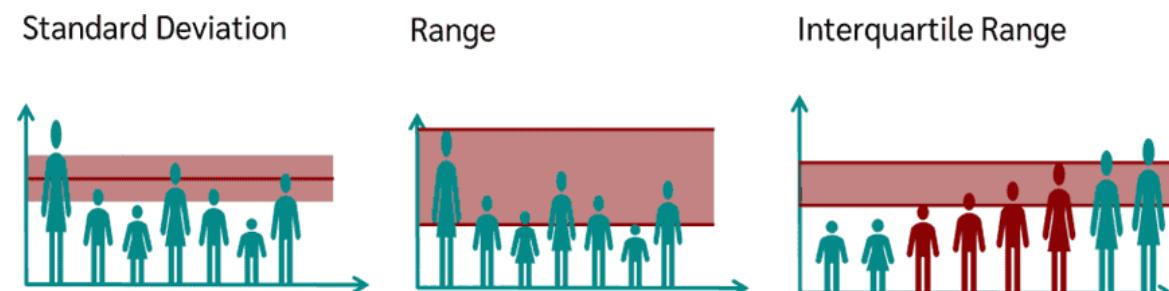
- Population variance:  $\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$
- Sample variance:  $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

## □ Quantiles: divide the data into 4 parts

- Q1 (25<sup>th</sup> percentile); Q2 (50th percentile or median), Q3 (75th percentile).

## □ Range/Inter Quantile Range (IQR)

- Range:  $x_{(n)} - x_{(1)}$
- IQR:  $Q3 - Q1$

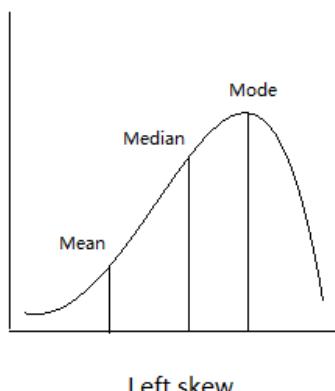


# Some summary statistics

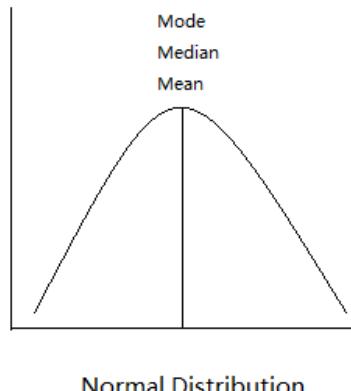
**Skewness:** measures the asymmetry of the distribution of data.

$$\text{Skewness} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{\frac{3}{2}}}$$

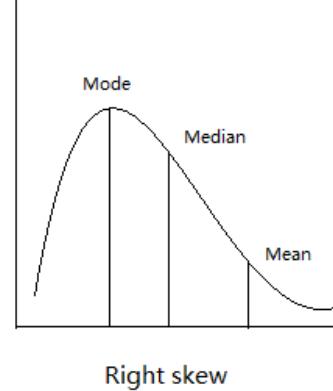
- Skewness = 0: Symmetric Distribution
- Skewness < 0: Negative Skewness (Left-Skewed)
- Skewness > 0: Positive Skewness (Right-Skewed)



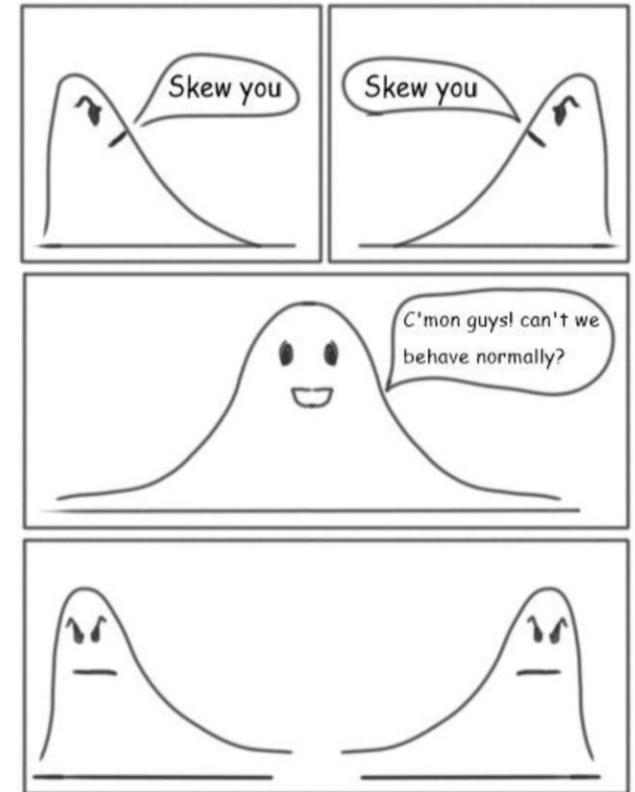
Left skew



Normal Distribution



Right skew

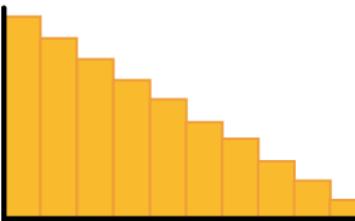


# Visualize distribution -- histogram

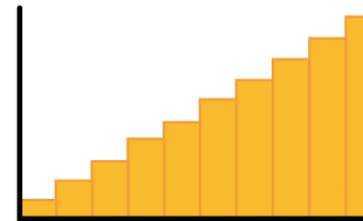
## Symmetric (normal) vs skewed and uniform distributions



**Normal distribution**  
(unimodal, symmetric,  
the “bell curve”)



**Right-skewed distribution**  
(Positively-skewed)



**Left-skewed distribution**  
(Negatively-skewed)

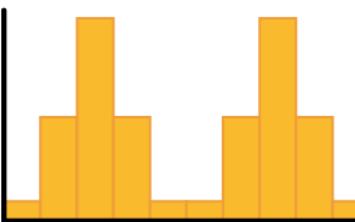


**Uniform distribution**  
(equal spread,  
no peaks)

## Unimodal vs bimodal distributions



**Normal distribution**  
(unimodal, symmetric,  
the “bell curve”)



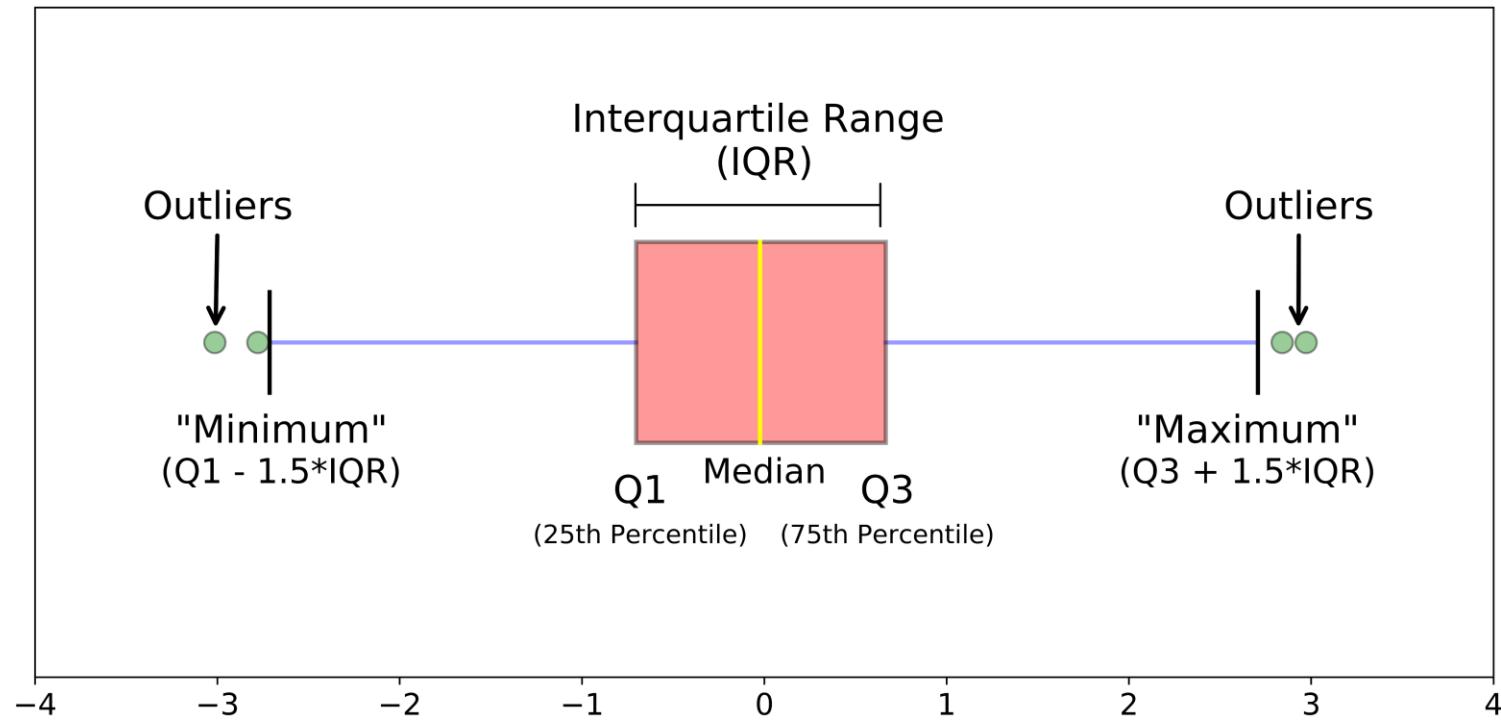
**Symmetric bimodal distribution**  
(two modes)



**Non-symmetric bimodal distribution**  
(two modes)

Always check the distribution

# Visualize distribution -- boxplot



# Thanks

Q & A